UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

FACULTADE DE MEDICINA E ODONTOLOXÍA

DEPARTAMENTO DE ANATOMÍA PATOLÓXICA E CIENCIAS FORENSES

La bioinformática

al servicio de la genómica

Memoria presentada para la obtención del grado de doctor por

Jorge Amigo Lechuga

Santiago de Compostela, septiembre 2012

El Doctor Ángel Carracedo Álvarez y el Doctor Antonio Salas Ellacuriaga, catedrático y profesor de Medicina Legal de la Facultade de Medicina e Odontoloxía de la Universidade de Santiago de Compostela, respectivamente,

CERTIFICAN

Que la presente memoria, que lleva por título "LA BIOINFORMÁTICA AL SERVICIO DE LA GENÓMICA", del licenciado en Ciencias Físicas por la Universidade de Santiago de Compostela Jorge Amigo Lechuga, ha sido realizada bajo nuestra dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para su presentación ante el Tribunal correspondiente.

Y para que así conste, firmamos la presente en Santiago de Compostela, a día 28 de septiembre de 2012

Fdo: Prof. Ángel Carracedo Álvarez. Fdo: Dr. Antonio Salas Ellacuriaga.

Fdo: D. Jorge Amigo Lechuga

Este trabajo ha sido financiado por un proyecto del Fondo de Investigación Sanitaria (PS09/02368) concedido a Ángel Carracedo, así como por proyectos del Ministerio de Ciencia e Innovación (SAF2008-02971 y SAF2011-26983) y de la Consellería de Cultura, Educación e Orden de la Xunta de Galicia (2012-PG180) concedidos a Antonio Salas.

Son muchas las personas a las que tendría que agradecer el poder haber llegado a la conclusión de este trabajo. En primer lugar a **mis padres**, que me inculcaron unos valores y pusieron a mi disposición una educación de alta calidad que me han ayudado enormemente a lo largo de mi vida. Por supuesto a **mi mujer**, sin cuyo ejemplo de constancia y tenacidad difícilmente hubiera sido capaz siquiera de terminar la carrera. Y también debo agradecerle a **Carlos Hernández** que me ayudase a decantarme por estudiar Ciencias Físicas en Santiago y que me permitiese entrar en el entorno universitario apenas unas semanas antes de aceptar una oferta en Madrid de una consultora canadiense.

Por el cambio de enfoque que supusieron sus consejos en mi concepción laboral, tengo que hacer una mención especial a **Ángel Carracedo** que, ayudado por la entonces reciente publicación de la secuencia del genoma humano, me convenció de la belleza de la genética humana. Su figura es aún a día de hoy una de mis mayores motivaciones para continuar investigando. Además, tuve por aquel entonces la suerte de conocer a uno de los firmantes de la anterior publicación, **Roderic Guigó**, quien me aportó la inspiración necesaria para dar el paso definitivo al mundo de la investigación al abrirme las puertas de su grupo apenas unos meses, tiempo suficiente para descubrir con fascinación el campo de la genómica computacional, y de reunir las fuerzas suficientes para ampliar mi formación en Inglaterra, lo que me permitió más tarde encajar perfectamente en el grupo de Ángel a mi vuelta.

A partir de ahí, todo son agradecimientos más cercanos, a gente con la que he trabajado horas y horas tratando de sacar proyectos adelante. Mi etapa en el CeGen comenzó con **Bea**, **María** e **Inés**, que me enseñaron con creces la exigencia del laboratorio y me ayudaron a aterrizar en el campo de la investigación molecular, y me permitió tratar de tú a tú con unos monstruos de la investigación de los que no paraba de aprender algo cada vez que me dirigía a ellos: desde las charlas en mis comienzos con **Clara, Ana** y **Celsa**, pasando por las conversaciones de despacho con **Xulio** y **Carolina**, hasta los intensos debates con **Susi** y **Xabi** o las siempre interesantes disertaciones de **Maviky** y **Pancho**.

Last but not least **Chris**, que no ha dejado nunca de proponerme proyectos interesantes de genética forense desde mi llegada al grupo, y por supuesto **Toño**, que con su constante disposición para el trabajo y su enfermiza capacidad de descubrir y sugerir nuevos retos sin duda me ha aportado muchos de los mayores placeres intelectuales vividos en este grupo.

A mi mujer y a mis hijos,

que lo son todo.

"Daría todo lo que sé por la mitad de lo que ignoro"

René Descartes (1596-1650)

I. INTRODUCCIÓN 1

1. Consideraciones previas 3

1.1. Sobre variabilidad, diversidad, y heredabilidad 4

1.2. Cuantificando la variabilidad 6

1.3. Medición de la variabilidad 9

2. Poblaciones humanas 15

2.1. Equilibrio Hardy-Weinberg 16

2.2. Linkage disequilibrium - LD 17

2.3. Mezcla génica poblacional 18

2.4. Ancestry informative markers – AIMs 18

2.5. Deriva génica 19

2.6. Selección 20

3. Grandes esfuerzos de recolección 23

3.1. Human Genome Project (HGP) 24

3.2. Perlegen Sciences, Inc. 25

3.3. Centre d'Etude du Polymorphisme Humain (CEPH) 26

3.4. International HapMap Project 27

3.5. 1000 Genomes 30

4. Acerca de la bioinformática 31

4.1. La bioinformática en su contexto 32

4.2. Aplicaciones de la bioinformática 33

4.3. Recursos de libre disposición 36

II. JUSTIFICACIÓN Y OBJETIVOS 43

III. RESULTADOS 49

The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project. 53

SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. 61

SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. 65

pop.STR - An online population frequency browser for established and new forensic STRs. 73

Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. 77

A reduced number of mtSNPs saturates mitochondrial DNA haplotype diversity of worldwide population groups. 91

Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. 101

Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and *in silico* binding site prediction. 119

Adaptive selection of an incretin gene in Eurasian populations. 135

ENGINES: exploring single nucleotide variation in entire human genomes. 149

Call for participation in the neurogenetics consortium within the Human Variome Project. 157

The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome. 165

GDF: Dealing with high-throughput genotyping multiplatform data for medical and population genetic applications. 181

IV. DISCUSIÓN 189

V. CONCLUSIONES 195

VI. REFERENCIAS 199

I. INTRODUCCIÓN

1. Consideraciones previas

Antiguamente, los científicos eran capaces de observar distintas características en las diversas especies, pero desconocían cómo se heredaban de una a otra generación. Incluso Darwin, que describió con gran detalle la importancia de la adaptación de dichas características específicas en la supervivencia de las especies, no llegó a saber que el origen de la mayoría de estos rasgos hereditarios pueden ser trazados hasta entidades persistentes llamadas genes, codificados en moléculas lineales de ácido desoxirribonucleico (ADN), y que varían no sólo entre especies sino también entre los miembros de una misma especie.

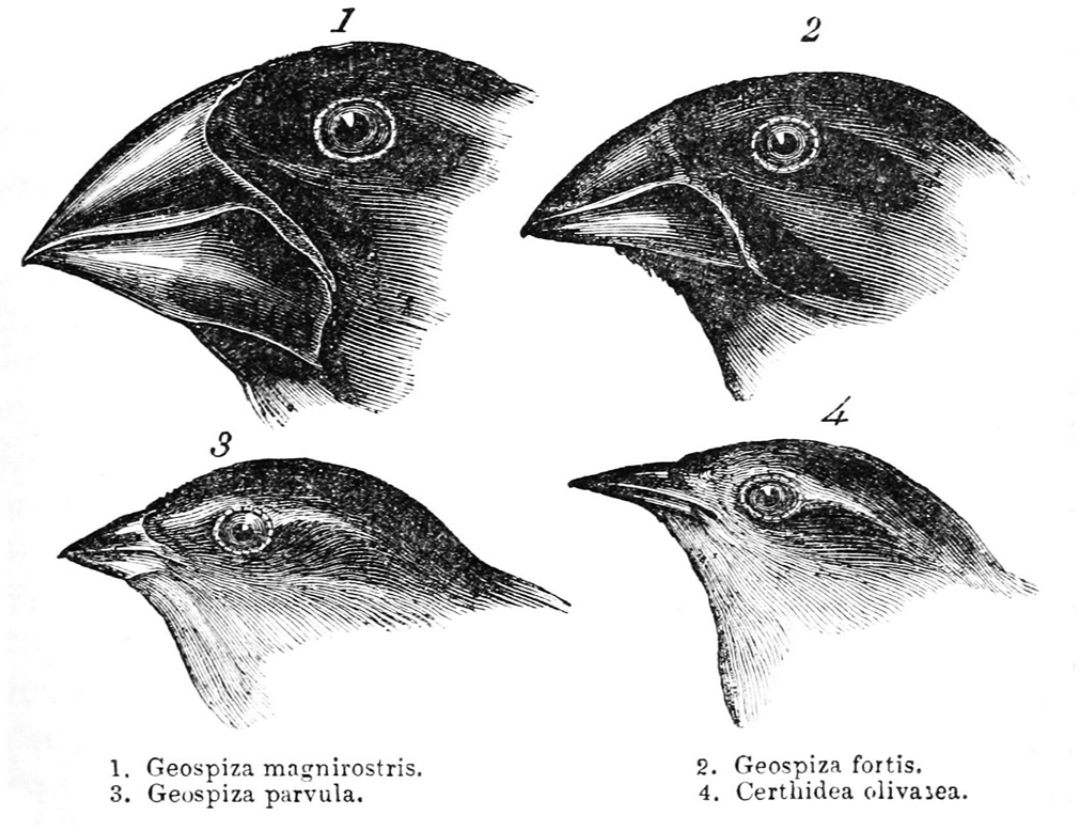


Figura 1: Pinzones de Darwin o pinzones de las Galápagos. Se denominan así a 14 especies diferentes, aunque estrechamente relacionadas, descubiertas por Charles Darwin en las Islas Galápagos durante su viaje en el Beagle. Todas ellas son de similares dimensiones, siendo sus mayores diferencias el tamaño y forma del pico, que se encuentra perfectamente adaptado a las diferentes fuentes de alimento. Fuente: http://www.wikipedia.org

1.1. Sobre variabilidad, diversidad, y heredabilidad

La variabilidad genética engloba a toda aquella variación en el material genético de una población o especie, incluyendo todos los genomas presentes en dicha especie ya sea nuclear, mitocondrial, ribosomal o el de otros orgánulos. Es la materia prima de la evolución y del cambio, ya que es sobre ella donde actúan los distintos procesos evolutivos. La manera de cuantificar la riqueza de esta variación es a través de la diversidad genética o diversidad intraespecífica, que viene dada por la cantidad de versiones distintas de los genes (alelos) y de su distribución, que a su vez es la base de las variaciones interindividuales (variedad de los genotipos). La proporción que esta diversidad aporta a la variación total observable de un individuo, en contraste con los efectos causados por la variación ambiental, se denomina heredabilidad [1].

Existen dos claras escalas de medida del cambio evolutivo. Mientras en el "Origen de las Especies" Charles Darwin estaba principalmente interesado en la evolución de las especies en tiempo geológico, lo que suele referirse como macroevolución, existen procesos que operan sobre la diversidad genética de una única especie en una escala de tiempo de generaciones, cuyo análisis suele referirse como microevolución. Para medir los procesos de microevolución que moldean la diversidad genética debemos estudiar los cambios en frecuencias de alelos y haplotipos presentes dentro de cada población, entendiendo por población un grupo de individuos de una misma especie y de sus descendientes que se cruzan entre sí, y que está sujeta a cambios evolutivos en los que subyacen cambios genéticos tales como la selección natural y deriva génica, que actúan principalmente disminuyendo su variabilidad eliminando las variaciones menos favorables o simplemente por azar, o la migración y mutación que actúan aumentándola al introducir variaciones nuevas.

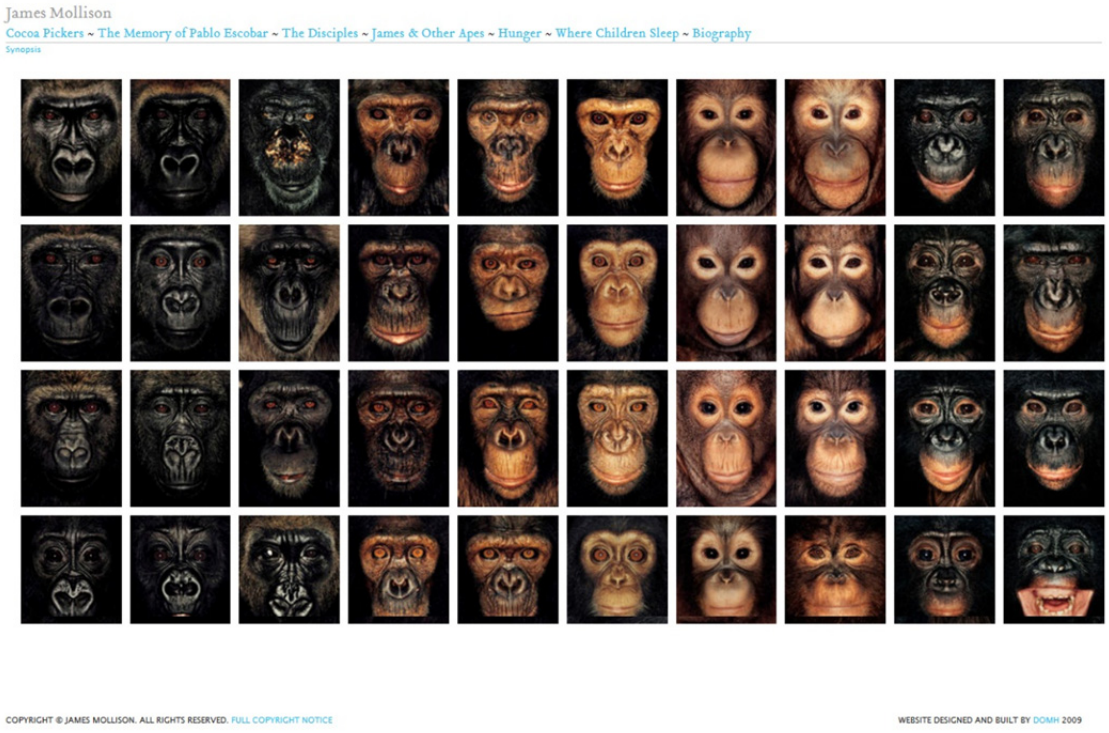


Figura 2: James & Other Apes, by James Mollison. Retratos fotográficos de distintos simios pertenecientes a distintas especies cercanas entre sí que muestran gráficamente la diversidad tanto interespecífica como intraespecífica. Fuente: http://www.jamesmollison.com

1.2. Cuantificando la variabilidad

La manera en que se mide la variabilidad genética es a través de marcadores genéticos, unos indicadores que podemos detectar, medir y cuantificar tanto dentro de un individuo como a nivel poblacional. Y aunque la mayor parte de los análisis de variación genética humana es hoy a nivel de ADN, y muy extensamente a nivel de secuencia, conviene recordar que hace apenas unas décadas, antes de la revolución del ADN recombinante de los años 70 y 80, se hacía a nivel de proteínas ya que eran lo único observable a través de geles de electroforesis, por no remontarnos al sistema ABO de grupos sanguíneos de principios del siglo XX [2].

A medida que la tecnología lo ha ido permitiendo se han ido definiendo distintos marcadores con distinta funcionalidad, aunque su finalidad siempre termina siendo la cuantificación de la variabilidad. Desde las más sencillas variaciones de un único nucleótido en la cadena de ADN (polimorfismo de secuencia única, o SNP) o pequeñas inserciones y deleciones en el genoma nuclear (indels), a las más complejas repeticiones en tándem (VNTRs) o reordenaciones estructurales. Todas ellas son variaciones que pueden ser detectadas y evaluadas a lo largo del genoma humano, y que permiten estudiar la variabilidad de dicho genoma tanto dentro del individuo como en un contexto poblacional.

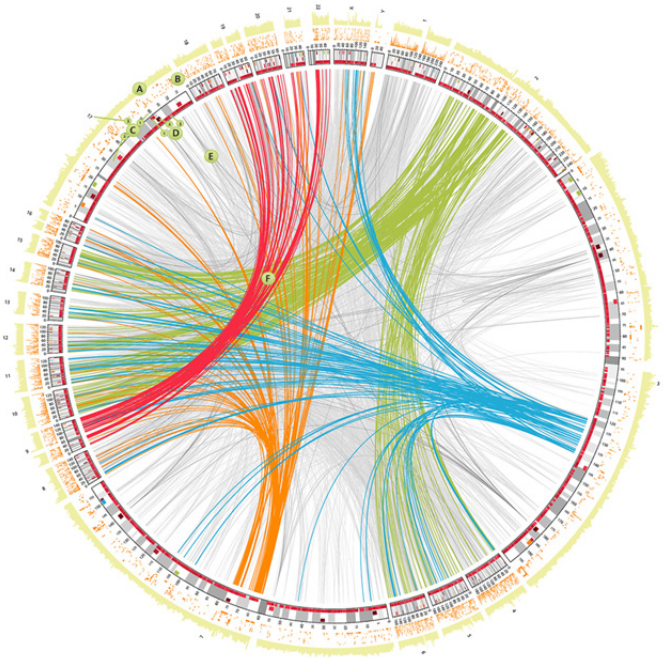


Figura 3: Representación gráfica de variantes usando Circos [3]. Los cromosomas se presentan circularmente mediante ideogramas etiquetados, pudiendo así mostrar la ubicación de los genes implicados en una enfermedad, regiones de alta similitud, y variaciones estructurales entre poblaciones. La figura muestra información del genoma humano, obtenida de la versión hg18 que mantiene la University of California Santa Cruz.

1.2.1. Polimorfismos, mutaciones y variantes

Una simple diferencia por la sustitución de una única base entre dos secuencias de ADN humano puede ser etiquetada de maneras diferentes, y desafortunadamente las reglas para aplicar dichas etiquetas no son siempre claras o consistentes. De hecho, y de manera un tanto arbitraria, se ha venido definiendo la existencia de un polimorfismo cuando al menos dos alelos están presentes en una población, y el alelo menor se encuentra a una frecuencia mayor del 1% [4]. A pesar de que todas las diferencias entre secuencias son debidas a mutaciones, el término *mutación* se reserva habitualmente para referenciar variaciones patogénicas, y por tanto se usa en contraste a *polimorfismo*. Aunque esta distinción de significado es particularmente prevalente entre genetistas clínicos, conlleva un claro riesgo potencial al ser difícil o incluso imposible saber con certeza si un cambio en una secuencia puede estar totalmente libre de acarrear un efecto fenotípico [5]. Es más, es posible encontrar mutaciones causantes de enfermedades con una frecuencia mayor del 1% en ciertas poblaciones, y por tanto podría encajar en la definición de polimorfismo. Ejemplo de esta situación son alguna mutación en europeos del gen CFTR responsable de la fibrosis quística, o mutaciones en algunas poblaciones africanas en el gen de la beta globina responsable de la anemia falciforme.

Un alelo con una frecuencia inferior al 1% suele llamarse *variante*. Claramente, como las frecuencias alélicas pueden variar ampliamente entre poblaciones, una variante de una población puede ser un polimorfismo en otra. Además, el término variante se suele usarse como genérico para incluir a polimorfismos y a mutaciones, ya que el término mutación tiene connotaciones negativas para el profano en la materia [6], hecho que ha llevado a discusiones acerca de la elección de términos alternativos cuando los científicos realizan labores de comunicación acerca de la diversidad genética al público general, en las que *variación*, *variante* o *alteración* se escogen preferentemente.

1.2.2. Short Tandem Repeats (STR)

Un tipo de variación genética muy dinámica y común en genomas eucariotas es la variación del número de copias de secuencias de ADN organizadas en tándem. Se denominan genéricamente VNTRs a estas variaciones en número de copias, pero en función del tamaño de dichas copias se clasifican principalmente en microsatélites si son unas pocas unidades de bases, minisatélites si son unas pocas decenas de bases, o simplemente satélites si son unos pocos millares de bases las que conforman la unidad repetida. Los microsatélites se conocen también como STRs, y aquellos que sirven como marcadores genéticos útiles tienen una frecuencia típica de repetición de entre 10 y 30 copias, y una tasa de mutación estimada en humanos de entre 10^{-3} y 10^{-4} por locus y por generación.

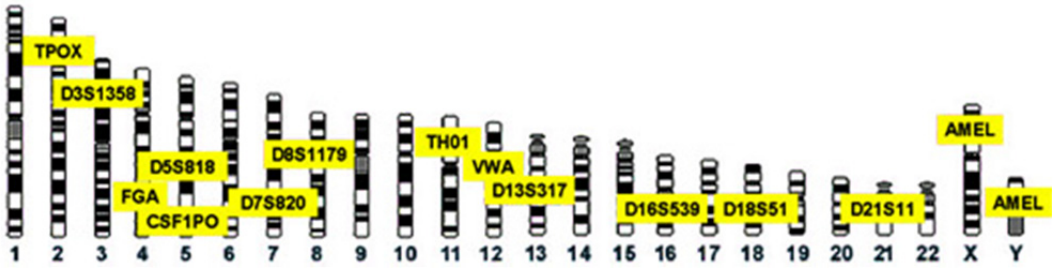


Figura 4: Localización de los STR empleados en el Combined DNA Index System (CODIS). A fin de poder indexar y buscar en su base de datos de ADN para la identificación humana, el Federal Bureau of Investigation (FBI) norteamericano utiliza 13 STRs ubicados por todo el genoma, más la amelogenina para determinar el sexo [7].

1.2.3. Single Nucleotide Polimorphism (SNP)

La variación más simple a nivel de secuencia de ADN es la sustitución de una única base (SNP), pudiendo ser una transición en caso de que el cambio sea de bases pirimidinas (C y T) o de bases purinas (A y G), o una transversión en caso de que el cambio sea entre ellas. También entran dentro de esta definición la inserción o la deleción de una única base, a pesar de que los mecanismos subyacentes a su generación así como su tratamiento analítico difieren de los de la sustitución de bases. La generación de sustituciones de bases surge fundamentalmente de la incorporación errónea de nucleótidos durante la replicación del ADN y de la mutagénesis causada por la modificación química de las bases o por el daño físico, y actualmente se estima que su tasa de mutación es de 1 cambio de base por cada 100 bases del genoma humano.

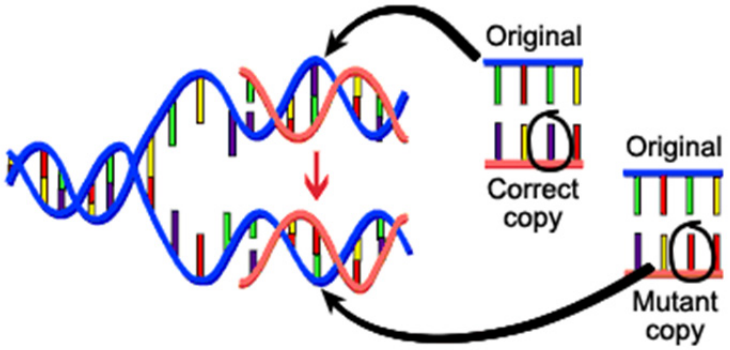


Figura 5: Mutación puntual de ADN. El efecto que produce el emparejamiento incorrecto de una base nucleotídica en el proceso de replicación de ADN es una alteración de la nueva secuencia, que podrá o no tener una repercusión funcional sobre el organismo. Fuente: http://www.wikipedia.org

1.3. Medición de la variabilidad

Los métodos de obtención de la información genética contenida en el ADN ha ido evolucionando lógicamente a lo largo de las últimas décadas, pero es precisamente en esta última década donde ha dado un salto no tanto cualitativo, ya que la secuenciación Sanger sigue siendo un estándar de calidad para la lectura de secuencias, como enormemente cuantitativo, ya que hemos pasado de investigar unas pocas variantes en unas pocas muestras a ser capaces de procesar miles de variantes en miles de muestras. Este enorme salto se ha producido a lo largo de una década, y aunque permanece en continua carrera hacia el aumento del rendimiento y el abaratamiento de costes, los dos tipos principales de técnicas que han revolucionado nuestra capacidad de investigación genómica han sido el genotipado de alto rendimiento y la ultrasecuenciación [8]. La primera puede llegar a permitir el análisis de miles de posiciones genómicas en miles de muestras, y la segunda nos está permitiendo analizar toda la variabilidad existente en las muestras de interés. La diferencia principal entre ambas radica en que el genotipado trabaja de una manera predefinida y estructurada, analizado solamente posiciones concretas del genoma elegidas con unos criterios definidos *a priori*, y la ultrasecuenciación lee toda la información de la secuencia de ADN analizada, detectando toda la variabilidad contenida en la región de interés o en todo el genoma si dicho interés así lo requiere.

1.3.1. Genotipado de alto rendimiento

El genotipo de una muestra se venía determinando a través de los métodos de secuenciación conocidos, lo cual aunque de alta fiabilidad resultaba tedioso cuando el número de variantes o de muestras a manejar dejaban de contarse con los dedos de las manos. La aparición de métodos más directos, que permitían trabajar con decenas de muestras a la vez y que podían determinar el contenido de decenas de posiciones concretas del genoma, permitió a la biología molecular saltar del análisis de rasgos o trastornos simples a estudiar fenotipos y enfermedades complejas, algo que antes resultaba prácticamente imposible por la baja eficiencia en coste de proceso.

Las tecnologías de genotipado han incluso evolucionado de manera exponencial con el tiempo. Desde su aparición hace apenas 10 años las capacidades tanto en manejo de número de muestras como en posiciones del genoma investigadas han ido en dramático aumento. Pronto se empezó a discriminar entre técnicas en función de su rendimiento, y lo que hace 5 años se consideraba alto rendimiento (muchos cientos o unos pocos miles de variantes genotipadas) pronto hubo que categorizarlo como medio rendimiento para distinguirlo de las distintas generaciones de chips de genotipado masivo que compañías como Affymetrix o Illumina han ido posicionando como estándares de referencia en el mercado. De hecho, en caso de existir la posibilidad, es una práctica habitual la inclusión de más de una plataforma en un proyecto de genotipado, ya que las tecnologías de muy alto rendimiento cubren un espectro amplísimo de posibilidades pero no son tan flexibles y adaptables como las de bajo y medio rendimiento. Éstas pueden usarse para cubrir los huecos que las anteriores dejan en las zonas de mayor interés para un proyecto concreto.

En la actualidad existen multitud de tecnologías de genotipado de alto rendimiento, aunque todas ellas tratan de determinar el contenido alélico de posiciones o regiones muy concretas prefijadas dentro del genoma a investigar. Aunque no en su totalidad, la mayor parte del trabajo de estas tecnologías, y su gran baza de cara a aumentar su rendimiento, se basa en la discriminación de alelos binarios, de tal manera que la plataforma de genotipado tendrá que elegir entre una de las dos formas conocidas del alelo investigado. Como tecnologías de análisis, difieren tanto en la reacción de discriminación alélica como en el método de detección de los productos de dicha reacción, incluso en el formato en el que tiene lugar la reacción, y su aplicación viene determinada no sólo por

el número de marcadores y muestras a analizar, sino también por el tipo de muestra, la flexibilidad del diseño o la posibilidad de automatización. Ejemplos de plataformas de genotipado son LightCycler, Taqman o Kaspar, que miden la transferencia de energía entre fluorocromos (FRET) sobre una fase homogénea, y SNaPshot, SNPlex, Genplex o las distintas opciones de Affymetrix e Illumina, todas ellas midiendo las intensidades de fluorescencia en fase sólida. A pesar de no competir en rendimiento con las grandes opciones de Affymetrix o Illumina, la tecnología MassARRAY de la compañía Sequenom, que mide los tiempos de vuelo de fragmentos ionizados en fase sólida mediante espectrometría de masas, resulta de gran utilidad dada su gran versatilidad y precisión, así como por su no predisposición a la detección de alelos binarios.

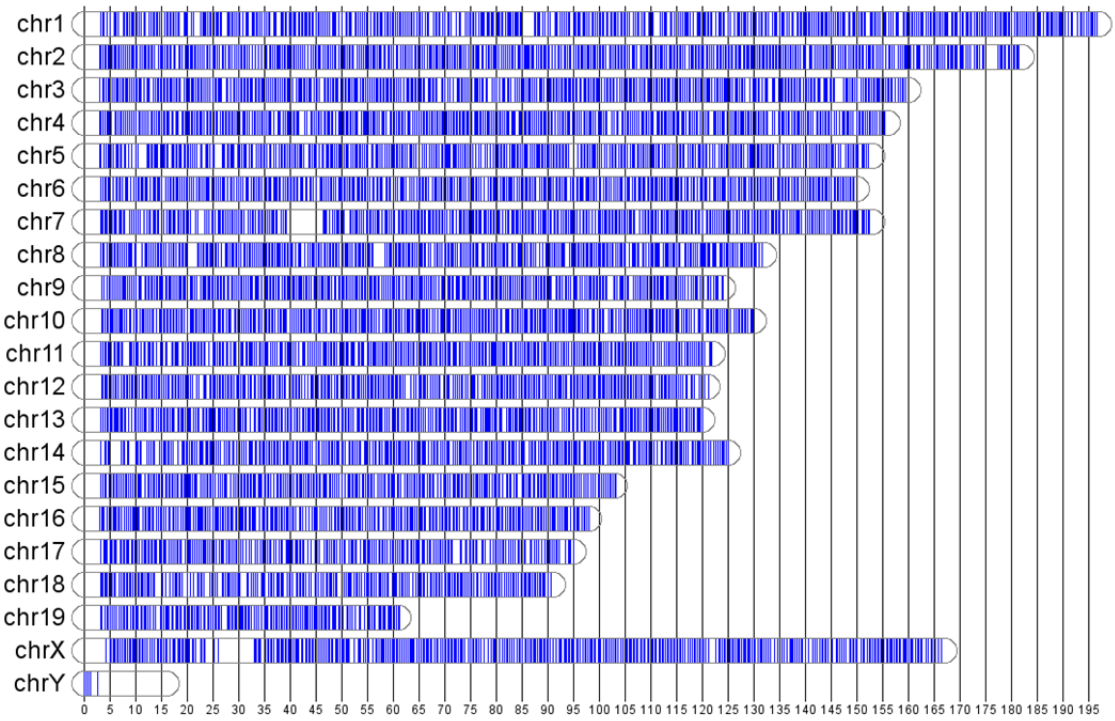


Figura 6: Distribución de los marcadores contenidos en el Mouse Universal Genotyping Array (MUGA). El MUGA es un chip diseñado para el genotipado de ratones, creado a partir de 7851 SNPs distribuidos a lo largo del genoma del ratón y espaciados entre sí por 325 kilobases, que utiliza la plataforma de genotipado Illumina Infinium [9].

1.3.2. Ultrasecuenciación

Desde finales de los 80, la secuenciación hoy llamada tradicional o secuenciación Sanger ha ido mejorando tanto en la calidad como en la cantidad de la secuencia obtenida a partir de una muestra, pudiéndose obtener en cada reacción varios cientos de nucleótidos. Dado que la mayor limitación para mejorar esta técnica venía dada por la dificultad creciente en la distinción de fragmentos de ADN cuyo tamaño difiere en un solo nucleótido, a principios de este siglo se empezó a optar por paralelizar masivamente todo proceso y mecanizarlo en la medida de lo posible en estaciones de ultrasecuenciación. A pesar de que la preparación de las muestras es más compleja y costosa, y que el análisis de los resultados requiere una importante dedicación bioinformática empleando computación de alto rendimiento, la aparición de esta tecnología posibilitó la enorme ampliación de las regiones de estudio en un genoma dentro de una misma reacción, así como la drástica reducción de los costes de secuenciación. En diez años se han podido reducir tanto los tiempos y los costes que si para obtener el genoma humano a través de secuenciación capilar se tardarían 3 años y costaría 300.000 dólares, ahora tardamos 1 día escaso por cerca de 1.000 dólares.

Al igual que en genotipado, existen multitud de técnicas de ultrasecuenciación, aunque todas ellas trabajan esencialmente igual: se divide el ADN en millones de pequeñas secuencias, se amplifican, y son éstas últimas las que se intentan leer en la plataforma para finalmente contextualizarlas mediante técnicas bioinformáticas de alineamiento y ensamblado. No se trata de un método directo como la secuenciación tradicional, sino de un ejercicio masivo de lectura en el que un alto porcentaje de esfuerzo se pierde (la eficiencia puede rondar el 40%, dependiendo de la técnica), pero la potencia de esta metodología es tan grande que aun así resulta muy útil y económicamente muy ventajosa. Algunos ejemplos de ultrasecuenciadores son el pionero 454 de Life Sciences (adquirida por Roche), el Genome Analyzer de Solexa (adquirida por Illumina), el SOLiD de Applied Biosystems (ahora parte de Life Technologies) o el más reciente Ion PGM Sequencer de Ion Torrent (adquirida por Life Technologies).

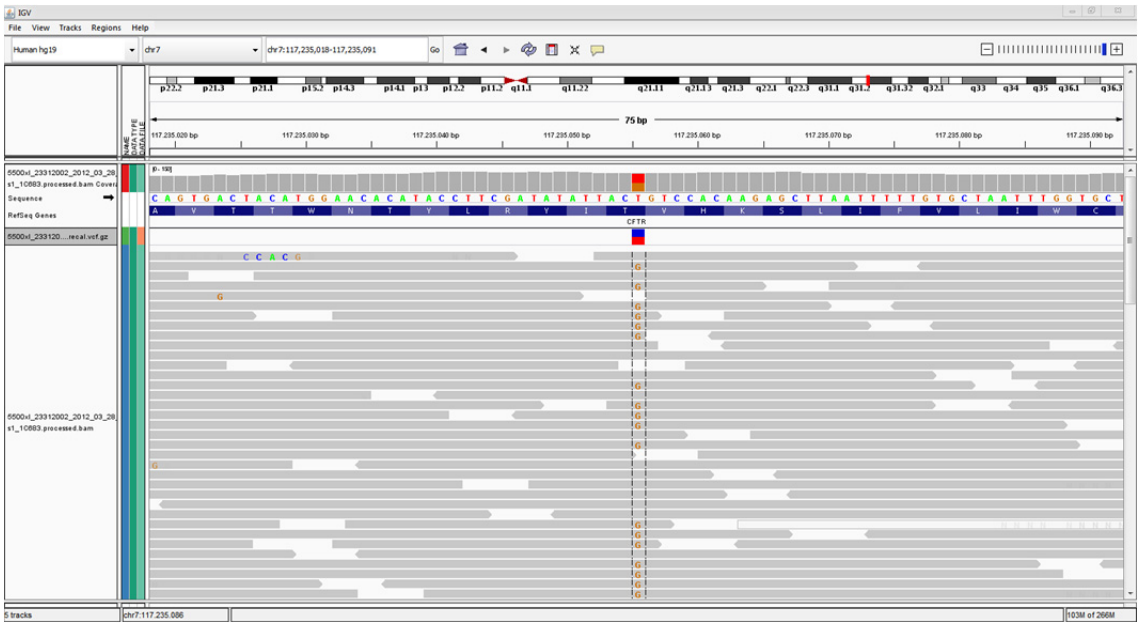


Figura 7: Visualización de resultados de ultrasecuenciación en el Integrative Genomics Viewer (IGV). Los millones de lecturas cortas provenientes del ultrasecuenciador se alinean y se presentan junto al genoma de referencia para mostrar la secuencia resultante, y los nucleótidos no coincidentes con la referencia se presentan con distintos colores junto a su evaluación como variantes reales de la muestra secuenciada [10].

Una parte crítica de la ultrasecuenciación es el análisis de los datos crudos ya que, así como la salida de una plataforma de genotipado sí informa directamente de los alelos obtenidos, la salida directa de un ultrasecuenciador son los millones de lecturas obtenidos, y no la secuencia final. Para obtener esta última se requiere de grandes recursos, tanto de almacenamiento como de computación, cuyos costes a día de hoy superan a los que se dedican estrictamente a la secuenciación en el laboratorio. Debido a que la capacidad de generación de secuencia crece ya a un ritmo superior al de la capacidad de procesamiento, la manera en la que se está tratando de lidiar con estas necesidades es mediante la generación de formatos de datos óptimos y algoritmos paralelos específicos que permitan hacer uso de la computación de altas prestaciones.

2. Poblaciones humanas

La genética poblacional llegó a madurar hacia 1930, principalmente con los trabajos de Ronald Fisher y Sewall Wright, incluso antes de conocerse el ADN como la unidad última de herencia. Ambos unieron los anteriormente incompatibles conceptos de selección natural de caracteres fenotípicos con los de la herencia mendeliana demostrando que alelos discretos podían estar fundamentados en rasgos continuos. Con el tiempo, la descripción de diversidad genotípica a nivel molecular y el hecho de que la selección no sirviera como el único proceso capaz de explicar los niveles de polimorfismos exigieron desarrollos sustanciales en la teoría de la genética poblacional. Hoy en día sabemos que los cambios de frecuencias alélicas a lo largo del tiempo son las pistas que nos permiten investigar los procesos evolutivos. Entendiendo los mecanismos por los que las fuerzas de la evolución actúan sobre estas frecuencias se pueden generar modelos matemáticos que se aproximen a la realidad, necesarios para entender la sutil interconexión entre dichas fuerzas así como para permitirnos inferir procesos pasados a partir de la diversidad actual.

2.1. Equilibrio Hardy-Weinberg

El mayor hito de la genética poblacional fue el de explicar cómo las frecuencias alélicas en una generación podían ser usadas para calcular las proporciones genotípicas en la siguiente generación en una población cruzándose al azar. En organismos diploides como los humanos, dos alelos *A* y *a* en el mismo locus y con frecuencia *p* y *q* respectivamente, pueden ser combinados para generar 3 genotipos: AA, Aa y aa. El principio de Hardy-Weinberg nos dice que en una población ideal deberíamos poder predecir las proporciones de los genotipos en la siguiente generación combinando los gametos simplemente al azar, de tal manera que la proporción de cada genotipo en la siguiente generación sería p^2 para *AA*, *2pq* para *Aa*, y q^2 para *aa*.

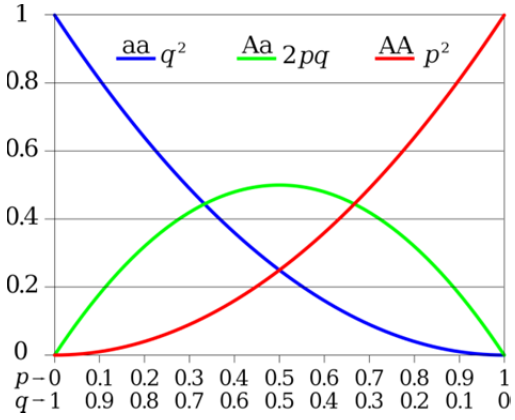


Figura 8: Principio de Hardy-Weinberg para dos alelos. El eje horizontal muestra las dos frecuencias alélicas *p* y *q*, el eje vertical muestra la frecuencia de los genotipos, y los tres posibles genotipos (AA, Aa y aa) se representan con las tres tendencias coloreadas. Fuente: http://www.wikipedia.org

Si las medidas esperables coinciden con las observadas se dice que la población se encuentra en equilibrio Hardy-Weinberg, y se considera que no hay signos de evolución si entendemos ésta como agente modificador de frecuencias alélicas, ya que para que se cumpla dicho equilibrio la población ideal debe cumplir ciertos requisitos, entre ellos el de que no esté sujeta a factores tales como selección, mutación o migración. Si por el contrario, las proporciones genotípicas no se encuentran en equilibrio Hardy-Weinberg se puede concluir razonablemente que la población sí está evolucionando, y que uno de los factores anteriores (o una combinación de ellos) está operando sobre dicha población.

2.2. Linkage disequilibrium - LD

La recombinación meiótica es una consecuencia de la reproducción sexual que favorece la habilidad de las poblaciones a adaptarse al medio a través de la combinación de alelos favorables en distintos *loci*. Esto genera nuevas combinaciones de alelos en la misma molécula de ADN, llamados haplotipos, y de esta manera se aumenta su diversidad. Mientras los alelos en distintos cromosomas se segregan al azar durante la meiosis, los alelos cercanos en un cromosoma no sufren recombinación de manera tan frecuente. Por ello, y a nivel poblacional, la recombinación puede ser estudiada investigando si la asociación entre alelos en distintos loci ocurre de manera más o menos frecuente que lo esperado al azar. Esta asociación no aleatoria se conoce como *desequilibrio de ligamiento* (LD).

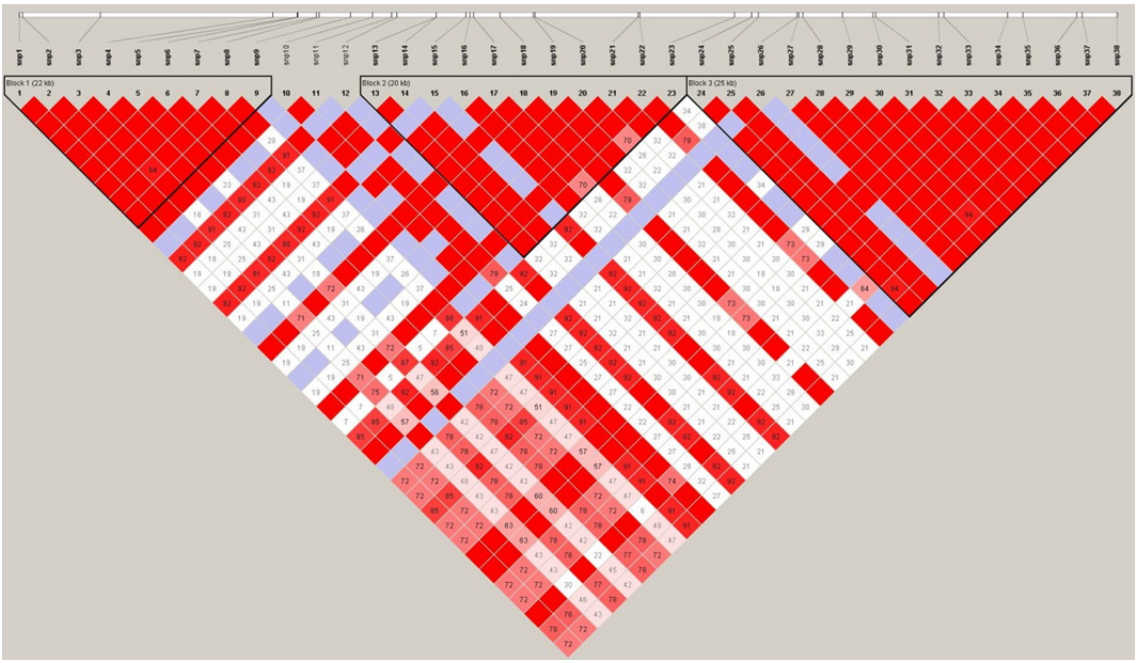


Figura 9: LD del gen GCH1 en hispanoamericanos [11]. Representación gráfica de los resultados de Haploview [12], donde la evaluación de la diferencia en frecuencia de los marcadores se hace dos a dos, lo que genera una representación en pirámide en la que los bloques de LD se destacan en rojo.

La manera más antigua y sencilla de medir LD es a través de la diferencia entre la frecuencia observada de un haplotipo de dos locus y la frecuencia esperada si los alelos segregaran al azar: si esta diferencia D es significativamente diferente de cero (evaluada con un test exacto de Fisher) entonces se dice que existe LD. Otras medidas, como $|D'|$ o r^2, se usan también

para cuantificar el LD ya que la dependencia que *D* tiene de las frecuencias alélicas hace que la comparativa de distintos valores de *D* sea de escasa utilidad. Tanto $|D'|$ o r^2 son valores normalizados y útiles en comparativas tales como estudios de asociación, en los que suele preferirse el valor de r^2 por más sensible para pequeños tamaños muestrales.

2.3. Mezcla génica poblacional

Las poblaciones no son entidades discretas, ya que intercambian individuos entre ellas, y esta mezcla genera poblaciones híbridas. Pero aunque las poblaciones vecinas frecuentemente realizan tales intercambios en un proceso continuo de migración bidireccional, el término mezcla génica se suele reservar a la formación de poblaciones híbridas por la mezcla de las poblaciones ancestrales que previamente habían estado aisladas unas de otras, y puede considerarse que dicha mezcla tuvo un comienzo en un evento temporal concreto cuando las poblaciones entraron en contacto por primera vez.

Cuando examinamos poblaciones actuales no detectamos solamente las proporciones de mezcla establecidas cuando las poblaciones originales se encontraron, sino la suma del flujo génico acumulado desde tal fecha hasta nuestros días, ya que la huella genética de tal mezcla hubo de ser modificada por la deriva, selección y mutación. Así, las consecuencias de la mezcla génica y el flujo génico pueden ser difíciles de ser distinguidos entre sí. Lógicamente, cuanto más distintas fuesen las poblaciones ancestrales entre sí más fácil será detectar y cuantificar la mezcla génica.

2.4. Ancestry informative markers – AIMs

El deseo de hacer predicciones acerca del origen poblacional ha sido estimulado no sólo por científicos forenses, sino también por aquellos interesados en la mezcla poblacional con fines epidemiológicos o para mapeado de genes asociados a una enfermedad usando LD. Los alelos presentes en una población pero no en otra se han llamado *marcadores privados* o *alelos población-específicos*, pero el interés radica en evaluar un mismo marcador en distintas poblaciones para la estimación de la *afiliación étnica* (o de la *ascendencia biológica*

como se prefirió denominar por las fuertes implicaciones culturales del término *etnia*) se determinaron como útiles aquellos marcadores autosomales binarios o multialélicos que presentasen grandes diferencias en frecuencia (>50%) entre una y otra población. Estos marcadores se conocen como *marcadores indicativos de ascendencia* (AIMs), y funcionan muy bien como indicador para la estimación de la proporción de mezcla entre poblaciones [13, 14].

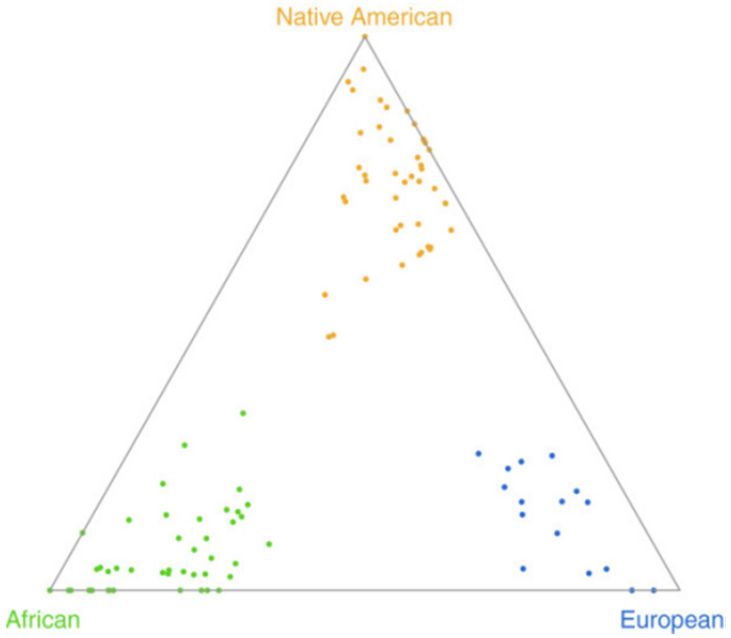


Figura 10: AIMs usados para inferir la ascendencia en poblaciones africanas, europeas o nativo americanas [15]. Los puntos representan AIMs individuales, y su ubicación representa su frecuencia en cada población ancestral. Disponiendo de suficientes marcadores genéticos, los genotipos individuales pueden ser comparados con las poblaciones ancestrales y asignárseles una probabilidad razonable de ascendencia a cada uno.

2.5. Deriva génica

Cada generación representa una muestra finita de la generación anterior, y por tanto la variación de las frecuencias alélicas entre ellas será debida exclusivamente a los procesos estocásticos de muestreo, lo que se conoce como *deriva génica aleatoria* [16], y es dependiente del tamaño de la población muestreada. Pero las poblaciones humanas no son ideales sino que se solapan, no tienen un tamaño constante y desde luego no están sometidas a un cruzamiento precisamente aleatorio. Por ello el concepto de *tamaño de población efectivo* nos permite comparar la cantidad de deriva génica que experimentan

distintas poblaciones, ya que representa el tamaño de una población ideal (no solapante, de tamaño constante y con cruzamientos aleatorios) con la misma cantidad de deriva génica que la población que se estudie, siendo casi siempre substancialmente menor que el tamaño real de la población. El propio tamaño efectivo sirve por tanto como medida de la deriva: a menor tamaño de población efectivo, mayor será la deriva génica.

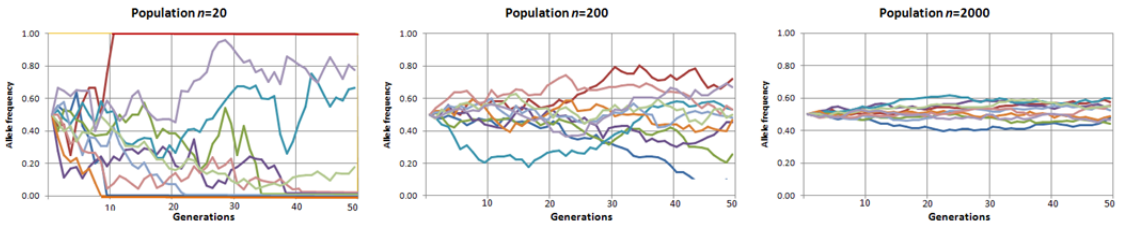


Figura 11: Efecto del tamaño poblacional en la deriva génica aleatoria. Comparación gráfica de los resultados de 10 simulaciones distintas del cambio aleatorio en la frecuencia de un alelo en poblaciones de 20, 200 y 2000 individuos a lo largo de 50 generaciones. Las poblaciones más pequeñas tienden a eliminar o a fijar el alelo simulado, mientras que las más grandes tienden a mantener estable. Fuente: http://www.wikipedia.org

El efecto que la deriva génica causa en las poblaciones es la reducción de la diversidad, ya que por los efectos aleatorios del muestreo la frecuencia de un alelo puede llegar a anularse (eliminación) o a extenderse por toda la población (fijación). Este efecto se acentúa con tamaños pequeños de población, donde un alelo emergente puede llegar a desaparecer o fijarse en unas pocas decenas de generaciones.

2.6. Selección

La selección natural, definida por Darwin y detallada por Fisher, es la reproducción diferencial de genotipos en generaciones sucesivas. La variación genotípica produce individuos con variables capacidades de supervivencia y reproducción en distintos ambientes, y la selección puede ocurrir en cualquier punto desde la formación del genotipo en la fertilización hasta la generación de progenie viable. La suma de los factores de selección (supervivencia, selección sexual, fertilidad y fecundidad) conforma la habilidad de un genotipo individual a sobrevivir y reproducirse, y es particularmente dependiente del ambiente.

Figura 12: Forma común y forma *carbonaria* de la mariposa del abedul. Ejemplo de melanismo industrial en el que la mutación de la forma común de este lepidóptero confirió a la forma melánica una ventaja mimética respecto a la forma clara con la llegada de la Revolución Industrial, especialmente en Inglaterra, ya que la atmósfera se llenó cada vez más de polvo de carbón que acabó por oscurecer las cortezas de los árboles en las regiones industriales. Fuente: http://www.wikipedia.org

El factor importante es la capacidad relativa de un genotipo comparado con otros genotipos compitiendo por los mismos recursos. Dicha capacidad relativa se mide a través del coeficiente de selección, que compara un genotipo con el más capaz en la población. Un coeficiente de selección de 0.1 representa una disminución en capacidad del 10% respecto al genotipo más viable. Las mutaciones que reducen la viabilidad de un genotipo dan lugar a una *selección negativa* o purificante, mientras que las mutaciones que la aumentan generan una *selección positiva*.

3. Grandes esfuerzos de recolección

El gran avance de la genómica de comienzos de siglo ha venido determinado por una serie de gigantescos proyectos con la implicación de consorcios públicos internacionales, empresas y centros de investigación, que no sólo han realizado esfuerzos titánicos por acometer un esfuerzo pionero en su momento, sino que su objetivo último era el de poner sus resultados a disposición de la comunidad científica. Precisamente, es esto último lo que ha permitido que hoy en día contemos con repositorios a nivel mundial de muy alta calidad, tanto por la información que contienen como por las herramientas y soporte ofrecidos para su consulta. Todos ellos coinciden en haber supuesto un desafío conceptual a nivel biológico, pero los retos que supusieron el análisis, almacenamiento y publicación de los resultados fueron también considerables. Sin el apoyo y la dedicación de una enorme cantidad de recursos bioinformáticos, tanto físicos como humanos, estos esfuerzos no hubieran resultado lo fructíferos que han demostrado ser.

3.1. Human Genome Project (HGP)

En 1990 comenzó formalmente un esfuerzo internacional, coordinado por el Departament of Energy (DOE) y el National Institute of Health (NIH) estadounidenses, y denominado Proyecto Genoma Humano. Originalmente se había planificado para 15 años, pero los avances tecnológicos durante esa etapa aceleraron la fecha de finalización al año 2003 [17]. Sus metas iniciales eran en su concepción muy ambiciosas, y para alcanzarlas también fue necesario estudiar el contenido genético de diversos organismos aparte del humano, desde la bacteria intestinal más conocida del ser humano (*Escherichia coli*) a la mosca de la fruta o el ratón de laboratorio: determinar la secuencia del ADN humano, identificar todos los genes contenidos en él, almacenar esta información en bases de datos, crear y mejorar las herramientas existentes para el análisis de datos, transferir la tecnología relacionada al sector privado, y hacerse cargo de los condicionantes éticos, legales y sociales que se generasen.

La presentación en sociedad de la secuencia del genoma humano se hizo en el año 2000, coincidiendo con el anuncio del esfuerzo privado paralelo que había sostenido Celera Genomics, y los resultados de esta versión funcional preliminar fueron publicados en el año 2001 tanto en *Nature* [18] como en *Science [19]*. Pero no fue hasta 2003 cuando se obtuvo la secuencia esencialmente completa del genoma humano [20], por encima del 92% de su totalidad, momento en el que el HGP se declaró finalizado y se publicaron las referencias que hoy se conocen bajo los códigos *NCBI36* y *hg18*, dependiendo si el organismo proveedor es el Centro National Center for Biotechnology Information (NCBI) o la University of California Santa Cruz (UCSC). Los esfuerzos de secuenciación siguieron en marcha, y en mayo 2006 se completó finalmente el secuenciado y ensamblado de todos los cromosomas. Fuera ya de este proyecto, bajo el auspicio del Genome Reference Consortium (GRC) se hizo pública en el año 2009 la versión más actualizada de la que se dispone, referencia conocida como *GRCh37* o *hg19*, que aporta una cobertura cercana al 99.8% del genoma humano completo.

Figura 13: Primeras publicaciones del genoma humano en revistas científicas. Los primeros análisis de la primera versión funcional de la secuencia del genoma humano se publicaron en sendas ediciones especiales de las revistas Nature y Science en febrero del 2001. Los artículos de Nature incluían se centraron en el esfuerzo público del Human Genome Project, mientras que los de Science se centraron en los resultados de la empresa Celera Genomics. Fuente: http://www.nature.com y http://www.sciencemag.org

3.2. Perlegen Sciences, Inc.

Affymetrix ha sido una empresa de referencia en el campo de los microarrays prácticamente desde su fundación en 1992, y a fin de poder centrarse la generación de datos de forma masiva creó en 2000 la filial Perlegen Sciences, Inc. [21]. El objetivo principal era el de hacerse cargo de todos los aspectos que la creación y recolección de estos datos masivos requerirían para la caracterización de variabilidad poblacional, tanto de marcadores genómicos como de expresión, dentro del proceso del descubrimiento de fármacos. Su misión era la de identificar patrones de variación genética entre individuos, para incorporar ese conocimiento en la atención al paciente y descubrir el fundamento genético de las enfermedades.

El trabajo de esta compañía fue usado como punto de partida para el proyecto HapMap, en el que colaboró muy activamente, y aunque la información hoy contenida en el proyecto supera con creces el ámbito inicial de Perlegen Sciences, Inc. los datos que hicieron públicos para 180 muestras representando 3 poblaciones mundiales (norteamericanos de origen africano, norteamericanos de origen europeo, y chinos Han) supusieron el primer repositorio masivo mundial de variabilidad humana.

3.3. Centre d'Etude du Polymorphisme Humain (CEPH)

El profesor Jean Dausset fundó en 1984 el CEPH como un laboratorio capaz de facilitar la distribución del ADN de 40 familias de referencia con el objetivo de coordinar la primera colaboración internacional para la creación de un mapa genómico. En 1993 se convierte en la Fundación Jean Dausset-CEPH, un centro de investigación sin ánimo de lucro en el que surgirán nuevas líneas de investigación, de entre las que destaca en 2002 la creación, en colaboración con el Human Genome Diversity Project (HGDP), de un recurso biológico creado a base de 1063 líneas celulares provenientes de 1050 individuos de 52 poblaciones mundiales distintas [22]. Dicho panel sigue siendo aún ahora una referencia biológica usada mundialmente para la investigación genética de las poblaciones humanas, e incluso ha sido refinado en distintos niveles de astringencia atendiendo a las relaciones familiares entre las muestras [23], lo cual facilita aún más su tratamiento estadístico.

La disposición de un panel de líneas celulares de referencia mundial, unido a la rápida evolución de la técnica de genotipado, llevó en 2007 a dos grupos de investigación distintos a realizar el mismo esfuerzo de procesar dicho panel con el chip de SNPs más denso disponible hasta la fecha, que era el de Illumina BeadStation [24, 25]. Ambos esfuerzos, a pesar de ser en gran medida solapantes, fueron muy beneficiosos para la comunidad científica ya que gracias al trabajo de las universidades de Stanford y de Michigan se dispuso de un repositorio masivo de variabilidad partiendo de dos fuentes altamente replicables y documentadas: el panel de muestras HGDP-CEPH y el chip de Illumina BeadStation.

3.4. International HapMap Project

El proyecto HapMap se formalizó en el año 2002 con el fin de desarrollar un mapa de haplotipos del genoma humano, ya que el conocimiento de dicho mapa permitiría encontrar genes y variaciones genéticas que afectasen a la salud [26-28]. La entonces reciente disponibilidad del genoma humano en su totalidad generó la idea de este proyecto, ya que una vez conocida su secuencia el paso lógico siguiente era el de estudiar su variabilidad, y en particular la capacidad de transmisión de las variantes en bloques haplotípicos. A través de la determinación de SNPs específicos que identificasen dichos bloques (tag SNPs) se pretende reducir la complejidad del estudio haplotípico al reducir el número de variantes a estudiar.

La secuencia de ADN de cada dos individuos es un 99.5% idéntica, pero sus diferencias pueden afectar enormemente a su riesgo individual de desarrollar una enfermedad. Un gran valor del proyecto HapMap es la reducción del número de SNPs requeridos para examinar todo el genoma buscando la asociación con su fenotipo de los 10 millones de SNPs que existen aproximadamente en cada persona medio millón de tag SNPs. Esto permite un examen del genoma mucho más eficiente, ya que el esfuerzo se optimiza al no tener que estudiar más SNPs de los necesarios sin comprometer con ello el tamaño del genoma a analizar.

Además de su uso en estudios de asociación, los resultados del proyecto son un recurso muy útil para estudiar los factores genéticos que contribuyen a la variación como respuesta a factores ambientales, susceptibilidad a infecciones, o efectividad a fármacos y vacunas. Todos estos estudios están basados en la asunción de que existen mayores frecuencias de los componentes genéticos contribuyentes en un grupo con cierta enfermedad o repuesta a un fármaco particular, que en un grupo similar de gente sin tal rasgo, lo que también se conoce como hipótesis de *variante común / enfermedad común*. Usando solamente los tag SNPs, los investigadores pueden encontrar regiones cromosómicas con diferentes distribuciones haplotípicas entre dos grupos de personas (casos y controles). Cada región se estudia entonces en detalle para descubrir qué variantes y qué genes contribuyen al rasgo estudiado, lo que aumentará la efectividad de cualquier intervención posterior.

Durante la Fase I del proyecto se genotipó un SNP común por cada 5000 bases del genoma, en 10 centros distintos y usando 5 diferentes, resultando algo más de 1 millón de SNPs genotipados en cerca de 200 muestras provenientes de 4 poblaciones: residentes de Utah con antepasados del norte y oeste de Europa, chinos Han de Pekín, japoneses de Tokio y yorubas de Ibadán. La calidad del genotipado era muy importante, analizándose usando muestras duplicadas y realizando controles de calidad periódicos en los centros. Pero para obtener suficientes SNPs para crear el mapa haplotípico completo, el consorcio tuvo que financiar un gran proyecto de resecuenciado para descubrir millones de SNPs adicionales como parte de la Fase II, así como añadir los resultados cedidos por las compañías Perlegen Sciences (más de 2 millones de SNPs) y Affymetrix (medio millón de SNPs), lo que en 2006 representaría más del triple de la variación conocida hasta el momento.

Acrónimo	Población	Grupo poblacional
ASW	African ancestry in Southwest USA	Africa
CEU	Utah residents with European ancestry, CEPH	Europe
CHB	Han Chinese in Beijing, China	East Asia
CHD	Chinese in Metropolitan Denver, Colorado	East Asia
GIH	Gujarati Indians in Houston, Texas	Central-South Asia
JPT	Japanese in Tokyo, Japan	East Asia
LWK	Luhya in Webuye, Kenya	Africa
MEX	Mexican ancestry in Los Angeles, California	America
MKK	Maasai in Kinyawa, Kenya	Africa
TSI	Toscans in Italy	Europe
YRI	Yoruba in Ibadan, Nigeria	Africa

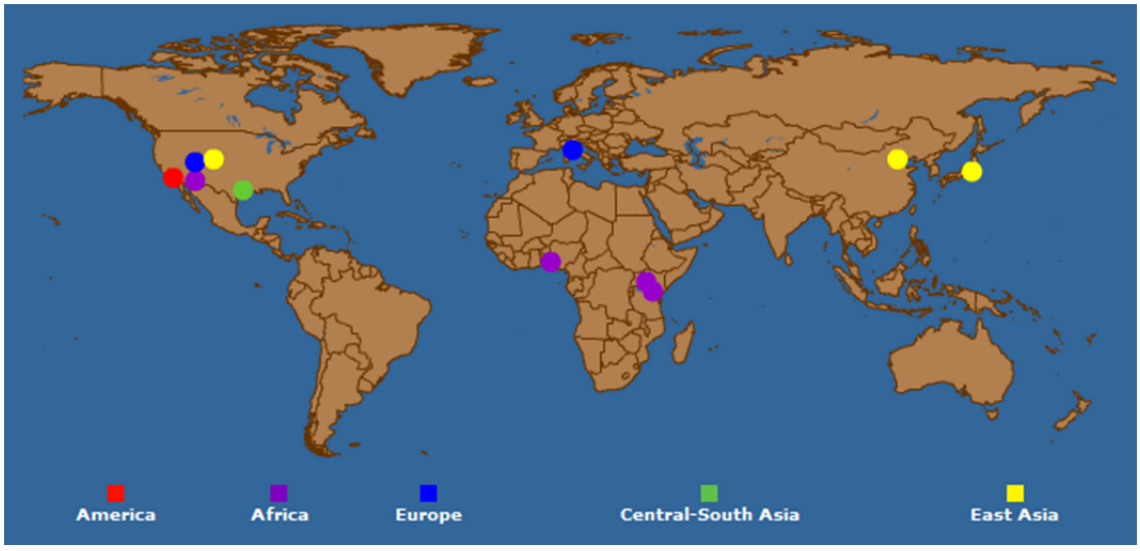


Figura 14: Poblaciones incluidas en el International HapMap Project. Listado de las 11 poblaciones incluidas en el proyecto, junto a sus acrónimos por los que son habitualmente denominadas y al grupo poblacional al que pertenecen, y su ubicación geográfica según su origen de muestreo. Fuente: http://spsmart.cesga.es

3.5. 1000 Genomes

Con la llegada de las técnicas de ultrasecuenciación y su drástica reducción de los costes de secuenciación surge en 2008 la iniciativa de crear una nueva y más completa referencia para la variabilidad humana usando como fuente la secuenciación de genomas humanos completos [29, 30]. Su principal objetivo era el de encontrar la mayor cantidad de variantes genéticas con una frecuencia en poblaciones inferior al 1%, lo que se conoce con el nombre de variantes raras, ya que por un lado podría complementar ampliamente la información conocida hasta la fecha, y por otro lado posibilitaría estudiar la llamada hipótesis de *variante rara / enfermedad común*. Esta hipótesis es una alternativa a la anterior *variante común / enfermedad común* que defiende que las variantes de baja frecuencia en la población actúan como principales factores genéticos de susceptibilidad las enfermedades comunes.

Los proyectos piloto que se realizaron para evaluar, entre otras cosas, las posibles variaciones entre distintos laboratorios y especialmente entre las distintas plataformas de ultrasecuenciación (no hay que olvidar que la tecnología era aún muy reciente cuando se concibió el proyecto) posibilitaron ya en 2010 disponer de un catálogo de variabilidad humana basada en 180 genomas completos [31]. Apenas un año más tarde el catálogo contaba ya más de 600 muestras, y actualmente cuenta con casi 1200 muestras de 13 poblaciones distintas, aunque el objetivo final es el de llegar a secuenciar aproximadamente 2500 genomas completos.

Debido a que las variantes genéticas en una región están generalmente asociadas con muchas otras de la misma región por patrones de desequilibrio de ligamiento, y aunque una variante asociada a una enfermedad puede ser un marcador para dicha región, los patrones de desequilibrio de ligamiento no permiten determinar qué variante o qué elemento genético en particular aporta la contribución causal al riesgo de la enfermedad. La información de este proyecto permite combinarse con estudios de asociación previos, aportando millones de variantes adicionales más allá de los genotipados directamente, lo que ayuda a localizar de manera más precisa las regiones asociadas a la enfermedad. Y una vez aisladas las regiones de interés, ya se dispone de información poblacional para las variantes en ellas contenidas, lo que permite no tener que realizar nuevos y grandes estudios de genotipado, al tiempo que identifica nuevas variantes sospechosas de ser incluidas en el estudio original.

4. Acerca de la bioinformática

Hemos sufrido en las últimas décadas una explosión masiva de la cantidad de datos biológicos disponibles debida en gran parte a los enormes avances en los campos de la biología molecular y de la genómica. La bioinformática, entendiendo el término en su sentido más amplio, no es más que la aplicación de algoritmos y de tecnologías de computación al manejo y análisis de dichos datos biológicos. Es un campo de investigación con clara vocación interdisciplinar, a caballo entre las ciencias biológicas y las computacionales, cuyo fin último es desvelar la información biológica contenida en las ingentes cantidades de datos generados para mejorar nuestro entendimiento de los fundamentos biológicos de los organismos. Hoy en día no se pueden entender la evolución de la biología molecular sin el apoyo constante de la bioinformática, desde el procesado de los datos a su almacenamiento, pasando por la extracción de la información en ellos contenida.

4.1. La bioinformática en su contexto

Tradicionalmente, la investigación en biología molecular se llevaba a cabo enteramente en el banco experimental del laboratorio, pero el enorme incremento de escala en la producción de datos ha necesitado incorporar computadores en este proceso de investigación. De hecho, el desarrollo de los métodos de análisis de secuencias ha sido fruto de las contribuciones de muchos investigadores provenientes de distintos campos científicos, desde técnicas de secuenciado y análisis de microarrays hasta la supercomputación o internet. La metodología que hoy en día es la base de consulta e investigación diaria en los laboratorios de genética podría datar sus comienzos en los años 80 con el esfuerzo de la comunidad científica por el almacenaje, indexación y consulta de las secuencias disponibles hasta la fecha, tanto de proteínas con la Protein Information Resource que acabó siendo la PIR-International Protein Sequence Database, como de ADN con la aparición de GenBank [32]. Precisamente, el término "bioinformática" empezó a ser usado en esta época para englobar distintas aplicaciones de la computación en el campo de las ciencias biológicas [33].

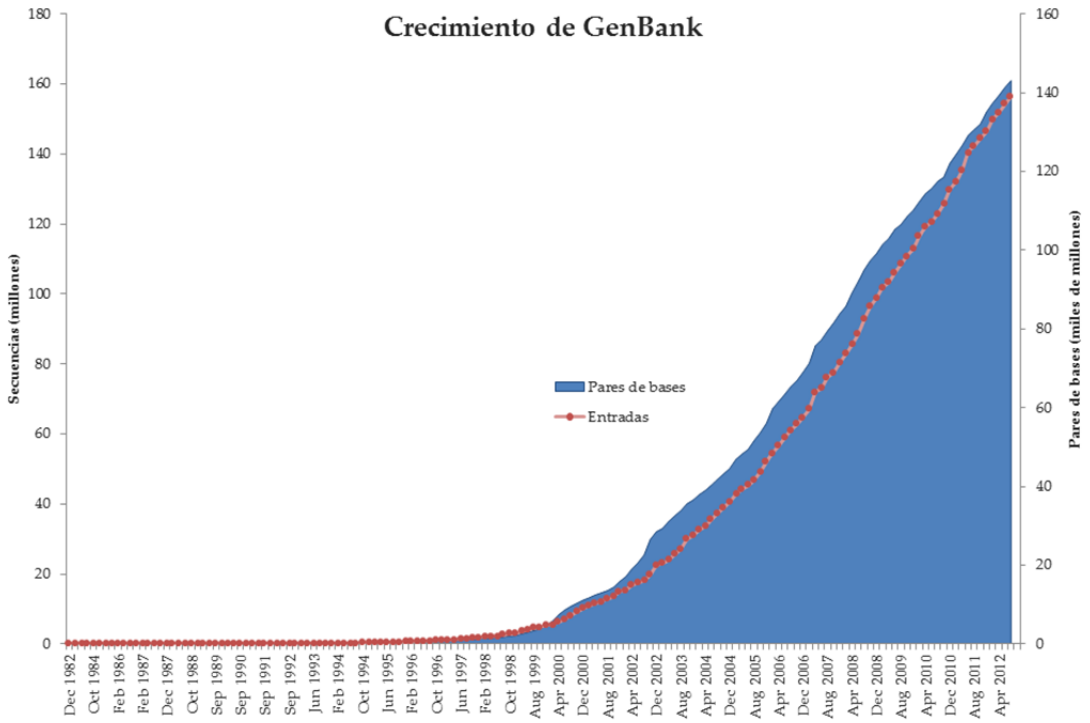


Figura 15: Evolución en el tiempo de los contenidos de GenBank [34]. El aumento de cantidad de información biológica disponible suele ilustrarse visualmente mediante la presentación de la cantidad de pares de bases y número de secuencias contenidas en GenBank.

La generación de secuencias, así como el subsiguiente almacenamiento, interpretación y análisis son tareas hoy en día enteramente dependientes de computadores. Pero el primer reto de la comunidad bioinformática actual es el propio almacenaje de esta cantidad masiva de datos, ya que debe hacerse de manera inteligente y eficiente. En ella recae la responsabilidad de proveer un acceso fácil y fiable a estos datos. Porque los datos en sí mismos carecen de significado antes de ser analizados, y el volumen de datos en el presente hace imposible la interpretación manual. Es por ello que herramientas computacionales muy incisivas han de ser desarrolladas para permitir la extracción de información biológica útil, y existen tres procesos biológicos centrales alrededor de los cuales se han de crear herramientas bioinformáticas:

- La secuencia de ADN determina la secuencia de la proteína.
- La secuencia de la proteína determina su estructura.
- La estructura de la proteína determina su función.

La integración de la información aprendida acerca de estos procesos biológicos fundamentales nos debería permitir alcanzar el fin último tan deseable de comprender completamente los organismos biológicos [35].

4.2. Aplicaciones de la bioinformática

La ciencia de la bioinformática tiene usos muy dispares a la par que beneficiosos en el mundo actual, incluyendo la medicina molecular, genómica microbiótica, agricultura y ganadería, y estudios comparativos [36]. Es más, la evolución de estas y otras ramas afines del saber es inimaginable hoy en día sin la aportación que la bioinformática supone, ya que en última instancia se encarga de tender el puente entre el vasto conocimiento anterior y la ingente cantidad de datos y recursos disponibles hoy en día.

4.2.1. Medicina molecular

El genoma humano tiene una gran importancia en los campos de la investigación biomédica y de la genética clínica ya que las enfermedades tienen, en mayor o menor medida, un componente genético. Éste podrá ser heredado (como en el caso de aproximadamente 4000 enfermedades hereditarias,

incluyendo la fibrosis quística o la corea de Huntington) o ser el resultado de la respuesta del cuerpo a un estrés ambiental que causa alteraciones en el genoma (como el cáncer, diabetes, enfermedades cardíacas,…). El conocimiento completo del genoma humano nos permite buscar los genes directamente asociados con diferentes enfermedades, tratando de descubrir o mejorar nuestro entendimiento acerca de las bases moleculares de estas enfermedades. Este nuevo conocimiento nos permitirá mejorar tratamientos, curas e incluso desarrollar test predictivos.

Hoy en día, todos los fármacos en el mercado en su conjunto actúan sobre aproximadamente 500 proteínas. Mejorando el conocimiento de los mecanismos de la enfermedad y usando herramientas computacionales para identificar y validad nuevas dianas de fármacos se pueden desarrollar medicinas más específicas que actúen sobre las causas y no meramente sobre los síntomas. Como beneficio adicional, estos fármacos altamente específicos prometen menos efectos secundarios que las medicinas actuales.

La medicina clínica se irá personalizando con el desarrollo del campo de la farmacogenómica, que estudia cómo la herencia genética de un individuo afecta a la respuesta de su cuerpo a los fármacos. Y como en el desarrollo de fármacos existe la necesidad de eliminar aquellos que presenten efectos adversos en un porcentaje de población significativo, potencialmente fármacos perfectamente viables para una cantidad de la población nunca llegarán al mercado. Hasta ahora los médicos han tenido que usar el método de prueba y error para encontrar el mejor fármaco para tratar a un paciente en particular, ya que aquellos que reportan los mismos síntomas clínicos pueden presentar una amplia gama de respuestas al mismo tratamiento. En el futuro, los clínicos tendrán la posibilidad de analizar el perfil genético de cada paciente y prescribirle no sólo el mejor tratamiento farmacológico disponible, sino también la dosis adecuada desde el principio.

Una vez descubiertos los detalles específicos de los mecanismos genéticos de las enfermedades se pueden desarrollar pruebas diagnósticas para medir la susceptibilidad de una persona a distintas enfermedades. Acciones preventivas como el cambio de hábitos diarios o comenzar un tratamiento en los estadios más iniciales en los que será más exitoso resultarán en grandes avances en la continua lucha contra la enfermedad.

En un futuro no muy lejano, el uso potencial de los genes mismos para tratar una enfermedad puede transformarse en una realidad. La terapia génica es una solución muy interesante para tratar, curar o hasta prever una enfermedad al modificar la expresión de los genes de una persona. Hoy en día es un campo tan verde como prometedor, con ensayos clínicos llevándose a cabo para distintos tipos de cáncer y otras enfermedades.

4.2.2. Genómica microbiana

Los microorganismos son ubicuos, es decir, que se encuentran en todas partes. Han sido encontrados sobreviviendo prósperamente en ambientes extremos de calor, frío, radiación, salinidad, acidez y presión. Están presentes en el medioambiente, en nuestros cuerpos, en el aire, en la comida y en el agua. Y tradicionalmente se han usado diversas propiedades de los microorganismos en la panadería, destilería y en la industria alimenticia en general. La llegada del genoma completo de estos microorganismos podría aumentar sus aplicaciones en el medioambiente, salud, energía o industria por ejemplo.

4.2.3. Agricultura y ganadería

El secuenciado de los genomas de plantas y animales debería reportar enormes beneficios para la comunidad agrícola. Las herramientas bioinformáticas pueden usarse para encontrar los genes en dichos genomas y elucidad sus funciones. El conocimiento genético específico podría entonces usarse para producir cultivos más fuertes, resistentes a sequías, enfermedades e insectos, así como para mejorar la calidad del ganado haciéndolo más saludable, resistente a enfermedades y más productivo.

4.2.4. Estudios comparativos

Analizar y comparar el material genético de diferentes especies es un método importante para estudiar las funciones de los genes, los mecanismos de las enfermedades hereditarias y la evolución de las especies. Las herramientas bioinformáticas pueden usarse para hacer comparaciones entre números, localizaciones y funciones bioquímicas de genes en diferentes organismos, aprovechando el conocimiento adquirido de unos para mejorar el entendimiento de otros. Los organismos que son adecuados para usarse en la investigación experimental se denominan organismos modelos, y poseen unas propiedades que los hacen ideales precisamente para fines de investigación,

incluyendo ciclos vitales cortos, ciclos reproductivos rápidos, fáciles de manejar, baratos y manejables a nivel genético. Para humanos, por ejemplo, suele usarse el ratón, ya que sus genomas se parecen bastante (más de un 98%) y además se han encontrado una gran correspondencia uno a uno de genes entre ambas especies. La manipulación del ratón a nivel molecular y las comparativas genómicas entre especies revelan información detallada de la funcionalidad de los genes humanos, su relación evolutiva, y los mecanismos moleculares de muchas enfermedades humanas.

4.3. Recursos de libre disposición

Existen multitud de bases de datos y de herramientas bioinformáticas de todo tipo disponibles hoy en día para la consulta y tratamiento de información relativa a la biología molecular. La oferta privada trata fundamentalmente de facilitar al máximo el manejo de la información a los expertos, ya que la complejidad tanto de la información como de las propias herramientas dificultan enormemente su uso por parte de los usuarios objetivos, en su mayoría biólogos o médicos con escasa formación bioinformática. La oferta pública es incluso más amplia, y suele ser suficiente para llevar a cabo la totalidad de una investigación, aunque suele implicar la necesidad de personal experto para su manejo y adolece de una absoluta disparidad en la calidad de su desarrollo o incluso en su no menos importante mantenimiento. De hecho suele ser un requisito impuesto por las más prestigiosas revistas del campo la exigencia del compromiso a una disponibilidad de dos años por parte del grupo desarrollador de una aportación a la publicación.

Al margen de esfuerzos particulares de grupos de investigación o empresas particulares, son de una considerable relevancia los ingentes esfuerzos que realizan el National Institute of Health (NIH) desde Estados Unidos a través del National Center for Biotechnology Information (NCBI) y el Welcome Trust desde el Reino Unido a través del Sanger Institute (WTSI) y del European Bioinformatics Institute (EBI), así como el consorcio europeo con base alemana que conforma el European Molecular Biology Laboratory (EMBL), ya que gracias a ellos existen de manera libre y gratuita datos y herramientas fundamentales para el avance de la biología molecular. Todos estos organismos citados tienen como *leitmotiv* no sólo la puesta a disposición de toda la comunidad científica de datos biológicos y herramientas para procesarlos, sino

también la documentación y formación en dichos recursos y su costosa puesta en marcha y mantenimiento, especialmente en un momento en el que el coste del análisis de los datos generados comienza a superar el de su propia generación.

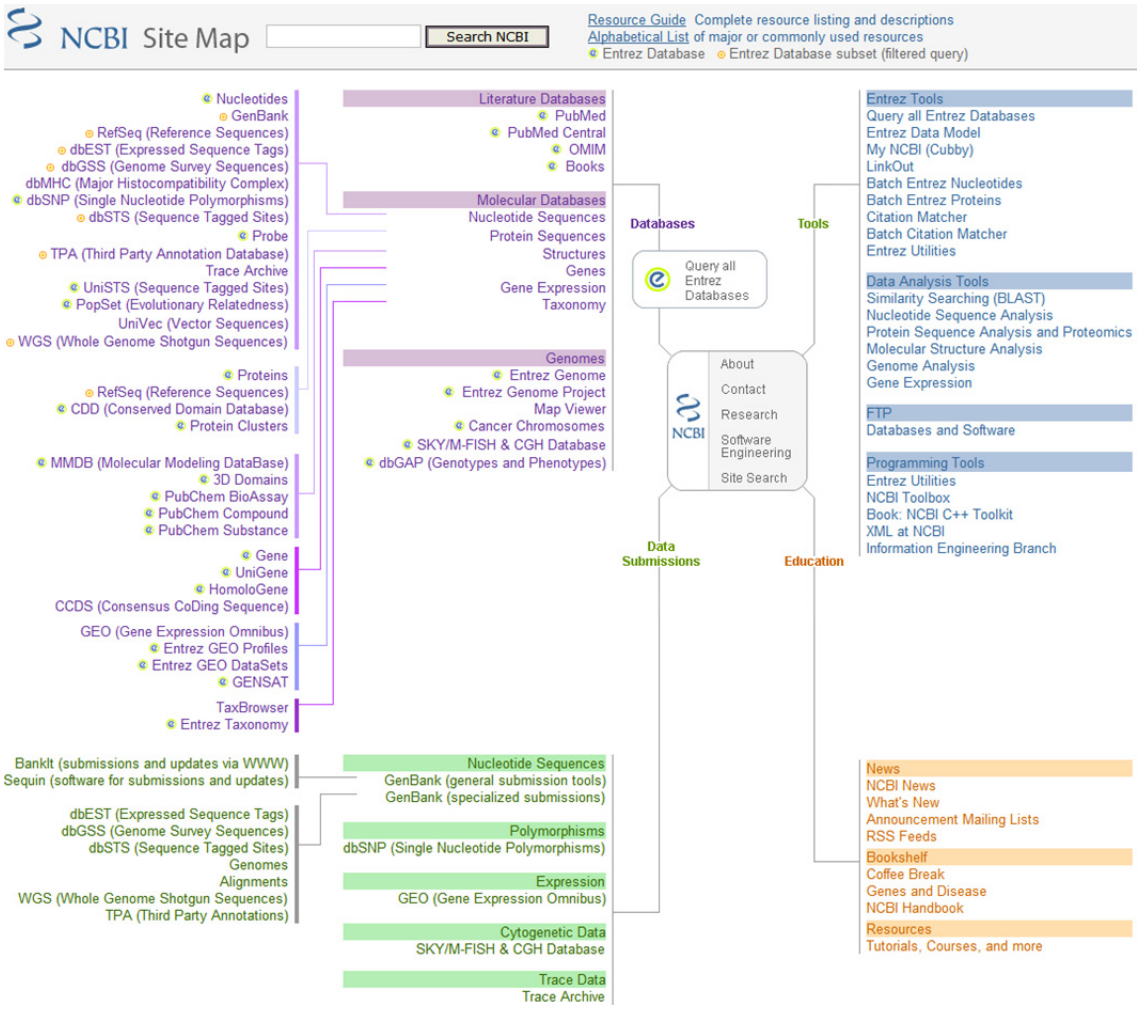


Figura 16: Mapa del portal web del NCBI. Inventario descriptivo de todos los recursos del NCBI disponibles en línea, agrupados por bases de datos accesibles por el sistema de consultas Entrez, herramientas diversas para el análisis genómico, envío de información para ser evaluada e incluida, y educación. Fuente: http://www.ncbi.nlm.nih.gov

Quizás lo que más ha cambiado en la última década, al margen de disponer de herramientas muy sofisticadas para el análisis de los datos existentes, ha sido la libre disposición de enormes repositorios biológicos de referencia. La aparición de internet a finales del siglo pasado y los esfuerzos internacionales por generar y hacer públicos cantidades de datos inimaginables

para un solo grupo de investigación, ha permitido a toda la comunidad científica avanzar a pasos más agigantados aún. Hoy en día no se puede entender la investigación sin la consulta a estos grandes repositorios antes, durante y al finalizar cada experimento, ya que son ahora parte esencial del diseño, ejecución y publicación de resultados. Pero para exprimir toda su potencial funcionalidad, los datos biológicos deben archivarse de manera consistente, y almacenarse de manera uniforme y eficiente. Estas bases de datos contienen información de un amplio espectro de áreas biológicas, desde las primarias o bases de datos de archivo, que se nutren de información y anotación de ADN y proteínas (secuencias, estructura y perfiles de expresión), a las secundarias o bases de datos derivadas, que gestionan los resultados del análisis de los recursos primarios (información de patrones de secuencia, variantes y mutaciones, relaciones evolutivas, o incluso información de la literatura haciendo uso de recursos bibliográficos). Es esencial que estas bases de datos sean fácilmente accesibles y que dispongan de un sistema de búsqueda intuitivo que permita a los investigadores obtener información muy específica sobre un tema biológico particular. Los resultados de las consultas deben entregarse de manera clara y consistente, idealmente acompañados de herramientas de visualización que faciliten su interpretación biológica.

La publicación Nucleic Acids Research (NAR) posee una extensa recopilación de las bases de datos disponibles más relevantes, funcionando así como un inventario de referencia que toda la comunidad científica pueda usar. Así mismo, anualmente publica una edición especial en la que se describen decenas de nuevas bases de datos que cubren distintas áreas de la biología molecular, así como otros tantos artículos describiendo las actualizaciones recientes de las que se hayan publicado previamente en la propia NAR y en otras publicaciones [37]. Entre todas ellas, dados los millones de usuarios que las usan y por su particular relevancia para esta memoria, destacan el mayor repositorio de referencia mundial de polimorfismos de base única (dbSNP) y la mayor herramienta de anotación automática de genomas metazoicos (Ensembl).

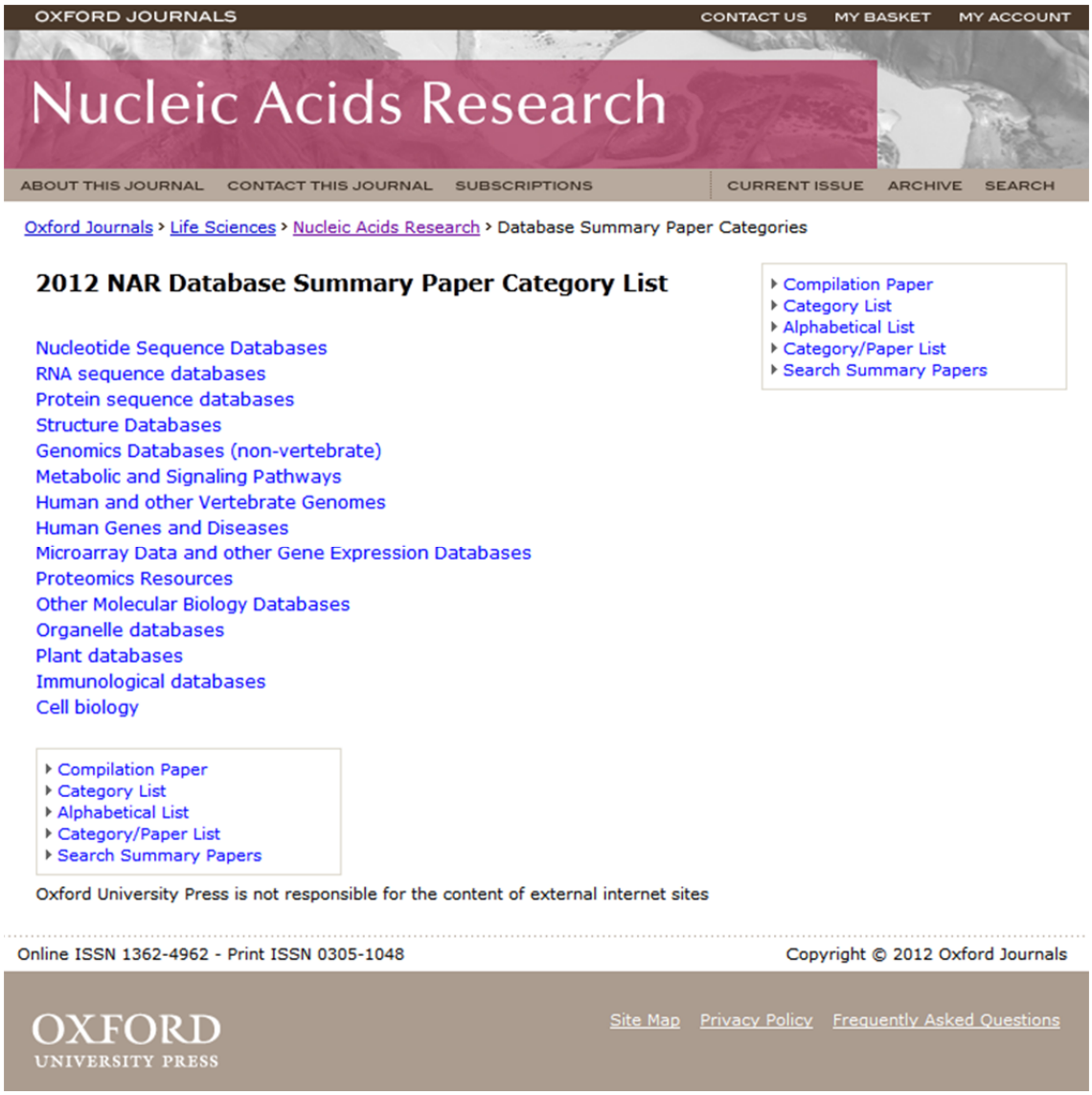


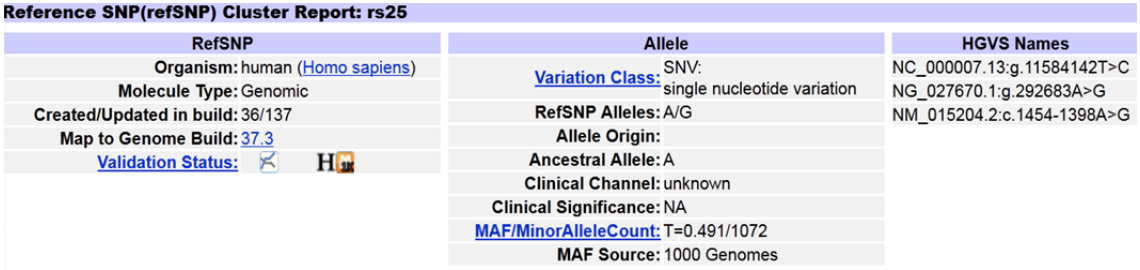
Figura 17: Portal de entrada al catálogo de bases de datos de la revista NAR. En la edición 2012 de este catálogo se añadieron 92 nuevas bases de datos y 100 artículos describiendo actualizaciones de otras ya existentes. Fuente: http://nar.oxfordjournals.org

4.3.1. dbSNP

La cantidad de recursos que el NCBI pone en línea a disposición del público son realmente amplios, tanto en número como en los ámbitos de conocimiento que cubren [38]. Desde su buque insignia GenBank®, que viene albergando desde hace más de 20 años más de 100 gigabases de secuencias nucleotídicas de más de 250,000 especies [39], pasando por la ayuda que presta PubMed a la investigación bibliográfica, disponiendo de más de 21 millones de citas provenientes de más de 24,000 publicaciones de ciencias de la vida, hasta

el alojamiento del genoma humano a través de la publicación de las distintas versiones actualizadas que el Genome Reference Consortium (GRC) se encarga de mantener.

La Database of Short Genetic Variations (dbSNP) es un repositorio de distintos tipos de variaciones genéticas cortas como SNPs, inserciones y deleciones, microsatélites y variaciones no polimórficas [40]. También almacena variaciones raras y comunes con sus genotipos y frecuencias alélicas, e incluye tanto variantes clínicamente significativas en humanos como polimorfismos benignos. Actualmente, en su versión NCBI dbSNP Build 137 de junio 2012, cuenta para el genoma humano con más de 53 millones de RefSNP clústers (o códigos *rs*, que son los códigos únicos de identificación de las variaciones una vez procesadas tras su envío), de los que más de 38 millones se encuentran validados, lo que representa una tasa de validación del 70% [41].



Reference SNP(refSNP) Cluster Report: rs25

RefSNP		Allele		HGVS Names
Organism:	human (Homo sapiens)	Variation Class:	SNV: single nucleotide variation	NC_000007.13:g.11584142T>C
Molecule Type:	Genomic	RefSNP Alleles:	A/G	NG_027670.1:g.292683A>G
Created/Updated in build:	36/137	Allele Origin:		NM_015204.2:c.1454-1398A>G
Map to Genome Build:	37.3	Ancestral Allele:	A	
Validation Status:		Clinical Channel:	unknown	
		Clinical Significance:	NA	
		MAF/MinorAlleleCount:	T=0.491/1072	
		MAF Source:	1000 Genomes	

Figura 18: Informe típico de dbSNP para una variación. El SNP rs25 es una variación del genoma humano, se encuentra validado, y su alelo menor T posee una frecuencia de 0.491 en más de mil muestras del proyecto 1000 Genomes. Los alelos que constan en RefSNP para esta posición son A/G debido a que se reportan respecto al bloque de ADN de referencia que está en sentido positivo, mientras que el SNP se ha reportado en sentido negativo y debería ser T/C. Fuente: http://www.ncbi.nlm.nih.gov

Originalmente fue creada para dar soporte al descubrimiento de polimorfismos a gran escala, como el del proyecto HapMap, pero enseguida fue considerado como repositorio mundial también para otros tipos de variaciones. Este hecho provocó que el término dbSNP pudiera llevar a engaño, ya que podría indicar que se trata de una base de datos exclusivamente de SNPs. De hecho, en los comienzos de la ultrasecuenciación, era habitual utilizar la información de dbSNP como indicativa de polimorfismos conocidos en la detección de variantes clínicas, lo cual pronto hubo de ser corregido con la llegada de grandes cantidades de variantes raras provenientes especialmente del proyecto 1000 Genomes. Así, el NCBI ha seguido manteniendo el acrónimo

de la base de datos, pero tenido que solventar este problema cambiando el título original "base de datos de polimorfismos de base única" por el de "base de datos de variaciones genéticas cortas", así como enfatizando atributos de la variación como la significación clínica o la frecuencia del alelo menor, o definiendo subconjuntos que poder usar razonablemente como sinónimos de polimorfismos o de variantes clínicamente significativas atendiendo precisamente a la frecuencia del alelo menor y a sus fuentes de origen.

4.3.2. Ensembl

Otra muestra de los enormes esfuerzos públicos por proveer de recursos genómicos a la comunidad científica es el del proyecto Ensembl que, aunque con enfoque particular en el genoma humano, contiene información de otros genomas cordados de muchos organismos modelos como el ratón, la rata o el pez cebra [42]. El proyecto comenzó en 1999 con el fin último de anotar automáticamente el genoma, integrar dicha anotación con otros datos biológicos disponibles, y publicar toda esta información en línea y de manera gratuita, algo que viene haciendo desde el 2000. Actualmente, en su versión Ensembl Release 68, disponen de información de 70 especies distintas [43].

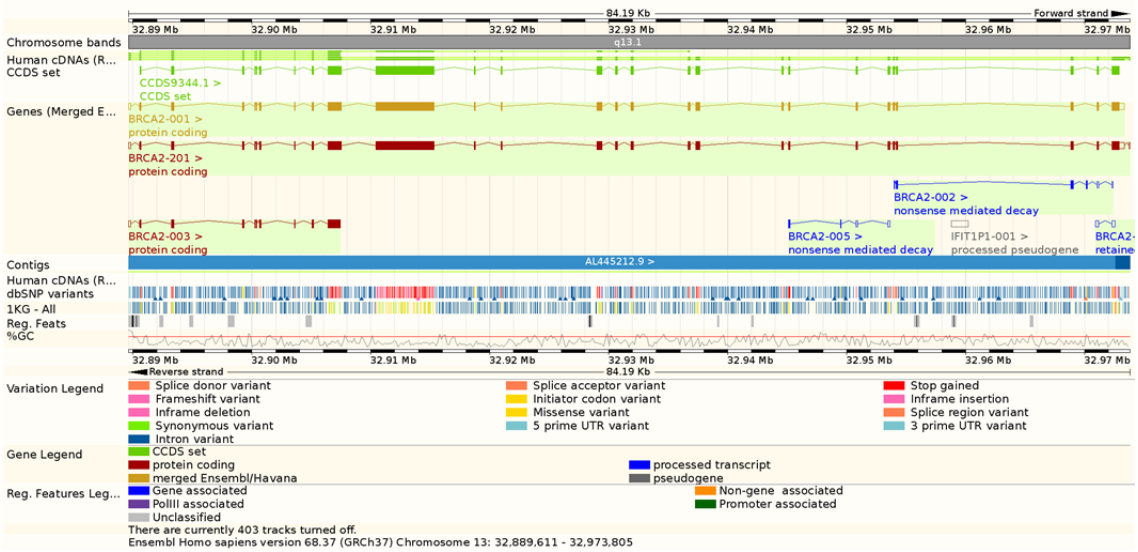


Figura 19: Vista típica del visor genómico Ensembl. Las distintas secciones horizontales de información o *tracks* que el usuario puede seleccionar bajo demanda, pudiendo mostrar entre las muchas opciones disponibles la región cromosómica consultada, los genes que contiene, o la información disponible de variación en dbSNP y en el proyecto 1000 Genomes. Fuente: http://www.ensembl.org

La cara más visible del proyecto es su visor genómico y su repositorio de recursos genómicos integrados, cuya densidad varía en función de la especie, siendo las de mayor completitud la de humano, ratón, rata y pez cebra, precisamente los genomas más consultados. Todas las especies disponen de anotaciones genéticas basadas en evidencias, así como de recursos de genómica comparativa, incluyendo alineamientos y relaciones de homología, ortología y paralogía. Todas estas anotaciones se integran con una enorme cantidad de fuentes externas de referencia, lo que convierte a Ensembl como un recurso integrador único. Además de todos los datos integrados en Ensembl y accesibles vía web, el proyecto también publica sus librerías de programación, lo que permite una interacción flexible y programática con sus datos para ser usados en el análisis genómico, tanto a través de la Ensembl API como del Ensembl BioMart.

II. JUSTIFICACIÓN Y OBJETIVOS

La publicación en los años 50 de la estructura del ADN, como génesis de la biología molecular moderna, coincidió en la misma década en la que aparecieron los primeros bocetos de una red de comunicaciones entre computadores que acabaría siendo lo que hoy conocemos como World Wide Web (WWW). No es difícil imaginar que los investigadores de estos campos tan distintos, aun pudiendo ser conscientes de los avances de los otros, jamás se imaginarían cómo ambas disciplinas se llegarían a entrelazar tan estrechamente. Pero no sólo el WWW, y en general las tecnologías de la información, son un ejemplo de un campo en principio apartado de la biología que acaba siendo crucial para el desarrollo de la misma. La evolución de la capacidad computacional que ha permitido aumentar el volumen de variables a analizar, los grandes avances en tecnologías de miniaturización que han posibilitado el manejo de materiales a escalas nanométricas, o incluso el gran aumento de la resolución de medida tanto en aparatos ópticos como electrónicos que capacitan al ser humano para caracterizar procesos con alta precisión y fiabilidad, son ejemplos de cómo distintas disciplinas científicas que nacen y crecen de manera relativamente aislada unas de otras acaban resultando fundamentales para el desarrollo y evolución de ellas mismas.

La biología había evolucionado a pasos de gigante gracias a los avances de la bioquímica y la robótica que, entre otras cosas, facilitaban enormemente la generación de datos de secuencias, pero se necesitaba extraer información de todos esos datos. Ya en la década de los 60 se entreveía dicha necesidad como un futuro cercano, pero a finales de los 70 con la llegada de la secuenciación Sanger se transformaron en vitales los desarrollos de nuevos algoritmos y de nuevos métodos para el archivado, difusión y consulta de toda la información generada.

Desde la década de los 70, donde comenzaron a surgir las primeras herramientas de conversión del código genético que obtenían la secuencia de aminoácidos a partir de la secuencia de ADN o los primeros algoritmos de alineamiento de secuencias, nuevos y variados métodos han ido apareciendo con el tiempo, parejos siempre al avance de la biología y de sus necesidades. Por ejemplo, no fue hasta la década de los 80 cuando comenzaron a aparecer las primeras bases de datos, sin las cuales el manejo de todos los datos y toda la información generados (es importante siempre distinguir entre datos e información), ya fuera a pequeña, media o gran escala, sería prácticamente imposible. Y en estas condiciones, en los comienzos de la década de los 90, se creó el Proyecto Genoma Humano. Y con él llegaron no sólo nuevos proyectos más ambiciosos de caracterización del genoma humano y de su variabilidad, sino también nuevas tecnologías de secuenciación capilar, de genotipado, y más recientemente de ultra-secuenciación. Todas estas nuevas tecnologías han ido apareciendo en un lapso de tiempo no mucho mayor a 15 años, y han ido obligando a los investigadores a entender la biología molecular desde perspectivas cada vez algo más distintas, y muy especialmente en el terreno del manejo de los datos.

El manejo eficaz tanto de los datos generados internamente como de la información externa disponible públicamente se convierte en un pilar básico de la investigación en la biología molecular. No sólo es necesario seguir planteando nuevas hipótesis biológicas que comprobar, sino que el propio desarrollo tecnológico que conlleva dicha comprobación debe ir paralelo a la investigación. De esta manera surgen por ejemplo bases de datos con información biológica de lo más variada: de secuencias, de proteínas, de variantes,… Es en este momento cuando la bioinformática se erige como la

gran solución horizontal que debe tratar de salvar la distancia entre la tecnología y el ser humano, entre la máquina y el investigador.

La presencia pública de muchas y muy variadas fuentes de información en línea hace necesaria su integración racional en los procesos de análisis. También resulta fundamental la generación y adaptación de algoritmos capaces de manejar grandes volúmenes de datos como los que resultan de las tecnologías de genotipado y de secuenciación. Los millones de variantes en los miles de muestras que hoy somos capaces de escrutar requieren de técnicas de programación adaptadas que entiendan el problema subyacente y permitan generar nuevo conocimiento a partir de él.

Este trabajo de tesis aborda distintos ámbitos de aplicación de técnicas bioinformáticas a la resolución de problemas surgidos del manejo, análisis, almacenamiento y consulta de grandes volúmenes de datos genómicos. Los objetivos generales que se han planteado para la elaboración de esta tesis han sido los siguientes:

- Procesar la información más básica de las tecnologías de genotipado de alto rendimiento, a fin de permitir obtener de manera rápida y sencilla una serie de parámetros y estadísticas básicas características de un experimento independientemente de la tecnología elegida.

- Facilitar la publicación y consulta de resultados de genotipado a baja y media escala, tanto de SNPs como de STRs, así como su interacción con los repositorios de variabilidad accesibles públicamente.

- Estudiar la viabilidad de gestionar localmente un repositorio propio de variabilidad humana basado en los recursos disponibles, tanto de información externa como de infraestructura interna.

- Transferir el conocimiento obtenido. Aportar herramientas existentes o soluciones *ad hoc* a los problemas que pueda presentar la investigación genética en el campo de la biología computacional.

III. RESULTADOS

Como resultado de la investigación realizada se presenta a continuación la producción científica generada. Se trata de un total de trece artículos de investigación publicados en diversas revistas científicas de ámbitos también diversos: desde la genética forense, pasando por la biología molecular, hasta obviamente la bioinformática. Todos estos artículos se han publicado a lo largo de estos cuatro últimos años, y se presentan en riguroso orden cronológico de publicación con el fin de reflejar la ejecución no lineal de las líneas de investigación seguidas.

The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project.

Amigo J, Phillips C, Lareu M, Carracedo A

International Journal of Legal Medicine. 06/2008

El explorador SNP*for*ID es una herramienta basada en web para la consulta y visualización de datos de frecuencias alélicas de polimorfismos de nucleótido único (SNPs) generados por el consorcio SNP*for*ID (http://www.snpforid.org/). A partir de este proyecto, se han generado a través del explorador paneles validados de SNPs para una variedad de aplicaciones forenses concentrándose en el grupo de 52 marcadores que conforman el conjunto de SNPs de identificación de tubo único. La interfaz web permite al visitante evaluar las frecuencias alélicas de los marcadores estudiados en todas las poblaciones disponibles usadas por SNP*for*ID para validar la variabilidad global de los SNPs. La interfaz se ha diseñado para ofrecer la utilidad de combinar poblaciones en grupos geográficos adecuados para la comparativa visual de poblaciones individuales o entre grupos definidos por el usuario y datos equivalentes de HapMap.

Int J Legal Med (2008) 122:435–440
DOI 10.1007/s00414-008-0233-7

TECHNICAL NOTE

The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project

Jorge Amigo · Christopher Phillips · Maviky Lareu · Ángel Carracedo

Received: 7 May 2007 / Accepted: 13 March 2008 / Published online: 20 May 2008
© Springer-Verlag 2008

Abstract The SNP*for*ID browser is a web-based tool for the query and visualization of the SNP allele frequency data generated by the SNP*for*ID consortium (http://www.snpforid.org/). From this project, validated panels of single nucleotide polymorphisms (SNPs) for a variety of forensic applications have been generated with the browser concentrating on the single-tube identification SNP set comprising 52 markers. A web interface allows the visitor to review the allele frequencies of the studied markers from all the available populations used by SNP*for*ID to validate global SNP variability. The interface has been designed to offer the useful facility of combining populations into appropriate geographic groups for visual comparison of populations individually or amongst user-defined groupings and with equivalent HapMap data.

Keywords SNP · SNP*for*ID · Online databases · Forensic allele frequency databases · HapMap

Accessibility: web access to this tool is granted at http://spsmart.cesga.es/snpforid.php

Electronic supplementary material The online version of this article (doi:10.1007/s00414-008-0233-7) contains supplementary material, which is available to authorized users.

J. Amigo (✉) · C. Phillips · Á. Carracedo
Spanish National Genotyping Center (CeGen) and Genomic Medicine Group, CIBERER,
University of Santiago de Compostela,
Santiago de Compostela, Spain
e-mail: jamigo@usc.es

C. Phillips
e-mail: chrisp@usc.es

M. Lareu · Á. Carracedo
Insitute of Legal Medicine, Genomic Medicine Group,
University of Santiago de Compostela,
Santiago de Compostela, Spain

Introduction

The SNP*for*ID consortium was set up in 2003 to develop single nucleotide polymorphisms (SNP) loci for use in human identification analysis: principally focused on forensic analysis but encompassing relationship testing (e.g., paternity analysis, confirmation of pedigree, etc.), enhanced prediction of geographic origin and medical sample identification. The main requirement from novel forensic marker sets, hitherto lacking in short tandem repeat loci (STRs), is the ability to successfully genotype highly degraded DNA without dropout: the differential loss of loci or alleles caused by PCR fragment sizes above ~125 bp or resulting from large differences in repeat number within a locus. For this reason, SNP*for*ID prioritized SNP sets that could be genotyped from amplified fragments generally below 100 bp and in multiplexes sufficiently large to provide equivalent, or better, discrimination power to the widely used 16-STR kits. A core 52 SNP multiplex has been developed for forensic analysis comprising loci primarily targeted from the p-arm and q-arm of each autosome [13]. This has been supplemented with SNP sets that allow the prediction of the geographic origin of a sample [9], enhanced characterization of the Y chromosome [3] and typing of haplotype-informative coding region SNPs in the mitochondrial genome [2].

An important aspect of the work of the consortium has been the promotion of an open source ethos for reporting the technical aspects of the SNP typing assays developed and the scientific findings together with the provision of tools to analyze SNP genotype data. The SNP*for*ID browser falls into the third category—an online tool that permits any researcher with genotyping data for the 52 SNPs in the forensic marker set to obtain allele frequency estimates from populations relevant to their own analyzes. The data is

presented in such a way that it is easy to collect and, when required, to combine allele frequency estimates from several populations into groups that better represent continental groups or geographic regions. HapMap allele frequency estimates from the four phase I study populations can also be listed to assist comparisons with the SNP*for*ID populations and as a benchmark for assessing the reliability of the estimates for each locus.

Two examples serve to illustrate the potential use of a frequency browser tool and highlight the flexibility of a combinational approach to reviewing SNP allele frequency data. In the first hypothetical example, a forensic laboratory in a nonurban region of northern Canada might wish to interpret a SNP profile by obtaining the frequency for the genotypes in both European and Inuit populations. Although Inuit data is available from the SNP*for*ID browser, Canadian European population data is lacking but could be adequately substituted with the combined European data readily obtained from the frequency page, allowing the investigator to report to court two appropriate cumulative frequency estimates for comparison. In the second, real case, a challenging paternity analysis of closely related individuals in Galicia (NW Spain) required SNP typing as a supplement to STR analysis. The investigator used the browser to compare and contrast Galician population estimates with various combinations of European populations and to obtain relevant frequency data permitting the assessment of the degree of local variation compared to European-wide patterns of variability for the 52 SNPs. In interpreting paternity analysis data involving related individuals, it is particularly important to gauge the degree of variability in the family investigated, the local population, and continent-wide to properly assess the significance of the genotypes and paternity indices obtained.

Data curation

Although it is easy to provide an open access website accessing the full set of population validation genotypes available, more power is provided by constructing a web tool that can read directly from a database of combinable data. Designing and programming a suitable search web tool with emphasis on visualization of allele frequencies became our main priority. From the start, it was decided that access to individual genotypes or sample profiles would not normally be required by forensic users and that data from multiple centers that can be combined or compared provides more flexibility. As such, each genotype is not particularly important as a single entity, but is considered as a whole when the allele frequency estimates are calculated from the query of joined databases. This does not preclude the possibility of SNP*for*ID centers genotyping standardized control samples such as those from the Coriell cell repositories (http://ccr.coriell.org/ccr/) or the positive control DNA supplied with standard forensic STR typing kits, then listing such profiles for each of the SNP sets developed by the consortium. Furthermore, the complete dataset of 52 SNP genotype profiles from all the populations listed (outside of HapMap) that underlie the allele frequency estimates are available as a flat text file download for each selected population, allowing user-defined analyses such as tests for independence or intrapopulation and interpopulation F_{st}.

The SNP*for*ID project represents the sum of efforts from six laboratories spread across Europe, so all the genotyping data generated required curation before being combined. A simple format database was created to form the basis for joining all available data and to allow for future developments that can also work from the same data. Data was indexed by sample and contained information of contributing laboratory, gender, population of origin (ascertained from the donors' declaration of their immediate ancestry), and 52 SNP genotypes. The curation process that checks data quality encompasses scrutiny of GeneMapper ID or Genotyper output from SNaPshot genotyping submissions made outside the SNP*for*ID laboratories plus assessment of Hardy–Weinberg equilibrium using chi-squared analysis together with F_{st} measurements comparing new populations with those of the same group.

A minor logistical problem in the initiation of the database was the collation and standardization of the data into a single repository. The binary nature of autosomal SNP data makes this process much easier than with multiple allele and haploid polymorphic loci utilized elsewhere in forensic science population databases like the Y Chromosome Haplotype Reference Database (YHRD, [10]) and Mitochondrial DNA Control Region Database (EMPOP, [7]). Therefore, all bases were inverted when necessary (e.g., CT base calls converted to AG) to match those listed in the Santa Cruz genome browser summary of dbSNP reference SNP data. Heterozygote genotypes were alphabetized and the locus listing order was, by previous convention, p-arm SNaPshot electrophoretic mobility (Auto1 SNPs) then q-arm (Auto2). Because all contributing laboratories used SNaPshot for the validation of populations, base standardization anticipates future submissions to the database from alternative genotyping platforms. To check genotyping quality, chi-squared analysis was made of the observed and expected genotype ratios in all populations having sufficient numbers of samples, although this had been previously performed on similar data to study interlaboratory concordance [13]. In addition, SNP*for*ID allele frequency estimates for African, European, and East Asian population groups were compared to those from the equivalent population panels of HapMap (termed Yoruba

from Ibadan, Nigeria [YRI]; CEPH Utah residents with European ancestry [CEU], and ASN, respectively, with ASN representing a panel of Chinese from Beijing [CHB] and Japanese from Tokyo [JPT] populations combined [1]).

Implementation

The web tool has been written in PHP and HTML, and it acts as an interface to the underlying database, an example of which is shown in Fig. 1. It was designed to go beyond text queries, and so certain graphical aids were developed to address this need. The first query point is a browsable world map allowing the visitor to locate each studied population and obtain frequency data with a single click. We used our own customized version of the DIY Map [5], a clickable zooming map written in Flash and configurable through an XML file providing the ability to not only spot the population locations and their population groups, but also to implement simple queries activated directly through clicks.

The graphical system of the data summary returned from the query provides visitors with a flexible and intuitive approach to the scrutiny of allele frequencies from single populations and in comparison to combinations of populations, enhanced to allow comparison of results using two different queries in parallel. This search system established itself as the main core of the application because all the possible queries that visitors were predicted to run had to be included together with the ability to preempt incorporation of future submissions of new populations or SNP sets to the database. As a result, the database is dynamically updated at the point in time each query is made, so the search page contains all current available data once it has been checked, curated, and incorporated. The same real-time updating process applies to the HapMap frequency data that is included in the data summary when available (48 out of the 52 SNPs have now been characterized by HapMap). In summary, the SNP data obtained from a query will always provide the most current frequency estimates for each SNP*for*ID and equivalent HapMap population: updated in real-time at the moment the query is made.

In keeping with the clean, easily interpreted pie chart summaries of SNP variability used successfully in the HapMap genome browser [14], we have mirrored the same approach in the pie charts used to visualize frequencies for each SNP*for*ID population or their combination, although actual allele frequencies are also listed as numeric values alongside the pie charts in the search return page. Charts display blue segments denoting the reference allele and red segments denoting the alternative allele with frequencies charted from 0.01 to 0.99. It is important to note two elements of the HapMap pie chart approach: (1) the reference allele segment is positioned counterintuitively on the left side of the zero point, i.e., from −3.6° (0.01 frequency) to −356° (0.99) and (2) triallelic SNPs that are now also in the browser as part of the ancestry-informative SNP sets from SNP*for*ID [8, 9] and were not included in the 1.1 million phase I SNPs characterized by HapMap. Therefore, the convention we propose to adopt for triallelic SNPs is to add a green segment for the third allele, denoting the least frequent allele observed in Africans and so likely to be the most recently derived substitution at the SNP.

Results

Depending on the options chosen for a search, the pie charts plotted in the query return page represent allele frequency estimates calculated from single populations or

L	15019	L-15019	M	Nigeria	AFRICA	CT	CT	AA	TT
L	16410	L-16410	M	Nigeria	AFRICA	CT	CC	AT	AA
L	16417	L-16417	M	Nigeria	AFRICA	TT	CT	AA	TT
L	16744	L-16744	M	Nigeria	AFRICA	CT	CC	AA	AA
L	16314	L-16314	M	Britain	EUROPE	TT	CC	AT	AT
L	16402	L-16402	M	Britain	EUROPE	CC	CT	TT	AT
L	16854	L-16854	M	Britain	EUROPE	CT	CC	AT	AT
L	16873	L-16873	M	Britain	EUROPE	TT	CC	TT	TT
L	17347	L-17347	M	Britain	EUROPE	CT	CT	AT	AA
L	17677	L-17677	M	Britain	EUROPE	CC	TT	AT	AT
L	17863	L-17863	M	Britain	EUROPE	CT	CC	AT	AT
L	18086	L-18086	M	Ireland	EUROPE	CT	CT	AT	AT
L	18123	L-18123	M	Ireland	EUROPE	TT	CC	AA	AA
L	18126	L-18126	M	Ireland	EUROPE	TT	CT	AA	AT
L	18147	L-18147	M	Ireland	EUROPE	CT	CC	TT	AT
L	18150	L-18150	M	Ireland	EUROPE	CC	CC	AA	TT
L	18244	L-18244	M	Ireland	EUROPE	TT	CT	AT	AT
L	18250	L-18250	M	Ireland	EUROPE	CT	CC	AT	AT
L	18274	L-18274	M	Ireland	EUROPE	CC	CT	AT	AT
L	18277	L-18277	M	Ireland	EUROPE	TT	TT	AA	AT
L	18280	L-18280	M	Ireland	EUROPE	CT	CT	AT	AT
L	18322	L-18322	M	Ireland	EUROPE	TT	CC	AT	AA
L	18999	L-18999	M	Britain	EUROPE	CT	TT	AA	TT
L	19021	L-19021	M	Ireland	EUROPE	CT	TT	AT	AA
L	19073	L-19073	M	Ireland	EUROPE	CC	CT	AA	AT

Fig. 1 Example snapshot from the joined SNP*for*ID database. Entry columns denote, from *left* to *right*, originating center; center sample ID; SNP*for*ID sample identifier; gender, population of origin; population group; and genotypes (A01–A54 in the same order of SNPs as search page top to bottom, allowing a direct transposition from a curated Excel file to the database)

their combinations as a single column for the search option plus multiple columns for up to four user-defined comparisons. Five population groups are summarized in a set of pie charts using the grouping of populations outlined in the search page listings that Fig. 2 shows. This grouping is based on a previous study of global variability that found a close match between geographic distribution of populations and genetic clustering using STRUCTURE to arrange populations into groups based on patterns of variability [11]. Using the same clustering algorithm for the 52 SNPs and 9 of the validation populations in the browser gave a broadly similar grouping within the confines of a much smaller range of loci and study populations (Fig. 3, K=4 in [13]). The separate listing of the South Asian population sample to those from Europe is a potentially contentious arrangement because populations from this region tended to cluster with other Eurasian populations from Europe, North Africa, and the Middle East in the Rosenberg study; however, the browser allows this population sample to be included with the six European populations or analyzed separately so the added flexibility provided is worth retaining, particularly as additional South Asian populations are likely to be sampled and submitted to the browser. This last point also illustrates the potential of a combine-and-compare approach in studying differences between populations because the pie charts provide an intuitive system for visualizing the contrasting allele frequency distributions found in some of the SNPs in the 52-SNP set. Such SNPs comprise about 10% of the full set and were chosen deliberately to provide indicators of geographic origin in the same way STR data can be used for this purpose [6]. Therefore, it seems likely that the use in the near future of dedicated sets of ancestry-informative SNP sets including those of SNP*for*ID [9] will also benefit from the system of allele frequency visualization adopted for this browser.

To statistically assess the goodness of fit of allele frequency estimates from SNP*for*ID and HapMap, an r^2 analysis was performed on appropriate population groupings matched to the HapMap study panels described above. ESM Fig. 1 presents an analysis of allele frequency estimate correlation between SNP*for*ID and HapMap genotyping for 48 of 52 SNPs analyzed in common. Goodness of fit between the paired datasets was assessed using r^2 analysis of appropriate SNP*for*ID study population groupings matched to the HapMap study panels: (a) European (EUR vs CEU), (b) East Asian (ASN vs combined CHB/JPT), and (c) African (AFR vs YRI). The listed r^2 values indicate good correlation of SNP*for*ID and HapMap frequency estimates for all loci and each pair of population groups.

As an illustration of the standard display features of the browser, a dataset of samples from Spain and Mozambique is illustrated in ESM Fig. 2 because both populations represent a data subset that can be readily compared to their continental-based population groups of Europeans and Africans, respectively. ESM Fig. 2 illustrates a complete query result for NW Spain and Mozambique with summary population-group pie charts showing allele frequency data for each SNP and the equivalent HapMap estimates when present. The SNP*for*ID population-group pie charts are designed to match the order of HapMap charts: EUR/CEU (SNP*for*ID European/HapMap CEPH European from Utah of northern and western European ancestry), ASN/CHB+JPT combined (SNP*for*ID East Asian/combined Chinese from

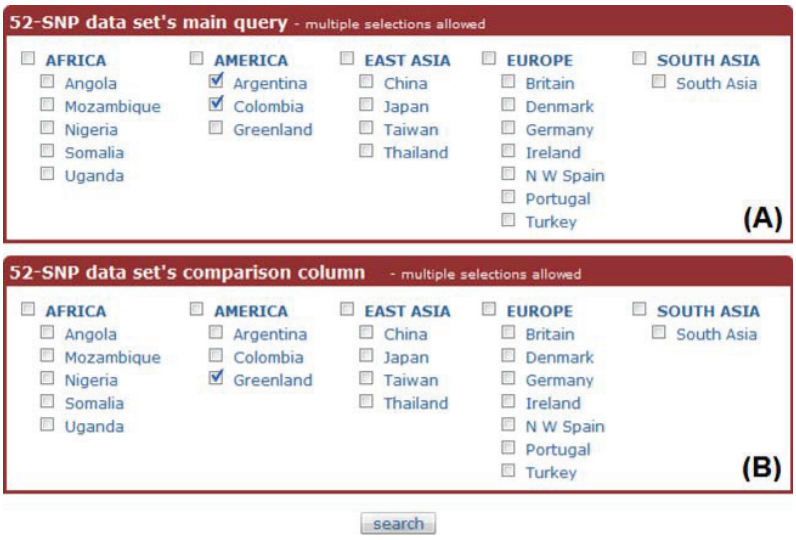


Fig. 2 **a** Search options available in the search page. Offset upper row tick-boxes allow combination of the listed populations of each region to create a full panel or population group. **b** Comparison options available in the search page. In each case, combinations can be tailored by the user to more closely match geographic distribution; in the example, ticking Argentina and Colombia in the search populations query and Greenland alone in the compare populations query permits comparison of North and South American population groups

Beijing and Japanese from Tokyo), AFR/YRI (SNP*for*ID African/Yoruba of Ibadan, Nigeria), plus SAS=SNP*for*ID South Asian and AME=SNP*for*ID American.

It is important to note that although all database profiles are complete, the sample number ranges from 7 (Japan) to 156 (Denmark) and clearly certain small population samples require interpretation with caution or exclusion altogether. The population data is structured in columns and the SNP data is structured in rows for all collated pie chart sets and corresponding full-frequency figures. These allele frequencies are shown numerically in columns under their corresponding genotyped base to four decimal places, and the pie charts are drawn to 1% allele frequency precision. A column of hyperlinks to dbSNP provides a convenient system for obtaining additional data for the individual SNP locus if required. The complete dataset of 52 SNP genotype profiles from all the populations listed (outside of HapMap) that underlie the allele frequency estimates are available as a flat text file download for each selected population, allowing user-defined analyses such as tests for independence or intrapopulation and interpopulation F_{st}.

Finally, at the time of writing, the website registered an average of 150 visits per month. The browser has been available to the public since December 2005 and has benefited in particular from links placed in the STRbase forensic marker information portal run by the National Institute of Standards and Technology (NIST, [12]) and the SNP*for*ID homepage (http://www.snpforid.org/).

Discussion

The SNP*for*ID browser represents a simple but highly effective visualization method to query and display the genotype data of the SNP*for*ID project. The format of the pie chart graphics also helps the researcher to quickly review the data, and the comparison with HapMap data as an external resource adds an appropriate system for confirming the precision of the allele frequency estimates given with both datasets being updated in real-time immediately before the display of the query results. This browser has been designed to be a web tool that can be rapidly accessed by the forensic practitioner requiring instant allele frequency data retrieval for a specific population plus a comparison at the same time with samples of global variability and is directly available at http://spsmart.cesga.es/snpforid.php.

Databases can fall into the trap of becoming static and out-of-date entities if they are not updated regularly. We have avoided this problem by recalculating allele frequency results at the moment a query has been submitted and by retrieving the current HapMap data at the same time. As well as ensuring all data displayed is the most current available, the dynamic system of data management we have adopted makes it easier to incorporate new data and to welcome submissions via e-mail from the worldwide forensic community (see contact information on the title page). This may represent a more efficient way to disseminate allele frequency data from an extending range of global populations than the conventional system of journal publication of allele frequency data. However, such an approach brings with it the problems of quality management more easily addressed in the curation of online haplotype loci databases mentioned previously (YHRD and EMPOP) where phylogenetic methods can be applied to check for typing errors. For this reason, we have decided to require scrutiny of raw genotyping data generated by contributing laboratories outside the SNP*for*ID consortium. We now include the ancestry-informative SNPs developed by SNP*for*ID [9] that supplement the identification SNP set. Ancestry-informative SNPs in particular benefit from the broadest range of shared population data because they show higher overall variability between populations. One favorable feature of autosomal SNP data in general is that relatively small population samples provide reliable allele frequency estimates. Therefore, submitting data to a shared database for SNPs of forensic interest should not represent a prohibitive amount of effort from those interested in validating these loci for forensic applications in their own laboratories.

Finally, we intend to allow for the possibility of linking allele frequency data to individual genotype profiles from widely used standard control samples such as the CEPH–HGDP panel of population samples [4] or the Coriell cell repositories control sample set. This would offer the simplest system for providing control profiles to help researchers that are establishing genotyping assays for the SNP*for*ID loci in their laboratories for the first time.

Acknowledgements The authors wish to thank Albert Vernon Smith and Lalitha Krishnan of the HapMap Project for their guidance in helping us link the browser to the HapMap SNP dataset, and Antonio Salas for his help with the genotyping quality assessment. We also would like to thank the Centro de Supercomputación de Galicia (CESGA) for their web hosting service and technical support. Funding from Xunta de Galicia: PGIDTIT06PXIB228195PR and Ministerio de Educación y Ciencia: proyecto BIO2006-06178 given to ML partially supported this work.

References

1. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P, The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
2. Brandstatter A, Salas A, Niederstatter H, Gassner C, Carracedo A, Parson W (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27:2541–2550

3. Brion M, Sanchez JJ, Balogh K et al (2005) Introduction of an single nucleotide polymorphism-based “major Y-chromosome haplogroup typing kit” suitable for predicting the geographical origin of male lineages. Electrophoresis 26:4411–4420
4. Cann HM, de Toma C, Cazes L et al (2002) A human genome diversity cell line panel. Science 296:261–262
5. Emerson J (2006) DIY Map: a clickable and zoomable map written in Flash. Available at http://www.backspace.com/mapapp/
6. Lowe AL, Urquhart A, Foreman LA, Evett IW (2001) Inferring ethnic origin by means of an STR profile. Forensic Sci Int 119:17–22
7. Parson W, Brandstatter A, Alonso A et al (2004) The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation results and perspectives. Forensic Sci Int 139:215–226
8. Phillips C, Lareu V, Salas A, Carracedo A (2004) Non binary single-nucleotide polymorphism markers. In: Doutremepuich C, Morling N (eds) Progress in forensic genetics, 10. Elsevier, Amsterdam, pp 30–32
9. Phillips C, Salas A, Sanchez JJ et al (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genetics 1:233–235
10. Roewer L, Krawczak M, Willuweit S et al (2001) Online reference database of European Y–chromosomal short tandem repeat (STR) haplotypes. Forensic Sci Int 118:106–113
11. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
12. Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res 29:320–322
13. Sanchez JJ, Phillips C, Borsting C et al (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis 27:1713–1724
14. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. Genome Res 15:1592–1593

SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data.

Morcillo-Suarez C, Alegre J, Sangros R, Gazave E, de Cid R, Milne R, Amigo J, Ferrer-Admetlla A, Moreno-Estrada A, Gardner M, Casals F, Pérez-Lezaun A, Comas D, Bosch E, Calafell F, Bertranpetit J, Navarro A

Bioinformatics (Oxford, England). 08/2008; 24(14):1643-4.

Los polimorfismos de nucleótido único (SNP) son el marcador más ampliamente utilizado en estudios para evaluar las asociaciones entre variantes genéticas y los rasgos complejos o enfermedades. También se están convirtiendo cada vez más importantes en el estudio de la evolución y de la historia de los seres humanos y otras especies. El análisis y tratamiento de los SNPs obtenidos gracias a tecnologías de alto rendimiento implica el uso de software costoso en tiempo y en dinero, diferente, complejo y generalmente de formato incompatible. SNPator es un programa de análisis de datos de SNPs fácil de usar y basado en la web que integra, entre muchos otros algoritmos, los pasos más comunes de un estudio de asociación de SNPs. Se libera al usuario de la necesidad de disponer de instalaciones informáticas de gran tamaño y un conocimiento en profundidad de la instalación del software genético y su manejo. Los datos genotípicos se leen directamente de los archivos de salida de las plataformas de genotipado habituales. Los datos fenotípicos de las muestras también pueden cargarse fácilmente. Mediante el uso de algoritmos integrados SNPator o llamando software genético estándar se pueden realizar muchos y diversos procedimientos de control de calidad y de análisis.

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 14 2008, pages 1643–1644
doi:10.1093/bioinformatics/btn241

Genetics and population analysis

SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data

Carlos Morcillo-Suarez[1,2,3], Josep Alegre[1,2,3], Ricardo Sangros[1,2,3], Elodie Gazave[1], Rafael de Cid[2,4], Roger Milne[2,5], Jorge Amigo[2,6], Anna Ferrer-Admetlla[1], Andrés Moreno-Estrada[1], Michelle Gardner[1], Ferran Casals[1], Anna Pérez-Lezaun[1,2], David Comas[1,7], Elena Bosch[1,7], Francesc Calafell[1,7], Jaume Bertranpetit[1,2,7] and Arcadi Navarro[1,2,3,7,8,*]

[1]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Barcelona, [2]National Genotyping Centre (CeGen), [3]Population Genomics Node (GNV8) National Institute for Bioinformatics (INB), [4]Genes and Disease Program, Center for Genomic Regulation (CRG), Barcelona, [5]Human Genetics Group, Human Cancer Genetics Program, Spanish National Cancer Centre, Madrid, [6]Unidade de Xenética, Facultad de Medicina, Santiago de Compostela, [7]CIBER en Epidemiologia y Salud Pública (CIBERESP) and [8]Institució Catalana de Recerca i Estudis Avançats, ICREA and Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Dr Aiguader 88, 08003 Barcelona, Spain

Received on February 21, 2008; revised and accepted on May 16, 2008

Advance Access publication May 30, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: Single nucleotide polymorphisms (SNPs) are the most widely used marker in studies to assess associations between genetic variants and complex traits or diseases. They are also becoming increasingly important in the study of the evolution and history of humans and other species. The analysis and processing of SNPs obtained thanks to high-throughput technologies imply the time consuming and costly use of different, complex and usually format-incompatible software. SNPator is a user-friendly web-based SNP data analysis suite that integrates, among many other algorithms, the most common steps of a SNP association study. It frees the user from the need to have large computer facilities and an in depth knowledge of genetic software installation and management. Genotype data is directly read from the output files of the usual genotyping platforms. Phenotypic data on the samples can also be easily uploaded. Many different quality control and analysis procedures can be performed either by using built-in SNPator algorithms or by calling standard genetic software.

Availability: Access is granted from the SNPator webpage http://www.snpator.org.

Contact: arcadi.navarro@upf.edu; bioinformatica.cegen@upf.edu

Supplementary information: Additional information, including tutorials and example datasets, is available from SNPator's webpage.

1 INTRODUCTION

The vast number of SNPs identified in the last few years and the development of high-throughput genotyping technologies have provided the opportunity for many research groups to undertake association studies of varying scales on a regular basis. SNP association studies have become crucial in the uncovering of genetic correlations of genomic variants with complex diseases, quantitative traits and physiological responses to drugs (e.g. Andrawiss, 2005). SNPs are also increasingly employed to study the history of populations and the evolution of species (e.g. Moreno-Estrada *et al.*, 2008; Tishkoff *et al.*, 2007).

In spite of the increasing popularity of SNP studies, processing and analyzing the huge amounts of data generated by genotyping technologies is still a burdensome and time consuming task. Hundreds of different software packages, most of them free for research purposes, have been developed to deal with particular problems and are available on the Web (http://linkage.rockefeller.edu/soft). Much time and effort is required, not only to identify the most appropriate algorithms and programs for each goal, but also to install them on local computers, to learn how they work or to give the appropriate format to input data. Within many genotyping projects, post-genotyping data management and analysis have become a bottleneck hindering the achievement of results. In order to help tackling these problems we have developed a web-based software solution called SNPator (for SNP analysis to results).

2 IMPLEMENTATION

2.1 Architecture and database features

The basic structure of SNPator consists of a central Linux server with MySQL and the PHP written application. This central node acts as a webserver and database manager. All the tasks and analyses that SNPator performs are coded in the form of WebServices that are executed remotely by computing servers and which can be called by external software other than SNPator.

Users can log into the application via web using a standard browser and introducing the usernames and passwords that can be obtained—without registration—from SNPator's webpage. Users

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

have different levels of privileges and can only access their own studies. A study is a working space—shared by as many users as necessary—where a set of data and all results generated from its analysis are stored. Each study starts with three types of data in highly customizable tables: a set of SNPs with related genomic information, a set of samples with population or phenotypic information and a set of genotypes. SNP and sample information can be easily uploaded using several methods, including, for SNPs, automatic upload from public databases such as dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) or HapMap (www.hapmap.org). Information on samples can include any sort of numerical, categorical or textual variables and can also be automatically uploaded after customizing the fields in the study. Genotypes can be uploaded directly from the output files generated by the most usual genotyping technologies (Illumina, Sequenom, SNPlex, etc.). All data within SNPator can be uploaded and downloaded in XML format to ease interaction with other software.

2.2 Quality control and analysis features

Once data have been uploaded, SNPator offers many quality control and analysis possibilities. Quality control options range from the detection of contradictory genotypes to the generation of graphical reports of uploaded plates. As to analysis, the simplest way in which SNPator can be used is to generate formatted data files ready to be used by other programs. Data can be downloaded into different formats ranging from ordered lists and matrices (to be imported into Excel or SPSS, for example) to input files for standard genetic software. Other analysis possibilities range from the simplest tests (such as Hardy Weinberg) to genomic overview measures (linkage disequilibrium, haplotype inference, population differentiation statistics, and others), disease-oriented analyses (allele or haplotype association tests, TDT and others) or multiple test corrections. Some analysis algorithms have been implemented as PHP scripts in SNPator, while others use standard external software that has been wrapped into WebServices.

Any action demanded by the user generates a job that will be sent to a queue and performed when resources become available. Most jobs will be performed immediately but those requiring more computational resources (haplotype estimations, for instance) will be put on hold while other such jobs are running. The appropriate screen provides users with information about the generation, execution, completion time and current status of their jobs. All the actions performed in SNPator generate results which are stored in a section called User Results. Results remain there as long as the user wants them and can be read, downloaded and even reused for further analysis in SNPator (in the case of workflows with more than one step). When launching an action, the user can ask to be sent an e-mail when this action is finished.

2.3 Filters and batch mode

SNPator implements several features that ease complex analysis procedures. First, users may define a set of criteria (filters) to select a subset of SNPs and samples from a study by means of easily created Boolean statements. The fraction of genotyping success of a SNP or sample can also be used as a criterion to set up a filter. When one of the filters is activated, all operations performed with SNPator will affect only the SNPs and samples selected in the filter and its genotypes.

Another feature which facilitates analysis is the Batch Mode in which several jobs can be simultaneously generated using as inputs different values in a field. If, for example, 'Sample Batch mode' for the field 'Population' is selected when running an allele frequency job, SNPator examines the 'Population' field, determines how many different values are there and runs as many 'allele frequency' tests as populations in that field.

2.4 System management

A web-based administration application has also been developed. Using it, it is possible to perform tasks such as managing the set of extant studies and the user privileges. It is also possible to obtain usage statistics by means of text or a graphical output. Such statistics include summaries of user logins and their actions, lists of currently running and waiting jobs, memory usage parameters and many others. This feature will be made generally available in future 'pre-packaged' versions of SNPator that users will be able to install on their own servers

3 APPLICATIONS AND USE TO DATE

SNPator is open to all users and it is currently the core application in the Spanish National Genotyping Center (www.cegen.org). CeGen is a nodal network of different genotyping facilities distributed in three different cities and created to allow scientists access to distinct genotyping technologies. Once samples are genotyped, data are uploaded into SNPator from the different platforms so that users can access them at a single point, add their own data and perform any analysis. External users can upload their own data by themselves. Over the last two years, SNPator has been used to perform, either in part or completely, more than 200 studies, ranging from association studies (Goertsches *et al.*, 2008) to population genetic analysis of genes or genome regions in different populations (Gardner *et al.*, 2007). SNPator differs from other packages in both its wide and ever-growing spectrum of possibilities and its extremely easy usage.

ACKNOWLEDGEMENTS

We are grateful to the CeGen coordination team for continuing support. We are indebted to the many users that have provided us with feedback about features to improve.

Funding: This work is funded by the National Institute for Bioinformatics and the National Genotyping Center, two platforms of Genoma España, and projects BFV2005 – 00243 to EB and BFU2006-15413-C02-01 to A.N.

Conflict of Interest: none declared.

REFERENCES

Andrawiss,M. (2005) First phase of HapMap project already helping drug discovery. *Nat. Rev. Drug Discov.*, **4**, 947.

Gardner,M. *et al.* (2007) Extreme individual marker F(ST)values do not imply population-specific selection in humans: the NRG1 example. *Hum. Genet.*, **121**, 759–762.

Goertsches,R. *et al.* (2008) Evidence for association of chromosome 10 open reading frame (C10orf27) gene polymorphisms and multiple sclerosis. *Mult. Scler.*, **14**, 412–414.

Moreno-Estrada,A. *et al.* (2008) Signatures of selection in the human olfactory receptor OR5I1 gene. *Mol. Biol. Evol.*, **25**, 144–154.

Tishkoff,S.A. *et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, **39**, 31–40.

SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access.

Amigo J, Salas A, Phillips C, Carracedo A

BMC Bioinformatics. 11/2008; 9(1):428.

En los últimos cinco años han aparecido grandes recursos en línea de variabilidad humana, especialmente HapMap, Perlegen y la fundación CEPH. Estas bases de datos de genotipos con información poblacional actúan como catálogos de la diversidad humana, y son ampliamente utilizados como fuentes de referencia para los estudios de genética de poblaciones. Aunque muchas conclusiones útiles pueden ser extraídas mediante la consulta de las bases de datos individualmente, la falta de flexibilidad para la combinación de los datos dentro de y entre cada una de las bases de datos no permite el cálculo de estadísticas clave de variabilidad poblacional.

Hemos desarrollado una nueva herramienta para el acceso y la combinación de grandes bases de datos genómicos de polimorfismos de un solo nucleótido (SNPs) para su uso generalizado en genética de poblaciones humanas: SPSmart (SNPs para Estudios Poblacionales). Una rápida serie de programas enlazados crea y mantiene un repositorio estático a partir de las bases de datos de genotipos con información poblacional más comúnmente usadas: los datos se extraen, se resumen en índices de referencia estadísticos estándar, y se almacena en una base de datos relacional que actualmente maneja $4x10^9$ genotipos y que se puede extender fácilmente a nuevas iniciativas de bases de datos. También hemos creado una interfaz web para el repositorio estático que permite la navegación de los datos subyacentes indexados por población y la combinación de las poblaciones, permitiendo la comparación intuitiva y directa de los grupos poblacionales. Toda la información servida está optimizada para visualizarla en la web, y la mayoría de los cálculos ya están pre-procesados en el repositorio estático para acelerar la exploración de datos y cualquier tratamiento computacional requerido.

En la práctica, SPSmart permite combinar poblaciones en grupos definidos por el usuario, mientras varias bases de datos se acceden y comparan en unos pocos pasos simples de una sola consulta. Realiza las consultas rápidamente y genera resúmenes gráficos sencillos de variabilidad de los SNPs de las poblaciones a través de la inspección visual de las frecuencias alélicas descritas en gráficos circulares estándar. Además, la descripción completa numérica de los datos se dispone en paneles de resultados estadísticos que incluyen métricas comunes de la genética de poblaciones como la heterocigosidad, *Fst* e *In*.

BMC Bioinformatics

BioMed Central

Software

Open Access

SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access

Jorge Amigo*[1], Antonio Salas[2], Christopher Phillips[1] and Ángel Carracedo[1,2]

Address: [1]Spanish National Genotyping Center (CeGen), Genomic Medicine Group, CIBERER, University of Santiago de Compostela, Galicia, Spain and [2]Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Galicia, Spain

Email: Jorge Amigo* - jorge.amigo@usc.es; Antonio Salas - antonio.salas@usc.es; Christopher Phillips - c.phillips@mac.com; Ángel Carracedo - angel.carracedo@usc.es

* Corresponding author

Published: 10 October 2008

BMC Bioinformatics 2008, **9**:428 doi:10.1186/1471-2105-9-428

Received: 15 May 2008
Accepted: 10 October 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/428

© 2008 Amigo et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: In the last five years large online resources of human variability have appeared, notably HapMap, Perlegen and the CEPH foundation. These databases of genotypes with population information act as catalogues of human diversity, and are widely used as reference sources for population genetics studies. Although many useful conclusions may be extracted by querying databases individually, the lack of flexibility for combining data from within and between each database does not allow the calculation of key population variability statistics.

Results: We have developed a novel tool for accessing and combining large-scale genomic databases of single nucleotide polymorphisms (SNPs) in widespread use in human population genetics: SPSmart (SNPs for Population Studies). A fast pipeline creates and maintains a data mart from the most commonly accessed databases of genotypes containing population information: data is mined, summarized into the standard statistical reference indices, and stored into a relational database that currently handles as many as 4×10^9 genotypes and that can be easily extended to new database initiatives. We have also built a web interface to the data mart that allows the browsing of underlying data indexed by population and the combining of populations, allowing intuitive and straightforward comparison of population groups. All the information served is optimized for web display, and most of the computations are already pre-processed in the data mart to speed up the data browsing and any computational treatment requested.

Conclusion: In practice, SPSmart allows populations to be combined into user-defined groups, while multiple databases can be accessed and compared in a few simple steps from a single query. It performs the queries rapidly and gives straightforward graphical summaries of SNP population variability through visual inspection of allele frequencies outlined in standard pie-chart format. In addition, full numerical description of the data is output in statistical results panels that include common population genetics metrics such as heterozygosity, *Fst* and *In*.

Background

The diverse nature of the major online SNP databases requires the researcher interested in population variability to query each in turn, to obtain allele frequencies, then to compile their own statistical indices for comparison of populations within and between databases. This task is

made more difficult by the different data formats of the results given by each database, whose focus does not always address the needs of researchers interested in population variability, and in some cases it may be necessary to download large data segments to run locally a specific population based analysis.

Because the large-scale SNP data repositories are heterogeneous, and in response to our own need for a graphical browser for complex and extensive SNP data where this was lacking, we developed a system to summarize genotypes from multiple populations quickly and easily. The system uses a fast pre-processing pipeline able to work with any population based SNP database and can bring together disparate information into more informative summaries of variability, locus data and statistical metrics.

Methods

SNP resources

Many online databases that catalogue human variability provide population information about the samples studied, notably HapMap [1,2], Perlegen [3] and the CEPH foundation [4,5]. For instance, data from the CEPH Foundation collating genotypes generated from the human genome diversity panel (HGDP) gives one of the most valuable resources for human population studies in terms of geographic coverage and samples analyzed (1056 samples from 51 diverse populations), with recent contributions releasing major quantities of genotypes, e.g. the Stanford University CEPH-HGDP SNP genotyping initiative has yielded 650,000 SNP genotypes in 971 samples [6]. However, the data is accessible only as flat text files of limited use for many of the needs of population and evolutionary genetics studies. The Michigan University CEPH-HGDP SNP genotyping initiative has replicated in large part that of Stanford, so both databases overlap significantly in SNPs and samples genotyped. Therefore these databases cannot be considered as fully independent when carrying out population genetics studies.

In contrast to the Stanford and Michigan databases, the HapMap Phase II database contains an extensive amount of common genetic variation characterized in just four population samples. One of the main aims of HapMap Phase III was to extend the genotyping to a wider range of populations comprising SNP data of 1,115 individuals from 11 populations. The Perlegen database is also extensive in terms of SNP number but limited in terms of populations studied.

Some SNP repositories have web-sites that allow the downloading of SNP genotypes and locus information (chromosome position, linkage disequilibrium, etc.). However none permit the comparison and re-combination of multiple populations or the computation of population genetics indices. The SPSmart addresses this gap in possible analysis approaches by allowing the user to make specific searches of SNP lists in chromosomal regions and/or genes and to make comparisons of SNP variation within and between each of the databases outlined on Table 1. In particular the option to compare SNP variability across different databases provides a valuable system for initiating SNP based population genetics studies.

Pre-processing the data

A common characteristic of the most widely accessed human population databases is infrequent or unpredictable update cycles. To remove the need of the user to check for updates we have implemented a fast pre-processing pipeline, able to work with any given SNP genotyping database that reports multiple populations, which can summarize information into the most useful statistical indices (allele frequencies, heterozygosity, *Fst* [7-9] and *In* [10]). Scripts generate a data mart from the pre-processed data of the most recent database build in multiple flat files and merges these with the latest dbSNP build (mid 2008: #129) to acquire additional SNP information such as chromosome, position, validation status, gene, reference allele, and ancestral allele derived from the Chimpanzee genome. Although each query would normally demand its own processing resources, pre-processing the data

Table 1: Main characteristics of the SNP databases currently accessed by SPSmart

Database	SNPs	Populations	Web Address
HapMap Phase II	4,098,495	4	http://hapmap.org/genotypes/
HapMap Phase III	1,614,792	11	http://hapmap.org/genotypes/
Perlegen	1,580,349	3	http://genome.perlegen.com/
Stanford HGDP	660,918	52	ftp://ftp.ceph.fr/hgdp_supp1/
Michigan HGDP	525,910	33	ftp://ftp.ceph.fr/hgdp_supp2/

BMC Bioinformatics 2008, **9**:428 http://www.biomedcentral.com/1471-2105/9/428

solves the major computing issues, so serving all these calculations through the web was the next logical step as shown on the workflow described on Figure 1.

All the SNP repositories that have been processed have their raw data freely available for bulk downloading. Their genotypes are compressed in plain-text files arranged in tables, differing only in the structure of those tables: Hapmap, Perlegen and the Stanford CEPH present their data in a SNP per row basis, with the samples in columns, while the Michigan CEPH data is arranged following the structure format (that is a sample in each couple of rows, with each SNP's allele 1 and allele 2 contained on the first and second sample line respectively.

The pre-processing engine has three major aims: (i) to rewrite the data into a more appropriate format for population combinations, (ii) to build all the possible summaries that may be requested by populations, and (iii) to merge the genotype data with dbSNP information. The output of the population pre-processing of any repository is a SNP list containing all the SNPs found in the database and files containing all the calculated statistical indexes per SNP and per population. The SNP list is used to retrieve additional dbSNP information through a collateral pre-processing engine, aiming to enrich the data mart. Placing the data into a relational database allows quick presenting of these pre-calculated results through the web interface, so a combination of those summaries for the requested population combinations is all that is required.

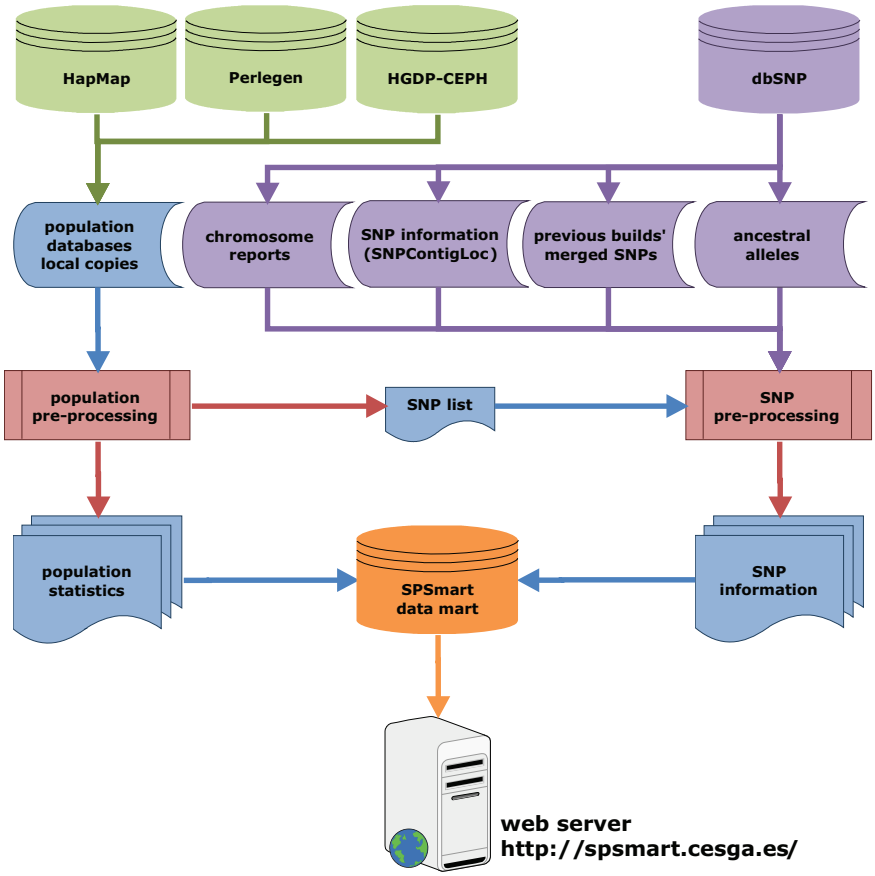


Figure 1
Flowchart of processes implemented in SPSmart. The underlying SPSmart processing engine is capable of dealing with virtually any database that contains genotypes grouped by populations. Any dataset is summarized into common populational statistical indexes, and then combined with dbSNP additional information in order to improve the online data browsing experience.

As the major population groups can be expected to be queried often their combinations are pre-processed, hence statistical parameters of the populations that constitute the group are pre-calculated and stored too.

Including a new dataset is fairly straightforward: the format of the new dataset is analysed and, if needed, the reading module of the population pre-processing script is adapted. Once the data is read, the data is internally structured in identical fashion to the other datasets and subsequent pre-processing is executed in the same way. Updating incorporated datasets is easier still since no script adaptation is required, just a new pre-processing run that takes from several minutes to a few hours depending on data size.

Programming languages

In order to satisfy the predicted needs demanded by SPSmart, a variety of programming languages were used in development. Perl was selected for all the pre-processing scripts, as it is recognized as one of the fastest programming languages for text-processing. The optimized regular expressions engine of Perl allows fast and reliable digesting of flat text files, so the resulting scripts are very powerful but undemanding in terms of resource consumption. To access the pre-calculated data mart and for presenting the data on each client, allowing user interaction through a web browser, the combination of PHP, MySQL and HTML was chosen, with due regard for common web standards such as CSS and XHTML, maximizing independence across different browsers. In addition Javascript was used to facilitate user input on the search section, to hide and show the results tabs, and for some minor design details. In combination the languages used produced a final web interface capable of rapid presentation of results while generating light pages for low bandwidth users.

Results

All the data is stored in a relational database such that each available genotyping dataset has a table of SNPs with their descriptive information, a table of genes and the SNPs present in them, population data for listing purposes, and a table per population and per population group containing their summarized statistics. We have also built a web interface to allow browsing the data mart and to structure queries by populations where users can select any combination of populations within a database to obtain sets of comparative data and statistical metrics.

Populations can be combined on the advanced search page with the possibility of comparing up to five different population combinations. The user should be aware that the sampling approaches and sample sizes of each repository are different and this must be taken into account when inter-population comparison are made, even if populations carry identical descriptions.

Tool usage

A SNP search can be performed by entering a list of SNPs, defining a chromosome region, or entering a gene name. The query reports results in a paginated manner comprising a frequencies tab for visual inspection of allele frequencies plus a statistics tab for a detailed and downloadable table containing statistical information for the queried SNPs for the defined population groupings.

Figures 2, 3 and 4 show three examples of queries that illustrate comparisons between databases and populations that illustrate the flexibility of comparing data from different sources with unified queries. Firstly (Figure 2), a comparison of population variability reported by Perlegen and Hapmap for SNP rs6824418 highlights a major discrepancy between these databases for Europeans: listed as an allele frequency of 0.065 in Perlegen (EUA) and 1 in Hapmap (CEU). Secondly (Figure 3), analysis of a fixed difference SNP such as rs2789823 provides a means to gauge European-African admixture in African Americans by comparing the reported variability in the Perlegen African American sample (AFA) with that of the Hapmap African sample of the Yoruba of Ibadan, Nigeria (YRI). Lastly (Figure 4), a SNP collected as a Native American informative ancestry marker (rs4698702) that has poor quality flanking sequence and cannot readily be genotyped may be substituted by performing a region search on chromosome 4 from 18230000 to 18250000 and locating a better alternative SNP (rs10012227) with a near identical frequency distribution due to association.

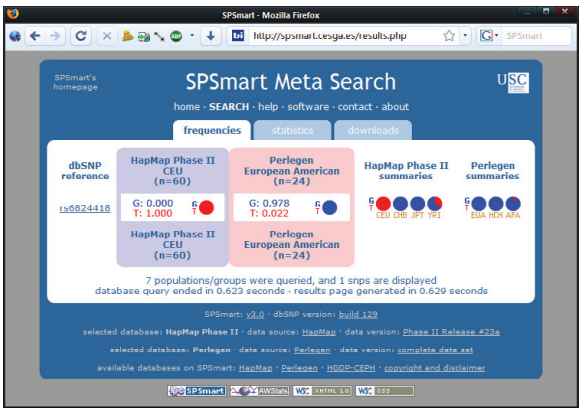


Figure 2
Finding discrepancies among databases. SNP rs6824418 data from Perlegen and HapMap indicating discrepant allele frequency estimates for populations EUA and CEU (European American and CEPH European respectively).

BMC Bioinformatics 2008, **9**:428 http://www.biomedcentral.com/1471-2105/9/428

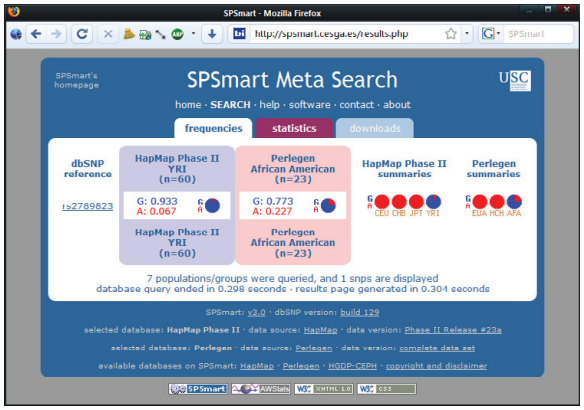


Figure 3
Comparing similar populations in different databases. SNP rs2789823 data from Perlegen and HapMap illustrating a fixed difference SNP that shows the degree of European:African admixture in the African American population sample of Perlegen (AFA) compared to the HapMap African population: the Yoruba of Ibadan, Nigeria (YRI).

Updating frequency

The frequency of updates of the available datasets and how these are incorporated in SPSmart depends largely on the databases themselves. If they are updated at their origin SPSmart can refresh the contents within a day of notification, however the main reference database of HapMap changes about twice a year, while Perlegen and CEPH databases are not expected to change at all. In addition, the SPSmart is designed in a way that allows easy implementation of new functionality. For instance, it is straightforward to implement new statistical indices as well as new filtering properties in response to user demands or changes in statistical approaches to population analysis reported in the literature.

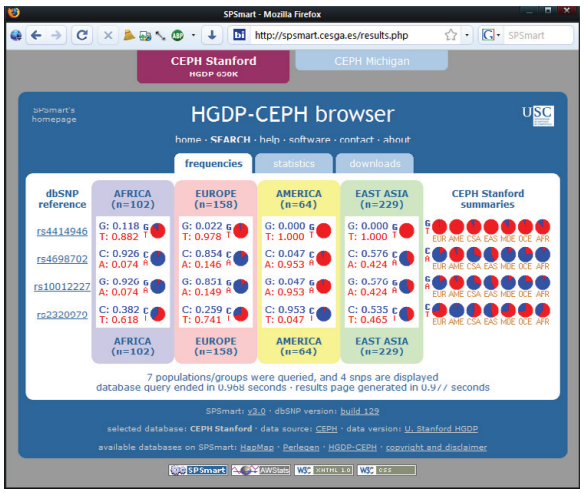


Figure 4
Inspecting a chromosome region. Using a chromosome region search to find an alternative SNP marker with improved quality flanking sequence in the same linkage disequilibrium block (rs10012227 as a better substitute for rs4698702).

Discussion

The major novelty of SPSmart is the ability to combine populations from within a database and to compare populations between different databases, then from both of these operations derive key population variability statistics.

The SPSmart engine http://spsmart.cesga.es/ provides several aids for population genetics and complex disease association studies; areas of research that can both be reliant on comparing SNP variability between populations. The SPSmart engine has successfully processed, and is currently running, a range of data sets encompassing: HapMap release #23a, the Stanford University and Michigan University CEPH-HGDP panels, and the Perlegen SNP data set. We aim to include any major new data sets that collate human SNP variability and implement an extended menu of statistical indices to further aid the population geneticist trying to make sense of the growing wealth of online SNP data.

Conclusion

There are a very large number of autosomal SNP genotypes freely available in the literature and databases. Each database resource presents its own storage procedures and formats and therefore it is difficult for a researcher to use and combine the data from these resources. To our knowledge, this is the first web tool that allows the combination of different datasets of human SNPs and population groups, and to compute statistical indices of interest for medical and population genetics investigations.

Availability and requirements

- Project name: SPSmart
- Project home page: http://spsmart.cesga.es/
- Operating system: Platform independent.
- Programming languages: Perl, PHP, SQL, HTML and JavaScript.
- Type of access: this web tool is freely available for non-commercial use.

Authors' contributions

JA carried out the design, programming and implementation of the software, and drafted the manuscript. AS, CP, and AC participated in the design of the software and the databases selection, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The grants from the Xunta de Galicia (PGIDIT06PXIB208079PR) and Fundación de Investigación Médica Mutua Madrileña awarded to AS partially supported this project. Thanks to Albert Vernon Smith, Lalitha Krishnan and Marcela K Tello-Ruiz of HapMap for their long-standing interest and support, and to Juan Villasuso and Natalia Costas of Centro de Supercomputación de Galicia (CESGA) for their web hosting service and their valuable technical support.

References

1. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437(7063):**1299-1320.
2. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site.** *Genome Res* 2005, **15(11):**1592-1593.
3. Peacock E, Whiteley P: **Perlegen sciences, inc.** *Pharmacogenomics* 2005, **6(4):**439-442.
4. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al.*: **A human genome diversity cell line panel.** *Science* 2002, **296(5566):**261-262.
5. Rosenberg NA: **Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives.** *Ann Hum Genet* 2006, **70(Pt 6):**841-847.
6. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, *et al.*: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319(5866):**1100-1104.
7. Gillespie JH: **Population genetics: a concise guide.** Baltimore, Md: The Johns Hopkins University Press; 1998.
8. Hartl DL, Clark AG: **Principles of population genetics.** 3rd edition. Sunderland, MA: Sinauer Associates; 1997.
9. Long JC: **The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics.** *Genetics* 1986, **112(3):**629-647.
10. Rosenberg NA, Li LM, Ward R, Pritchard JK: **Informativeness of genetic markers for inference of ancestry.** *Am J Hum Genet* 2003, **73(6):**1402-1422.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

pop.STR - An online population frequency browser for established and new forensic STRs.

Amigo J, Phillips C, Salas T, Fernández Formoso L, Carracedo A, Lareu M

Forensic Science International: Genetics Supplement Series. 01/2009; 2(1):361-362.

Como parte de la salida del consorcio original SNP*for*ID creamos una base de datos de frecuencias alélicas en línea de libre acceso para polimorfismos de un solo nucleótido (SNPs) que llevó al desarrollo del explorador de SNPs SPSmart. Este sistema permite la combinación de las poblaciones hasta en 5 grupos definidos por el usuario para generar métricas poblacionales de interés forense (por ejemplo, la heterocigosidad y las estadísticas *F*) a partir de comparaciones poblacionales pre-calculadas. Los primeros 52 marcadores SNP*for*ID se extendieron a los 650.000 SNPs tipados sobre el panel CEPH de diversidad genómica humana (CEPH_HGDP) por las universidades de Stanford y Michigan en 51 poblaciones, más 3 millones de SNPs de la fase III de HapMap y Perlegen de 11 y 3 poblaciones respectivamente.

Hemos adaptado el marco de los datos de SPSmart y sus algoritmos para crear el explorador pop.STR que permite el análisis de frecuencias alélicas de STRs forenses de forma idéntica a SPSmart. Basamos los primeros datos en estudios internos de frecuencia de los 15 STRs de AmpflSTR Identifiler 1 y los 5 nuevos STRs ESS en las mismas 52 poblaciones del CEPH-HGDP utilizado por los estudios de SNPs de Stanford y Michigan.

Forensic Science International: Genetics Supplement Series 2 (2009) 361–362

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/FSIGSS

Research article

pop.STR—An online population frequency browser for established and new forensic STRs

Jorge Amigo [a], Christopher Phillips [a,b,*], Toño Salas [a,b], Luís Fernandez Formoso [b], Ángel Carracedo [a,b], Maviky Lareu [b]

[a] *Genomics Medicine Group, CIBERER, University of Santiago de Compostela, Santiago de Compostela, Spain*
[b] *Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain*

ARTICLE INFO

Article history:
Received 25 August 2009
Accepted 27 August 2009

Keywords:
STR
Identifiler
ESS
MiniFiler
Online databases
Population genetics
SPSmart

ABSTRACT

We recently produced allele frequency data for 20 forensic STRs in more than 50 worldwide populations. The STRs characterized include 5 new European Standard Set (ESS) STRs where novel low frequency and intermediate-repeat genotypes found were confirmed by sequence analysis. Data for the 20 STRs has been collated into an open-access online frequency browser at: http://spsmart.cesga.es/popstr.php that allows users to combine populations into groups to generate re-calculated allele frequency estimates from the merged genotype data. The flexibility to combine populations in this way and the graphical summaries provided for each marker's allele frequencies offers the forensic analyst an informative system to consult STR variability in a global range of populations.

© 2009 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

As part of the original SNP*for*ID Consortium output [1] we created an open-access online allele frequency database for single nucleotide polymorphisms (SNPs) that led to the development of the SPSmart SNP browser [2–4]. This system allows the combination of populations into 1–5 user-defined groupings while generating population metrics of forensic interest (e.g. heterozygosity and *F*-statistics) from pre-calculated population comparisons. The initial 52 SNP*for*ID markers were extended to 650,000 SNPs from the Stanford and Michigan University's CEPH human genome diversity panel (CEPH-HGDP) studies of 51 populations, plus 3 million SNPs from HapMap phase III and Perlegen data repositories of 11 and 3 populations respectively.

We have adapted the SPSmart data framework and algorithms to create the *pop.STR* browser which allows the analysis of forensic STR allele frequencies in identical fashion to SPSmart. We based the first data build on in-house frequency studies of the 15 STRs of AmpFlSTR Identifiler® and the 5 new ESS STRs in the same 52 populations of the CEPH-HGDP used by the Stanford and Michigan SNP studies.

2. Results

With the plethora of STR combinations now available we decided to present a simple homepage menu of sets, beginning with the 6 Applied Biosystems AmpFlSTR kits: Identifiler®; ESS; MiniFiler®, SGM Plus®, and Profiler Plus®. This forms a selection system so users can choose the appropriate set from the 20 STRs currently in *pop.STR*.

Compared to a graphic summary of the binary allele variation of SNPs as simple pie charts in SPSmart, the presentation of the multiple alleles of STRs within a web-page format creates a considerable challenge—particularly those of STRs such as FGA or D21S11. We used a simple 'medal ribbon' chart per population group with singleton and rare alleles shown as clear yellow segments amongst darker colours, with the allele frequencies themselves listed in a separate downloadable page. An example output for FGA is shown in Fig. 1 for African and European population groupings. Although conventional bar-chart options will be available in *pop.STR* output, we feel these are not well suited to extensive allele ranges or multiple population comparisons.

* Corresponding author at: Genomics Medicine Group, CIBERER, University of Santiago de Compostela, Calle San Fransisco S/N, 15705 Santiago de Compostela, Spain. Tel.: +34 981 582 327; fax: +34 981 580 336.
E-mail address: c.phillips@mac.com (C. Phillips).

1875-1768/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigss.2009.08.178

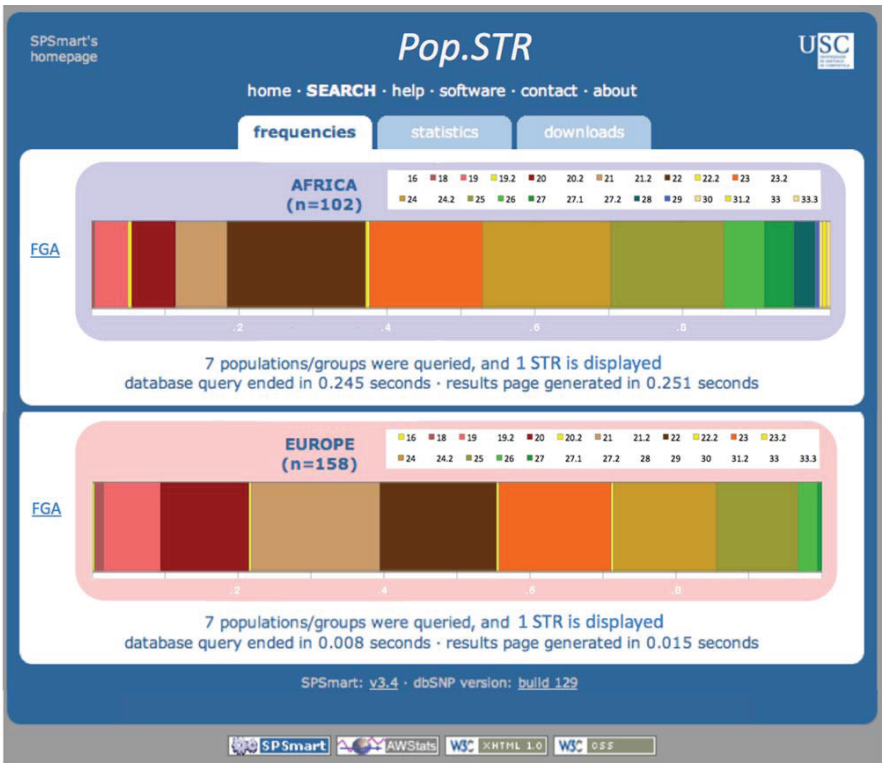


Fig. 1. Graphic representation of allele frequency variation in the STR FGA for groupings of 7 African and 7 European populations from CEPH. Singleton or rare alleles share the same yellow or light orange labels to contrast with more common alleles in each case.

3. Discussion

A potential issue with plans to extend the current population scope of *pop.STR* is the paucity of information for the 5 ESS markers: D1S1656; D2S441; D10S1248; D12S391 and D22S1045. These loci await comprehensive population surveys and we aim to both collate data for the new STRs as well as add published allele frequencies for the established STRs of Identifiler® to expand the coverage of *pop.STR* still further.

Conflict of interest

None.

References

[1] http://www.snpforid.org/.
[2] J. Amigo, C. Phillips, M. Lareu, Á. Carracedo, The SNP*for*ID browser: an online tool for query and display of frequency data from the SNP*for*ID project, Int. J. Legal Med. 122 (2008) 435–440.
[3] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinform. 9 (2008) 428–433.
[4] J. Amigo, C. Phillips, A. Salas, Á. Carracedo, Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes, BMC Bioinform. 10 (Suppl. 3) (2009) S5.

Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes.

Amigo J, Phillips C, Salas A, Carracedo A

BMC Bioinformatics. 02/2009; 10 Suppl 3:S5.

Están disponibles gratuitamente bases de datos que contienen grandes cantidades de SNPs (polimorfismos de un solo nucleótido) para los investigadores interesados en aplicaciones de genética médica y/o poblacional. Si bien muchos de estos repositorios de SNPs han implementado herramientas de recuperación de datos para uso general de extracción, ellas solas no pueden cubrir todo el espectro de necesidades de la mayoría de los estudios de genética médica y de poblaciones.

Para resolver esta limitación, hemos construido repositorios estáticos personalizados internamente de los datos crudos proporcionados por las mayores bases de datos públicas. En particular, para análisis de genética de poblaciones basada en genotipos hemos construido un conjunto de secuencias de comandos de procesamiento de datos que lidian con los datos no procesados procedentes de las bases de datos más importantes de variación de SNPs (por ejemplo HapMap, Perlegen), dividiéndolos en genotipos individuales para luego agruparlos en poblaciones, y fusionarlos con información adicional descriptiva complementaria extraída de dbSNP. Esto permite no sólo la estandarización y la normalización interna de los datos de genotipos obtenidos de repositorios diferentes, sino también el cálculo de los índices estadísticos desde simples estimaciones de frecuencia alélicas a pruebas más elaboradas de diferenciación genética entre poblaciones, junto con la capacidad de combinar las muestras de poblaciones de diferentes bases de datos.

El presente estudio demuestra la viabilidad de la aplicación de secuencias de comandos para el manejo de bases de datos extensas de genotipos de SNPs con un bajo coste computacional, tratando ciertas cuestiones complejas que surgen de la naturaleza divergente y la configuración de los repositorios de SNPs más populares. La información contenida en estas bases de datos también se puede enriquecer con información adicional obtenida de otras bases de datos complementarias, con el fin de construir un repositorio estático dedicado. La actualización de la estructura de datos es sencilla, así como la fácil inclusión de nuevos datos externos y el cálculo de los índices estadísticos suplementarios de interés.

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes

Jorge Amigo*[1], Christopher Phillips[2], Antonio Salas[2] and Ángel Carracedo[1,2]

Address: [1]Spanish National Genotyping Center (CeGen), Genomic Medicine Group, CIBERER, University of Santiago de Compostela, Galicia, Spain and [2]Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Galicia, Spain

Email: Jorge Amigo* - jorge.amigo@usc.es; Christopher Phillips - c.phillips@mac.com; Antonio Salas - antonio.salas@usc.es; Ángel Carracedo - angel.carracedo@usc.es

* Corresponding author

from Second International Workshop on Data and Text Mining in Bioinformatics (DTMBio) 2008
Napa Valley, CA, USA. 30 October 2008

Published: 19 March 2009

BMC Bioinformatics 2009, **10**(Suppl 3):S5 doi:10.1186/1471-2105-10-S3-S5

This article is available from: http://www.biomedcentral.com/1471-2105/10/S3/S5

© 2009 Amigo et al; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Databases containing very large amounts of SNP (Single Nucleotide Polymorphism) data are now freely available for researchers interested in medical and/or population genetics applications. While many of these SNP repositories have implemented data retrieval tools for general-purpose mining, these alone cannot cover the broad spectrum of needs of most medical and population genetics studies.

Results: To address this limitation, we have built in-house customized data marts from the raw data provided by the largest public databases. In particular, for population genetics analysis based on genotypes we have built a set of data processing scripts that deal with raw data coming from the major SNP variation databases (e.g. HapMap, Perlegen), stripping them into single genotypes and then grouping them into populations, then merged with additional complementary descriptive information extracted from dbSNP. This allows not only in-house standardization and normalization of the genotyping data retrieved from different repositories, but also the calculation of statistical indices from simple allele frequency estimates to more elaborate genetic differentiation tests within populations, together with the ability to combine population samples from different databases.

Conclusion: The present study demonstrates the viability of implementing scripts for handling extensive datasets of SNP genotypes with low computational costs, dealing with certain complex issues that arise from the divergent nature and configuration of the most popular SNP repositories. The information contained in these databases can also be enriched with additional information obtained from other complementary databases, in order to build a dedicated data mart. Updating the data structure is straightforward, as well as permitting easy implementation of new external data and the computation of supplementary statistical indices of interest.

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

Background

Many areas of study in genetics, such as human population genetics, are based on genomic diversity, and this variability can only be measured reliably by studying large amounts of data. These studies are only realistically available to big organizations and institutions, and their resulting databases become important data resources for many other genetics projects. Therefore the ability of individual researchers to browse large databases such as HapMap http://www.hapmap.org/ or CEPH http://www.cephb.fr/en/cephdb/ is critical meaning any improvement in data management can be as valuable as the data itself.

The availability of different repositories of human variation represents an aid for researchers on one hand, but an inherent obstacle to their thoughtful combination on the other. Merging data from different databases, even if very similar, represents a major challenge for most users. An important aspect of online data obtained for population genetics studies is that not all databases reference the same material, with each database accessing different populations with their own samples and sample size, so often populations with the same description must be treated separately.

Data marts

A common trend in the field of data repositories is the adoption of data marts, comprising specialized subsets of entire databases designed specifically to answer focused questions [1]. Data marts benefit from a streamlining of the dataset, which avoids querying more data than is needed. This exploits the data stored in a repository, but can use unique structures or summary statistics generated specifically for an area of research. Thus, data marts benefit from the existence of a broadly based database, are less general than a repository, but provide more effective and efficient support for tailored uses of the data.

The use of these data structures is indicated in enterprise-wide data, when operated by departments whose database structures are subject to occasional modifications [2]. The same idea can be ported to any database structure, since it can integrate and consolidate all relevant data into a single data mart without high operational overheads.

Our implementation consists of a large-scale rewriting of all the databases of interest in which we prepare the data to be queried for population genetics purposes, standardizing and normalizing their formats into a common and simplified structure while enriching the data mart with complementary information.

Large genotyping databases

With the current availability and quality of online genome databases it is increasingly feasible to conduct population genetics research using *in-silico* resources [3] as an adjunct to the traditional strategy of sampling populations of interest and genotyping a range of polymorphic markers. Population genetics studies are not co-incidental to the characterization of the human genome or analysis of complex disease but are critical in informing how such analyses should be properly framed with reference to the level of susceptibility, the particular allele frequency distributions and the demographic history shown by a population. Autosomal SNPs, while individually less informative *per se* in population variability terms than e.g. mitochondrial and Y-chromosome loci or autosomal microsatellites, benefit from being densely distributed and well characterized at the sequence and functional level. The characterization of the population variability of SNPs is now catching up with information about their genomic role or their ability to provide landmarks for association studies, promoted in large part by detected differences in linkage disequilibrium patterns between population groups or in admixed populations [4,5]. The evolution of HapMap has illustrated the increased emphasis on extending large-scale genomic projects towards a broader scope of populations studied rather than loci genotyped. HapMap Phase III has almost tripled the study populations from four to eleven while the SNPs studied have been consolidated more than expanded.

Text parsing

The parsing of large amounts of data has been a core approach in bioinformatics from the very beginning. In fact, programming and scripting languages with optimized pattern matching capabilities have been available for a long time (notable examples include Perl and Python), and the use of their built-in regular expressions makes it easier to deal with large numbers of extensive plain text files [6,7]. Current text-mining approaches benefit from these algorithms, which are flexible yet powerful.

All the main public genetics databases provide compressed-format dumps of their data for in-house processing, so once the raw data of interest is available as text files it only requires some familiarity with their format to inspect the required fields from each respective data dump. Although the amount of information to be processed does not generally represents a limitation as the parsing process will be completely automated, efficient programming allows best use of computer resources.

Methods

By building a data mart for population genetics we aimed to improve population data management regardless of size, while consolidating data from different sources by including a number of complex, pre-calculated fields, data structures, and function libraries [8]. Our main goal is to provide a flexible and reliable single repository where the

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

major databases of this field of study can be represented, to form the basis for creating custom queries both within and between each database.

Population based data resources

Many online databases cataloguing human variability provide population information about the samples studied, notably HapMap [9,10], Perlegen [11] and the CEPH foundation [12]. They also provide the raw data that underlines each online database for downloading and local analysis. We have chosen the raw data from the above repositories to be included in our data mart: the stable Hapmap Phase II release 24 and the preliminary released Phase III version, the Stanford and Michigan University CEPH-HGDP (Human Genetic Diversity Panel) SNP genotyping data (although the two datasets are significantly overlapping in SNPs and samples [13]), and the Perlegen dataset. Figures 1, 2 and 3 outline the genotype data of each database, showing the overall amount of data to be managed when building a query.

The datamart created is supplemented with dbSNP [14] data to map all the above databases to the same common reference. This overcomes issues of databases being mapped to different dbSNP builds, and automatically prepares the mart for incorporated any future SNP databases. Table 1 shows the overall number of SNP codes shared among all the processed databases.

Data format analysis

Scrutiny of the publicly available core population-based SNP databases indicates similarities that all share: they are all dumped in plain text arranged by columns, and these columns are divided into descriptive data plus the genotypes themselves. We can use the descriptive information to include as much detail in the data mart as required, but the main aim of processing these files is to read them at the genotype level and to store genotyping calculations into appropriate variables.

Hapmap, Stanford CEPH and Perlegen: tabulated format

Hapmap, Stanford and Perlegen use a similar format for their raw data, comprising genotyped individuals samples *versus* SNPs table, and they only differ in the character used to separate the columns (HapMap Phase II, III and Perlegen use blanks, Stanford tabulations) plus the amount of descriptive columns to characterize each SNP line. Once the amount of descriptive columns is stated, it

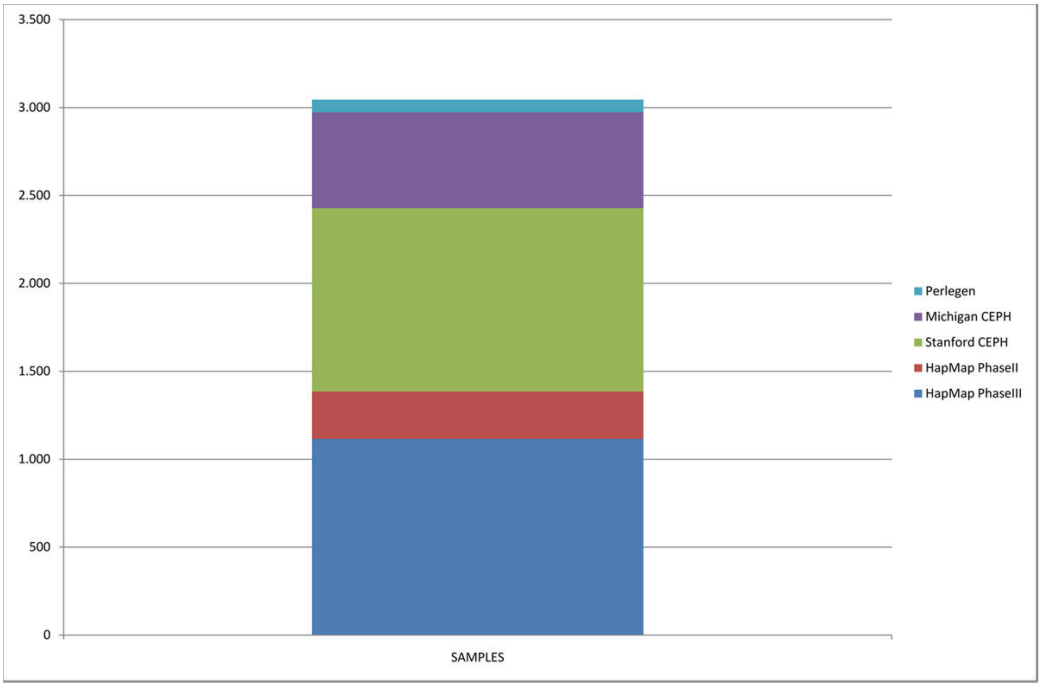


Figure 1

Number of samples present on the data mart. There are 3045 samples represented on our repository. The distribution of the number of samples per database vary from the most ambitious ones such as HapMap Phase III and the Stanford HGDP that contain over 1000 samples each, to others with less variation representation such as Perlegen, with only 71 samples on it.

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

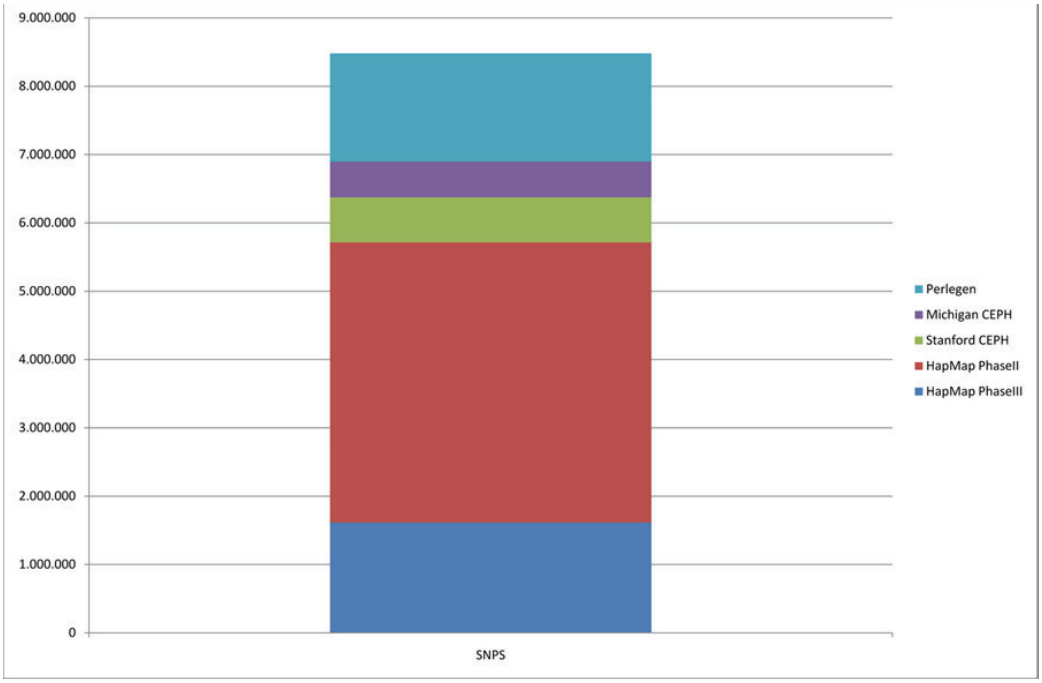


Figure 2
Number of SNPs present on the data mart. Around 8.5 × 10^6 SNPs are processed from the different databases, although these SNPs are not independent. Considering the SNP codes sharing presented on Table 1, where the HapMap Phase II database is the major SNP contributor, the number of distinct SNPs represented on the data mart is close to 4.5 × 10^6.

is possible to jump to the first column of genotypes and read them in full. The format used by the Stanford CEPH comprises samples *versus* SNPs table, without additional information. In contrast, Perlegen provides some extra columns such as chromosome position or available alleles, but does not refer directly to reference SNP codes but to internal ones requiring an auxiliary translation file. Finally, HapMap goes further by providing Perlegen's additional data with additional columns such as strand information or the genotyping protocol used for single SNPs.

Michigan CEPH: structure format

The reference data from the Michigan University is formatted following the requirements of the population stratification analysis software Structure [15], which comprises several header lines containing the SNP list amongst other information, then pairs of lines for each sample containing the first and second allele per SNP in the first and second line of the pair respectively. Parsing this database is therefore completely different from the rest. Once this SNPs *versus* samples data structure can be processed, along with the converse samples *versus* SNPs structure, any upcoming database of genotypes will presumably only require a slight adaptation of either structure reading module, making this system very flexible in terms of data mart expansion.

Design of scripting variables

The biggest challenge of the parsing script design is to allow the data structures to be as versatile as needed but consuming as little computational resources as possible; specifically, in terms of processor running time and memory required. Once the genotypes are highlighted in each file format, the script should store as much relevant information from them as possible for extensive later use. For this reason, it is important to reduce to the minimum the indexing level of the hashes used in the script making them fast to build up and query while minimizing demands on memory.

The data itself is already contained in the raw compressed data files, so the proposed data mart will only contain metadata extracted and calculated from them, such as summarizing counts and percentages. For this reason, all the counts in the script are internally structured in hashes, which are indexed by population and re-used for each chromosome. In this way the script optimizes the mem-

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

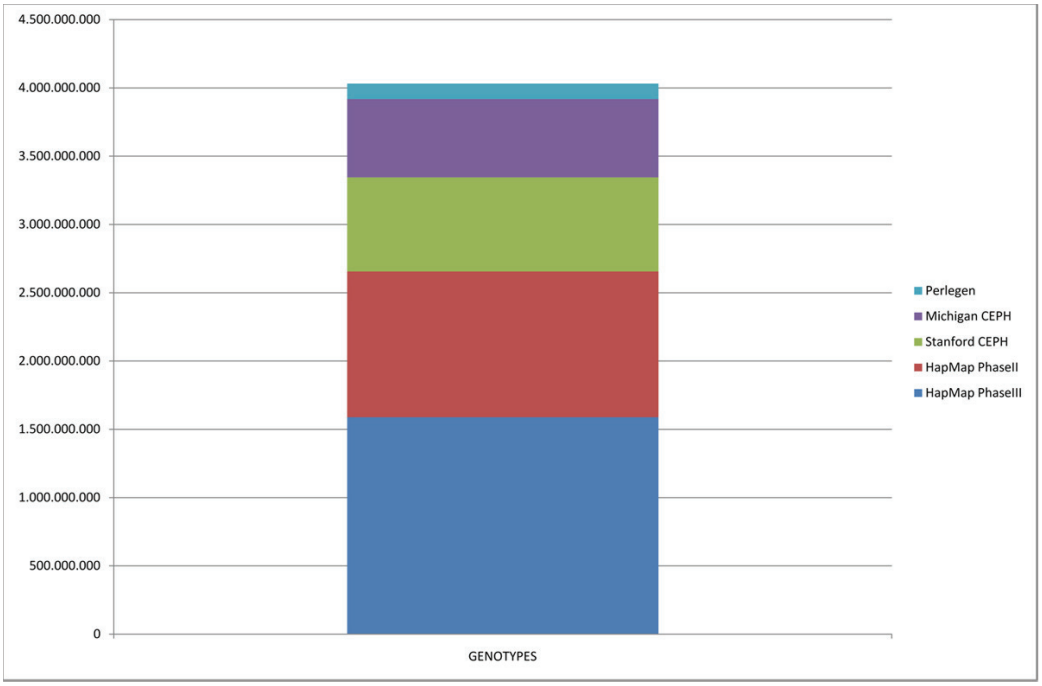


Figure 3
Number of genotypes present on the data mart. A total of above 4 × 10^9 genotypes are summarized on our data mart. Although the number of samples on Perlegen is not very high, its SNP coverage is, transforming this database along with both HapMap phases into the major genotyping contributors with over 10^9 genotypes each.

ory consumption, and at the same time allows structuring of results into populations. By storing this metadata, which can be as extensive as desired, we have constructed a very detailed data mart queried independently of the original data and fully focused on our field of interest.

Results and discussion

All processed data is placed into a MySQL database to contain all statistical results simply indexed by SNP code. The main challenge of the data mart design is the formation of a global design that allows the combination of SNP resources with different stuctures. It involves processing large SNP databases that require an efficient data indexing to minimize access times and memory requirements while retaining the versatility of the created scripts for new databases.

Although some databases may contain extensive additional information about SNP loci, it is worth noting that we focused on genotypes alone so the data files indicated on Table 2 represent the minimum number of files needed to build the population data mart described. Therefore files contain raw genotypes, SNP code translations (Perlegen data dumps contain internal codes only)

Table 1: Shared SNPs among the different databases.

	dbSNP	HapMap II	HapMap III	Perlegen	Stanford
HapMap II	4097825				
HapMap III	1611772	1549224			
Perlegen	1585334	1267374	682386		
Stanford	660823	660060	658947	294956	
Michigan	525859	525307	525011	242910	525909

The number of common SNP codes has been taken from direct inspection of the raw data, after being mapped to the same reference SNP code by merging the dbSNP information. The numbers shown are the SNP codes that match among all the databases processed: dbSNP build 129, HapMap Phase II and III, Perlegen, and the human genome diversity panels from the universities of Stanford and Michigan.

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

Table 2: Raw data resources needed for the data mart creation.

DATABASE	RESOURCE
dbSNP	reference alleles from b129_SNPContigLoc_36_3.bcp.gz at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/ ancestral alleles from SNPAncestralAllele.bcp.gz and Allele.bcp.gz at ftp://ftp.ncbi.nih.gov/snp/database/organism_data/human_9606/ and ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/shared_data/ chromosome positions, validation status and loci from reports at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/chr_rpts/ merged snps from RsMergeArch.bcp.gz at ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/
HapMap II	http://ftp.hapmap.org/genotypes/2008-10_phaseII/fwd_strand/non-redundant/
HapMap III	http://ftp.hapmap.org/genotypes/2008-07_phaseIII/hapmap_format/forward/
Perlegen	http://genome.perlegen.com/browser/download.html
Stanford	ftp://ftp.cephb.fr/hgdp_supp1/
Michigan	ftp://ftp.cephb.fr/hgdp_supp2/GENO/

Processing each database requires its own raw data. The following are all the file sets, indicated by database, that were processed in order to build the described data mart. All of them are publicly available online.

or information about the samples familial relationships with others in the same set, required when building independent statistics.

dbSNP, as a reference database

In the first instance mapping information is taken from dbSNP to form the reference template for other databases. The data for each SNP is obtained by parsing the files described in Table 2 to generate a list of SNPs per chromosome with descriptive information from dbSNP, such as the ancestral allele, to characterize each locus.

Processing the dbSNP database is perforrmed once per build and takes ~8 hours on a standard computer. The data is then merged with the SNP list of population databases included in the data mart, taking 10 to 15 minutes per database. This process is run when a population database or dbSNP is updated.

Unifying chromosome mapping and SNP codes

There are two main problems when trying to compare the same SNP information from different databases: firstly, although a SNP may be named equally in multiple repositories its chromosome location may not coincide due to mapping changes between dbSNP versions; secondlt the SNP may just be named differently. The first issue will only affect queries by location, but it can be easily solved by always using the chromosome location from a chosen dbSNP build, not necessarily the last one, as consistency is the only requirement. However use of different SNP codes to refer to the same locus requires translating them into a common reference, either because of using internal SNP codes as Perlegen does, or because of being mapped to an older dbSNP build not reflecting the latest SNP label merges or renames.

The logical way to solve both problems is to map all the databases to the most recent dbSNP build. This will not only permit multiple chromosome positions, but also allows the data mart to contain updated SNP codes. By parsing the locations from the chromosome reports of the last dbSNP build and merging information from previous builds, we generate a mapping reference to use with the SNP lists from each processed database ready for placing into the data mart.

The oriented reference allele

Although the major issues for SNP comparison are addressed, we also wanted to include a system to unify the strand interogated by the reported genotyping assay,. Although the strand information was part of the dbSNP raw data, a proportion of SNPs in repositories were genotyped on the complementary strand and required a mapping reference for allele calls. Therefore we opted to use the reference allele. The reference allele is arbitrary when working with genotypes, but it is still used to sort the genotyping alleles. So from the reference allele the direction is discerned and adjusted appropriately in each database. This orientation reference can be used to adjust the reporting of alleles from different databases that detect opposite strands.

Data mart creation and structure

The set of scripts designed in the present study is able to process the major SNP databases and to generate a normalized data mart for them all, using relatively few resources. The most critical script processes the raw data from each database, as it has to be powerful but flexible. The script must read databases in the given format and calculate several statistical indices.

There are two main reading modules to handle samples *versus* SNPs or SNPs *versus* samples formats, and generate data uniform data structure. The statistical module follows and creates all the statistical summaries, from the simplest allele frequency estimates to more complex met-

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

rics of population differentiation, by building simple internal counts and summarizing them at the end. Finally, a writing module is in charge of generating a CSV file per population plus a list of the SNPs and the populations processed.

Once all the summarized data is written on these CSV files, a small script merges the SNP lists of each database with the additional SNP descriptive data from dbSNP. The merging script generates extra CSV files if relevant, such as the SNP codes merged or SNPs removed after comparison to dbSNP. The CSV files are loaded into a MySQL database by another script that generates the SQL commands to create each table definition, with SNP codes indexed to speed up any later inspection.

Maintaining the data mart

The frequency of updates of the databases currently accessed is very low while dbSNP updates annually. HapMap data is rebuilt twice a year in contrast to Perlegen and the two CEPH databases, which appear to be static. Therefore new HapMap releases invlove running our complete pipeline (~2 hours on a standard computer), but a new dbSNP release requires only the merging script on each database SNP list, and updating only the SNP tables of the data mart (~1 hour for all the databases present).

The interdependency of each database is outlined in Figure 4, where only the HapMap Phase III substructure of the data mart is shown. Each database replicates this structure, illustrating how compartmentalized the data mart is. Therefore it would be easy to add a new SNP database or to update existing ones.

We have implemented and summarized the most common population statistical indices. If new statistical indices are required the script processing the raw data needs to be updated, the statistical module would require modification, and the whole set of databases re-processed to reflect these changes. This represents a major update effort, as the entire data mart has to be rewritten, but in fact only requires a day of processing due to the flexibility the processing pipeline developed.

Consumption of resources

One of the main aims of this project was to develop a tool for extracting the most relevant data from large SNP databases in such a way that a non-expert user can successfully complete the task using a standard computer. Firstly we focused on the memory requirements so the variables structure was designed to be as simple as possible, and secondly we optimized the main internal loops present in the script enabling the running time to be reduced to a minimum. This optimization led to the results displayed in Figures 5 and 6, indicating that all five major reference databases are processed in just 12 hours in total on a standard computer (although these are completely independent tasks), and that the maximum amount of free RAM needed for the computer is 1.8 GB (due mainly to the combination of a large number of samples and populations in the Stanford and Michigan CEPH data). Without considering the data that has to be extracted from dbSNP to be used as the mapping reference, the total number of genotypes currently contained in the data mart is above 4×10^9. The total disk space needed is 16 GB, which is relatively small considering the size of the databases contained, and that half of that size is dedicated to the storing of the raw genotypes retained for user downloads.

Posterior data mart use

The creation of a specific and smaller repository from larger ones was motivated by the need to avoid processing irrelevant data present in many repositories, as well as fully controlling its format and structure. We relied on text mining approaches when processing large variation repositories in order to obtain all available genotyping data for each SNP, and then summarizing that information to store it in a lean yet flexible data mart.

As an illustration of the marts use, a researcher might want to study the admixture of European and African populations in SNP rs2789823, amongst others, by querying all the variation repositories available. Normally this would mean browsing each database in turn while adapting to their different interfaces and data formats, and annotating the relevant information. Our tool alternatively mines available information for the SNP, and pre-calculates the relevant statistical indices that allow interpretation of the SNP variability. Therefore only the populations need to be selected. In the example given, our datamart rapidly creates output that indicates the Perlegen African American population at rs2789823 has a high degree of European admixture when compared with the HapMap African population (Yoruba of Ibadan, Nigeria).

Once the data has been summarized and organized, the next logical step is to build custom tools to query the new data structure and generate statistical metrics. The web-based tool SPSmart [16] has been designed with the aim of exploiting the previously generated data. It is therefore an online interface for the data mart built from the previously described reference databases, and is mainly focused to meet the routine analysis demands of population geneticists. These include comparing populations from different databases, inspecting allele frequencies across current available population databases, or studying the genetic differentiation amongst various combinations of populations.

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

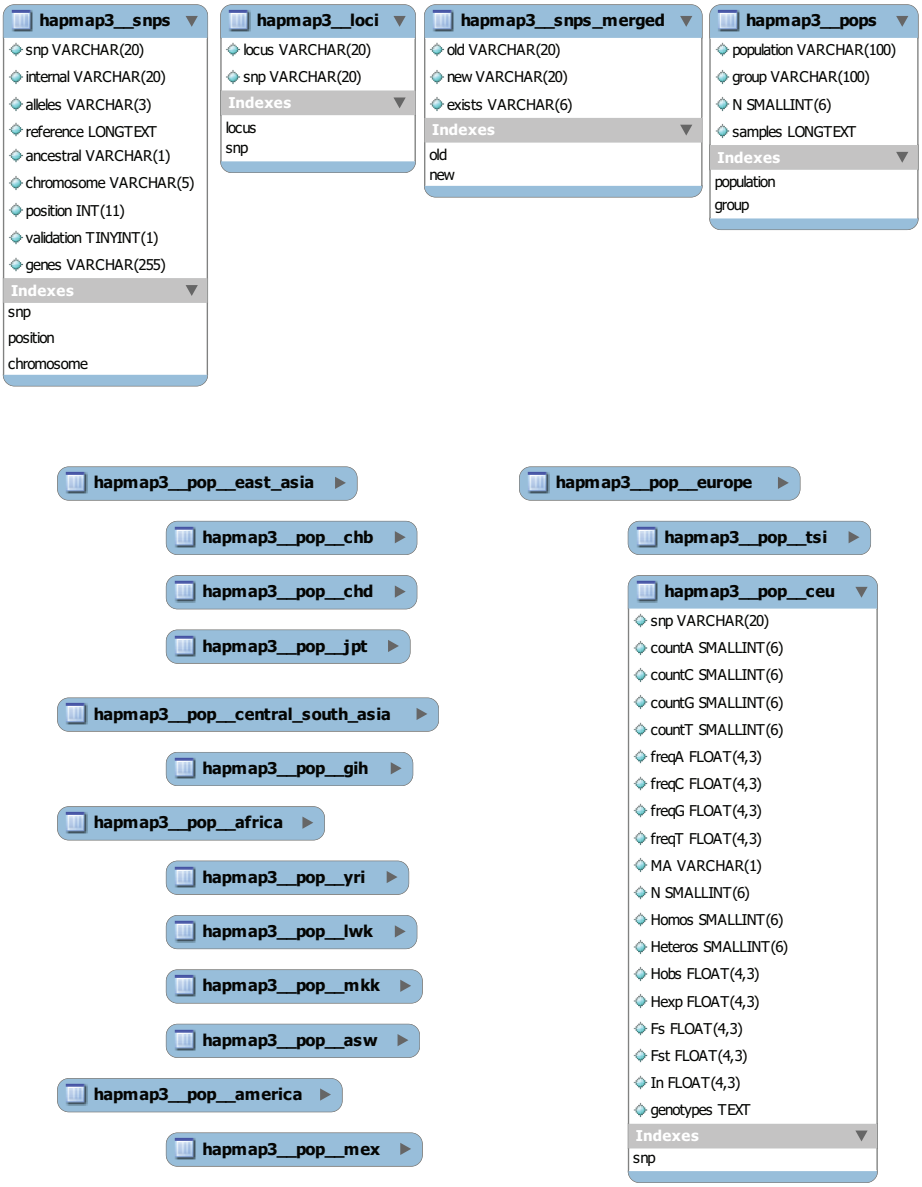


Figure 4
Data mart tables for the HapMap Phase III database. Each database summarized is present on the data mart as a set of tables containing descriptive SNP information and population specific calculations. Every database will have all the table structures expanded at the top of the image, and the amount of the population specific ones shown with the "__pop__" label will depend on the amount of populations covered by the database. Only the CEU population table structure has been expanded on the image, but the rest of the population tables share the same structure that allows filling each population SNP with all the available counts and calculations performed by the raw data processing script.

Future work

Since processing each database is completely independent from the rest, we can distribute the work through a parallel computer or through a grid system. Due to the large size of the raw data to be processed, currently around 2 GB of compressed text files, we have chosen the first option in

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

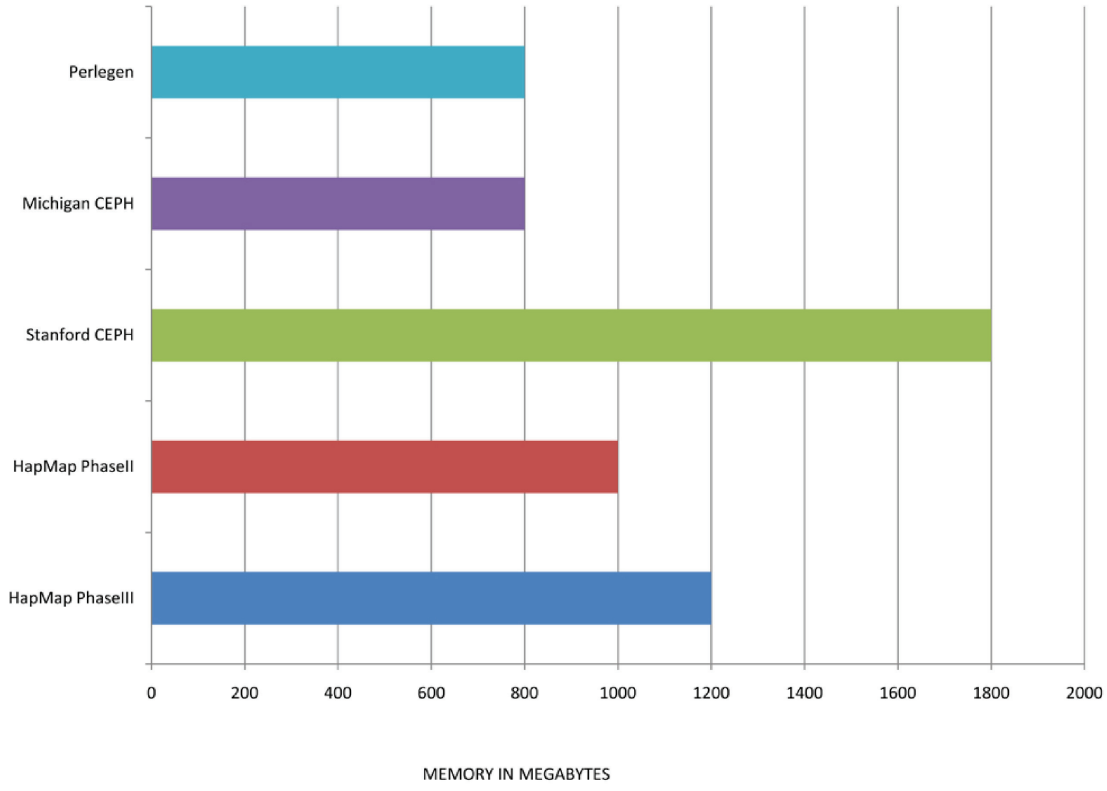


Figure 5
Memory needed to process each database. The memory required to deal with different databases depends not only on their number of samples and SNPs, but also on the raw data files structure. Although not more than 1 GB of memory has been enough for most the databases, the Stanford data needed some more due to its high population coverage. The fact of containing so many samples and representing so many populations on single files per chromosome forced the processing script to store plenty of indexed information that demanded high computational resources. The optimized design of the variables, along with the strict memory handling of the script, minimized this issue never requiring more than 2 GB.

order to minimize the latencies that data transfers through the network may cause. We are currently implementing our pipeline on a shared memory node system with SMP NUMA architecture available at the Supercomputing Centre of Galicia (CESGA; http://www.cesga.es/). We can take advantage of the fact that the CESGA also hosts our data mart and the networking among the different machines is optimal. Our first tests show that this type of implementation is fully reliable, as we are obtaining similar benchmarking results compared to local runs, and our goal is to build a static pipeline structure on this supercomputer that would not only dramatically reduce our dependency on the network for large data uploads when updating any database, but also have a dedicated machine for our needs.

We have designed the data mart for handling high-throughput SNP genotyping data in such way that allows easy expansion, not only in terms of the databases accessed, but also in terms of new statistical indices that will be of interest to researchers. Thus, new repositories can be added to the data mart structure simply by adapting the reading module, while implementation of new statistics can easily be accomplished by adding the necessary formulae to the data and writing module of the processing script.

Conclusion

There is a wide range of autosomal SNP genotypes resources freely available in public databases, each presenting their own storage procedures and formats. Due to this lack of homogeneity it is difficult to adapt to each

BMC Bioinformatics 2009, **10**(Suppl 3):S5 http://www.biomedcentral.com/1471-2105/10/S3/S5

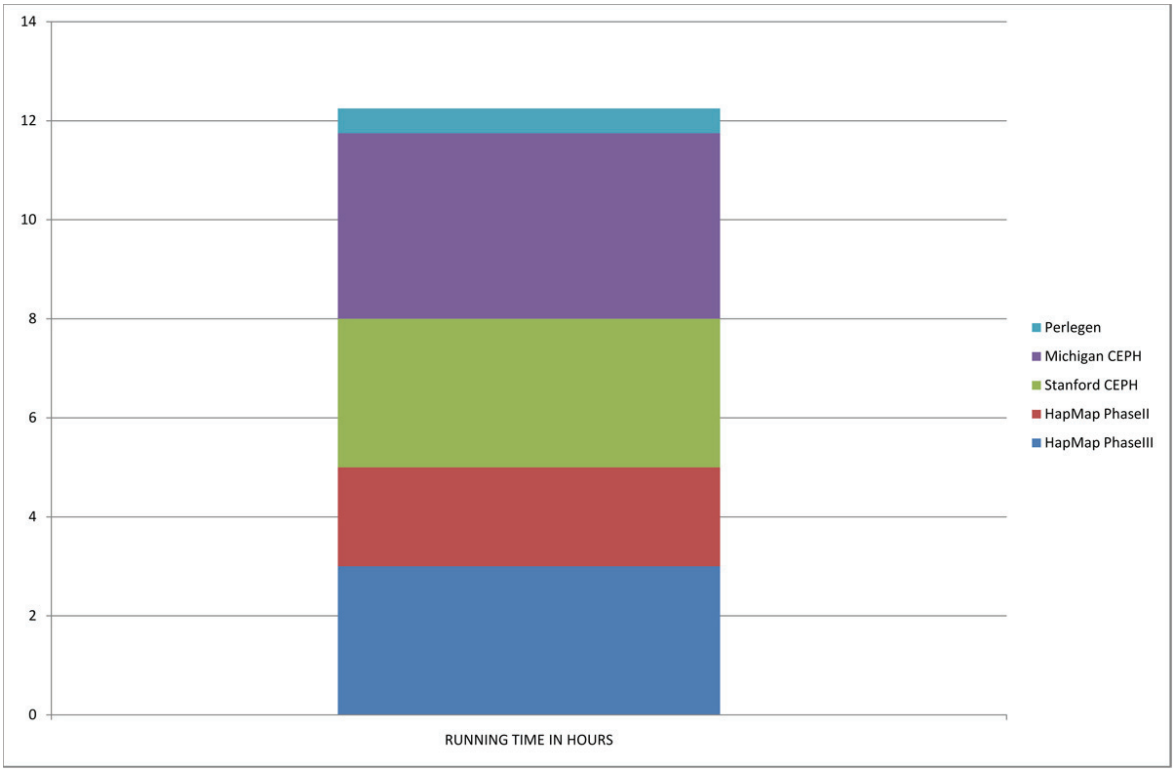


Figure 6
Databases' processing times. Cumulative time is presented, taking 12 hours to deal with all the available databases, although each task is independent from the others and therefore can be run in parallel. The maximum time would then be the 4 hours that the Michigan data needs to be processed.

database interface requirements and, with the software currently available, it is impossible to combine such disparate results for meta-analysis. Here we have shown that it is viable and highly efficient to work directly with the raw data of each repository to build data mart tailored to population genetics needs that uses in-house computational resources.

Adapting these major variation repositories in such a lean and versatile manner is a novel and ambitious approach to SNP based population genetics analysis, as it deals with a vast amount of information but is able to generate a flexible resource to obtain population statistics of any population or custom population group. Once the raw data is pre-processed, it is relatively easy to compute new statistical indices of interest and where new inter-population comparisons can be made. In addition, the strategy presented here allows the direct combination of different SNP genotyping repositories in a straightforward manner.

Availability and requirements

• Project name: SPSmart

• Project home page: http://spsmart.cesga.es/software.php

• Operating system: Platform independent.

• Programming languages: Perl and SQL.

• Type of access: all the scripts provided to generate the described data mart are freely available for non-commercial use.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JA carried out the design and implementation of the described data mart, as well as the programming of the

text parsing engine, and drafted the manuscript. AS, CP and AC participated in the design of the software and the database selection, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Our thanks to the helpdesks of the Foundation Jean Dausset – Centre d'Etude du Polymorphisme Humain, dbSNP and HapMap for their valuable guidance and advice on the available raw data. Two grants from the Fundación de Investigación Médica Mutua Madrileña awarded to AS partially supported this project.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 3, 2009: Second International Workshop on Data and Text Mining in Bioinformatics (DTMBio) 2008. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S3.

References

1. McNamee LA, Launsby BD, Frisse ME, Lehmann R, Ebker K: **Scaling an expert system data mart: more facilities in real-time.** *Proc AMIA Symp* 1998:498-502.
2. Arnrich B, Walter J, Albert A, Ennker J, Ritter H: **Data mart based research in heart surgery: challenges and benefit.** *Stud Health Technol Inform* 2004, **107(Pt 1):**8-12.
3. Phillips C: **Online resources for SNP analysis: a review and route map.** *Mol Biotechnol* 2007, **35(1):**65-97.
4. Rosenberg NA: **Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives.** *Ann Hum Genet* 2006, **70(Pt 6):**841-847.
5. Smith MW, O'Brien SJ: **Mapping by admixture linkage disequilibrium: advances, limitations and guidelines.** *Nat Rev Genet* 2005, **6(8):**623-632.
6. Apostolico A, Galil Z: **Pattern matching algorithms.** *New York: Oxford University Press*; 1997.
7. Dougherty D: **Sed & awk.** *Sebastopol, CA: O'Reilly*; 1990.
8. Isken MW, Littig SJ, West M: **A data mart for operations analysis.** *J Healthc Inf Manag* 2001, **15(2):**143-153.
9. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437(7063):**1299-1320.
10. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site.** *Genome Res* 2005, **15(11):**1592-1593.
11. Peacock E, Whiteley P: **Perlegen sciences, inc.** *Pharmacogenomics* 2005, **6(4):**439-442.
12. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, *et al.*: **A human genome diversity cell line panel.** *Science* 2002, **296(5566):**261-262.
13. Bandelt HJ, Yao YG, Richards MB, Salas A: **The brave new era of human genetic testing.** *Bioessays* 2008, **30(11–12):**1246-1251.
14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic acids research* 2001, **29(1):**308-311.
15. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155(2):**945-959.
16. Amigo J, Salas A, Phillips C, Carracedo A: **SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access.** *BMC Bioinformatics* 2008, **9(1):**428.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

A reduced number of mtSNPs saturates mitochondrial DNA haplotype diversity of worldwide population groups.

Salas A, Amigo J

PloS One. 01/2010; 5(5):e10218.

Los altos niveles de variación que caracterizan a la molécula de ADN mitocondrial (mtDNA) son debidos en última instancia a su alta tasa media de mutación. Es más, la variación del mtDNA está estructurada profundamente en diferentes poblaciones y grupos étnicos. Existe un interés creciente en la selección de un número reducido de polimorfismos de un solo nucleótido del mtDNA (mtSNPs) que aporte el máximo grado de poder de discriminación en una población dada. Las aplicaciones de un panel de mtSNPs seleccionado abarcan desde estudios antropológicos y médicos a la casuística de la genética forense. Este estudio propone un nuevo método basado en la simulación que explora la capacidad de distintos paneles de mtSNPs para obtener los máximos niveles de poder de discriminación. El método explora subconjuntos de mtSNPs de diferentes tamaños elegidos al azar de un panel preseleccionado de mtSNPs basado en la frecuencia. Más de 2.000 genomas completos que representan a los tres principales grupos de población humana continental (África, Europa y Asia) y dos poblaciones mezcladas ("afroamericanos" e "hispanos") han sido recogidos a partir de GenBank y de la literatura, y se utilizaron como conjuntos de entrenamiento. La diversidad haplotípica fue medida para cada combinación de mtSNPs y en comparación con los paneles mtSNPs existentes disponibles en la literatura. Los datos indican que sólo un número reducido de mtSNPs entre 6 y 22 son necesarias para tener en cuenta el 95% de la diversidad haplotípica máxima de una muestra de población dada. Sin embargo, sólo una pequeña proporción de los mejores mtSNPs están compartidos entre las poblaciones, lo que indica que no hay un sistema perfecto de mtSNPs "universales" adecuados para todos los contextos poblacionales. El poder de discriminación proporcionada por estos mtSNPs es mucho mayor que la potencia de los paneles de mtSNPs propuestos en la literatura hasta la fecha. Algunas combinaciones de mtSNPs también producen altos valores de diversidad en poblaciones mezcladas. El enfoque computacional propuesto para explorar combinaciones de mtSNPs que optimizan el poder de discriminación de un conjunto dado de mtSNPs es más eficiente que los anteriores enfoques empíricos. En contraste con los resultados precedentes, los resultados parecen indicar que sólo se necesitan unos pocos mtSNPs para alcanzar altos niveles de poder de discriminación en una población, independientemente de su fondo ancestral.

A Reduced Number of mtSNPs Saturates Mitochondrial DNA Haplotype Diversity of Worldwide Population Groups

Antonio Salas[1]*, **Jorge Amigo**[1,2]

1 Unidade de Xenética, Departamento de Anatomía Patolóxica e Ciencias Forenses, and Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain, **2** Grupo de Medicina Xenómica, Universidade de Santiago de Compostela, Galicia, Spain

Abstract

Background: The high levels of variation characterising the mitochondrial DNA (mtDNA) molecule are due ultimately to its high average mutation rate; moreover, mtDNA variation is deeply structured in different populations and ethnic groups. There is growing interest in selecting a reduced number of mtDNA single nucleotide polymorphisms (mtSNPs) that account for the maximum level of discrimination power in a given population. Applications of the selected mtSNP panel range from anthropologic and medical studies to forensic genetic casework.

Methodology/Principal Findings: This study proposes a new simulation-based method that explores the ability of different mtSNP panels to yield the maximum levels of discrimination power. The method explores subsets of mtSNPs of different sizes randomly chosen from a preselected panel of mtSNPs based on frequency. More than 2,000 complete genomes representing three main continental human population groups (Africa, Europe, and Asia) and two admixed populations (''African-Americans'' and ''Hispanics'') were collected from GenBank and the literature, and were used as training sets. Haplotype diversity was measured for each combination of mtSNP and compared with existing mtSNP panels available in the literature. The data indicates that only a reduced number of mtSNPs ranging from six to 22 are needed to account for 95% of the maximum haplotype diversity of a given population sample. However, only a small proportion of the best mtSNPs are shared between populations, indicating that there is not a perfect set of ''universal'' mtSNPs suitable for all population contexts. The discrimination power provided by these mtSNPs is much higher than the power of the mtSNP panels proposed in the literature to date. Some mtSNP combinations also yield high diversity values in admixed populations.

Conclusions/Significance: The proposed computational approach for exploring combinations of mtSNPs that optimise the discrimination power of a given set of mtSNPs is more efficient than previous empirical approaches. In contrast to precedent findings, the results seem to indicate that only few mtSNPs are needed to reach high levels of discrimination power in a population, independently of its ancestral background.

Citation: Salas A, Amigo J (2010) A Reduced Number of mtSNPs Saturates Mitochondrial DNA Haplotype Diversity of Worldwide Population Groups. PLoS ONE 5(5): e10218. doi:10.1371/journal.pone.0010218

Editor: Vincent Macaulay, University of Glasgow, United Kingdom

Received January 14, 2010; **Accepted** March 22, 2010; **Published** May 3, 2010

Copyright: © 2010 Salas, Amigo. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Two grants from Fundacion de Investigacion Medica Mutua Madrile a (2006/CL370 and 2008/CL444), a grant Grupos Emerxentes from Xunta de Galicia (2008/XA122) and ''Fondos Feder'', and a grant from the Ministerio de Ciencia e Innovacion (SAF2008-02971) awarded to A. Salas supported this project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: antonio.salas@usc.es

Introduction

Variations in human mtDNA molecules have been deeply investigated in several fields of biomedical research such as forensic genetics, molecular anthropology, and disease studies [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]. There are 100 to 10,000 copies of the mtDNA genome per cell, and each of them consists of circular molecules of about 16,569 base pairs (bps). The mtDNA is inherited exclusively from the mother. Each mtDNA genome can be divided into two main parts: the control and the coding region. The control region occupies about 1,200 bp of the molecule and contains, among other, regulatory elements related to the replication of mtDNA or gene expression. It is usually divided into two segments characterised by their high (average) mutation rate, namely the first and second hypervariable segments (HVS-I/II). The coding region encodes 37 densely packed genes (13 for proteins, 22 for transfer RNA [tRNA], and two for subunits of ribosomal RNA [rRNA]) that are needed to maintain the correct function of the mitochondrion.

The study of mtDNA variability has been approached using different methodologies. Analysis of restriction fragment length polymorphism (RFLP) sites and screening procedures such as heteroduplex analysis (HD) and single strand conformation polymorphisms (SSCP) have been extensively used in the past and are still used by many laboratories [2,20,21]. Since the mid-1990s, sequencing the HVS-I (and sporadically the HVS-II),

coupled with the analysis of selected RFLP sites, has been the most common strategy for analysing mtDNA variation. Nowadays, there are more than 130,000 HVS-I mtDNAs from different human population groups reported in the literature.

The interest in analysing complete mtDNA genomes is growing as indicated by the more than 6,700 complete genomes available in the literature and in GenBank, most of them reported in the past five years. Sequencing complete genomes is now benefiting from improvements to the sequencing chemistry and the higher sophistication of automatic sequencers. However, this analysis is technically complex and costly and, therefore, unfeasible not only for most biomedical applications at high-throughput scales, but also for those applications depending on low quality or small amounts of DNA (e.g., forensic samples). This is the main reason why the majority of laboratories just target the control region (usually the HVS-I), complemented by analysing selected coding region SNPs. The traditional screening approaches (RFLP, SSCP) for genotyping mtSNPs are now being replaced by lower cost mini-sequencing techniques that allow multiplexing several polymorphisms in single reactions [22,23,24,25]. In particular, forensic geneticists are especially interested in developing mini-sequencing assays that interrogate a reduce numbers of mtSNPs per reaction (e.g., multiplexes of five to 20 mtSNPs), because the technique can perform well with degraded or low copy number samples [26]; many evidentiary samples only allow a single PCR reaction, and high-throughput mtSNP techniques are unsuitable for sub-optimal samples [27]. Ideally, targeted mtSNPs should retain the maximum level of discrimination power in a single mtDNA test and, therefore, a careful selection of mtSNPs for mini-sequencing assays is a key step in the process. Generally, this selection is based on phylogenetic criteria, where mtSNPs are chosen from the phylogeny to represent the main branches of the phylogenetic tree (which define haplogroups); however, this strategy does not necessarily optimise the discrimination power of a particular set of mtSNPs. Alternatively, mtSNPs can be selected according to their mutation rate; mtSNPs with the highest mutation rates tend to yield the maximum diversity values. However, current positional mutation rates are under suspicion because the methods employed to compute them could be flawed [28,29], and only recently a new proposal of site-specific mutation rates has been published aimed at overcoming the problems of past approaches [29]. It is impossible to decide *a priori* which one of these two approaches is more efficient for maximizing discrimination power. It is possible that a combination of both rationales could perform better. On the other hand, a particular combination of mtSNPs could yield good results in a specific population context, but might be unsuitable in a different population group (say Europeans *versus* Native Americans). Ideally, we could also envisage selecting a universal mtSNP panel that could generate reasonable discrimination power independently of the population group considered. Moreover, this panel would be particularly useful when dealing with highly admixed populations (e.g., the US, South American admixed populations, highly cosmopolitan cities).

With the latter applications in mind, we aimed to explore the optimum combinations of mtSNPs needed to maximize the diversity values of a given population group, taking into account the premise that multiplexing techniques (excluding high-throughput platforms) only allow genotyping a moderate number of mtSNPs (usually a maximum of 20 to 30). The method employed here is based on an algorithm that allows the exploration of the full set of combinations arising from a given set of known mtSNPs. We then evaluate the combinations yielding the highest values of diversity and the best candidate mtSNPs within these combinations. Various biomedical applications will also be discussed.

Material and Methods

Complete genome database

The database used for the simulation experiments was built based on the following criteria:

a) Selecting complete genomes from the literature and GenBank capable of being used as proxies of human population samples representing main continental regions. This criterion filters out those available complete genomes that have been analysed based on phylogenetic criteria or, in particular, patients in (mtDNA) disease studies [1,30,31,32,33,34,35,36];

b) The main population groups should be represented by at least 300 complete genomes such that most of the considered 'speedy' mutations (*sensu* [37]) can be polymorphic with a minimum allele frequency (MAF) >5% in the whole database; and

c) The compiled database should represent at least three main continental groups, namely, Africans, Asians, and Europeans.

According to these criteria, we collected the following datasets: (i) $N = 309$ from [5], representing the African subset; (ii) $N = 672$ Japanese to represent the Asian subset [38]; and (iii) $N = 241$ from [24] and $N = 192$ [39] representing the European subset. In addition, $N = 326$ individuals belonging to the American haplogroups A2, B2, C1, D1, and X2a (see [3,10] and references therein) were also collected for representing a Native American subset. Apart from the mentioned relatively homogenous population groups, two admixed population samples from the US, 'African-American' and 'Hispanic' datasets ($N = 140$ and $N = 125$ respectively) from [40], were also used in the simulation experiments.

We are aware of potential sequencing errors affecting some of these datasets and have tried to disregard suspicious complete genomes or use corrected versions of the reported datasets (not necessarily those available in GenBank) [29,41]. For instance, some of the Tanaka's complete genomes were corrected in Kong et al. [42] but the flawed versions of these genomes are still in GenBank [41].

As usual, mtSNPs are referred using the revised Cambridge Reference Sequence or rCRS [43].

Panel of SNPs

The approach followed in this study is based on exploring all (or a subset of all) possible combinations that arise from combining a given set of n mtSNPs candidates taken m at a time (in what follows, m value). Given the fact that the number of mtSNP variants considered in this study is above 3,200, it is computationally impossible to explore the entire universe of combinations arising from such a large number of variants (for instance, the number of possible combinations $C(100,20)$ is $>10^{22}$). To overcome this problem, we selected the 394 mtSNPs from the whole dataset fulfilling a MAF >0.05. By definition this criterion eliminates rare mtSNPs (MAF $\leq$0.05) under the premise that these SNPs cannot substantially contribute to increase diversity levels of populations. These variants are mainly 'private' to single genomes (singleton mutations usually located at the tips of the mtDNA phylogenies); therefore, these variants cannot be extrapolated to independent or larger samples due to ascertainment bias. On the contrary (see above), the selected mtSNPs (and in particular those that better contribute to increase the discrimination power) are polymorphic in different human populations; either (i) because they mutated before the divergence of major population groups, or (ii) because they have a high mutation rate (mutational hotspots). Variants that are known to be problematic from a genotyping point of view were disregarded (16182C, 16183C, 16193+C, variants around 310, length variation around positions 523–524, etc.).

Programming

For the panel of mtSNPs considered and all possible m values, we computed two diversity indices in every population dataset: (a) the 'normalised' number of haplotypes, defined as $H = h/N$, where h represents the number of different haplotypes in the dataset defined by a given m and N is the dataset sample size; and (b) the haplotype diversity, defined as $HD = 1 - \sum_{i}^{n} p_i^2$, where p_i is the haplotype frequency of the i haplotype resulting from each m combination of mtSNPs. We are aware to the fact that H does not scale out the dependence on sample size [44,45,46], and therefore the results on H cannot be extrapolated across samples of different sizes.

The main drawback of this study is the extremely high computational cost needed to explore the entire universe of all possible mtSNP combinations. We run a parallelisable script on a shared memory node system with the SMP NUMA architecture and a cluster containing over 2500 processors named *Finis Terrae*, which is available at the Supercomputing Centre of Galicia (CESGA; http://www.cesga.es/). Considering that studying each of the m values is completely independent from the rest, we dispatched each individual analysis run among the available cores of the cluster, reducing the overall running time to the largest m value analysis as if performed alone.

We ran up to 10,000 iterations for each m. We recorded the values of H and HD and derived their mean and standard deviation values. We also recorded the maximum values of diversity obtained for each m among the full number of iterations.

For every dataset, we also computed the maximum diversity values (H_{MAX} and HD_{MAX}) that could be obtained considering the complete genome information.

Results

Considerations about computational limitations

The large amount of mtSNPs considered in the different mtSNP panels tested made it computationally unfeasible to explore all possible combinations of n mtSNPs taken m at a time. We, therefore, developed an algorithm that creates subsets of m mtSNPs by sampling without replacement and equal probability each studied panel, while recording the best mtSNP combinations for every m value (the ones giving the maximum diversity values of H and HD). Other sampling strategies were used, such as those based on a stepwise mtSNP selection (find the best single mtSNP and then the best one to add it, and so on) but the one presented here was the most efficient in terms of maximizing the values of H and HD (data not shown). However, we first assessed the impact of analysing only a given number of combinations (iterations) of mtSNPs on the estimation of the parameters of interest. We observed that the trend in H or HD values with different m values was reasonably stable when exploring at least 10,000 mtSNP combinations (Figures S1, S2, S3, S4, S5). Based on this simulation, we decided to run 10,000 iterations for every population scenario considered in this study.

We are aware that this approach will probably prevent capture of the best candidate combination of mtSNPs (the one providing the highest real value of H and HD), but this is not the main goal of the simulations. The simulation approach could be considered a success if the best combination of mtSNPs obtained from it performed much better than traditional procedures. On the other hand, note that the standard deviation values of H and HD for the different m values (Figure 1) were very low, indirectly indicating

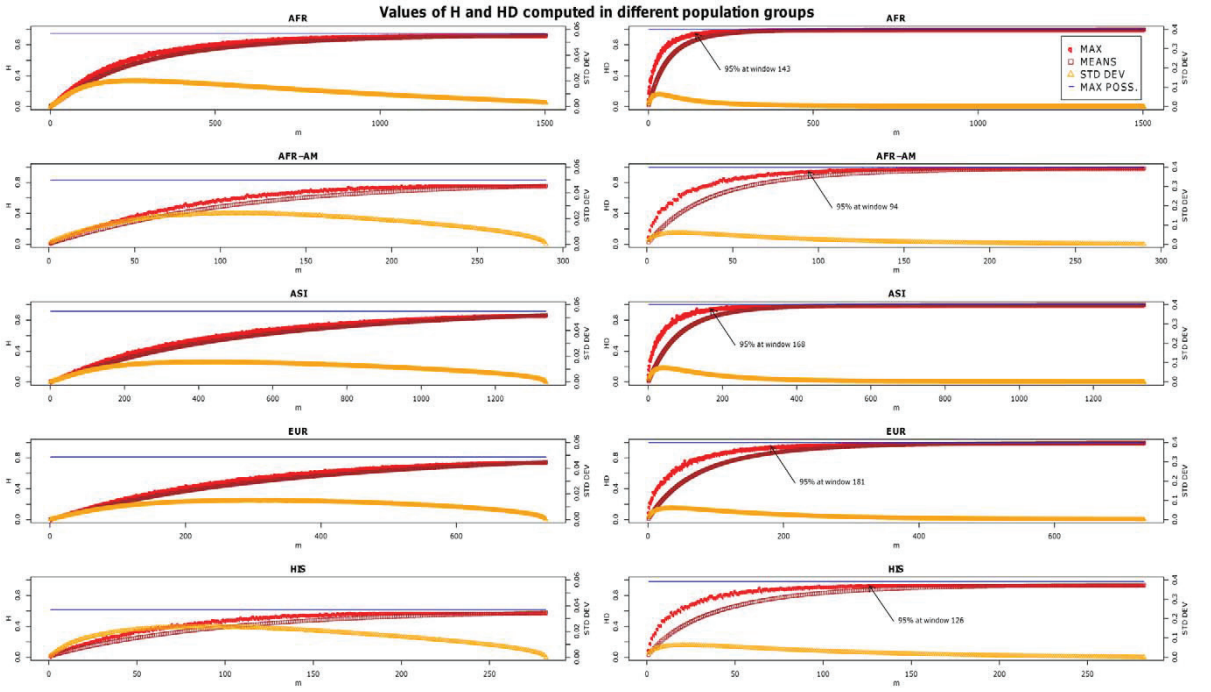


Figure 1. Values of *H* (left panel) and *HD* (right panel) computed in different population groups. The x-axis represents the numberofmtSNP considered. The top right legend indicates the colour codes for the maximum values of *H* and *HD* (max), the means, standard deviations, and maximum possible value of *H* and *HD* (computed assuming the full information provided by the complete genomes in each population dataset). Codes for populations are AFR = Africans, AFR-AM = 'African-Americans', ASI = Asians, EUR = Europeans, and HIS = 'Hispanics'.
doi:10.1371/journal.pone.0010218.g001

not only that these values do not fluctuate significantly on different iterations, but also that the best values should be close to the ones obtained using 10,000 iterations.

Performance of different mtSNP combinations on diversity values

The shapes of the *H* and *HD* curves in Figure 1 respond to the same expected phenomenon: increased values of *m* are accompanied by progressive increments in the values of *H* and *HD*. However, the *m* values needed to reach a *plateau* are different for *H* and *HD*. While *H* grows slowly by increasing the number of mtSNPs analysed, *HD* reaches its maximum quickly at lower mtSNP *m* values. Thus, somehow unexpectedly, the *HD* curves (Figures 1) indicate that only a small number of mtSNPs is required to saturate the diversity values to 95% of the HD_{MAX} independently of the population database employed (Figures 1 and Table 1). In other words, the addition of new mtSNPs beyond a certain point does not apparently contribute to an increase in discrimination power.

The mathematical difference between H_{MAX} and HD_{MAX} and the mean values of *H* and *HD* are more pronounced for intermediate *m* values (Figures 1) because these *m* values admit more different combinations of mtSNPs, thereby yielding a wider range of *H* and *HD* values. Concordant with this observation is the shape of the standard deviations, indicating that the maximum fluctuations of *H* and *HD* values also occur around these intermediate *m* values.

Evaluation of the best candidate mtSNPs

All the mtSNPs were ranked in a list according to the number of times each of them appears in a combination that maximizes the *HD* values; the first position being assigned to the one that appears most often (Table S1).

To some extent, the best mtSNPs in the different population groups considered in this study overlap (Table S1 and Table S2). For instance, transition T16519C appears to be the best mtSNP in virtually all the different population groups. Other well-known mutational hotspots, T152C, T195C, T16189C, and T16311C, occupy different positions in the ranking (which in part mirrors the stochastic nature of the simulation approach used in the present study), but all are listed among the top 15 top mtSNPs in the three main continental groups when looking at the *HD* values (Table 2 and S1).

The scores in Table S1 are useful for those analysts (e.g. forensic geneticists) wishing to empirically design a test panel of mtSNPs in a given population context. Thus, for instance, from Figures 1 (and Table 1), we know that nine mtSNPs sufficiently account for 95% of the haplotype diversity in a given African sample. From Table S1, it can be observed that the top nine polymorphisms in the African sample are for the values on *HD* (sorted by their position in the ranking): T16519C, T152C, T16189C, A189G, T16093C, G16129A, T195C, T16311C, and G143A.

Diversity accounted by existing mtSNP panels and the simulation-based approach

A number of different mtSNP panels have been proposed in the literature to date, and these provide a suitable framework for evaluating the efficiency of the simulation-based approach used in this study.

Table 1 summarises the values of *H* and *HD* for the different mtSNP panels proposed in the literature and those used in this study, all of them evaluated using the same collections of complete genomes (see M&M). Some of the panels reported in the literature were designed for specific population groups, mainly Europeans and Asians, so these panels behave worse in population samples with different genetic backgrounds. For instance, the panel of 45 mtSNPs proposed by Brandstätter et al. [25] was designed to increase the discrimination power within the typically European haplogroup H, but the discrimination power of these mtSNPs in Asians or Native Americans is much lower (Table 1). The opposite example is the mtSNP panel provided in Álvarez-Iglesias et al. [47].

The results summarised in Table 1 clearly indicate that almost all the panels proposed in the literature yield lower values of *H* and *HD* than the mtSNPs combinations inferred from the present study (with the caveat that *H* is not appropriate for inter-population comparisons). In addition, the number of mtSNPs needed to reach 95% of the H_{MAX} and HD_{MAX} are very low for most of the population groups (ranging from 10 to 22); while all the panels available in the literature yield values of diversity well below 95% (Table 1).

Evolutionary nature of the most discriminating mtSNPs

Not surprisingly, the top five mtSNPs that better contribute to increase the diversity values in the different population groups are non-synonymous or are located in non-coding or un-translated regions of the mtDNA molecule (Table 2). In other words, these mtSNPs are almost polymorphic in different human population groups (universal) because they do not seem to be as subjected to the effect of stabilizing or purifying natural selection as coding region variants. These top mtSNPs are in fact located in the control region (non-coding), are transitions, and roughly match the top mutational hotspots reported by [29] (Table 2).

Discussion

Several fields of research (population, medical, and forensic genetic) are interested in the analysis of mtSNPs for several genetic applications. Some of these applications rest on the ability of a selected (low) number of mtSNPs providing a high discrimination power when analysing human population or casework evidentiary samples. We have used a simulation-based approach to evaluate the discrimination power of mtSNPs in different population contexts, including samples representing the main continental regions and admixed population groups.

Our simulated approach indicates that no more than a dozen mtSNPs is sufficient to account for ~95% of the maximum level of *HD* diversity for almost all population groups. However, admixed populations, such as 'African-Americans', need to double this amount (~22 mtSNPs) to reach similar values of diversity. The top mtSNP variants are mutational hotspots mainly located in the control region. In addition, only a small proportion of the best mtSNPs are shared between population groups, indicating thatthere is not a perfect set of 'universal' mtSNPs suitable for all population contexts.

Today, there are two well-known strategies commonly used for selecting mtSNPs aimed to account for the highest levels of diversity in population groups, namely, the phylogenetic-based approach and mtSNP selection based on mutational hotspots. The results of the present study seem to indicate however that these traditional strategies perform worse than the simulation-based approach developed in the present study, and that in reality, it is a combination of highly mutable mtSNPs and haplogroup diagnostic sites that optimized the ability of a given mtSNP panels to account for the highest levels of diversity.

The strategy proposed in this study is particularly relevant for forensic studies, where small panels of mtSNPs are frequently

Table 1. Diversity values (*H* and *HD*) for the SNPs considered in different articles published in the literature and comparison with those obtained in the present study.

		AFR			AFR-AM			ASI			EUR			HIS			ALL		
Study	**n° SNPs**	**N**	**H**	**HD**	**N**	**H**	**HD**	**N**	**H**	**HD**	**N**	**H**	**HD**	**N**	**H**	**HD**	**N**	**H**	**HD**
Brandstätter et al. (2003)	16	9	0.0485	0.5660	6	0.0571	0.3830	6	0.0223	0.7153	15	0.0647	0.9155	10	0.0640	0.5533	15	0.0270	0.7991
Vallone et al. (2004)	11	4	0.0189	0.5217	2	0.0214	0.4033	5	0.0104	0.7508	11	0.0416	0.8242	3	0.0400	0.6485	11	0.0138	0.7343
Quintáns et al. (2004)	17	10	0.0404	0.1789	3	0.0286	0.1099	7	0.0134	0.7446	16	0.0416	0.9075	8	0.0640	0.7830	16	0.0224	0.8661
Umetsu et al. (2005)	36	15	0.0674	0.7471	8	0.0714	0.6737	31	0.0670	0.9453	21	0.0600	0.8975	14	0.1040	0.8252	36	0.0563	0.9583
Grignani et al. (2005)	16	7	0.0216	0.2530	1	0.0143	0.0826	5	0.0104	0.5214	13	0.0370	0.6030	2	0.0240	0.3435	13	0.0103	0.4723
Brandstätter et al. (2006)	45	21	0.0755	0.7191	6	0.0500	0.5253	13	0.0298	0.7313	34	0.0878	0.8803	6	0.0560	0.5415	40	0.0419	0.7843
Wiesbauer et al. (2006)	10	6	0.0296	0.3764	3	0.0286	0.0564	4	0.0119	0.4848	10	0.0370	0.8424	7	0.0640	0.6805	10	0.0155	0.6868
Lee et al. (2006)	22	9	0.0350	0.2713	4	0.0357	0.1497	21	0.0446	0.9225	7	0.0208	0.5636	9	0.0800	0.8210	22	0.0281	0.8875
Álvarez-Iglesias et al. (2006)	32	15	0.0755	0.6019	5	0.0429	0.1876	29	0.0670	0.8630	13	0.0393	0.5692	19	0.0960	0.8068	32	0.0534	0.9052
Coble et al. (2004)	59	31	0.2399	0.9410	14	0.1429	0.8558	30	0.0848	0.8889	57	0.1894	0.9463	12	0.1520	0.8810	57	0.1275	0.9504
Endicott et al. (2006)	20	7	0.0216	0.3326	2	0.0214	0.1220	5	0.0089	0.0666	1	0.0046	0.0046	1	0.0160	0.0160	11	0.0069	0.1160
Köhnmenn et al. (2008)	22	13	0.0836	0.7832	7	0.0714	0.6091	9	0.0387	0.8504	22	0.0993	0.9428	11	0.0800	0.7631	22	0.0477	0.8971
Wu et al. (2008)	10	9	0.0512	0.6350	3	0.0286	0.3952	10	0.0253	0.8171	9	0.0208	0.5736	9	0.0640	0.7621	10	0.0213	0.8771
Watkins et al. (2008)	32	16	0.0701	0.5872	5	0.0429	0.4317	19	0.0298	0.7736	21	0.0439	0.7730	15	0.0960	0.7866	28	0.0373	0.9036
Rosa et al. (2008)	19	11	0.0755	0.6174	6	0.0571	0.3725	12	0.0461	0.8394	18	0.0600	0.8978	12	0.0880	0.7930	19	0.0442	0.8811
Present study	9/22/11/10/10	184	0.8679	0.9990	189	0.5714	0.9885	100	0.5774	0.9950	88	0.4111	0.9909	97	0.4960	0.9772	195	0.5325	0.9977
Maximum possible values	–	–	0.9569	0.9997	–	0.9286	0.9990	–	0.9435	0.9998	–	0.8545	0.9989	–	0.7280	0.9876	–	0.9075	0.9998

Thus, the mtSNPs reported in the different published panels are used to compute *HD* and *H* in the complete genomes considered in the present study. In the row labelled as ''Present study'' we indicate the number of mtSNPs needed to cover at least 95% of the maximum *HD* in the different population groups; numbers in the second column separated by slash correspond to AFR (Africans), AFR-AM ('African-Americans'), ASI (Asians), EUR (Europeans), and HIS ('Hispanics'), respectively. The final column headed as ALL refer to the values after lumping the full sub-sets of complete genomes. The bottom row indicates the maximum possible values of *H* and *HD* for the different population groups considering the whole genome information.

doi:10.1371/journal.pone.0010218.t001

Table 2. Excerpt of the data in Table S1 showing the top five mtSNP that are shared between the top 15 mtSNPs in the three main continental groups.

Position	AFR	AFR-AM	ASI	EUR	HIS	ALL	rCRS	Variant	MapLocus	MR
16519	1	1	1	1	1	1	T	C	MT-DLOOP1	1
152	2	3	4	2	5	2	T	C	MT-DLOOP2	2
16189	3	7	2	6	14	3	T	C	MT-DLOOP1	6
16129	6	115	3	8	6	4	G	A	MT-DLOOP1	7
195	7	44	9	14	0	8	T	C	MT-DLOOP2	5

MR (mutational ranking) column refers to the position of these variants in the list of relative site-specific mutation rates as reported in [29]. Other legends are as in Table S1.

doi:10.1371/journal.pone.0010218.t002

demanded in routine casework, but we can also foresee other biomedical applications. For instance, mtSNP panels could be used to evaluate mtDNA instability in studies on tumours [48,49,50], where patients belong to different population groups.

Some caveats should be added about the potential bias arising from the limited number of complete genomes considered in this study. Some potential ascertainment bias could arise when computing *HD* using a limited number of complete genomes. The reproducibility of these results will only be possible with the availability of independent complete genome datasets. A bootstrap strategy could be used instead, but the high computational demands of this procedure makes it unfeasible. However, it can tentatively be said that there are few considerations that allow us to predict reproducible results when applied to independent complete genome datasets. These considerations are that (a) the best mtSNPs have by definition an MAF above 5%, (b) various of the best mtSNPs overlap in different population groups, and (c) most of the best mtSNPs have a high mutation rate. Therefore, these mtSNPs appear in different parts of the worldwide phylogeny and are not restricted to any particular population or ethnic group.

This study has demonstrated that future proposals for mtSNP panels aimed at obtaining a high discrimination power could be considered in the light of the simulation approach proposed here. The phylogenetic approach, although essential for most mtDNA studies (e.g. [19,51,52,53]), is probably not the best tool for predicting the discrimination power of a particular set of a mtSNP panel, but can still be useful for understanding the biological nature of a selected panel of mtSNPs and assist in its design. The mtSNP panels proposed in the literature do not perform as well as those suggested by this study.

Supporting Information

Figure S1 Effect of size iteration (number of mtDNA combinations explored from the full universe of possible combinations) for the estimation of H and HD in the African dataset. Only the mtSNPs overlapping in all the population datasets were used.
Found at: doi:10.1371/journal.pone.0010218.s001 (0.31 MB TIF)

Figure S2 Effect of size iteration for the estimation of H and HD in the 'African-American' dataset. Only the mtSNPs overlapping in all the population datasets were used.
Found at: doi:10.1371/journal.pone.0010218.s002 (0.25 MB TIF)

Figure S3 Effect of size iteration for the estimation of H and HD in the Asian dataset. Only the mtSNPs overlapping in all the population datasets were used.
Found at: doi:10.1371/journal.pone.0010218.s003 (0.29 MB TIF)

Figure S4 Effect of size iteration for the estimation of H and HD in the European dataset. Only the mtSNPs overlapping in all the population datasets were used.
Found at: doi:10.1371/journal.pone.0010218.s004 (0.27 MB TIF)

Figure S5 Effect of size iteration for the estimation of H and HD in the 'Hispanic' dataset. Only the mtSNPs overlapping in all the population datasets were used.
Found at: doi:10.1371/journal.pone.0010218.s005 (0.24 MB TIF)

Table S1 Scores of all the mtSNPs in the different population samples (AFR = Africans, AFR-AM = "African-Americans", ASI = Asians, EUR = Europeans, and HIS = "Hispanics"; ALL = all the complete genomes considered as a single group) based on HD. First, the number of times a particular mtSNP shows up when computing the maximum values of HD in every iteration and m values is recorded. The mtSNPs are sorted from those that appear more times in the different iterations to those that never appear. The final score is given according to their relative position in this ranking; from 1 (received by the best mtSNP) to n (number of mtSNP in each panel). The columns indicate, in this order, the position of the mtSNP according to the rCRS [48], the mutational change (all are transitions unless a suffix indicates a transversion or an indel specified), the population sample set (as indicated above), the rCRS variant, the nature of the mutational change (transition, transvertion, or indel), the MapLocus, a shorthand of the locus, the locus description, the coding or non-coding condition of the mutational change, the changed-codon, the amino acid change, the relative position of the mtSNP within genes, the relative position of the mtSNP within codons, and their synonymous or non-synonymous condition.
Found at: doi:10.1371/journal.pone.0010218.s006 (0.12 MB XLS)

Table S2 Scores of all the mtSNPs in the different population samples based on H. More details in legend of Table S1.
Found at: doi:10.1371/journal.pone.0010218.s007 (0.12 MB XLS)

Acknowledgments

We would like to acknowledge the CESGA for their supercomputing availability and support and Yong-Gang Yao and an anonymous reviewer for critically reading and commenting on the manuscript.

Author Contributions

Conceived and designed the experiments: AS JA. Performed the experiments: AS JA. Analyzed the data: AS JA. Contributed reagents/materials/analysis tools: AS JA. Wrote the paper: AS JA.

References

1. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034–1036.
2. Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, et al. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64: 232–249.
3. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, et al. (2009) Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. Curr Biol 19: 1–8.
4. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences from the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152–1171.
5. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, et al. (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82: 1130–1140.
6. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H-J (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22: 339–345.
7. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, et al. (2002) The making of the African mtDNA landscape. Am J Hum Genet 71: 1082–1111.
8. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276.
9. Baudouin SV, Saunders D, Tiangyou W, Elson JL, Poynter J, et al. (2005) Mitochondrial DNA and survival after sepsis: a prospective study. Lancet 366: 2118–2121.
10. Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, et al. (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3: e1764.
11. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, et al. (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752–770.
12. Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, et al. (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72: 313–332.
13. Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, et al. (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137–1152.
14. Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, et al. (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69: 844–852.
15. Chinnery PF, Howell N, Andrews RM, Turnbull DM (1999) Clinical mitochondrial genetics. J Med Genet 36: 425–436.
16. Chinnery PF, Howell N, Andrews RM, Turnbull DM (1999) Mitochondrial DNA analysis: polymorphisms and pathogenicity. J Med Genet 36: 505–510.
17. Chinnery PF, Johnson MA, Wardell TM, Singh-Kler R, Hayes C, et al. (2000) The epidemiology of pathogenic mitochondrial DNA mutations. Ann Neurol 48: 188–193.
18. Carelli V, Giordano C, d'Amati G (2003) Pathogenic expression of homoplasmic mtDNA mutations needs a complex nuclear-mitochondrial interaction. Trends Genet 19: 257–262.
19. Salas A, Bandelt H-J, Macaulay V, Richards MB (2007) Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int 168: 1–13.
20. Barros F, Lareu MV, Salas A, Carracedo A (1997) Rapid and enhanced detection of mitochondrial DNA variation using single-strand conformation analysis of superposed restriction enzyme fragments from polymerase chain reaction-amplified products. Electrophoresis 18: 52–54.
21. Salas A, Rasmussen EM, Lareu MV, Morling N, Carracedo Á (2001) Fluorescent SSCP of overlapping fragments (FSSCP-OF): a highly sensitive method for the screening of mitochondrial DNA variation. Forensic Sci Int 124: 97–103.
22. Álvarez-Iglesias V, Barros F, Carracedo Á, Salas A (2008) Minisequencing mitochondrial DNA pathogenic mutations. BMC Med Genet 9: 26.
23. Quintáns B, Álvarez-Iglesias V, Salas A, Phillips C, Lareu MV, et al. (2004) Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int 140: 251–257.
24. Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, et al. (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 118: 137–146.
25. Brandstätter A, Salas A, Niederstätter H, Gassner C, Carracedo Á, et al. (2006) Dissection of mitochondrial superhaplogroup H using coding region SNPs. Electrophoresis 27: 2541–2550.
26. Mosquera-Miguel A, Álvarez-Iglesias V, Lareu MV, Carracedo Á, Salas A (2009) Testing the performance of mtSNP minisequencing in forensic samples. Forensic Sci Int Genet 3: 261–264.
27. Cerezo M, Černý V, Carracedo Á, Salas A (2009) Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. Electrophoresis 30: 3665–3673.
28. Bandelt H-J, Kong Q-P, Richards M, Macaulay V (2006) Estimation of mutation rates and coalescence times: some caveats. In: Bandelt H-J, Richards M, Macaulay V, eds. Human mitochondrial DNA and the evolution of Homo sapiens. Berlin: Springer-Verlag. pp 47–90.
29. Soares P, Ermini L, Thomson N, Mormina M, Rito T, et al. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84: 740–759.
30. Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, et al. (2008) Climate change and postglacial human dispersals in southeast Asia. Mol Biol Evol 25: 1209–1218.
31. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, et al. (2005) Saami and Berbers–an unexpected mitochondrial DNA link. Am J Hum Genet 76: 883–886.
32. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, et al. (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 314: 1767–1770.
33. Kivisild T, Tolk H-V, Parik J, Wang Y, Papiha SS, et al. (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19: 1737–1751.
34. Brisighelli F, Capelli C, Álvarez-Iglesias V, Onofri V, Paoli G, et al. (2009) The Etruscan timeline: a recent Anatolian connection. Eur J Hum Genet 17: 693–696.
35. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, et al. (2009) New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS ONE 4: e5112.
36. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, et al. (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci U S A 105: 1596–1601.
37. Bandelt H-J, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150–1160.
38. Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, et al. (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14: 1832–1850.
39. Finnilä S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68: 1475–1484.
40. Just RS, Diegoli TM, Saunier JL, Irwin JA, Parsons TJ (2008) Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. Forensic Sci Int Genet 2: e45–48.
41. Yao Y-G, Salas A, Logan I, Bandelt H-J (2009) mtDNA data-mining in GenBank needs surveying. Am J Hum Genet in press.
42. Kong Q-P, Salas A, Sun C, Fuku N, Tanaka M, et al. (2008) Distilling artificial recombinants from large sets of complete mtDNA genomes. PLoS ONE 3: e3016.
43. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.
44. Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoretical Population Biology 3: 87–112.
45. Egeland T, Salas A (2008) Statistical evaluation of haploid genetic evidence. TOForensicSJ 1: 4–11.
46. Egeland T, Salas A (2008) Estimating haplotype frequency and coverage of databases. PLoS ONE 3: e3988.
47. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44–55.
48. Salas A, Yao Y-G, Macaulay V, Vega A, Carracedo Á, et al. (2005) A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2: e296.
49. Vega A, Salas A, Gamborino E, Sobrido MJ, Macaulay V, et al. (2004) mtDNA mutations in tumors of the central nervous system reflect the neutral evolution of mtDNA in populations. Oncogene 23: 1314–1320.
50. Cerezo M, Bandelt H-J, Martin-Guerrero I, Ardanaz M, Vega A, et al. (2009) High mitochondrial DNA stability in B-cell chronic lymphocytic leukemia. PLoS One 4: e7902.
51. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335: 891–899.
52. Yao Y-G, Bravi CM, Bandelt H-J (2004) A call for mtDNA data quality control in forensic science. Forensic Sci Int 141: 1–6.
53. Bandelt H-J, Olivieri A, Bravi C, Yao Y-G, Torroni A, et al. (2007) 'Distorted' mitochondrial DNA sequences in schizophrenic patients. Eur J Hum Genet 15: 400–402; author reply 402-404.

Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel.

Phillips C, Fernandez-Formoso L, Garcia-Magariños M, Porras L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Freire-Aradas A, Gomez-Carballa A, Mosquera-Miguel A, Carracedo Á, Lareu MV

Forensic Science International. Genetics. 03/2010; 5(3):155-69.

El panel CEPH de líneas celulares de diversidad genómica humana (CEPH-HGDP) de 51 poblaciones distribuidas globalmente se utilizó para analizar los patrones de variabilidad de 20 STRs de identificación humana básica. Los marcadores analizados comprendían los 15 STRs de Identifiler, una de las combinaciones de STRs más utilizada en forense, además de cinco recientemente introducidos STRs del grupo europeo estándar (ESS): D1S1656, D2S441, D10S1248, D12S391 y D22S1045. De los genotipos obtenidos para los STRs ESS hemos identificado alelos raros, intermedios o fuera de escala que no habían sido previamente reportados para estos *loci*. Ejemplos de los nuevos alelos de los STRs ESS encontrados fueron caracterizados a través del análisis de secuencias. Esto reveló una amplia variación de repeticiones estructurales en tres STRs ESS, con D12S391 mostrando una variabilidad particularmente alta de AGAT y AGAC en unidades de repetición en tándem. La distribución geográfica global de las muestras del panel CEPH permitió estudiar en detalle el alcance de la subestructura mostrada por los 20 STRs entre poblaciones y entre los grupos poblacionales de origen. Se hizo una evaluación de la capacidad informativa forense de los nuevos STRs ESS en comparación con los *loci* que reemplazarán: CSF1PO, D5S818, D7S820, D13S317 y TPOX, con resultados que muestran una mejora clara del poder de discriminación usados en combinaciones que los genotipos de los *loci* ESS nuevos. También se midió la capacidad de Identifiler y STRs ESS para inferir la ascendencia de las muestras CEPH-PDGH y demostrar que los STRs forenses usados en grandes combinaciones tienen el potencial para diferenciar los grupos poblacionales más importantes, pero sólo con la suficiente fiabilidad cuando se utilizan con otros marcadores informativos de ascendencia tales como polimorfismos de un solo nucleótido. Finalmente comprobamos la posible asociación de la vinculación entre las dos combinaciones de STRs ESS estrechamente posicionadas en el cromosoma 12: vWA y D12S391 mediante el examen de pares de genotipos de conjunto completo de datos CEPH.

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel

C. Phillips [a,b,*], L. Fernandez-Formoso [a], M. Garcia-Magariños [a], L. Porras [a], T. Tvedebrink [c], J. Amigo [b], M. Fondevila [a], A. Gomez-Tato [d], J. Alvarez-Dios [d], A. Freire-Aradas [a], A. Gomez-Carballa [a], A. Mosquera-Miguel [a], Á. Carracedo [a,b], M.V. Lareu [a]

[a] Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[b] Genomics Medicine Group, CIBERER, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain
[c] Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark
[d] Faculty of Mathematics, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Article history:
Received 10 December 2009
Received in revised form 2 February 2010
Accepted 6 February 2010
Available online xxx

Keywords:
Short tandem repeat
STR
ESS
CEPH human genome diversity panel
Theta
Population substructure

ABSTRACT

The CEPH human genome diversity cell line panel (CEPH-HGDP) of 51 globally distributed populations was used to analyze patterns of variability in 20 core human identification STRs. The markers typed comprised the 15 STRs of Identifiler, one of the most widely used forensic STR multiplexes, plus five recently introduced European Standard Set (ESS) STRs: D1S1656, D2S441, D10S1248, D12S391 and D22S1045. From the genotypes obtained for the ESS STRs we identified rare, intermediate or off-ladder alleles that had not been previously reported for these loci. Examples of novel ESS STR alleles found were characterized by sequence analysis. This revealed extensive repeat structure variation in three ESS STRs, with D12S391 showing particularly high variability for tandem runs of AGAT and AGAC repeat units. The global geographic distribution of the CEPH panel samples gave an opportunity to study in detail the extent of substructure shown by the 20 STRs amongst populations and between their parent population groups. An assessment was made of the forensic informativeness of the new ESS STRs compared to the loci they will replace: CSF1PO, D5S818, D7S820, D13S317 and TPOX, with results showing a clear enhancement of discrimination power using multiplexes that genotype the new ESS loci. We also measured the ability of Identifiler and ESS STRs to infer the ancestry of the CEPH-HGDP samples and demonstrate that forensic STRs in large multiplexes have the potential to differentiate the major population groups but only with sufficient reliability when used with other ancestry-informative markers such as single nucleotide polymorphisms. Finally we checked for possible association by linkage between the two ESS multiplex STRs closely positioned on chromosome-12: vWA and D12S391 by examining paired genotypes from the complete CEPH data set.

© 2010 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The widely used forensic markers of the AmpF/STR Identifiler® multiplex (Applied Biosystems: AB, Foster City, US) comprise amelogenin plus 15 autosomal STRs, that are also combined in the smaller scale AmpF/STR multiplexes of SGM Plus®, CoFiler®, Profiler Plus® and MiniFiler™. The forensic community has studied these STRs in numerous population samples but poor coverage is still evident in Oceania and to a lesser extent in the continental regions of South Asia, East Asia and the Americas [1].

The CEPH human genome diversity panel (CEPH-HGDP) contains just over a thousand individuals from African, European, North African/Middle Eastern, Central-South Asian, East Asian, Native American and Oceanian populations [2]. The identification of duplicates and first-degree relatives within the sample set reduces the CEPH panel to 971 individuals if second-degree relatives such as cousin pairs are retained [3], further reduced to 952 with the removal of second-degree relatives (termed H971 and H952 subsets respectively). We used the H971 subset to analyze the variation of the 15 STRs of Identifiler and supplemented this study with a further five loci, included in the European Standard Set or ESS STRs. The ESS loci have been recently adopted in Europe as markers able to improve on the discriminatory power and performance of five CODIS STRs in SGM Plus and Identifiler, specifically: CSF1PO, D5S818, D7S820, D13S317 and TPOX [4]. Two ESS STRs: D1S1656 and D12S391 have been chosen to improve the

* Corresponding author at: Forensic Genetics Unit, Institute of Legal Medicine, University of Santiago de Compostela, Santiago de Compostela, Galicia, Spain. Tel.: +34 981 582 327; fax: +34 981 580 336.
E-mail address: c.phillips@mac.com (C. Phillips).

1872-4973/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigen.2010.02.003

overall informativeness of an optimum 15 STR core set for European use and the other three: D2S441, D10S1248 and D22S1045 add enhanced genotyping success when analyzing highly degraded DNA by amplifying fragments well below current average amplicon sizes.

Until now the five new ESS markers lack extensive population studies and more importantly have not been fully characterized at the sequence level in non-European populations. Allele frequency analysis of the ESS loci in the broadly distributed CEPH populations provided the best opportunity to find and sequence rare or population specific alleles found in intermediate positions to the common repeats or outside current reference ladder size ranges in these five STRs. As the PCR primers used in the commercial kits currently being developed for the new STR combinations have not been published we made use of primer designs we originally developed when establishing two of the ESS STRs: D1S1656 and D12S391 [5,6]. This had the advantage that we could utilize previously optimized designs and PCR conditions plus comprehensive sequenced allelic ladders for the two STRs. Similarly the three short-amplicon ESS STRs were typed using the primer designs of the National Institute of Standards and Technology (NIST) where these STRs had been originally developed and characterized [7]. We combined the new ESS loci into a simple stand-alone 5-plex permitting straightforward typing of population samples to supplement existing surveys based on Identifiler or SGM Plus STRs (the full ESS multiplex comprises the 10 SGM Plus loci plus the above five). The same 5-plex assay can add further discrimination power to criminal casework samples previously typed with any of the established kits and has proved to be both robust and sensitive when analyzing challenging DNA [8].

The geographic breadth of the CEPH panel sampling gave the opportunity to explore various aspects of the new ESS loci of interest to forensic laboratories waiting to assess the usefulness of these STRs. As well as sequence analysis of rare alleles we were able to examine the following: (i) the predicted improvement in discrimination power of ESS STRs in each population group; (ii) the ancestry informativeness of all 20 STRs using a standard population reference panel with confirmed ancestral origins; (iii) the potential association by linkage between the two most closely positioned STRs in the full ESS set: vWA and D12S391, separated by 6.37 Mb on chromosome-12; and (iv) the extent of substructure amongst CEPH populations, some of which are known to exhibit very high levels of stratification.

In order to allow easy scrutiny of the extensive allele frequency data (based on more than 19,000 genotypes) generated by this study we reorganized an existing open-access SNP allele frequency browser [9] into a dedicated site named *pop.STR* [10] to accommodate the STR data. This allele frequency information is now freely available at: http://spsmart.cesga.es/popstr.php website. The help section of the *pop.STR* website provides a summary plot outlining the marker composition of each forensic multiplex from Applied Biosystems. Finally it is important to stress that intact individual STR profiles (i.e. with each genotype linked together per individual) were not retained in any format, to protect the privacy of the CEPH-HGDP donors.

2. Materials and methods

2.1. Population samples and STR genotyping

Although the CEPH-HGDP has a total of 1064 samples originating from 51 populations we used the H971 subset which excludes individuals previously identified (3) as duplicates, atypical samples (one mixed ancestry sample each from Africa and East Asia) or close relatives comprising parent-offspring and sib pairs. The H971 subset includes 19 cousin pairs removed from the other H952 subset. One widely recognized problem with the CEPH-HGDP is the disparity in population sample sizes, ranging from 6 African San of Namibia to 50 Middle East Palestinians of Central Israel. This means a proportion of CEPH populations cannot provide realistic allele frequencies given the broad range of alleles normally observed in the forensic STRs studied here. For this reason population designations were only used to identify the existence of specific STR alleles: i.e. those confined to one population alone or to analyze stratification in populations compared to their parent groups. For all other purposes populations were combined into continental-based groups which had been previously established [11] with the following composite populations, sample sizes and labels: 6 African (112 AFR), 8 European (158 EUR), 4 North African/Middle Eastern (170 ME), 9 Central-South Asian (204 S ASN), 17 East Asian (232 E ASN), 2 Oceanian (30 OCE) and 5 American (65 AME).

Identifiler STRs were genotyped with the commercial kit from AB and standard recommended amplification conditions except using 10 μl final PCR volume. The five new ESS loci were genotyped using previously optimized primers outlined in Table 1, with a single modification made to the originally reported reverse PCR primer of D22S1045 to address peak quality problems which lengthened the amplicons for this STR. The PCR reactants and cycling conditions were briefly: 1 μl DNA (0.5 ng/μl) 1 μl primer mix (ratios in Table 1), 3 μl H_2O, 5 μl Qiagen Multiplex PCR Master Mix (Cat. No. 206143), 95 °C × 15 min pre-denaturation, then 34 cycles of: 94 °C × 30 s, 58 °C × 90 s, 72 °C × 90 s, then a final elongation of 72 °C × 10 min.

Table 1
Genotyping and sequence primers used for the ESS 5-plex STR analyses. Non-specific mobility shift tails in lower case.

STR	Dye	Genotyping primers	Ratio in PCR primer mix[a]	Observed repeat numbers	Amplicon span (with tails)	Sequencing primers
D10S1248	6-FAM	TTAATGAATTGAACAAATGAGTGAG	0.75 μl	7	78 (79)	CTCTGTATCCCACCCCTG
		gCAACTCTGGTTGTATTGTCTTCAT		19	126 (127)	AAAGCAAACCTGAGCATTAGCC
D1S1656	6-FAM	GTGTTGCTCAAGGGTCAACT	0.75 μl	8	117 (135)	CCATATAAGTTCAAGCCTGTGTT
		ctctctctctctccttGAGAAATAGAATCACTAGGGA		19.3	164 (182)	GAGAAATAGAATCACTAGGGA
D12S391	6-FAM	AACAGGATCAATGGATGCAT	0.75 μl	12	197	AGAGACTGTATTAGTAAGGCTTC
		TGGCTTTTAGACCTGGACTG		27.2	259	TGGCTTTTAGACCTGGACTG
D2S441	VIC	CTGTGGCTCATCTATGAAAACTT	0.72 μl	8	76 (77)	CTGAGCCCTAATGCACCCA
		gAAGTGGCTGTGGTGTTATGAT		17	112 (113)	gAAGTGGCTGTGGTGTTATGAT
D22S1045	NED	ATTTTCCCCGATGATAGTAGTCT	0.72 μl	9	104	AGCTGCTATGGGGGCTAGATT
		CGGCACAGTGTGAGTGATCAC[b]		19	134	CGGCACAGTGTGAGTGATCAC

[a] Volume of 100 μM primer stocks.
[b] Sequence primer design used for genotyping PCR to reduce peak artifacts.

2.2. Sequence analysis of new ESS STRs and development of allelic ladders

The unlabelled primers used to sequence selected ESS STR alleles are listed in Table 1. In most cases these annealed outside of the genotyping sequence spans in order to detect possible deletions and sequence changes close to the repeat regions if any commercial kit genotyping primers gave slightly longer products than our own. Rare or atypical alleles were isolated from heterozygous genotypes with the largest size interval or from homozygotes for common alleles by excising product bands located on a silver-stained polyacrylamide fractionation gel (*T*: 9%, *C*: 5%). The alleles of interest were eluted by incubation of the gel segment in 50 ml of 20% Chelex™ overnight, followed by three freeze–thaw cycles, then centrifugation for 4 min at 13,000 rpm. The resulting purified supernatant was sequenced using standard AB Big-dye™ protocols.

Once characterized by sequencing, mainly homozygote alleles were combined into ladders for the ESS 5-plex by collating as many of the common repeats as possible. Existing ladders for D1S1656 and D12S391 were enhanced with additional alleles while the three short-amplicon ESS STR allelic adders were built de novo.

2.3. Calculation of forensic informativeness metrics

The power of discrimination denoted by random match probability (RMP), power of exclusion from paternity (PE) and typical paternity index (TPI) were calculated for each STR in the seven population groups using the Promega PowerStats Excel macro [12]. To assess the relative informativeness of different multiplexes, the cumulative RMP and PE plus the combined TPI for different STR multiplexes were also estimated using in-house Excel calculators that are available upon request.

2.4. Assessment of ancestry informativeness of the STR allele frequency distributions

Although generally less informative for the purpose than uniparental variability or ancestry-informative marker single nucleotide polymorphisms (AIM-SNPs), autosomal STRs have been used to attempt an inference of ancestry for a profile [13,14]. Furthermore it is useful to know how realistic it is to infer the ancestry of standard STR profiles as a potential tool to help direct police investigations. We measured the ancestry informativeness of the 20 STRs using three different approaches.

Firstly we assessed three genetic distance metrics for each marker from the 10 pairwise comparisons of the five major population groups: (i) a simple cumulative allele frequency differential: δ_C [15], plus two values specifically designed to gauge population distances with STR variation: (ii) average square distance (ASD) and (iii) $\delta\mu^2$, as described by Goldstein et al. [16,17] and based on the stepwise mutation model expected to occur with micro-satellites in populations sharing common ancestral alleles. Of these, $\delta\mu^2$ represents the most appropriate measure as it only focuses on the squared difference between the mean repeat alleles. Data processing and computations were programmed and carried out using R (http://www.r-project.org) with the underlying code available upon request. This analysis allowed an assessment of the relative divergence between population groups of each STR in turn.

Secondly we assessed the ability of the STR genotypes to cluster the CEPH panel samples as a full set, followed by a reduced set (excluding certain closely related populations that occupy a large geographic distribution), applying the widely used *structure* Bayesian grouping algorithm [11]. A previous study by Rosenberg et al. [18] suggested that for the most ancestry-informative STRs ~40 loci can provide a grouping of the CEPH panel populations almost as thorough as that obtained with the full parent set of 377 loci [18: Fig. 4]. We made the same analysis with the two 15-plex sets and all 20 loci to see if an optimum grouping could be obtained assigning the CEPH panel to five clusters (K:5) corresponding to the five major population groups (AFR, EUR, E ASN, OCE and AME). To obtain the clearest patterns of group membership from *structure* we reduced the study population complexity by excluding admixed populations from the Middle East and South Central Asian (ME and S ASN). This can help to overcome the disruptive effects of attempting to separate geographically close populations (in this case European Eurasians with non-European Eurasians) that can show divergent but poorly differentiated variability when limited numbers of markers are analyzed. The relative ability of each forensic STR set to differentiate ancestries with *structure* was analyzed by measuring the group misclassification rate at K:5, recorded as those samples with less than 0.5 group membership proportions in their cluster. Although *structure* is primarily a grouping algorithm rather than a classification system, the rate of incorrect clustering in each population group provides a means to assess the ancestry assignment error that can be expected when using a likelihood ratio analysis: the favored approach to obtain the probability that an unknown profile originates from a particular population group [13,14,19]. Lastly, *structure* provides a gauge of the genetic divergence amongst the analyzed populations given as net distances between the clusters, in effect providing a simple measure of the relative population divergences obtained with each STR set.

Thirdly we made a formal Bayesian analysis of ancestry of the CEPH panel by applying a standard likelihood ratio analysis that equates classification probability to the cumulative STR profile frequency using each group's allele frequency estimates. Differentiations were made of three, five and seven population groupings, comprising: AFR, E ASN and EUR; these three plus AME and OCE, and; these five plus S ASN and ME. Classification error rates were measured using reclassification of STR profiles to obtain the percentage apparent error for each STR set and group differentiation. We used in-house R based calculators available upon request.

2.5. Analysis of potential association by linkage of vWA and D12S391 alleles

The ESS multiplex combines two closely positioned STRs on chromosome 12: the established marker vWA and the new D12S391. The loci are sited at nucleotide positions: 6,093,104 and 12,449,930 respectively [20: NCBI genome build 37.1] – a separation of ~6.37 Mb which is closer than any other commonly used same-chromosome STR pair. So assumptions of independence must be properly tested ahead of using these two STRs to routinely construct a cumulative frequency for ESS multiplex profiles. We made two complementary tests for association between vWA and D12S391 using the whole set of CEPH genotype pairs (i.e. disregarding population labels) to maximize the power of the tests.

Firstly, we used the χ^2 test of independence between D12S391 and vWA allele frequency distributions. As a large proportion of these combinations, involving the rarest alleles, had zero observations or very low frequencies they created considerable potential for error in χ^2 tests. Therefore we grouped certain allele classes into logical sets that had comparable repeat sizes, e.g. the rare vWA 11 and 13 repeats were combined with the common 14 repeats. Secondly, we performed simulations based on the CEPH panel allele frequency estimates for vWA and D12 to generate 10,000 simulated sample sets in order to comprehensively measure the accuracy of the *p*-value for the χ^2 test of independence. This allowed a comparison with the *p*-values obtained from the original χ^2 tests of the observed CEPH two-genotype combinations.

2.6. Estimation of substructure amongst the 51 CEPH populations

A common challenge to the presentation of STR data in court is the problem that allele frequency estimates used to construct a profile frequency are based on population-group estimates likely to be unrepresentative of smaller, often isolated, subpopulations from which a defendant may originate. Therefore without suitable adjustment for the effect of stratification allele frequencies derived from a survey of a population group tend to underestimate those of a subpopulation and the rarity of the profile is artificially exaggerated [21,22]. A commonly applied correction factor for population stratification is theta: Θ, derived from F statistics [23,24], and a standard rule-of-thumb adjustment value is 0.1. This results in STR allele frequencies being adjusted upwards by ~10% to allow for the tendency of stratification to make those alleles more common in a subpopulation.

The Karitiana and Surui from Amazonia [25], and to a lesser extent, Papua New Guinean and Bourgainvilian from the SW Pacific region represent CEPH populations with higher than average relatedness detected amongst their samples, even in the adjusted H971 and H952 subsets [3]. As a result these populations are likely to show systematic differences in allele frequencies compared to their parent population-group as a whole. Therefore analysis of the CEPH panel STR variability gave an opportunity to assess the extent of substructure amongst the CEPH populations which in many cases are likely to show much higher levels than a forensic practitioner would normally expect to encounter.

One problem with examining substructure in many CEPH populations is very small sample sizes (excluding detected first-degree relatives, Karitiana and Surui have 8 and 14 individuals respectively) that will not be representative of the actual population structure. Therefore we decided to measure Θ between all 51 CEPH populations as well as between the population groups previously defined for these populations [11], placing them into one of the seven groups of AFR, EUR, E ASN, AME, OCE, S ASN and ME. Estimates of Θ for each STR were made using maximum likelihood estimation [26] and Weir and Hill's method of moment likelihood estimation [24].

3. Results and discussion

3.1. STR allele frequencies

Seven population group summary allele frequencies for the 20 STRs studied are listed in Table 2. The rare alleles observed in each group are marked as superscripts, ranging from singletons to three observations of an allele. Wherever possible these alleles were confirmed by sequence analysis of at least one example. The individual population allele frequency data generated from this study are listed at the *pop.STR* browser directly accessible at: http://spsmart.cesga.es/popstr.php with the option to choose from five forensic STR multiplexes (and their combinations) and query up to five user-defined groupings of populations. Several population statistical metrics are automatically generated from the data query including observed and expected heterozygosities, F_s and F_{st} (i.e. within-populations and within-groups fixation indices respectively) plus the population divergence metric I_n [18,19]. Example *pop.STR* frequency charts for D8S1179 in CEPH population groupings: East Asian and European plus D2S441 in African and Oceanian are shown in Fig. 1.

3.2. Sequence analysis of rare alleles

The repeat structures observed in the three new ESS STRs with compound repeat variation are outlined in Fig. 2. The other two ESS loci: D10S1248 and D22S1045 showed no sequence or repeat structure variation amongst 2–3 examples of each allele across the observed ranges of 7–19 and 9–19 repeats respectively. In contrast short-amplicon STR D2S441 displayed a series of repeat variants and two commonly observed SNPs but was the least comprehensively characterized by sequencing. Alleles 8, 9, 13, 13.3 and 15 listed in STRbase were observed but not sequenced while the intermediate allele 14.3 was not observed. Fig. 2 shows that a TTTA repeat occurs at a fixed position relative to the terminal repeat in the longest alleles of D2S441. A TCCA repeat was found next to the 3 bp repeat in two sequences, both Africans homozygous for the 12.3 repeat with this motif also homozygous, suggesting an association between this particular repeat motif and the 12.3 allele in Africans. Both repeat unit SNPs in D2S4441 detailed in Fig. 2 were seen regularly across a range of alleles and populations.

The complex ESS STR D1S1656 showed 1 short allele of 8 repeats additional to those in STRbase but the longest alleles this database lists: 20, 20.3 and 21 were not observed in the CEPH panel populations. Several examples of each of the D1S1656 alleles, except 11, were sequenced but no SNP variation was found. We observed two simple departures from the regular repeat patterns outlined in STRbase: one variant position for the TGA intermediate repeat and the absence amongst four of the seven shortest alleles of the terminal TAGG repeat.

The most complex repeat patterns were shown by D12S391. A novel 2 bp AT intermediate repeat allele with variable positions was observed exclusively amongst the longer repeat alleles of Mbuti Pygmies. The widely observed 3 bp GAT intermediate repeat allele found in the shorter repeat range 18.3–22.3 was also found to be variable in position, though we did not observe the 20.3 or 21.3 repeats listed in STRbase, nor a rare 17.3 reported in Europe [27]. A large number of alleles were sequenced for D12S391, except examples of 15, 20 and 26 repeat alleles, but no SNPs were found. However by far the most variability was found in the relative lengths of the AGAT-AGAC tandem repeat runs that in effect take the form of two STR spans in a tandem array. Scrutiny of a wide range of sequences from D12S391 homozygotes we analyzed revealed variable numbers of AGAT repeats independent of the proximal AGAC repeats so the great majority were heterozygous for a given length of overlapping AGAT/AGAC spans and created complex sequence peak patterns. This indicates that D12S391 has a much larger allelic range when the AGAT and AGAC tandem repeat lengths are treated as two variables hidden from the normal total-amplicon length estimation using electrophoresis. Therefore 8–21 AGAT repeats and 5–11 AGAC repeats potentially create 98 combinations each making a unique allele, increased further when the GAT intermediate repeat is included. While the AGAT-AGAC length ratios are not detectable by capillary electrophoresis means, base composition analysis using ion-pair reversed-phase high-performance liquid chromatography and electrospray ionization quadrupole time-of-flight mass spectrometry (ICEMS) has already been demonstrated to be highly effective at characterizing nucleotide variability within the STR repeats of Identifiler [28]. So the range of alleles detectable in D12S391 could potentially be increased fivefold by using typing systems sensitive to variation in the base composition of STR repeats.

To conclude, the repeat structures we detected in the three new ESS STRs showing compound repeat structures can be summarized in verbose format as follows:

D2S441; $[TCTA]_4$ $[TCA]_{0-1}$ $[TCCA]_{0-1}$ $[TCTA]_{2-10}$ $[TTTA]_{0-1}$ $[TCTA]_2$
D1S1656; $[TAGA]_{1-4}$ $[TGA]_{0-1}$ $[TAGA]_{7-14}$ $[TAGG]_{0-1}$ $[TG]_5$
D12S391; $[AGAT]_{1-5}$ $[GAT]_{0-1}$ $[AGAT]_{2-10}$ $[AT]_{0-1}$ $[AGAT]_{1-11}$ $[AGAC]_{5-11}$ $[AGAT]_{0-1}$

Table 2
Summary allele frequencies for seven CEPH population groups. Rare alleles are shown as superscript numbers denoting singletons, 2 or 3 observations per group. Light grey headers denote the five new ESS STRs, dark grey headers the five CODIS STRs in SGM Plus/Identifiler that they replace.

CSF1PO	EUR	AFR	E ASN	AME	OCE	S ASN	ME
6		0.010[2]					
7	0.003[1]	0.064	0.011				0.003[1]
8	0.007[2]	0.084			0.053[3]		0.009[3]
9	0.036	0.035	0.040	0.009[1]	0.026[1]	0.010	0.006[2]
10	0.288	0.287	0.230	0.091	0.263	0.273	0.284
11	0.333	0.213	0.234	0.336	0.105	0.322	0.297
12	0.278	0.272	0.404	0.491	0.447	0.348	0.344
13	0.049	0.025	0.067	0.073	0.105	0.044	0.050
14	0.003[1]	0.010[2]	0.013			0.003[1]	0.006[2]
15	0.003[1]						

D2S1338	EUR	AFR	E ASN	AME	OCE	S ASN	ME
11	0.006[2]						
15	0.003[1]						
16	0.055	0.069	0.004[2]			0.008	0.063
17	0.244	0.064	0.092	0.100	0.060[3]	0.088	0.266
18	0.071	0.079	0.109	0.036	0.040[2]	0.106	0.100
19	0.120	0.178	0.203	0.255	0.440	0.162	0.087
20	0.146	0.139	0.132	0.082	0.100	0.162	0.188
21	0.026	0.153	0.029	0.027	0.100	0.028	0.019
22	0.026	0.124	0.042	0.118	0.040[2]	0.077	0.056
23	0.097	0.079	0.161	0.364	0.140	0.162	0.100
24	0.101	0.059	0.158	0.018[2]	0.080	0.121	0.072
25	0.084	0.020	0.063			0.072	0.050
26	0.019	0.035	0.007			0.013	

D3S1358	EUR	AFR	E ASN	AME	OCE	S ASN	ME
6		0.005[1]					
12		0.005[1]					
13			0.004[2]	0.017[2]		0.003[1]	
14	0.094	0.099	0.031	0.034		0.078	0.035
15	0.294	0.332	0.395	0.543	0.308	0.373	0.258
16	0.242	0.356	0.308	0.293	0.423	0.233	0.292
17	0.206	0.158	0.194	0.095	0.135	0.218	0.270
18	0.148	0.045	0.060	0.017[2]	0.135	0.080	0.132
19	0.016		0.007			0.016	0.013

D5S818	EUR	AFR	E ASN	AME	OCE	S ASN	ME
7	0.003[1]		0.022	0.186			
8	0.003[1]	0.074			0.019[1]		0.019
9	0.048	0.005[1]	0.080	0.042	0.019[1]	0.070	0.044
10	0.074	0.069	0.182	0.110	0.346	0.111	0.134
11	0.345	0.193	0.318	0.458	0.346	0.335	0.278
12	0.326	0.371	0.260	0.195	0.173	0.312	0.275
13	0.181	0.252	0.131	0.008[1]	0.096	0.162	0.216
14	0.019	0.025	0.007			0.010	0.034
15		0.005[1]					
16		0.005[1]					

D7S820	EUR	AFR	E ASN	AME	OCE	S ASN	ME
6		0.005[1]					
7	0.013	0.015		0.010[1]		0.039	0.022
7.3		0.015					
8	0.190	0.218	0.174	0.010[1]	0.194	0.211	0.105
8.1		0.010[2]					
9	0.152	0.109	0.038	0.030	0.028[2]	0.073	0.108
10	0.274	0.307	0.201	0.110	0.222	0.247	0.322
10.3		0.005[1]		0.010[1]			
11	0.203	0.218	0.348	0.480	0.278	0.245	0.268
11.3		0.005[1]					
12	0.129	0.079	0.199	0.270	0.222	0.164	0.143
13	0.035	0.005[1]	0.038	0.080	0.056	0.021	0.029
14	0.003[1]	0.010[2]					0.003[1]

FGA	EUR	AFR	E ASN	AME	OCE	S ASN	ME
16	0.003[1]						
17							0.006[2]
18	0.013	0.005[1]	0.021			0.016	0.006[2]
18.2							0.003[1]
19	0.077	0.035	0.034	0.073	0.100	0.042	0.047
19.2		0.005[1]				0.003[1]	0.003[1]
20	0.123	0.055	0.073	0.073	0.025[1]	0.083	0.069
20.2	0.003[1]		0.002[1]				
21	0.174	0.060	0.117	0.094	0.075	0.128	0.182
21.2			0.002[1]				0.009[3]
22	0.165	0.190	0.147	0.104	0.100	0.164	0.164
22.2	0.003[1]	0.010[2]	0.005[2]			0.016	
23	0.152	0.160	0.225	0.073	0.150	0.130	0.195
23.2	0.003[1]		0.005[2]			0.005[2]	
24	0.148	0.170	0.197	0.177	0.250	0.216	0.170
24.2			0.009			0.010	
24.3						0.003[1]	
25	0.100	0.155	0.078	0.188	0.200	0.141	0.091
25.2			0.007			0.010	
26	0.029	0.060	0.064	0.177	0.050[3]	0.023	0.038
27	0.006[2]	0.045	0.007	0.021[2]		0.005[2]	0.003[1]
27.1					0.025[1]		
27.2			0.002[1]				
28		0.030	0.005[2]			0.003[1]	0.009[3]
28.1					0.025[1]		
29		0.005[1]		0.010[1]		0.003[1]	0.003[1]
30		0.005[1]					
31.2		0.005[1]					
33				0.010[1]			
33.3		0.005[1]					

TH01	EUR	AFR	E ASN	AME	OCE	S ASN	ME
6	0.263	0.069	0.109	0.397	0.327	0.206	0.272
7	0.140	0.347	0.307	0.517	0.077	0.186	0.144
8	0.114	0.347	0.071		0.423	0.157	0.100
9	0.221	0.158	0.416	0.009[1]	0.154	0.250	0.297
9.3	0.256	0.035	0.071	0.078		0.198	0.153
10	0.003[1]	0.025	0.027		0.019[1]	0.003[1]	0.034
11	0.003[1]	0.010[2]					
12		0.010[2]					

TPOX	EUR	AFR	E ASN	AME	OCE	S ASN	ME
6		0.084				0.003[1]	0.006[2]
7	0.003[1]	0.045		0.009[1]			0.006[2]
8	0.523	0.257	0.498	0.518	0.120	0.436	0.494
9	0.097	0.272	0.120		0.380	0.088	0.163
10	0.103	0.144	0.044	0.026	0.120	0.080	0.091
11	0.258	0.183	0.309	0.228	0.340	0.369	0.216
12	0.016	0.015[3]	0.027	0.219	0.040[2]	0.026	0.025
13			0.002[1]				

vWA	EUR	AFR	E ASN	AME	OCE	S ASN	ME
11		0.005[1]					
13	0.003[1]				0.020[1]		
14	0.119	0.084	0.233	0.009[1]		0.119	0.103
15	0.129	0.158	0.011	0.017[2]	0.020[1]	0.078	0.131
16	0.194	0.302	0.144	0.414	0.360	0.189	0.272
17	0.258	0.129	0.313	0.250	0.240	0.303	0.219
18	0.190	0.168	0.198	0.224	0.260	0.199	0.209
19	0.103	0.099	0.076	0.078	0.080	0.091	0.056
20	0.003[1]	0.035	0.024		0.020[1]	0.021	0.009[3]
21		0.015[3]		0.009[1]			
22		0.005[1]					

D8S1179	EUR	AFR	E ASN	AME	OCE	S ASN	ME
8	0.013					0.016	0.006[2]
9	0.019		0.002[1]				0.006[2]
10	0.126	0.010[2]	0.131	0.167	0.019[1]	0.140	0.081
11	0.052	0.045	0.091	0.053	0.038[2]	0.075	0.059
12	0.132	0.124	0.118	0.096	0.115	0.091	0.147
13	0.300	0.149	0.247	0.289	0.212	0.223	0.203
14	0.206	0.342	0.196	0.333	0.327	0.231	0.222
15	0.129	0.233	0.144	0.044	0.231	0.153	0.206
16	0.019	0.084	0.053	0.018[2]	0.038[2]	0.060	0.053
17	0.003[1]	0.015	0.013		0.019[1]	0.008[2]	0.016
18			0.004[2]			0.005[1]	

D13S317	EUR	AFR	E ASN	AME	OCE	S ASN	ME
7		0.005[1]					
8	0.108	0.030	0.270	0.034	0.396	0.157	0.103
9	0.095	0.020	0.143	0.414	0.063[3]	0.028	0.050
10	0.069	0.045	0.145	0.172	0.042[2]	0.082	0.059
11	0.317	0.257	0.250	0.138	0.167	0.296	0.269
12	0.261	0.450	0.143	0.121	0.208	0.312	0.409
13	0.114	0.153	0.038	0.086	0.125	0.085	0.078
14	0.036	0.040	0.011	0.034		0.034	0.031
15						0.005[2]	

D16S539	EUR	AFR	E ASN	AME	OCE	S ASN	ME
8	0.010	0.025	0.002[1]		0.020[1]	0.028	0.013
9	0.106	0.207	0.276	0.196	0.020[1]	0.144	0.142
10	0.039	0.141	0.130	0.304	0.140	0.116	0.063
11	0.319	0.268	0.231	0.179	0.240	0.361	0.326
12	0.313	0.258	0.258	0.250	0.280	0.229	0.259
13	0.158	0.086	0.076	0.071	0.220	0.119	0.184
14	0.045	0.015[3]	0.027		0.080	0.003[1]	0.009[3]
15	0.010						0.003[1]

D18S51	EUR	AFR	E ASN	AME	OCE	S ASN	ME
9						0.003[1]	
10	0.013		0.002[1]			0.022	0.003[1]
11	0.010[3]	0.010[2]	0.007			0.022	0.022
11.2							0.003[1]
12	0.147	0.051	0.045	0.093		0.077	0.173
12.2							0.006[2]
13	0.120	0.030	0.217	0.130	0.088	0.118	0.110
13.1			0.002[1]				
13.2		0.010[2]					
14	0.203	0.056	0.167	0.231	0.147	0.236	0.154
15	0.137	0.131	0.179	0.167	0.206	0.195	0.116
16	0.140	0.131	0.109	0.065	0.029[2]	0.159	0.129
17	0.127	0.232	0.081	0.176	0.235	0.047	0.148
18	0.067	0.136	0.057	0.111	0.088	0.052	0.069
19	0.020	0.141	0.045		0.088	0.044	0.044
20	0.010[3]	0.035	0.041	0.009[1]	0.059	0.011	0.016
21	0.007[2]	0.030	0.023	0.009[1]	0.029[2]	0.008	0.006[2]
22			0.018		0.029[2]		
23		0.005[1]	0.002[1]	0.009[1]		0.003[1]	
24						0.003[1]	
25			0.005[2]				

D22S1045	EUR	AFR	E ASN	AME	OCE	S ASN	ME
9	0.006[2]					0.005[2]	
10		0.088	0.002[1]			0.005[2]	0.010
11	0.103	0.157	0.270	0.016[2]	0.021[1]	0.233	0.129
12	0.022	0.029	0.002[1]			0.003[1]	0.043
13	0.003[1]	0.020	0.004[2]		0.063[3]	0.003[1]	0.003[1]
14	0.048	0.064	0.027	0.016[2]		0.080	0.050
15	0.381	0.186	0.252	0.582	0.479	0.380	0.411
16	0.359	0.167	0.243	0.344	0.313	0.225	0.288
17	0.077	0.275	0.179	0.041	0.104	0.060	0.050
18		0.010[2]	0.015		0.021[1]	0.007[3]	0.017
19		0.005[1]	0.004[2]				

D1S1561	EUR	AFR	E ASN	AME	OCE	S ASN	ME
8						0.013	
9						0.010	
10		0.025	0.002[1]			0.013	0.003[1]
11	0.062	0.059	0.062		0.063[3]	0.123	0.076
12	0.156	0.049	0.051	0.025[3]	0.063[3]	0.083	0.126
13	0.045	0.172	0.098	0.057	0.104	0.116	0.086
13.3		0.005[1]					
14	0.114	0.225	0.071	0.139	0.083	0.101	0.093
14.3		0.010[2]	0.002[1]				
15	0.127	0.137	0.278	0.156	0.083	0.191	0.156
15.3	0.045	0.010[2]			0.021[1]	0.013	0.043
16	0.146	0.147	0.220	0.131	0.271	0.166	0.209
16.3	0.071	0.083	0.007	0.049		0.038	0.053
17	0.039	0.044	0.089	0.033	0.208	0.050	0.073
17.3	0.149	0.025	0.062	0.279	0.021[1]	0.053	0.056
18	0.010		0.011	0.025[3]	0.063	0.008[3]	0.003[1]
18.3	0.026	0.010[2]	0.038	0.098		0.020	0.017
19.3	0.010		0.009	0.008[1]	0.021[1]	0.005[2]	0.007[2]

D2S441	EUR	AFR	E ASN	AME	OCE	S ASN	ME
8	0.013	0.005[1]			0.042[2]	0.003[1]	
9	0.003[1]	0.005[1]	0.009	0.008[1]			0.003[1]
10	0.163	0.029	0.215	0.730	0.729	0.220	0.116
11	0.298	0.319	0.394	0.156	0.146	0.453	0.334
11.3	0.080	0.093	0.049			0.043	0.073
12	0.032	0.142	0.157	0.025[3]	0.083	0.065	0.050
12.3		0.088					
13	0.045	0.044	0.035			0.022	0.036
13.3	0.026	0.020	0.002[1]			0.003[1]	
14	0.279	0.225	0.133	0.082		0.163	0.368
15	0.048	0.025	0.007[3]			0.030	0.017
16	0.013						0.003[1]
17		0.005[1]					

D10S1248	EUR	AFR	E ASN	AME	OCE	S ASN	ME
7		0.005[1]					
8		0.005[1]		0.016[2]			
9		0.005[1]					0.007[2]
10		0.005[1]					
11	0.003[1]	0.025		0.008[1]		0.010	0.010[3]
12	0.038	0.123	0.058	0.057	0.022[1]	0.033	0.040
13	0.244	0.235	0.347	0.164	0.109	0.161	0.215
14	0.298	0.275	0.262	0.475	0.109	0.286	0.318
15	0.244	0.201	0.229	0.205	0.239	0.261	0.228
16	0.144	0.098	0.076	0.066	0.478	0.191	0.119
17	0.026	0.020	0.027	0.008[1]	0.043[2]	0.055	0.063
18	0.003[1]	0.005[1]	0.002[1]				
19						0.003[1]	

D19S433	EUR	AFR	E ASN	AME	OCE	S ASN	ME
9		0.005[1]					
10		0.109					
11	0.006[2]	0.069				0.010	
11.2				0.017[2]			
12	0.100	0.099	0.036	0.025[3]		0.067	0.100
12.2		0.020	0.007	0.025[3]		0.010	
13	0.232	0.238	0.315	0.237	0.327	0.223	0.244
13.2	0.016	0.079	0.045	0.263	0.058[3]	0.013	0.019
14	0.394	0.193	0.254	0.144	0.442	0.295	0.231
1.2	0.019	0.054	0.100	0.085	0.096	0.070	0.034
15	0.123	0.040	0.056	0.102	0.038[2]	0.098	0.147
15.2	0.055	0.050	0.145	0.076	0.038[2]	0.106	0.113
16	0.029	0.015[3]	0.016	0.025[3]		0.060	0.072
16.2	0.023	0.030	0.027			0.039	0.034
17						0.008[3]	
17.2	0.003[1]						0.006[2]

D21S11	EUR	AFR	E ASN	AME	OCE	S ASN	ME
25.3		0.005[1]					
26		0.015			0.019[1]	0.013	
27	0.039	0.030				0.010	0.016
28	0.145	0.173	0.036	0.098	0.038[2]	0.129	0.147
28.2			0.009			0.003[1]	
29	0.242	0.163	0.260	0.205	0.308	0.206	0.191
29.2			0.002[1]			0.003[1]	
30	0.248	0.173	0.325	0.143	0.269	0.206	0.228
30.2	0.042	0.010[2]	0.018	0.045	0.058	0.041	0.031
30.3		0.005[1]					
31	0.045	0.139	0.105	0.036	0.077	0.062	0.078
31.2	0.113	0.040	0.063	0.116	0.077	0.129	0.134
32	0.010		0.025	0.054	0.019[1]	0.013	
32.2	0.087	0.084	0.108	0.205	0.058	0.126	0.138
33		0.010[2]	0.002[1]				
33.2	0.026	0.025	0.045	0.089	0.038[2]	0.046	0.034
34		0.045					
34.1		0.005[1]					
34.2			0.002[1]	0.009[1]		0.013	
35		0.064					0.003[1]
35.2	0.003[1]						
36		0.005[1]					
36.1		0.005[1]					
37		0.005[1]					
37.2					0.019[1]		
38.2					0.019[1]		

D12S391	EUR	AFR	E ASN	AME	OCE	S ASN	ME
14						0.005[2]	
15	0.029	0.064	0.022			0.007[3]	0.017
16	0.019	0.059	0.002[1]	0.033	0.104	0.013	0.010[3]
17	0.096	0.181	0.100	0.025[3]	0.042	0.113	0.070
17.3	0.026	0.005[1]	0.002[1]			0.015	0.007[2]
18	0.163	0.240	0.218	0.067	0.167	0.207	0.222
18.3	0.045	0.005[1]	0.002[1]			0.025	0.020
19	0.106	0.181	0.224	0.392	0.313	0.145	0.109
19.3	0.013					0.005[2]	0.013
20	0.119	0.103	0.176	0.333	0.125	0.113	0.139
21	0.119	0.049	0.102	0.092	0.125	0.102	0.116
22	0.119	0.015[3]	0.080	0.025[3]	0.021	0.117	0.136
22.3						0.003[1]	
23	0.096	0.039	0.040	0.033	0.083	0.075	0.096
24	0.035	0.020	0.020			0.040	0.036
24.2		0.005[1]					
25	0.010	0.005[1]	0.007			0.015	0.007[2]
25.2		0.010[2]					
26	0.006[2]	0.005[1]			0.021		0.003[1]
26.2		0.010[2]					
27			0.004[2]				
27.2		0.005[1]					

3.3. Forensic informativeness of the STR multiplexes

Table 3 gives the cumulative random match probability, cumulative paternity exclusion probability and combined typical paternity indices for four STR sets in the seven population groups studied with the small-scale MiniFiler multiplex of 8 STRs included for reference. The ESS STRs show approximately one order of magnitude greater random match probability compared to the Identifiler STRs in five population groups with two groups: AME and OCE showing near identical levels of forensic identification informativeness between both marker sets. Therefore in EUR, AFR, E ASN, S ASN and ME population groups the five new ESS STRs

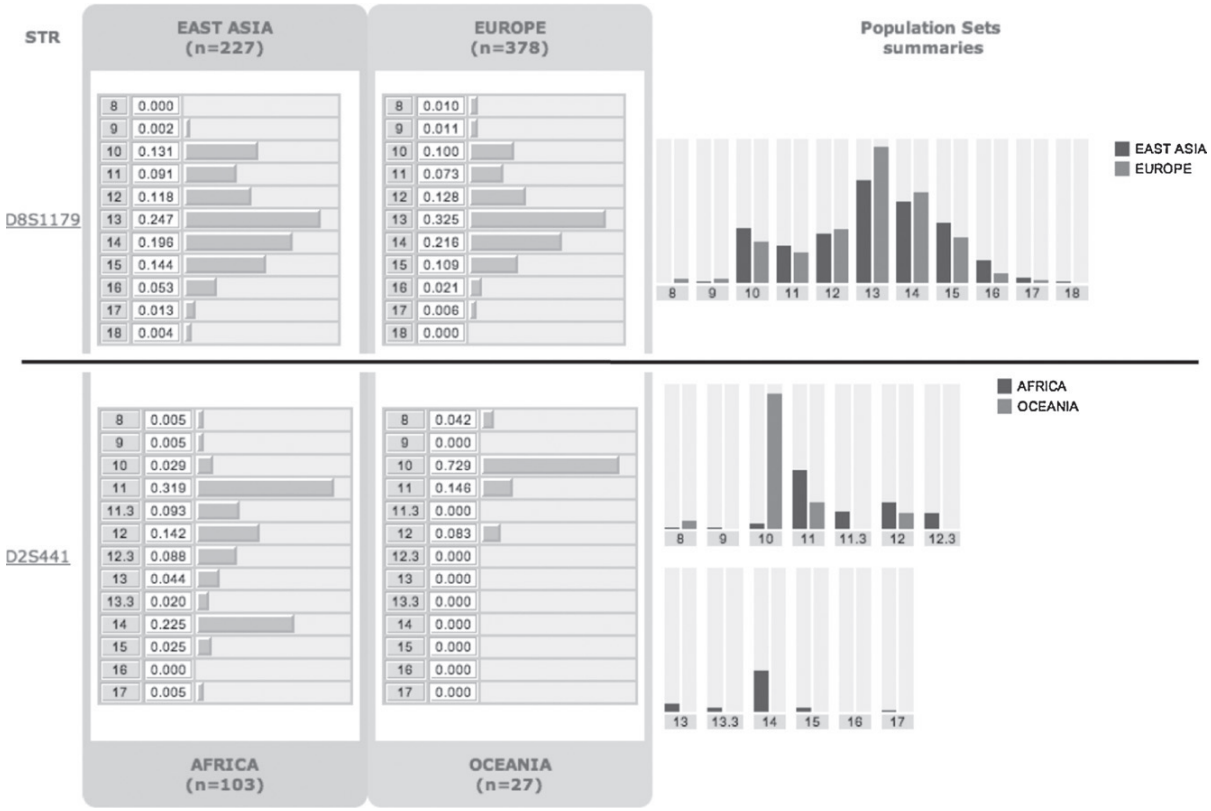


Fig. 1. Example of *pop.STR* allele frequency output for D8S1179 and D2S441. Population comparisons show the least and most divergent STR variability respectively found in the CEPH panel.

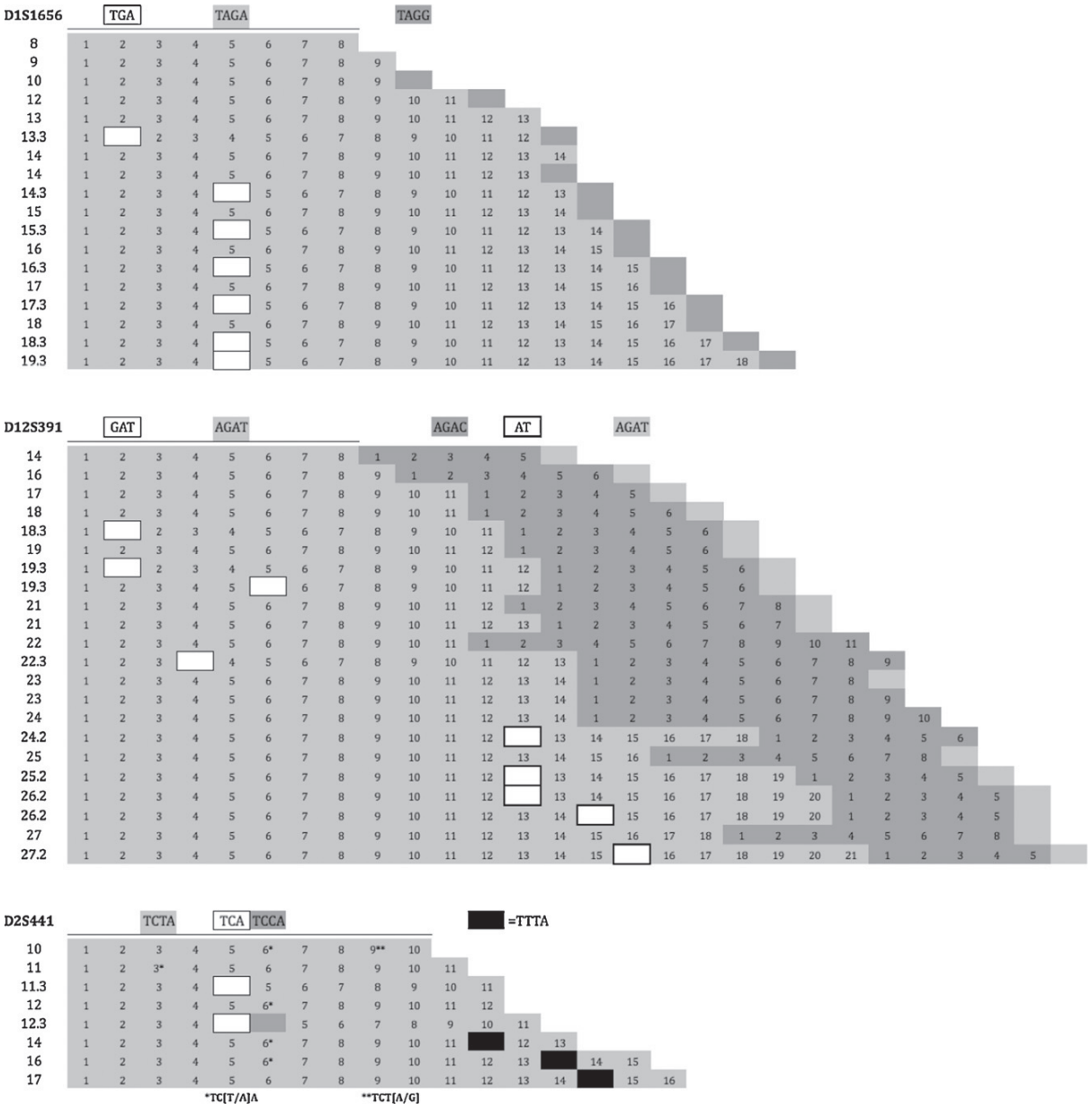


Fig. 2. Summary repeat structure maps of the three compound STRs characterized by sequence analysis. Top horizontal line denotes the core repeat unit and smallest tandem repeat length observed in each STR. The positions of two SNPs in the core repeat unit of D2S441 are marked with asterisks.

provide a 10-fold increase in discrimination power over the five Identifiler STRs they supersede. With the exception of the Native American group, cumulative paternity exclusion probability and combined typical paternity indices are all marginally improved with use of ESS loci in relationship testing compared to Identifiler.

3.4. Ancestry informativeness of the 20 STRs

The divergence values: δ_C, average distance squared and $\delta\mu^2$ obtained from 10 pairwise population-group comparisons are listed in supplementary Table S1. The $\delta\mu^2$ values are listed separately with STRs ranked in order of descending divergence. Although ESS STRs predominate at the top of the table with some of the highest average population divergences, this measure does not directly reflect the degree of variability in each STR, with the relatively uninformative D2S441 giving a much higher divergence than most other loci. Fig. 1 shows the two pairs of allele frequency distributions for the highest and lowest divergence values obtained from STRs that occupy opposite extremes of allele variability, illustrating that divergence is dictated by contrasting allele frequencies between populations not by overall levels of polymorphism. The other two measures of population-group divergence gave comparable relative ranking of STRs. For comparison with a large set of ancestry-informative STRs we calculated the equivalent $\delta\mu^2$ values for the same CEPH population group comparisons based on 783 STRs studied by Ramachandran et al. [29]. These markers give a summary average $\delta\mu^2$ value of 1.446 (i.e. averaged across 10 group comparisons and 783 loci). Therefore all but one of the forensic STRs characterized (D2S441 with an average $\delta\mu^2$ of 1.627) are less informative for ancestry than a

Table 3
Forensic informativeness values for four STR sets in seven CEPH population groups. RMP: cumulative random match probability, PE: cumulative paternity exclusion probability, TPI: combined typical paternity index.

Cumulative values		EUR	AFR	E ASN	AME	OCE	S ASN	ME
	RMP	9.5E+16	2.2E+17	6.2E+16	3.2E+14	2.3E+14	1.8E+17	1.5E+17
Identifiler	PE	1.3E+06	4.5E+06	3.3E+05	1.1E+04	1.9E+04	1.6E+05	2.3E+05
	TPI	1.2E+06	4.5E+06	2.7E+05	4.6E+03	8.5E+03	1.2E+05	1.9E+05
	RMP	1.2E+18	6.1E+18	4.5E+17	3.2E+14	1.4E+14	2.1E+18	1.1E+18
ESS	PE	5.7E+06	6.6E+06	6.8E+05	1.1E+04	4.5E+04	2.8E+05	7.6E+05
	TPI	5.9E+06	6.5E+06	6.1E+05	3.0E+03	1.9E+04	2.3E+05	7.0E+05
	RMP	1.2E+23	4.2E+23	3.0E+22	3.4E+18	4.0E+18	1.5E+23	8.4E+22
20 STRs	PE	2.0E+08	3.9E+08	2.4E+07	9.5E+04	5.0E+05	9.0E+06	2.5E+07
	TPI	1.9E+08	3.7E+08	1.9E+07	1.7E+04	1.4E+05	6.5E+06	2.0E+07
	RMP	3.1E+09	5.5E+09	4.9E+09	2.6E+08	4.0E+07	4.2E+09	3.0E+09
MiniFiler	PE	8.7E+01	1.2E+02	8.4E+01	1.3E+01	2.0E+01	7.6E+01	5.5E+01
	TPI	3.1E+03	1.3E+04	2.2E+03	1.8E+02	1.7E+02	6.4E+02	1.1E+03

comprehensive set of STRs chosen specifically to analyze population variability.

Results of the *structure* analyses gave several different types of data with which to assess the ancestry informativeness of STR sets that can be summarized as: (i) triangle plots of three-way group comparisons where close proximity to a vertex denotes high probability of group membership and therefore a reliable indication of that ancestry; (ii) a measurement of misclassification error from individuals with less than 0.5 (50%) membership of their true population group of origin; (iii) cluster plots where individual group membership proportions are represented by $n = K$ color segments in columns and where informative markers for ancestry in sufficient numbers tend to produce clean plots with pre-assigned population groups corresponding well to the groupings set by K; and (iv) summary values including group-wide cluster membership proportions and a genetic divergence matrix based on overall net distance values.

All three-way comparison triangle plots were made for the seven group comparisons to assess use of *structure* for ancestry analysis of the full CEPH panel (supplementary Fig. S1). Triangle plots summarize sets of three likelihood ratios from three-group membership probabilities with each vertex representing a maximum probability of 1. The comparison of EUR, ME and S ASN populations showed no differentiation with a clear lack of discernable separation of individuals from these groups. Triangle plots for this comparison with Identifiler and ESS STR sets are shown in Fig. 3. It can be concluded that the CEPH populations of Europe, Middle East and South Asia, essentially components of a larger trans-continental Eurasian group, do not exhibit well-differentiated allele variation in the 20 forensic STRs studied. This observation largely reflects the patterns found in much larger panels of both STRs [11] and SNPs [30] indicating the lack of divergence we observed is a characteristic of the populations not a limitation of the marker sets themselves. Therefore ME and S ASN individuals were excluded from follow up *structure* analyses to help produce the clearest population differentiations. In contrast the most clearly differentiated CEPH population groups in *structure* were AFR, EUR and AME with the plots in Fig. 3 showing a large majority of individuals positioned close to the correct vertex using both 15-plex sets.

A measure of the efficiency of ancestry inference using *structure* can be obtained from individual group membership values provided for each run. These are listed for each CEPH sample in supplementary Table S2 for the reduced CEPH panel. Summary group-wide membership proportions are shown in Table 4. Assuming a group membership proportion of 0.5 or more to the correct cluster infers a successful ancestry assignment we recorded individuals showing this value in an alternative group as erroneous assignments and those with less than 0.5 membership across all groups as not assigned. The proportions of CEPH individuals correctly and incorrectly assigned or not assigned are summarized in Table 5 for each STR set. The values in Table 5 indicate ancestry assignment error using group membership proportions ranged from 1.7% to 12.4% (both these values using Identifiler STRs) when attempts to group OCE samples are ignored – a differentiation that failed for ESS and 20 STR sets. This represents a general error rate of ~10% since all error rates except one were below this level. Furthermore a majority of group membership proportions in the incorrectly assigned samples were substantially lower than the median values of those correctly assigned. Therefore in addition to a membership proportion >0.5 the value itself can guide the interpretation of an assignment. We recorded the number of incorrectly assigned individuals with group membership proportions >0.9 (ESS and 20 STRs, 4 groups) or >0.7 (Identifiler, 5 groups) as a more realistic indication of the rate of erroneous assignment of unknown samples (summarized in supplementary Table S2). With these criteria there is a significant drop in the error rate to mainly 1–2% with one outlier error rate of 3.8%, but a consequent drop in the number of individuals that can be assigned when using these higher minimum membership proportions of 0.7 or 0.9. As would be expected data from all 20 STRs gives the most classification success, while the ESS 15-plex gives better overall performance than Identifiler for AFR, EUR, and AME differentiations although this might relate in part to the failure of ESS loci to infer the ancestry of any CEPH OCE samples. In summary, the group membership proportions data suggests that in routine use as ancestry indicators STRs have potential to provide a relatively low error system only if strict group membership proportion criteria are used with a consequent lower assignment rate of ~50–80%. Overall, the ESS 15-plex appears to give a reasonably practical balance between success and error rates for a four-group comparison with the advantage that it is a single multiplex likely to be broadly adopted in the near future. The summary population distance metrics from *structure* are given in supplementary Table S3. Values indicate that forensic STRs lack sufficient divergence to adequately differentiate OCE populations from other groups and the same lack of divergence affects the specific differentiation of EUR and E ASN population groups.

The misclassification rates obtained from a Bayesian analysis of the CEPH panel STR profiles for 3, 5 and 7 population group differentiations are given in supplementary Table S4. The three-

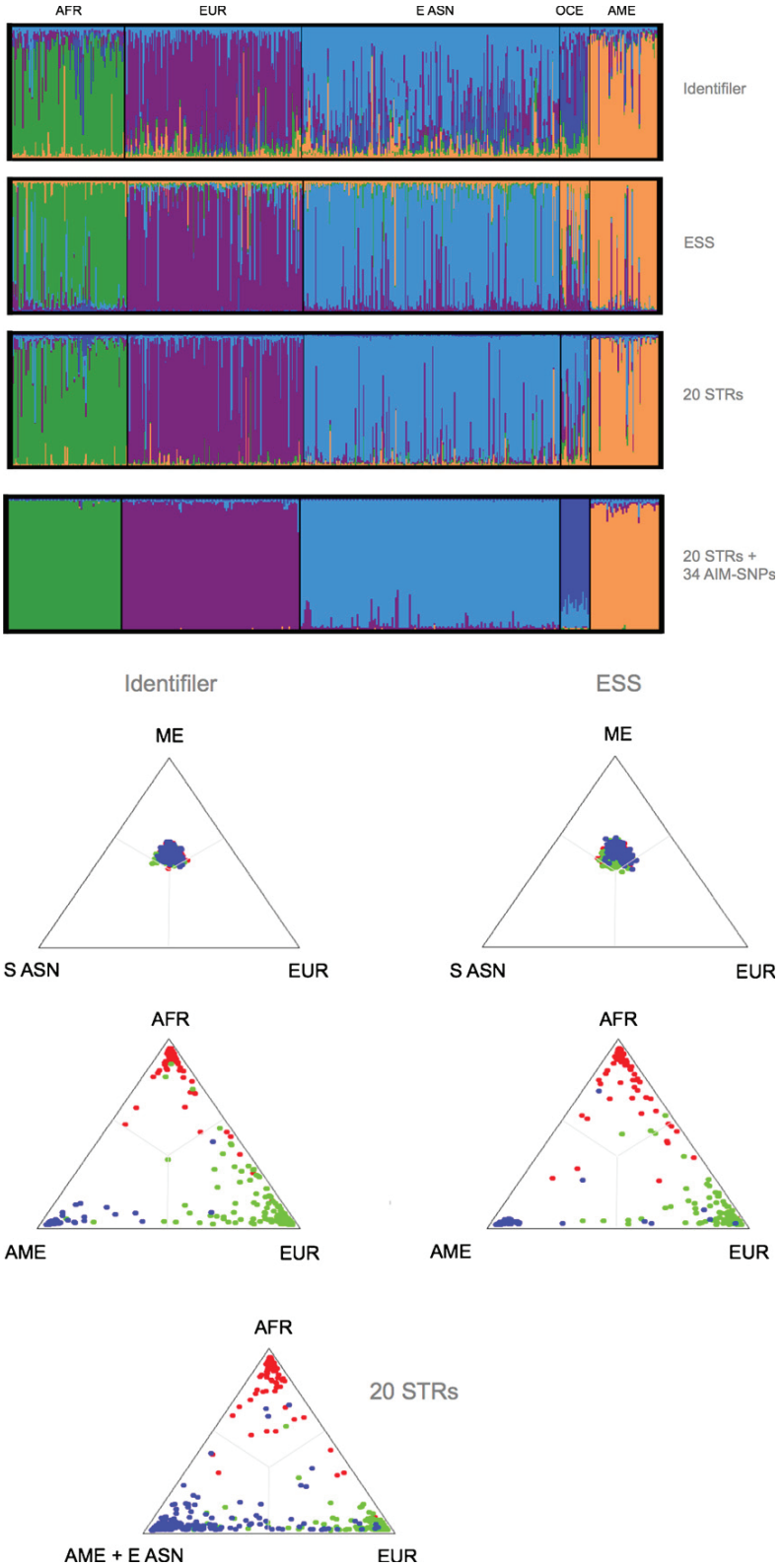


Fig. 3. Cluster plots at K:5 from *structure* analysis of 3 STR sets plus 20 STRs with 34 AIM-SNPs. Triangle plots show the least differentiated and most differentiated three-way population group comparisons.

Table 4
Structure analysis: average membership proportions of pre-defined population groups in five clusters at an optimum grouping of K:5 for Identifiler, ESS and all 20 STRs. Values in bold indicate clusters correctly matched to pre-defined groups.

	1	2	3	4	5	*n*
Identifiler						
Africa	**0.7**	0.084	0.059	0.111	0.046	101
Europe	0.061	**0.662**	0.123	0.093	0.06	155
East Asia	0.039	0.119	**0.579**	0.188	0.074	225
Oceania	0.101	0.108	0.121	**0.576**	0.094	26
America	0.02	0.048	0.056	0.059	**0.817**	59
ESS						
Africa	**0.783**	0.066	0.096	0.023	0.032	
Europe	0.035	**0.795**	0.112	0.016	0.042	
East Asia	0.041	0.102	**0.809**	0.008	0.041	
Oceania	0.091	0.219	0.257	0.024	0.41	
America	0.024	0.072	0.045	0.005	**0.854**	
20 STRs						
Africa	**0.845**	0.062	0.051	0.023	0.02	
Europe	0.025	**0.83**	0.112	0.006	0.025	
East Asia	0.019	0.079	**0.87**	0.004	0.027	
Oceania	0.105	0.253	0.509	0.014	0.119	
America	0.013	0.042	0.058	0.006	**0.881**	

group differentiation between Africans, Europeans and East Asian populations shown in Table 6 reflects what could be considered a reasonable starting point for a broad-brush ancestry assignment. In this differentiation the ESS loci provide a better overall error rate of 15% compared to 20–25% for Identifiler though Africans are better differentiated with these STRs. As would be expected, using genotypes from all 20 STRs gives the best classification error reduced to ~12%. An overall ancestry assignment error rate with the Bayesian classifier can be summarized as: 12–15% using ESS STRs to make a three-group differentiation and 15% using 20 STRs to make a five group differentiation that includes Americans and Oceanians. The classification error we observed using Identifiler is considerably higher than the summary error rate of ≤10% found in a recent study of the same STRs [14] although these Bayesian analyses were limited to pairwise group comparisons and did not include African populations.

Table 6
Reclassification of reduced CEPH panel samples using a Bayesian classifier. Bold values denote % success using each STR set for a simple three-way group differentiation.

	% classified into each group		
	AFR	E ASN	EUR
Identifiler			
CEPH Africans	**93.2**	3.88	2.91
CEPH E Asians	7.49	**79.74**	12.78
CEPH Europeans	7.01	17.2	**75.8**
ESS			
CEPH Africans	**82.52**	12.62	4.85
CEPH E Asians	4.41	**84.58**	11.01
CEPH Europeans	3.18	11.46	**85.35**
20 STRs			
CEPH Africans	**89.32**	7.77	2.91
CEPH E Asians	2.2	**84.58**	10.13
CEPH Europeans	2.55	9.55	**87.9**

Finally to test the viability of improving classification success of forensic STR sets with dedicated ancestry-informative SNPs we performed identical *structure* and Bayesian analyses with genotypes from the 20 STRs plus a previously developed 34plex ancestry-informative SNP panel [19]. Using both marker sets combined to classify the reduced CEPH panel resulted in error free assignments and group membership proportions greater than 95% in most cases (above 70% in all cases). It is notable that this represents an improvement of the SNP based inference of ancestry, as well as an improvement to the use of STRs alone, as the 34plex SNP assay previously incorrectly classified five Europeans (i.e. ~3% error for EUR) that are now correctly assigned. Furthermore we did not originally attempt a differentiation of American and Oceanian populations with the SNP panel alone but the successful use of

Table 5
Structure analysis: rates of ancestry assignment for each STR set using group membership proportions of 0.5 or more to denote correct assignment. Samples not assigned had proportions less than 0.5 in all groups. All OCE were incorrectly assigned for ESS and 20 STR sets. MMP: mean membership proportions.

STR set	Group (*n*)	Not assigned	Correctly assigned	MMP correctly assigned	Incorrectly assigned	MMP incorrectly assigned
Identifiler	AFR	10	83	0.826	8	0.605
	(101)	9.9%	82.2%		7.9%	
	EUR	25	115	0.831	15	0.597
	(155)	16.1%	74.2%		9.7%	
	E ASN	43	154	0.722	28	0.578
	(225)	19.1%	68.4%		12.4%	
	OCE	3	19	0.722	4	0.575
	(26)	11.5%	73.1%		15.4%	
	AME	5	53	0.9	1	0.696
	(59)	8.5%	89.8%		1.7%	
ESS	AFR	6	87	0.918	8	0.715
		5.9%	86.1%		7.9%	
	EUR	6	135	0.925	14	0.692
		3.9%	87.1%		9.0%	
	E ASN	10	196	0.925	19	0.687
		4.4%	87.1%		8.4%	
	AME	3	54	0.963	2	0.755
		5.1%	91.5%		3.4%	
20 STRs	AFR	3	94	0.931	4	0.6375
		3.0%	93.1%		4.0%	
	EUR	3	137	0.943	15	0.7155
		1.9%	88.4%		9.7%	
	E ASN	5	210	0.947	10	0.719
		2.2%	93.3%		4.4%	
	AME	2	54	0.972	3	0.684
		3.4%	91.5%		5.1%	

combined STRs and AIM-SNPs suggests this is a viable option. The cluster plot at K:5 from *structure* analysis of the 20 + 34 marker set is included in Fig. 3.

Therefore our assessment of STRs using *structure* and an equivalent Bayesian classifier suggests that STR data from routine casework could form part of an ancestry prediction system but would only reach an adequate level of ancestry informativeness when combined with other population differentiating markers, such as AIM-SNPs. An advantage of *structure* is the ability to combine both types of allelic data in the same analysis. A disadvantage is that reference population genotypes are required so laboratories interested in this analysis would need to build their own reference sets since we have protected the privacy of the CEPH donors by releasing summary allele frequency data only. For this reason we are currently adapting our in-house online Bayesian classifier [19] to work with user-defined STR and SNP allele frequencies and/or those of CEPH to provide ancestry assignments for genotype profiles of both marker types.

3.5. Tests for association by linkage of vWA and D12S391

For the unadjusted χ^2 tests of independence of vWA and D12 the p-value was 0 – rejecting a null hypothesis of independence, but as described the χ^2 test is not reliable in this case. In contrast, selective grouping of rare alleles into the nearest common repeat gave a p-value of 0.9188 strongly indicating marker independence. Grouped allele simulations gave a proportion of p-values smaller than the first test of 0.999 confirming the validity of the original χ^2 p-value.

Therefore despite some limitations imposed on a standard association test approach by a large number of rare alleles in both the chromosome-12 STRs, we found no evidence of association as a result of their close physical proximity. The handling of vWA and D12 genotypes from closely related individuals, as is normally the case in relationship testing, will need more thorough analysis with the reporting of these STR data as non-independent diplotypes most likely to be the optimum approach.

3.6. Analysis of Θ estimates

Table 7 lists the range of Θ estimates between populations (Table 7A) and between groups (Table 7B), for each of the 20 STRs. Fig. 4 plots the likelihood of the estimates as a function of Θ for each STR. Estimates of Θ from the two different methods used were generally concordant, so in all but three cases the method of moment estimates fell within the 95% maximum likelihood confidence intervals listed in Table 7 and denoted by the intersections of solid and dotted lines in the plots. Furthermore in STRs with narrower allele ranges both Θ estimates were very similar suggesting such STRs give consistent measures of substructure even when sample sizes may be below ideal numbers of individuals.

The average Θ values across 20 loci shown in Table 7 indicate that the commonly used correction factor of $\Theta = 0.1$ is highly conservative as it represents more than three times the average value derived from our analysis of a wide range of populations, separated in this panel by global-scale geographic distances. To further compare the actual Θ values obtained in 20 STRs to a general purpose 0.1 correction factor, we performed cluster analysis of the Θ values for populations and for groups. Supplementary Table S5 summarizes the cluster analysis groupings and shows that, with the exception of the outlying marker TH01 which gave values of $\Theta = 0.07$ (populations) and $\Theta = 0.086$ (groups), all other STRs fall into two main sets of loci with Θ values of 0.02 or 0.04. Therefore applying the standard forensic Θ correction factor represents a very conservative approach to adjustment for population stratification. Furthermore, our Θ estimates are comparable to a previous detailed study of substructure amongst a large set of Australian population samples using Identifiler STRs by Walsh et al. [31]. This study found a maximum Θ of 0.02 between Europeans and the most isolated Native Australian population analyzed: unadmixed Aboriginals from the Northern Territories, but also some outlying values amongst certain very stratified Native Australian populations as high as $\Theta = 0.06$.

In summary, we find no evidence from substructure analysis comparing the isolated CEPH populations with the other panel populations to suggest high levels of endogamy will inflate Θ values to any level close to the commonly applied $\Theta = 0.1$ correction factor. Therefore continued use of this value remains a highly conservative approach to profile frequency adjustment acting in favor of the defendant [32].

Table 7A
Theta estimates for 20 STRs amongst 51 strata (CEPH populations) using maximum likelihood estimation (MLE) and method of moment likelihood estimation (MoM). MLE LCI/UCI: 95% lower/upper confidence intervals, SE: standard error.

STR	MLE	MLE LCI	MLE UCI LCI	MoM	MoM SE
D1S1656	0.02550	0.01836	0.03443	0.02710	0.01057
D2S441	0.06381	0.04722	0.08486	0.08263	0.01975
D10S1248	0.01877	0.01032	0.02984	0.02708	0.01057
D12S391	0.01870	0.01253	0.02637	0.02732	0.01060
D22S1045	0.04734	0.03361	0.06485	0.05654	0.01558
D8S1179	0.01998	0.01199	0.03037	0.02671	0.01050
D21S11	0.02074	0.01389	0.02924	0.02107	0.00951
D7S820	0.02559	0.01612	0.03813	0.03087	0.01136
CSF1PO	0.02499	0.01499	0.03873	0.02626	0.01048
D3S1358	0.01830	0.00970	0.03001	0.01908	0.00914
TH01	0.07823	0.05789	0.10404	0.07914	0.01921
D13S317	0.05401	0.03889	0.07292	0.05441	0.01525
D16S539	0.02061	0.01175	0.03235	0.02398	0.01004
D2S1338	0.03538	0.02618	0.04635	0.04329	0.01339
D19S433	0.03293	0.02404	0.04379	0.03353	0.01169
vWA	0.01879	0.01052	0.02968	0.02199	0.00966
TPOX	0.04578	0.03067	0.06638	0.04238	0.01322
D18S51	0.02145	0.01437	0.03024	0.02478	0.01037
D5S818	0.03560	0.02385	0.05091	0.03331	0.01164
FGA	0.01136	0.00625	0.01783	0.01294	0.00819
Average	0.03189			0.03572	
Maximum	0.07823			0.08263	

Table 7B
Theta estimates for 20 STRs amongst 7 population group strata.

STR	MLE	MLE LCI	MLE UCI	MoM	SE MoM
D1S1656	0.01574	0.00976	0.02519	0.02040	0.01375
D2S441	0.05911	0.03547	0.09931	0.07549	0.04226
D10S1248	0.01117	0.00514	0.02324	0.01948	0.01325
D12S391	0.01299	0.00767	0.02144	0.01854	0.01272
D22S1045	0.02985	0.01731	0.05283	0.04040	0.02450
D8S1179	0.00868	0.00391	0.01762	0.01076	0.00840
D21S11	0.01158	0.00695	0.01910	0.01260	0.00943
D7S820	0.01724	0.00862	0.03363	0.01957	0.01336
CSF1PO	0.01422	0.00697	0.02900	0.01359	0.01001
D3S1358	0.01202	0.00517	0.02656	0.01533	0.01095
TH01	0.08617	0.04998	0.14994	0.06991	0.03953
D13S317	0.04030	0.02377	0.06971	0.04886	0.02892
D16S539	0.01691	0.00846	0.03349	0.01846	0.01269
D2S1338	0.02312	0.01410	0.03768	0.02700	0.01736
D19S433	0.02232	0.01384	0.03638	0.02171	0.01446
vWA	0.02182	0.01178	0.04059	0.02191	0.01458
TPOX	0.04682	0.02434	0.09082	0.03659	0.02249
D18S51	0.01207	0.00712	0.02004	0.01863	0.01287
D5S818	0.03354	0.01845	0.06205	0.02021	0.01364
FGA	0.00384	0.00173	0.00733	0.00721	0.00648
Average	0.02498			0.02683	
Maximum	0.08617			0.07549	

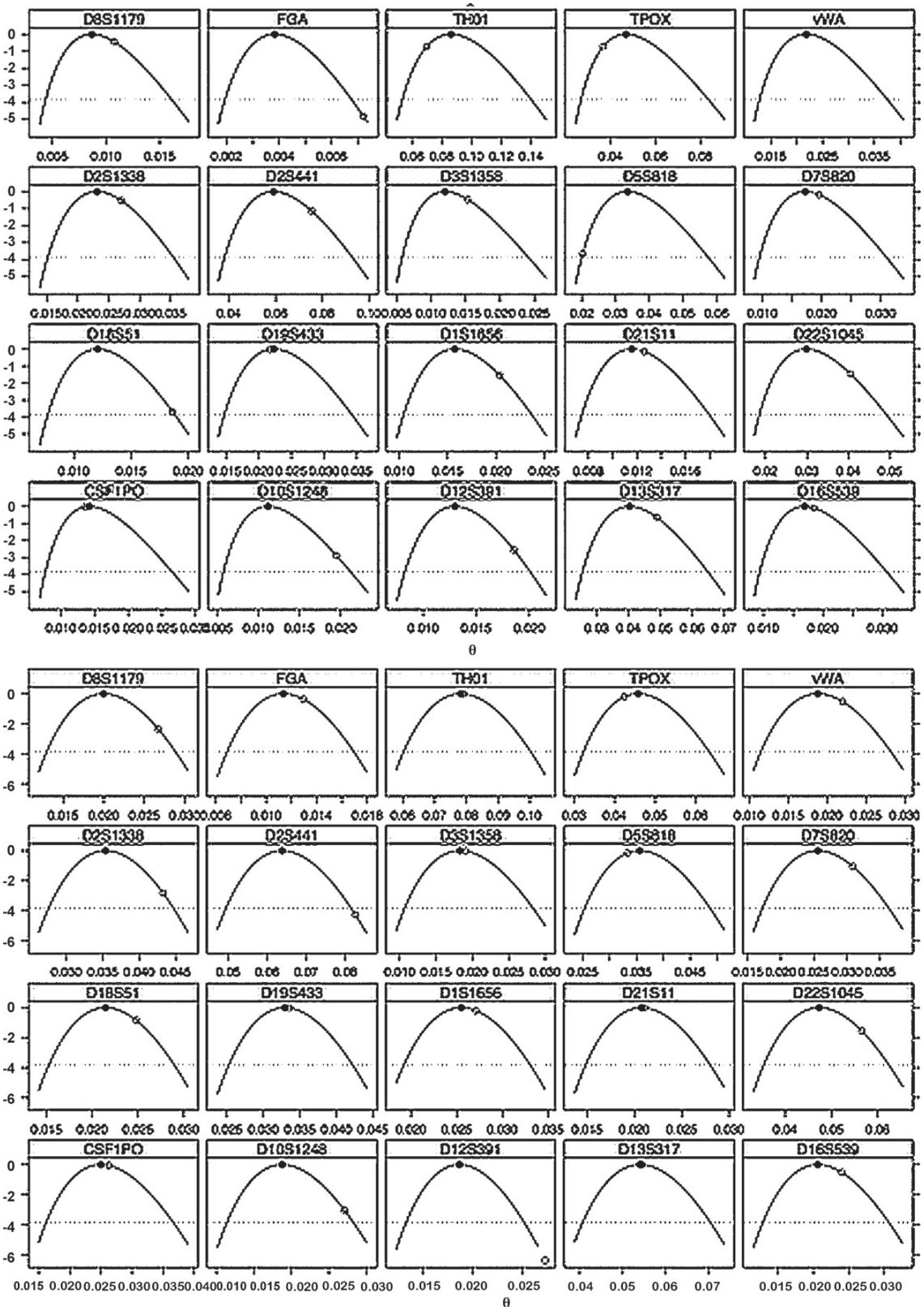


Fig. 4. Estimates of theta for 20 STRs using 51 CEPH populations as substructure indicators: between 51 populations and between the 7 parent groups. Plots show the −log likelihood as a function of theta. The solid circle denotes the maximum likelihood estimation and the open circle denotes Weir and Hill's method of moment estimation. The intersection of the solid and dotted lines indicate a 95% maximum likelihood confidence interval.

4. Conclusions

Analysis of a broadly based population panel such as the CEPH-HGDP allows an assessment of marker variability across a wider geographic scope than is normally possible and we believe is an important preamble to the introduction of the new ESS STRs. The geographic focus of a large majority of reported population studies for the established forensic STRs has tended to reflect those countries with national DNA databases and high-throughput laboratories. In the main these are European and North American and this will continue to be the case with the ESS loci chosen specifically for use in Europe. For this reason we decided to initiate a thorough population survey of the five new ESS loci alongside those of Identifiler rather than wait for kits to become available. This has the benefit of accelerating the assessment of the informativeness of the new ESS 15-plex for identification and paternity analyses. This study has shown that the STRs of the ESS 15-plex will satisfy their intended purpose and provide improved discrimination in nearly all populations compared to the STRs of the commonly used multiplexes they will begin to replace in routine use.

Sequence analysis of the ESS loci indicates that, in addition to better allele variability, the new STRs will provide extended discrimination from nucleotide variation in the repeats. In particular D12S391 has the potential to provide considerably more discrimination from the estimation of the ratio of AGAT and AGAC repeats in the core tandem repeat motifs of this locus by using ICEMS technology to infer the base composition of alleles. Of the short-amplicon ESS loci, D2S441 also shows some sequence complexity that could be exploited for improved discrimination in the future.

Assessment of the ability of STRs to assign ancestry showed that forensic micro-satellites are potentially informative for this purpose. Because the assignment error averages 12–15% for much of the allelic variability shown by the 20 STRs we would continue to advocate the use of dedicated AIM-SNP sets for forensic ancestry analysis [19] and do not believe ancestry inference of DNA profiles alone is a viable strategy. However the combination of highly informative AIM-SNP panels with the STR data from most standard forensic analyses improves the strength of assignments and considerably reduces the assignment error found when using either marker set alone. Combining STRs and a 34plex AIM-SNP set provided an error free classification of the CEPH panel into the five major population groups and has the advantage of combining STR data generated in nearly all forensic cases with a single SNP multiplex.

Amongst the most practical findings from this study for the routine reporting of existing Identifiler profiles or new ESS profiles, is the confirmation that a Θ substructure correction factor of 0.1 represents a highly conservative figure, applicable to all STRs and an adjustment that will always act in favor of a defendant, whether or not they originate from an isolated and potentially endogamous subpopulation. Very few, if any, individuals are likely to be members of populations more stratified than those of the CEPH Karitiana and Surui [25].

An additional benefit from proceeding with development of a stand-alone 5-plex to achieve this study has been its successful use in routine forensic typing where it has proved to be a valuable and sensitive additional multiplex to AB MiniFiler™ when analyzing degraded DNA [8], or to Identifiler and Powerplex 16 when additional STRs are necessary to resolve difficult relationship testing scenarios [33]. Therefore it seems likely that all the STR genotype data generated in this study will be useful in the immediate future as kits are combined for specific applications. Moreover we aim to collate Identifiler and ESS data from other laboratories to improve the coverage and sample sizes of the allele frequencies listed in the *pop.STR* database from this study.

Acknowledgements

M.V. Lareu is supported by Xunta de Galicia, Spain: Fund PGIDTIT06P-XIB228195PR and Ministerio de Educación y Ciencia, Spain: project BIO2006-06178. M. Garcia-Magariños is supported by Ministerio de Educación y Ciencia, Spain: project MTM2008-001661.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2010.02.003.

References

[1] ALFRED allele frequency database: http://alfred.med.yale.edu/alfred/.
[2] H.M. Cann, C. de Tomas, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, et al., A human genome diversity cell line panel, Science 296 (2002) 261–262.
[3] N.A. Rosenberg, Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives, Ann. Hum. Genet. 70 (2006) 841–847.
[4] P. Gill, L. Fereday, N. Morling, P.M. Schneider, New multiplexes for Europe-Amendments and clarification of strategic development, Forensic Sci. Int. 163 (2006) 155–157.
[5] M.V. Lareu, C. Pestoni, M. Schurenkamp, S. Rand, B. Brinkmann, Á. Carracedo, A highly variable STR at the D12S391 locus, Int. J. Legal Med. 109 (1996) 134–138.
[6] M.V. Lareu, S. Barral, A. Salas, C. Pestoni, Á. Carracedo, Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus, Int. J. Legal Med. 111 (1998) 244–247.
[7] J.M. Butler, Y. Shen, B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA, J. Forensic Sci. 48 (2003) 1054–1064.
[8] C. Phillips, A. Barbaro, L. Fernandez Formoso, D. Ballard, D. Syndercombe Court, Á. Carracedo, M. Lareu, Development and validation of a next generation STR ESS-pentaplex, Forensic Sci. Int. Genet. Suppl. 2 (2009) 25–26.
[9] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinformatics 9 (2008) 428.
[10] J. Amigo, C. Phillips, A. Salas, L. Fernandez Formoso, Á. Carracedo, M. Lareu, pop.STR—an online population frequency browser for established and new, Forensic Sci. Int. Genet. Suppl. 2 (2009) 361–362. , doi:10.1016/j.fsigss.2009.08.178.
[11] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381–2385.
[12] Promega Powerstats download page: http://www.promega.com/geneticidtools/powerstats/.
[13] A.L. Lowe, A. Urquhart, L.A. Foreman, I.W. Evett, Inferring ethnic origin by means of an STR profile, Forensic Sci. Int. 119 (2001) 17–22.
[14] M. Graydon, F. Cholette, L.K. Ng, Inferring ethnicity using 15 autosomal STR loci – comparisons among populations of similar and distinctly different physical traits, Forensic Sci. Int. Genet. 3 (2009) 251–254.
[15] I. Halder, B.Z. Yang, H.R. Kranzler, M.B. Stein, M.D. Shriver, J. Gelernter, Measurement of admixture proportions and description of admixture structure in different U.S. populations, Hum. Mutat. 30 (2009) 1299–1309.
[16] D.B. Goldstein, A. Ruiz-Linares, L.L. Cavalli-Sforza, M.W. Feldman, An evaluation of genetic distances for use with microsatellite loci, Genetics 92 (1995) 6723–6727.
[17] D.B. Goldstein, A. Ruiz-Linares, L.L. Cavalli-Sforza, M.W. Feldman, Genetic absolute dating based on microsatellites and the origin of modern humans, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 6723–6727.
[18] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.
[19] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, Á. Carracedo, The SNP*for*ID Consortium, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (2007) 273–280.
[20] vWA data: http://www.ncbi.nlm.nih.gov/genome/sts/sts.cgi?uid=240641; D12 data: http://www.ncbi.nlm.nih.gov/genome/sts/sts.cgi?uid=2703.
[21] D.J. Balding, R.A. Nichols, A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, Genetica 96 (1995) 3–12.
[22] B.S. Weir, The rarity of DNA profiles, Ann. Appl. Stat. 1 (2007) 358–370.
[23] B.S. Weir, C.C. Cockerham, Estimating *F*-statistics for the analysis of population structure, Evolution 38 (1984) 1358–1370.
[24] B.S. Weir, W.G. Hill, Estimating *F*-statistics, Ann. Rev. Genet. 36 (2002) 721–750.
[25] ALFRED population descriptions: The Karitianas make up a very small Amazonian basin population that is composed of less than 200 people (1994 est.) who live in a single village on a reservation in Brazil's Rondonia Province. The approximately 800 Rondonian Surui, or Paiter, live in several small villages scattered along the

border between the Brazilian provinces of Mato Grosso and Rondonia which was in isolation from the outside world until 1969.
[26] T. Tverdebrink, Overdispersion in allelic counts and θ-correction in forensic genetics, Ann. Appl. Stat., 2010, submitted for publication.
[27] D. Schmid, K. Anslinger, B. Rolf, Allele frequencies of the ACTBP2, D18S51, D8S1132, D12S391, D2S1360, D3S1744, D5S2500, D7S1517, D10S2325 and D21S2055 loci in a German population sample, Forensic Sci. Int. 151 (2005) 303–305.
[28] H. Oberacher, F. Pitterl, G. Huber, H. Niederstätter, M. Steinlechner, W. Parson, Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry, Hum. Mutat. 29 (2008) 427–432.
[29] S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman, L.L. Cavalli-Sforza, Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 15942–15947.
[30] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100–1104.
[31] S.J. Walsh, R.J. Mitchell, F. Torpy, J.S. Buckleton, Use of subpopulation data in Australian forensic DNA casework, Forensic Sci. Int. Genet. 1 (2007) 238–246.
[32] J.S. Buckleton, J.M. Curran, S.J. Walsh, How reliable is the subpopulation model in DNA testimony? Forensic Sci. Int. 157 (2006) 144–148.
[33] C. Phillips, M. Fondevila, M. García-Magariños, A. Rodriguez, A. Salas, Á. Carracedo, M.V. Lareu, Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers, Forensic Sci. Int. Genet. 2 (2008) 198–204.

Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and *in silico* binding site prediction.

Ortiz-Barahona A, Villar D, Pescador N, Amigo J, del Peso L

Nucleic Acids Research. 04/2010; 38(7):2332-45.

La respuesta de transcripción impulsada por el factor de inducción de hipoxia (HIF) es central para la adaptación a la restricción de oxígeno. Por lo tanto, la completa identificación de las dianas de HIF es esencial para la comprensión de las respuestas celulares a la hipoxia. Aquí se describe una estrategia computacional basada en la combinación de rastros filogenéticos y meta-análisis de perfiles de transcripción para la identificación de genes diana HIF. La comparación de los candidatos resultantes con el ya publicado HIF1A inmunoprecipitación de cromatina en todo el genoma indica una alta sensibilidad (78%) y especificidad (97,8%). Para validar nuestra estrategia, hemos realizado HIF1A inmunoprecipitación de cromatina en un conjunto de supuestos objetivos. Nuestros resultados confirman la robustez de la estrategia computacional para predecir sitios de unión de HIF y revelan varias dianas novedosas de HIF, incluyendo el factor de transcripción de silenciamiento co-represor del silenciador de RE1 (RCOR2). Además, el alineamiento de los polimorfismos descritos con sitios de unión predichos para HIF identificó varios polimorfismos de nucleótido único (SNPs) que puedan alterar la unión de HIF. Como prueba de concepto, demostramos que el SNP rs17004038, que se alinea con un *locus* de un elemento funcional de respuesta a la hipoxia en el factor de inhibición de migración de macrófagos (MIF), previene la inducción de este gen por la hipoxia. En conjunto, nuestros resultados muestran que la estrategia propuesta es una poderosa herramienta para la identificación de dianas directas de HIF que amplía nuestro conocimiento de la adaptación celular a la hipoxia y proporciona pistas sobre la variación inter-individual en esta respuesta.

2332–2345 *Nucleic Acids Research, 2010, Vol. 38, No. 7* *Published online 8 January 2010*
doi:10.1093/nar/gkp1205

Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and *in silico* binding site prediction

Amaya Ortiz-Barahona[1], Diego Villar[1], Nuria Pescador[1], Jorge Amigo[2] and Luis del Peso[1,*]

[1]Department of Biochemistry, Universidad Autónoma de Madrid-Instituto de Investigaciones Biomédicas CSIC, Madrid and [2]Spanish National Genotyping Center (CeGen), Genomic Medicine Group, CIBERER, Universidad de Santiago de Compostela, Galicia, Spain

Received November 4, 2009; Revised December 10, 2009; Accepted December 11, 2009

ABSTRACT

The transcriptional response driven by Hypoxia-inducible factor (HIF) is central to the adaptation to oxygen restriction. Hence, the complete identification of HIF targets is essential for understanding the cellular responses to hypoxia. Herein we describe a computational strategy based on the combination of phylogenetic footprinting and transcription profiling meta-analysis for the identification of HIF-target genes. Comparison of the resulting candidates with published HIF1a genome-wide chromatin immunoprecipitation indicates a high sensitivity (78%) and specificity (97.8%). To validate our strategy, we performed HIF1a chromatin immunoprecipitation on a set of putative targets. Our results confirm the robustness of the computational strategy in predicting HIF-binding sites and reveal several novel HIF targets, including RE1-silencing transcription factor co-repressor (RCOR2). In addition, mapping of described polymorphisms to the predicted HIF-binding sites identified several single-nucleotide polymorphisms (SNPs) that could alter HIF binding. As a proof of principle, we demonstrate that SNP rs17004038, mapping to a functional hypoxia response element in the macrophage migration inhibitory factor (MIF) locus, prevents induction of this gene by hypoxia. Altogether, our results show that the proposed strategy is a powerful tool for the identification of HIF direct targets that expands our knowledge of the cellular adaptation to hypoxia and provides cues on the inter-individual variation in this response.

INTRODUCTION

Cells respond to chronic hypoxia by altering their gene expression pattern to optimize metabolic oxygen consumption, maintain energy balance and restore oxygen supply. Many of the genes involved in this adaptive response are directly regulated by the hypoxia-inducible factor (HIF) (1), a transcription factor that is activated when oxygen tension drops. HIF is a heterodimer composed of an oxygen-regulated alpha subunit (HIFα) (2) and a constitutively expressed beta subunit (HIFβ, also known as Aryl receptor nuclear translocator, ARNT) (3) that partners with a number of basic-helix–loop–helix transcription factors. Oxygen affects both HIFα half-life (4) and transactivation (5). In normoxia, HIFα is hydroxylated at two proline residues (6,7) by a family of dioxygenases (EGL nine homologs, EGLNs) that require oxygen as cosubstrate (8,9). This posttranslational modification labels HIFα for proteosomal degradation, as the proline-hydroxilated form is recognized by an E3-ubiquitin ligase complex that contains the VHL tumor suppressor (10). In addition, another dioxygenase (factor inhibiting HIF, FIH) catalyzes the oxygen-dependent hydroxilation of an asparagine residue, located in the C-terminal transactivation domain, preventing its

*To whom correspondence should be addressed. Tel: +34 91 585 4440; Fax: +34 91 585 4400; Email: luis.peso@uam.es; lpeso@iib.uam.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

interaction with the p300 coativator and blunting HIFα transcriptional activity (11–13). In hypoxia, all these hydroxylation reactions become compromised, due to the reduced availability of oxygen, resulting in HIFα stabilization and recruitment of coactivators, such as p300. Thus, under hypoxia, HIF accumulation allows its interaction with HIFβ and its binding to the RCGTG motif, known as hypoxia response element (HRE), within regulatory regions of its target genes. There are three genes encoding for HIFα subunits: HIF1α, HIF2α (also known as EPAS) and HIF3α. HIF1α and HIF2α have been extensively studied, while HIF3a remains poorly characterized. The regulation of HIF1α and 2α by hypoxia is similar and both bind to the same core motif (1). However, recent evidence indicates that these transcription factors induce overlapping but not identical sets of genes (14,15), suggesting nonredundant functions for HIF1α and HIF2α.

Given the central role of HIF in the transcriptional response to hypoxia, the characterization of HIF target genes provides critical insights into the adaptations required to cope with reduced oxygen tension. Over a hundred HIF-targets have been described (1) as the result of research efforts focused on individual genes. These studies revealed that many of the genes regulated by hypoxia are involved in the reprogramming of cellular metabolism and restoration of oxygen supply. More recently, a number of studies described the effect of hypoxia in the transcriptome by means of gene expression profiling. These studies, covering a wide range of cell types and conditions (16–26), revealed a large number of novel potential targets. Although undoubtedly relevant, a major drawback of this approach is that it cannot distinguish between direct and secondary HIF targets. In addition, no attempts have been made to combine the results of all these studies. Such integrative studies, or meta-analysis, have higher statistical power to detect relevant effects than single studies and provide a generalization to the individual experiments. In fact, several works (27) have demonstrated that the application of meta-analysis to multiple independent gene expression data sets leads to the identification of sets of significant, differentially expressed genes, void of the artifacts of individual studies. Finally, two recent reports (28,29) coupled transcript profiling and chromatin immunoprecipitation (ChIP) followed by hybridization to genomic tiling microarrays (ChIP–Chip) to identify direct HIF targets. A comparative analysis is needed to reveal the extent of overlap between conclusions of both studies and also whether further studies are required. Thus, in spite of intense research efforts, the complete characterization of HIF targets is still unresolved.

In silico identification of transcription-factor-binding sites (TFBS) is a powerful tool to complement experimental identification of transcription factor targets (30). These methods rely on the comparison of candidate sequences to a position-specific scoring matrix (PSSM) constructed by alignment of known binding sites for the transcription factor of interest. HIF binds to a short, but extremely well-conserved [A/G]CGTG motif. Conservation of other positions outside this motif is controversial: while some studies suggest that some positions show a base distribution significantly different from random expectation (1,31,32), other studies fail to find conservation outside the core RCGTG (28,29). Nevertheless, the low information content of most PSSMs, including that of HIF, and the size of mammalian genomes result in a large number of potential hits across the genome. Since conserved noncoding sequences (CNS) are enriched in *cis*-regulatory elements (33,34), a successful approach to reduce the number of spurious hits is to restrict the search for TFBS to these regions. The identification of CNSs, based on multiple species alignment of noncoding genomic sequences, reveals evolutionarily conserved regions (phylogenetic footprinting) that may have been selected during evolution due their regulatory or structural function. The algorithm PhastCons implements a two state hidden Markov model that provides a score value that reflects the conservation of each base of a reference genome within a multiple species alignment (35). Therefore, potential regulatory regions can be inferred from PhastCons elements (groups of adjacent nucleotides with a significant conservation score).

Recent works (36–38) have demonstrated that the combination of gene expression data and TFBS prediction is a powerful tool for the identification of transcription factor target genes. In the present study, we applied a probabilistic model that integrates the evidence for the regulation of each particular gene by hypoxia (transcript profiling meta-analysis) and the presence of high-scoring HIF-binding sites (HBSs) for the identification of novel HIF targets. The application of this strategy results in a list of 216 predicted targets, most of them not previously reported as regulated by hypoxia. We tested the accuracy of our strategy by experimentally validating several of the identified HBSs by ChIP–quantitative polymerase chain reaction (qPCR). Moreover, we demonstrated that RCOR2, one of the borderline targets identified, is indeed a HIF-target gene. In addition, the strategy reported herein provides the coordinates for several hundred potential HBSs. We propose that, in addition to the identification of HIF-target genes, this information can be useful to identify genome variants within the population that could have an altered hypoxic response. As a proof of principle, we found that one of these variants has a major impact on the hypoxic induction of macrophage migration inhibitory factor (MIF). Given the relevance of hypoxia in pathologies, such as cancer and cardiovascular disease, an altered response to hypoxia could be among the underlying causes explaining different clinical courses and/or response to treatments.

MATERIALS AND METHODS

Cell culture and hypoxic conditions

The cell lines HeLa, HepaC1/4 and HepG2 were maintained in Dulbecco's modified Eagle medium, while HepaC1 and HepaC4 cell lines were grown in MEM-α medium. In all cases, the culture medium was supplemented with 100 U/ml penicillin, 100 μg/ml streptomycin and 10% (v/v) fetal bovine serum. Cells were grown at 37°C in a humidified atmosphere containing 5% CO_2.

For hypoxia treatments, cells were grown at the indicated oxygen concentration in a Whitley hypoxystation (don Whitley Scientific, UK).

Plasmid construction

Human genomic DNA extracted from HeLa cells was used as template for PCR amplification of MIF and RCOR2 promoter regions using primers 1 + 2 and 7 + 8 (Supplementary Table S1), respectively. Reporter constructs were generated by cloning the PCR products into the pGL3-Basic plasmid (Invitrogen). The identity of all constructs was verified by sequencing. The mutant HRE and single-nucleotide polymorphism (SNP) constructs were generated by site-directed mutagenesis, employing PCR QuikChange Site-direct mutagenesis kit (Stratagene). Primers harboring the desired mutation were 2 + 3 (HREmut MIF), 9 + 10 (HREmut RCOR2) and 5 + 6 (SNP-HREmut MIF), respectively (Supplementary Table S1).

Reporter assays

Cells were plated in six-well plates 24 h prior transfection. Each plate was transfected with a DNA mixture containing 0.9 μg (HeLa cells) or 1.9 μg (HepG2 cells) of the indicated reporter plasmid and 0.1 μg of a plasmid encoding the *Renilla* firefly luciferase under the control of a SV40 promoter. 12–13 h after transfection, cells were replated in 24-well plates and then transferred to hypoxic conditions (1% oxygen) or left under normoxic conditions for 24 h. Subsequently, firefly and renilla luciferase activities were determined using a dual luciferase system (Promega, Madison, WI, USA). In order to correct for transfection efficiency, the luciferase activity was normalized to the *Renilla* luciferase activity. Each experimental condition was assayed in duplicate.

ChIP assays

For ChIP assays, HeLa cells were grown on 10-cm plates until they reached 85% confluence, at which point they were exposed to hypoxia (1% oxygen) or left under normoxic conditions for 6 h. Subsequently, cells were fixed for 12 min at 4°C by adding formaldehyde to culture media to final concentration of 1% (v/v). Cross-linking was stopped by the addition of glycine (0.125 M final). The cells were washed with cold phosphate-buffered saline (PBS) and then lysed by scraping in 1 ml of lysis buffer [1% sodium dodecyl sulfate (SDS), 10 mM EDTA, 50 mM Tris/HCl, pH 8.1 and a protease inhibitor cocktail, Roche]. Cell lysates were incubated on ice for 10 min and then sonicated to shear the DNA to fragments between 200 and 1500 bp. After the removal of the insoluble material by centrifugation, 50 μl of each sample was removed and stored (input), while 100 μl were diluted in 1-ml immunoprecipitation buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl and 20 mM Tris/HCl, pH 8.1). The lysates were precleared with 200 μg of a Salmon Sperm DNA/Protein A agarose 50% slurry (Upstate Biotechnology, Lake Placid, NY, USA) for 1 h at 4°C. The samples were then immunoprecipitated twice, initially with whole rabbit serum for 6 h [immunoglobulin G (IgG) control] and then overnight at 4°C with a polyclonal anti-HIF1 alpha antiserum (Abcam, ab2185). Immunocomplexes were recovered by the addition of 400 μg of Salmon Sperm DNA/Protein A agarose 50% slurry to the samples that were then sequentially washed for 15 min in TSE I (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris/HCl, pH 8.1 and 150 mM NaCl), TSE II (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris/HCl, pH 8.1 and 500 mM NaCl) and buffer III (0.25 M LiCl, 1% NP-40, 1% deoxycholate, 1 mM EDTA and 10 mM Tris/HCl, pH 8.1). Finally, the complexes were washed twice with TE buffer (10 mM Tris, pH 8.0 and 1 mM EDTA) and extracted twice with a buffer containing 1% SDS and 0.1 M $NaHCO_3$. The eluates were pooled, and cross-linking was reversed by the addition of 200 mM NaCl (final concentration) and overnight incubation at 65°C. The proteins were removed by the addition of proteinase K (30 μg/sample) for 2 h at 42°C, and the DNA was purified by phenol–chloroform extraction and ethanol precipitation. Immunoprecipitated DNA was amplified by qPCR using the primers (11–48) indicated in Supplementary Table S1.

RNA extraction and qPCR

Cells were harvested in 1 ml of Ultraspec reagent (Biotecx). Subsequently RNA was reverse-transcribed to cDNA (Improm-II reverse transcriptase; Promega).

q-PCR was performed with the LC FastStart DNA master SYBR GreenI kit (Roche Applied Science) and in a Light Cycler system (Roche Applied Science) using the indicated primers (Supplementary Table SI). Data were analyzed with Light Cycler software version 3.5.28 (Idaho Technology). For each sample, duplicate determinations were made and the gene copy number was normalized to the amount of β-actin.

Meta-analysis of gene profiling data sets

For the meta-analysis, we downloaded 16 independent experiments from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) (39) database (Supplementary Table S2). For those experiments analyzing more than two conditions (for example the effect of hypoxia and HIF overexpression), we generated independent data sets for each comparison. Thus, we generated 19 data sets from the 16 experiments (Supplementary Table S2). In all the cases, untreated normoxic cells were used as reference. For each data set, we calculated the mean for each probe values in the biological replicates. Probes with null values were discarded. Then, for each probe, we calculated the effect of treatments (hypoxia, hypoxia mimetics or HIF expression) as the logarithm of the ratio of the means of treated and control samples. Finally, individual log-ratios values were normalized by subtraction of the mean of all the log-ratios across the data set and division by their standard deviation.

For the meta-analysis, each gene locus was treated independently and tested for the null hypothesis that no

gene was modulated by treatments. To this end, the normalized log-ratio values of all probes (across all data sets) mapping to the gene locus under consideration were compared to zero using one-sample *t*-test. The resulting *P*-values were corrected for multiple testing by applying false discovery rate. Genes with adjusted *P*-values <0.01 were considered significantly regulated by treatments. Custom Perl scripts were used for the analysis, complemented with R-based extensions for the statistical calculations.

Identification and scoring of HBSs

For the identification of HBSs, we restricted our search to noncoding regions of genes. To this end, we considered all the RefSeq transcripts encoded by each locus and identified the intronic and untranslated regions within by projecting all transcripts, therefore excluding transcript-specific coding sequences. In addition, for each locus, we selected a 5-kb region upstream the transcription start site, TSS. When necessary, the upstream region was trimmed to avoid overlap with adjacent loci. For genes with several TSS, we selected the outermost TSS to define the 5-kb upstream region (the region upstream the remaining TSS is considered as part of the intronic regions). After the localization of all noncoding regions, we identified mammal or vertebrate PhastCons elements (35) within. Adjacent PhastCons elements were fused if more than 50% of the sequence in the resulting fused region was conserved. We refer to these PhastCons elements located in noncoding regions as conserved noncoding sequences (33) or CNSs. Then, we identified conserved RCGTG motifs within these CNSs. A motif was considered conserved when it was present at least in four mammals, including human and mouse. Sequences lacking conserved RCGTG motifs were discarded as potential HBSs. Finally, sequences containing a conserved motif were scored according to a PSSM. For the generation of this matrix, we selected 23 well-characterized HIF-binding sequences corresponding to 22 HIF-target genes, together with an orthologous sequence (Supplementary Table S4). Then, we used a chi-squared test to determine those positions with an observed distribution of residues significantly different to that expected by chance. This analysis revealed that, in addition to the RCGTG motif, some positions from −1 to +17 (being the R residue of the core motif the position +1) showed a significantly skewed distribution ($P < 0.01$). For each position, we calculated the log-odds ratio of the observed frequencies of each nucleotide over the background frequency found for that nucleotide. The background frequency was obtained from the counts of each nucleotide in the CNSs: A, 0.275; C, 0.223; G, 0.229; T, 0.273. The log-odds ratios were arranged in a 4×18 matrix (a column per position and a row for each nucleotide) so that the score for the nucleotide i at position j is:

$$S_{i,j} = \log_2\left(\frac{\text{freq}_{i,j}^{\text{observed}}}{\text{freq}_i^{\text{background}}}\right)$$

To calculate the score for the whole sequence (S) we added the individual scores for each position. Since not all positions had the same information content (Supplementary Table S4), the contribution of each position to the final score was weighted by the information content of the position (I_j):

$$S = \sum_j I_j * S_{i,j}$$

The information content for each position j was calculated from the Shannon entropy:

$$I_j = -\sum \text{freq}_i^{\text{background}} * \log_2\left(\text{freq}_i^{\text{background}}\right) - \sum \text{freq}_{i,j}^{\text{observed}} * \log_2\left(\text{freq}_{i,j}^{\text{observed}}\right)$$

The RefSeq coordinates, PhastCons coordinates and the alignments corresponding to the identified CNSs were downloaded from the UCSC genomic browser (http://genome.ucsc.edu/index.html) (40,41). All coordinates correspond to the hg18 human genome assembly. The analysis was performed with custom scripts written in Perl.

Classification of genes as HIF target/nontarget

To classify any given gene as a HIF target or nontarget, we calculated the relative likelihood that the gene belongs to any of these two groups. To this end, we constructed models that, given the fold induction of the gene and associated *P*-value according to our meta-analysis and the score of the potential HBSs found within the gene locus, assign a probability to the gene in each of the two states. Then, the relative likelihood of being a HIF target is the ratio (odds ratio) between the probabilities according to each model. In the HIF-target model (T), we fitted the distribution of fold induction ($f_t^f(x)$) and HBS score ($f_t^s(y)$) values for the set of well-characterized HIF-target genes (Supplementary Table S4) to a normal (Gauss) density function. For the nontarget (Background, B) model, we assumed that most of the genes in the genome are not regulated by HIF, thus calculated the Gaussian density functions describing the distribution of fold induction and HBS score values for all the analyzed genes [$f_b^f(x)$ and $f_b^s(y)$, respectively]. Then, the probability of a gene being a HIF target given its fold induction (and *P*-value) and HBS score, $P(x,y,p|T)$, is the product of functions describing HBS score and fold induction:

$$P(x,y,p|T) = f_t^s(\max(y_i)) * \left(f_t^f(x) + \left(p * f_b^f(x)\right)\right)$$

where x and p are the meta-analysis values for the fold induction and associated *P*-value respectively and max (y_i) is the maximum score of the HBSs found within the gene locus. Similarly the probability of the gene being a nontarget (background) is:

$$P(x,y,p|B) = f_b^s(\max(y_i)) * \left(f_b^f(x) + \left(p * f_t^f(x)\right)\right)$$

2336 Nucleic Acids Research, 2010, Vol. 38, No. 7

Finally, the ratio of these two likelihoods (odds ratio) represent the relative probability of being a HIF-target gene:

$$\frac{P(x,y,p|T)}{P(x,y,p|B)},$$

for simplicity we refer to this ratio as P_T/P_B ratio.

In these expressions, the contribution of the fold induction value ($f^f(x)$) to the probability was weighted by the *P*-value associated to the mean so that when the fold induction is not reliable (for large *P*-values approaching 1) its contribution to the probability is very similar for target and nontarget:

$$f_t^f(x) + \left(p * f_b^f(x)\right) \approx f_b^f(x) + \left(p * f_t^f(x)\right)$$

In this case, the classification (odds ratio) is just based on the score value:

$$\frac{P(x,y,p|T)}{P(x,y,p|B)} \approx \frac{f_t^s(\max(y_i))}{f_b^s(\max(y_i))}$$

RESULTS

Meta-analysis of gene expression profile data sets from cells exposed to hypoxia

In order to identify HBSs, we designed a strategy based on the intersection of two independent approaches: (i) identification of hypoxia-modulated genes through the analysis of multiple gene expression data sets from publicly available databases (transcription profiling meta-analysis); (ii) identification of evolutionarily conserved HIF-binding motifs within potential *cis*-regulatory regions (phylogentic footprinting). For the first approach, we selected 19 data sets from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) (39) corresponding to 16 independent experiments that analyzed the gene expression profile of cells exposed to hypoxia or hypoxia mimetics (for simplicity, we refer to them as hypoxia herein). We processed each of these data sets to calculate the $\log_2$ of the hypoxia/normoxia ratio (log-ratio) for each probe and then considered as significantly regulated by hypoxia those probes whose log ratio was >1.96 or >2.6 SD above or below the data set mean. A gene was considered modulated when at least one of its probes was significantly up- or downregulated. This analysis revealed that only a small group of genes was induced by hypoxia across all the experiments (Figure 1A and Supplementary Table S3). In addition, even when a relatively relaxed criterion (>1.96 SD from the mean) was used to ascribe genes to the downregulated group, no gene was found consistently repressed by hypoxia in all the experiments (Figure 1A). On the other hand, the number of nonredundant genes modulated by hypoxia increased rapidly with the number of experiments (Figure 1B). When taking into consideration all 19 data sets, we

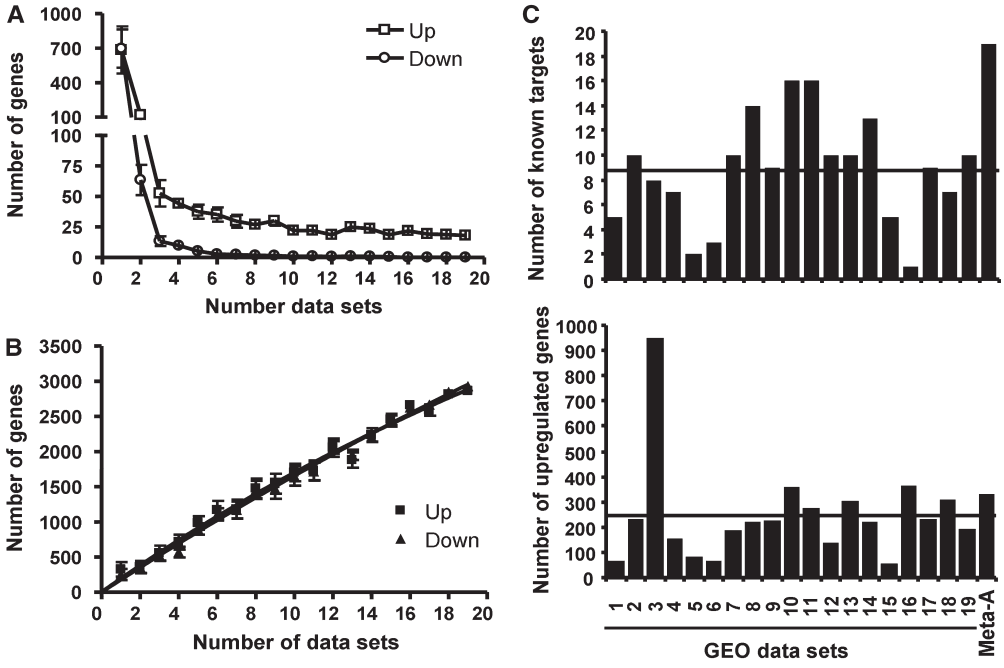


Figure 1. Comparison of individual gene profiling studies versus meta-analysis. The indicated number of data sets (number of data sets) was randomly selected out from the 19 GEO tables without replacement. The number of genes whose expression was 1.96 SD from the mean in all (**A**) or at least in one (**B**) of the selected data set was recorded and the procedure repeated 10 times. The graph represents the mean number of recorded genes and error bars the standard deviation. (**C**) For each individual data set (1 to 19, see Supplementary Table S2), the genes showing a fold induction ratio >2.6 SD above the mean were considered upregulated. In the case of the meta-analysis (Meta-A), genes with a corrected *P*-value <0.01 and mean fold induction positive were considered upregulated. The graph represents the number of known target genes (according to ref. 1) represented in the upregulated group in each case (upper graph) together with the total number of upregulated genes (lower graph). The horizontal lines in each graph represent the average number of known and upregulated genes across the 19 data sets.

found a total number of 2864 up- and 2929 downregulated nonredundant genes that together account for 49.6% of the genes represented in these data sets. These results indicate that the simple intersection of results from individual experiments is too restrictive, while their combination results in excessive noise, highlighting the need for a statistical analysis of the combined data sets. To this end, we treated each gene as an independent hypothesis, compared to the null hypothesis that the gene is not modulated by hypoxia and thus the mean value of the log-ratios of all its probes is 0. For each gene, we obtained the value of the log-ratio for all associated probes across all the data sets, calculated their mean (mean fold induction) and used one-sample *t*-test to ask whether it differed significantly from 0. After correction of the resulting *P*-values for multiple testing (false discovery rate), we selected genes with a *P*-value below 0.01 as regulated by hypoxia. This analysis resulted in a total of 259 (2.22%) genes induced and 191 (1.64%) genes repressed by hypoxia out of 11 673 genes represented in all GEO data sets. As a crude measure of the meta-analysis performance, we looked for known HIF targets (1) in the set of upregulated genes identified in each independent study or in our meta-analysis. As shown in Figure 1C, the meta-analysis excelled the performance of the individual studies, recovering a higher number of known targets than any of them. In addition, the increased sensitivity did not seem to be accompanied by a reduction of specificity, since the total number of upregulated genes identified by the meta-analysis was not different to the average number identified in individual studies.

In-silico identification of HBSs

For the prediction of genome-wide HIF binding positions, we searched for the occurrence of RCGTG motifs in the human genome. Since CNSs are genomic regions enriched in *cis*-regulatory elements, we restricted our search to these regions in order to increase the chances of finding relevant motifs and to reduce the number of spurious hits. In addition, we only considered RCGTG motifs that were conserved in, at least, four species, including mouse. For each locus, we defined CNSs as PhastCons elements mapping to introns, untranslated regions and promoter regions upstream of the TSS. This search resulted in 9458 potential HBSs (conserved RCGTG motifs) distributed across 3980 genetic loci (34.1% of the analyzed genes). We found no conserved HBSs in the remaining 7693 (65.9%) gene loci analyzed.

Integration of the meta-analysis results and the presence of conserved HBSs showed that, as expected, the proportion of genes upregulated by hypoxia that contained at least one conserved HBS was significantly higher than expected by chance ($P = 3.2 \times 10^{-14}$, Table 1). In contrast, we found no significant association between the presence of an HBS and downregulation of the gene by hypoxia ($P = 0.42$, Table 1). Since we found no evidence for a direct role of HIF on downregulation of gene expression, we focused on genes upregulated by hypoxia throughout the study.

Table 1. The presence of an HBS correlates with transcriptional upregulation but not with repression by hypoxia

	HBS+		HBS−		*P*-value
	Obsr.	Expc.	Obsr.	Expc.	
Upregulated	145	88	113	170	3.18×10^{-14}
Nonregulated	3766	3823	7459	7402	
Downregulated	69	64	121	126	0.42
Nonregulated	3766	3771	7459	7454	

Genes were classified as upregulated [mean log2(hypoxia/normoxia) > 0 and $P < 0.01$], downregulated [mean log2(hypoxia/normoxia) < 0 and $P < 0.01$] or nonregulated ($P > 0.01$).
The number of genes in each group with at least one potential HBS (HBS+) or none (HBS−) is shown (Obsr.).
The number of genes expected by chance in each group is also shown (Expc.).
The significance of the difference between the observed versus expected frequencies was calculated by a Chi-squared test and the resulting *P*-value is shown (*P*-value).

Scoring of HBSs

The alignment of a set of well-characterized HREs reveals that, in addition to the core RCGTG motif, other positions present a distribution of bases significantly different to that expected by chance (Supplementary Table S4 and ref. 31). Thus, we decided to use this information to infer functional HBSs. An 18-residue-long PSSM was generated based on the alignment of 46 sequences (see Supplementary Table S4 and 'Materials and Methods' section for details), and subsequently used it to assign a score value to each of the identified HBSs. In order to assess the ability of this score to discriminate functional HBSs, we studied the distribution of scores for HBSs recently identified by genome-wide ChIP–Chip (28,29). A comparative analysis reveals that there is very little overlap among the HIF-binding regions reported in these two studies (Figure 2A), probably because of the different cell lines/experimental conditions, data analysis and array platforms used in each work. Thus, we only considered the overlapping group of genes (Supplementary Table S5) as a reliable set of HIF targets. Figure 2B shows the distribution of scores of our predicted HBSs that map to any of the published HIF-binding regions (28,29). For comparison, we also plotted the score distribution for all the HBSs identified within CNSs across the genome (control). These results show that the score distribution for experimentally determined HBSs is shifted toward higher score values. In particular, the scores of the HBSs mapping to the regions identified in both reports (highly reliable HBSs) are higher than those of control genes, and the mean score for this group is significantly different from that of the controls (Figure 2B). It is worth pointing out that only four (GAPDH, LDHA, PGK1 and TF) of the 39 regions common to both ChIP–chip studies are coincident with the HRE regions used to construct the scoring matrix. Thus, the results shown in Figure 2B are not due to overfitting of our PSSM matrix to a specific set of HBSs.

Our search often predicted several HBSs for a given locus (Supplementary Table S6). Therefore, in order to

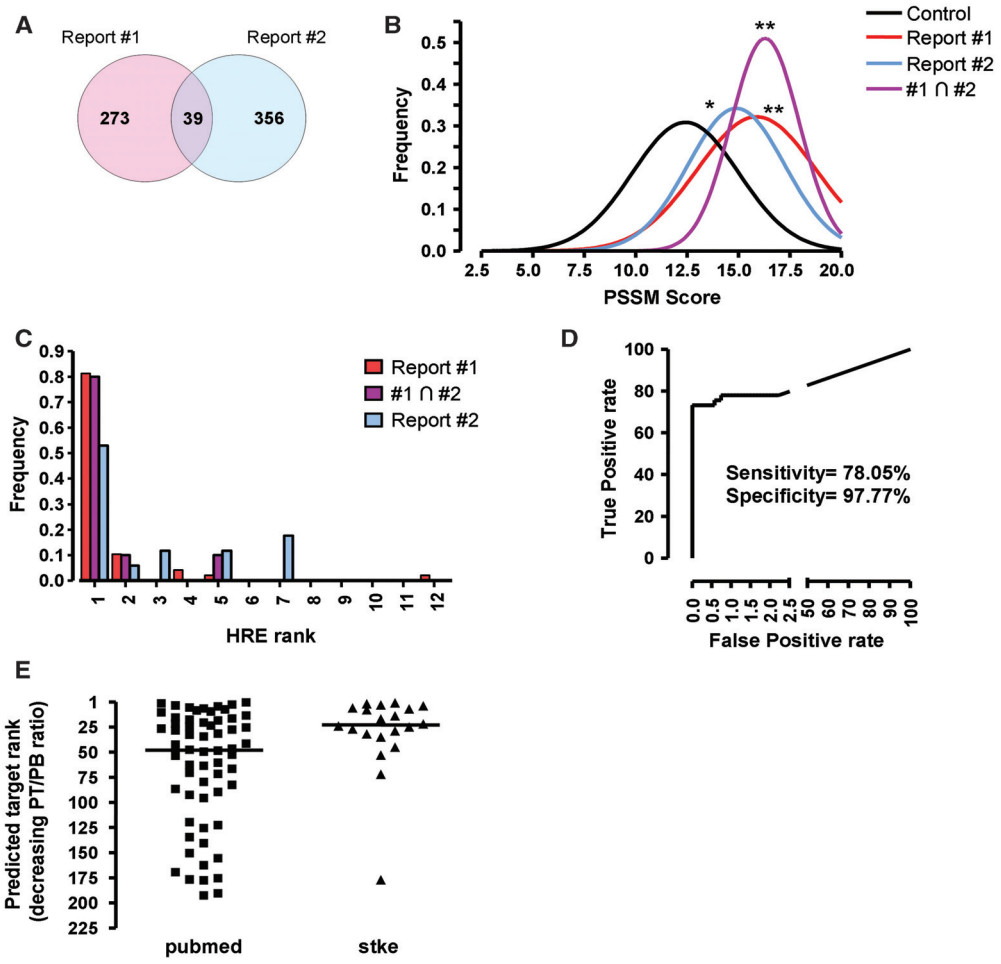


Figure 2. High HBS scores correlate with functional HIF-binding sites. (**A**) Venn diagram showing the number of overlapping HIF-binding sites identified by ChIP–chip in two published reports (ref. 29, Report #1; ref. 28, Report #2). (**B**) The scores of HBSs identified by our strategy were discretized (binning size 0.5 U) and their frequency distribution was calculated and adjusted to a Gauss curve by nonlinear fitting. The graph shows the resulting curves for all the HBSs identified across the genome (control), the HBSs mapping to HIF-binding regions identified by ChIP–chip in each report (Report #1, Report #2) or those HBS in regions common to both reports (#1∩#2). The scores in each group were compared (ANOVA) and statistically significant differences with the control group are indicated by asterisks (*, $P < 0.01$; **, $P < 0.001$). (**C**) The potential HBSs identified for each gene were ranked according to their score in decreasing order (rank 1 corresponds to the highest scoring HBS) and the rank of the predicted HBSs mapping to HIF-binding sites was recorded. The figure shows the rank frequency distribution for predicted HBSs mapping to HIF-binding regions identified by ChIP–Chip in each report (Report #1, Report #2) or regions common to both reports (#1∩#2). (**D**) Receiver operating characteristic (ROC) curve of known positive/negative (see text) targets versus prediction using a P_T/P_B ratio of 6.5 as threshold to classify genes as potential targets. (**E**) Genes identified as potential targets (P_T/P_B ratio >6.5) were sorted in decreasing P_T/P_B ratio order. The graph represents the rank of known HIF targets, according to ref. 1 (Stke) or a bibliographic search (PubMed), within the predicted target list. Horizontal line represents the median of each group.

further test the relevance of the score value, we next studied which one of the HBSs identified for each locus mapped to the sites bound by HIF according to ChIP–chip studies (28,29). To this end, the HBSs identified for each gene were ranked according to their score (rank 1 corresponded to the HBS with the highest score for a given locus) and the number of predicted HBSs of each rank that were coincident with an experimentally determined HBS was represented (Figure 2C). The data show that, in most cases, the experimentally validated HBS for a given locus matched the predicted HBS of highest score value. Altogether, the results in Figure 2B and C support that the HBS score is a good predictor of functionality.

Probabilistic model of integrated binding site and gene expression data

When combined, the two approaches described above resulted in a list (Supplementary Table S6) in which each gene had associated parameters reflecting the magnitude of its modulation by hypoxia (fold induction and associated *P*-value) and one or several potential HBSs

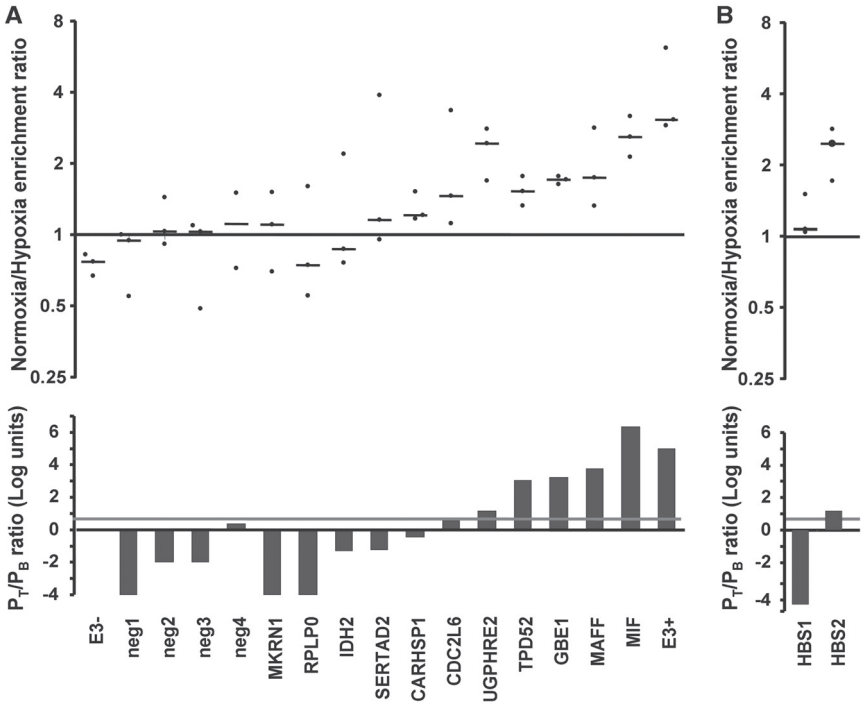


Figure 3. Experimental validation of HIF binding to predicted sites. HeLa cells were exposed to normoxia or hypoxia (1% oxygen) for 6 h. After treatments, cells were processed for chromatin immunoprecipitation using antibodies directed to HIF1a. The binding of HIF1a to the predicted HBS within the indicated genes (**A**) was determined by qPCR. In the case of UGP2, HIF binding to two conserved HBSs was tested (**B**). The graph shows the ratio of the immunoprecipitated material in hypoxia over normoxia. The results from three independent experiments (black circles) and their median (line) are shown. In order to normalize data from the three independent experiments, the hypoxia/normoxia ratio is represented as fold over the mean value obtained for all the negative controls in each experiment. Neg1, IRS4; neg 2, STT3A; neg 3, HIVEP; neg4, LTBP1. The binding of HIF1a to the HRE within EGLN3 enhancer (E3+) or to a nonfunctional RCGTG within EGLN3 locus (E3–) were used as internal controls (ref. 31). For comparison, P_T/P_B ratio (in logarithmic scale) for each target is shown (bottom histogram), along with the threshold value of 6.5 (grey line).

mapping to regulatory regions within the locus, each of them having an associated score value. Our goal was to use this information to calculate a measure of the relative likelihood that the gene is an HIF target, as opposed of being nondirectly regulated by HIF (background). To this end, we constructed models that assign a probability to the gene in each of the two cases, and obtained the ratio of the two probabilities (odds ratio), we refer to this odds ratio as P_T/P_B ratio. In order to determine the optimum value for the P_T/P_B ratio for maximum sensitivity and specificity of gene classification, we used receiver operating characteristic (ROC) curve analysis (Figure 2D). For this analysis, the common set of 39 HIF targets from ChIP–Chip studies (Figure 2A and Supplementary Table S5) was used as known true targets. On the other hand, we selected genes that, while presenting conserved HBSs, were not induced by hypoxia (>30 probes in all data sets, mean fold induction between -0.3 and $+0.3$ and $P > 0.5$) as negative set. According to this analysis, a P_T/P_B ratio >6.5 resulted in an optimum sensitivity of 78.05% and a selectivity of 97.77%. Thus, we calculated the P_T/P_B ratio for all genes represented in GEO data sets and classified them as HIF targets (P_T/P_B ratio >6.5) or nontargets (P_T/P_B ratio $\leq$6.5 or lack of HBSs). Through this strategy, we predicted 216 HIF-target genes (Supplementary Table S6). Among them, 20 were previously known as HIF targets (1) and for 44 additional genes, some bibliographic evidence for their regulation by hypoxia was found (Supplementary Table S6). The remaining 152 genes are, to our knowledge, novel potential targets. The representation of the position of known target genes in our ranked list of predicted targets (Figure 2E) shows that they cluster toward the top positions [median values of 23 and 48, for the known targets from (1) and PubMed, respectively]. Thus, the P_T/P_B ratio accurately represents the probability of being an HIF target.

Experimental validation of model predictions

In order to evaluate the accuracy of our predictions we exposed HeLa cells to normoxia or hypoxia (1% oxygen) for 12 h and determined HIF1a binding to a set of predicted HBSs by ChIP–qPCR. For this purpose, we randomly selected six genes (RPLP0, MAFF, IDH2, SERTAD2, TPD52 and CARHSP1) among those that, according to our meta-analysis, were significantly upregulated by hypoxia ($P < 0.01$). To simplify the validation, we restricted our selection to genes having a single potential HBS. In addition, we included in this analysis the HBSs identified for the UDP-glucose

Pyrophosphorylase (UGP2) gene due to its potential role in glycogen metabolism (manuscript submitted for publication) and the HREs identified within MIF and CDC2L6 genes because of their inter-individual variation (see subsequent discussion). Finally, we also included five HBSs motifs located in genes that were not induced by hypoxia in any of the GEO data sets (IRS4, STT3A, HIVEP1 and LTBP1) as a negative control group to estimate the background hypoxic/normoxic enrichment ratio. After the treatments, cells were processed for ChIP using an antibody directed to HIF1a. Then, we determined the quantity of each of the target sequences in the HIF1a immunoprecipitate by qPCR and expressed it as percentage of the input of immunoprecipitated chromatin. Finally, we calculated the ratio for the enrichment in samples exposed to hypoxia over normoxic samples. As expected, the hypoxic/normoxic enrichment ratio for the negative controls was close to 1 (Figure 3A). In addition, we found a high enrichment ratio for the HBS in the EGLN3 enhancer (31), used as positive control. Importantly, in spite of the high variability observed for the biological replicates, we found a good correlation between target prediction (Figure 3A bottom histogram) and experimental determination of HIF binding (Figure 3A, top graph). In general, only the candidates above the threshold (UGP2, TPD52, GBE1, MAFF1 and MIF) showed a consistent positive enrichment ratio in all three independent experiments. In contrast, the genes predicted as negative (MKRN1, RPLP0, IDH2 and SERTAD2) showed a pattern similar to that of negative controls. There were two exceptions, CARHSP1 and CDC2L2, that did not behave as predicted. It is important to note, however, that in both cases the P_T/P_B ratio was close to the threshold value of 6.5.

To further test the strength of our classifier, we determined HIF binding to the two potential HBSs identified within the UGP2 gene, HBS_1 and HBS_2, located in chromosome 2 at positions 63 922 445 and 63 923 300, respectively. According to their P_T/P_B ratios, 3×10^{-5} and 14.4, respectively, only one of them (HBS_2) was classified as an HBS. In agreement with our prediction, only HBS_2, but not HBS_1, was consistently found in HIF1a immunoprecipitates (Figure 3B). Collectively, these results confirm the accuracy of our predictions.

Next, we decided to determine the biological relevance of the identified HBSs. To this end, we focused on RE1-silencing transcription factor co-repressor 2 (RCOR2) because it was classified as positive by our strategy with a P_T/P_B ratio close to the threshold (9.63). In addition, it was not found significantly upregulated by hypoxia in the gene expression profiling meta-analysis (fold induction = 0.44 and $P = 0.399$). Thus, the classification of RCOR2 as a true target was mainly based on the score value of its HBS (see 'Materials and Methods' section).

First, we investigated HIF1a binding to the potential HBS by ChIP–qPCR. As shown in Figure 4A, chromatin from the target region was enriched in HIF1a-immunoprecipitates from cells exposed to hypoxia. The enrichment was similar to that observed for the EGLN3 enhancer and was not observed when a control IgG was used for the immunoprecipitation (Figure 4A). To determine whether HIF1a binding to this site had a functional effect, we measured levels of RCOR2 mRNA in cells exposed to hypoxia by qPCR. Figure 4B shows that RCOR2 mRNA was induced by hypoxia in several cell lines. Moreover, the induction of RCOR2 was dependent on functional HIF since it was observed in HepaC1 cells, but not in HepaC1 derivate lacking HIFβ (42). Finally, we generated a reporter construct (pGL3-RCOR2) by cloning the RCOR2 promoter region, containing the putative HRE, upstream a firefly luciferase gene. Transfection of this construct into HepG2 cells demonstrated that the promoter activity was induced by hypoxia (Figure 4C). The induction was of similar magnitude to that observed for other HIF-regulated regions such as VEGFA promoter and EGLN3 enhancer (Figure 4C). Importantly, the mutation of the putative HRE completely abolishes the regulation of RCOR2 promoter by hypoxia (Figure 4C). Thus, RCOR2 is a novel hypoxia regulated gene whose induction under low oxygen is dependent on HIF activity and the presence of the HRE identified by our computational strategy. In addition, these results further support the relevance of our HIF-target predictions and show the robustness of our approach even for borderline cases.

Identification of SNPs that interfere with the response to hypoxia

The adaptation to hypoxia is largely dependent on HIF-mediated gene expression. Therefore, the identification of SNPs mapping to HBSs could reveal individuals with an altered response to hypoxia. The strategy described above provided a catalog of genome-wide HBSs (Supplementary Table S6). Thus, we decided to use this information to search for SNPs mapping to these sites. We retrieved from the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) entries whose genomic coordinates were coincident with the RCGTG motifs identified by our computational strategy. This search resulted in 146 SNP mapping to HBS motifs (Supplementary Table S7). Among them, 12 corresponded to sites within potential HIF targets (Table 2). We focused on rs17004038 and rs10624 because they were validated SNPs and were located within the only HRE identified for MIF and CDC2L6 genes, respectively. As shown in Figure 3, both sites bound HIF *in vivo*; thus, rs17004038 and rs10624 map to functional HBSs. In order to investigate the biological effect of these polymorphisms, we cloned the wild type MIF promoter region or the C→A variant, corresponding to the SNP rs17004038, upstream a firefly luciferase gene and performed reporter assays with these constructs. As previously reported (43), MIF promoter region (WT) was robustly induced by hypoxia (Figure 5). Mutation of the HBS identified by our strategy completely abrogated luciferase induction (Figure 5, mutHRE), demonstrating its role in the transcriptional upregulation of MIF promoter and in agreement with its binding to HIF (Figure 3). Importantly, the variant allele C→A was not upregulated by hypoxia and its behavior was indistinguishable from the mutant HRE (Figure 5). In contrast to the strong effect on the regulation by hypoxia,

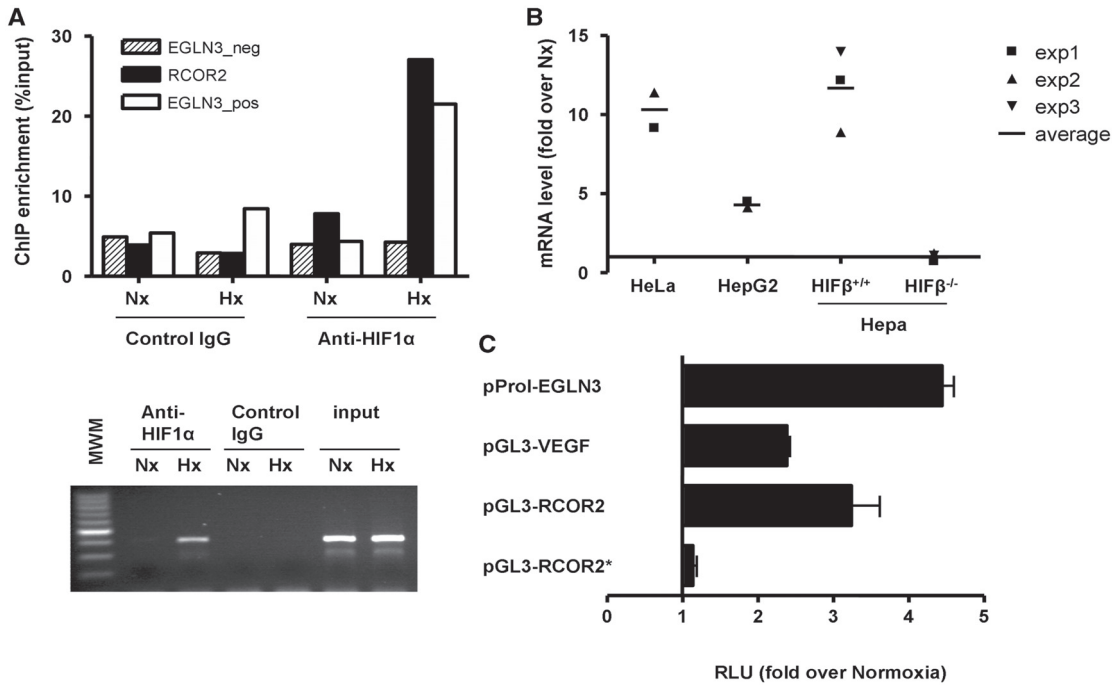


Figure 4. RCOR2 is a HIF-target gene. (**A**) HeLa cells were exposed to normoxia (Nx) or hypoxia (Hx, 1% oxygen) for 12 h. After treatments, cells were processed for chromatin immunoprecipitation using antibodies directed to HIF1a (anti-HIF1a) or control immunoglobulins (control IgG). The binding of HIF1a to the predicted HRE within RCOR2, to the HRE within EGLN3 enhancer (EGLN3_pos) or to a nonfunctional RCGTG within EGLN3 locus (EGLN3_neg) were determined by quantitative (upper panel) and semi-quantitative PCR (RCOR2, lower panel). MWM, molecular weight marker. (**B**) HeLa, HepG2 and Hepa C1/C4 cells were exposed to normoxia or hypoxia for 12 h and the level of RCOR2 mRNA was determined by quantitative PCR. The amount of each mRNA in samples was normalized to the content of β-actin mRNA in the same sample. The graph represents the fold values of hypoxic over normoxic mRNA levels normalized to the value of 1 (horizontal axis). Data represents the values from three independent experiments and their average (horizontal bar). (**C**) HepG2 cells were transfected with a reporter plasmid containing RCOR2 promoter region (−1770 to −795) upstream a luciferase reporter gene. Where indicated (asterisk) the consensus HRE sequence (ACGT) was mutated to TAGC. For comparison, reporter constructs containing the EGLN3 enhancer and VEGF promoter were included. The graphs represent the corrected luciferase activity values of each hypoxic sample over the luciferase activity obtained in normoxic cells. Data shown are a representative experiment out of three independent determinations.

Table 2. SNP mapping to RCGTG motifs within potential HIF-targets

SNP_ID	Validation	Allele	SNP_HBS	Gene	Chr	Coordin	Max_HBS
rs17004038	cluster, freq	C/A	_ACGT	MIF	22	24 236 591	YES
rs17152486	freq, hapmap	C/T	A_GTG	FLJ23834	7	105 671 891	YES
rs3758554	freq	C/G	CA_GC	LDB1	10	103 874 672	NO
rs16943318	cluster, freq, hapmap	G/A	CAC_C	RORA	15	61 209 971	NO
rs13358075	hapmap	T/A	ACG_G	SPOCK1	5	136 347 100	NO
rs58433430	NA	C/T	A_GTG	KLHL14	18	30 265 315	YES
rs2901215	NA	T/G	ACG_G	TIAL1	10	121 356 578	NO
rs56306258	NA	A/T	_CGTG	ANKRD12	18	9 136 756	NO
rs56033752	NA	G/A	AC_TG	DLG2	11	83 898 942	NO
rs34394782	NA	C/A/T	CACG_	CITED2	6	139 695 473	NO
rs34476700	NA	G/-	_CGTG	SENP3	17	7 463 409	NO
rs56298217	NA	G/A	GC_TG	CLK3	15	74 914 649	NO

The '_' symbol indicates the position of the SNP within the potential HIF-binding site. Max_HBS indicates whether the SNP maps to the highest scoring HBS identified for each locus.
Note that only one of the possible alleles generates an RCGTG motif.

this point mutation had no effect on the basal activity of the promoter showing an activity 1.16 ± 0.46 (mean ± SD) over wild-type promoter ($P = 0.85$). These results demonstrate that some polymorphic variants have a dramatic effect on gene regulation by hypoxia.

DISCUSSION

The identification of the complete set of genes directly regulated by HIF is essential to fully understand the array of cellular responses activated during adaptation

2342 Nucleic Acids Research, 2010, Vol. 38, No. 7

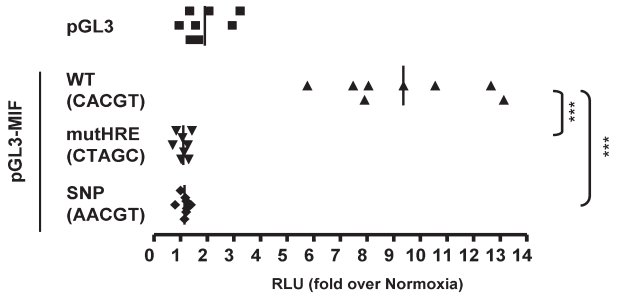


Figure 5. The allelic variant C/A (rs17004038) abrogates MIF induction by hypoxia. HeLa cells transfected with a reporter plasmid containing MIF genomic region (−31 to +108) upstream a luciferase reporter gene. Where indicated the consensus HRE sequence (CACGT) was mutated to CTAGC (mutHRE) or to AACGT (SNP). The graph represents the corrected luciferase activity values of each construct in cells exposed to hypoxia over the luciferase activity obtained in normoxic cells. Data show the results for eight experiments and its mean value (vertical line). Statistically significant differences with control group (WT) are indicated by asterisks (***$p < 0.001$).

to hypoxia. In this task, gene expression and TFBS data generated by high-throughput tools are fundamental. However, comparison between different studies (Figures 1A and 2A) reveals little overlap in the results, probably not only because of the particularities in the response to hypoxia in each particular system, but also because of the intrinsic noise associated to these techniques. Herein we describe a novel probabilistic strategy that integrates the rich information contained in gene expression profiling databases with classic bioinformatic approaches to predict TFBSs. The evaluation of this strategy, using published ChIP–chip data as a benchmark, indicates that it has a low error rate while retaining a high sensitivity. In agreement, our experimental validation revealed that five out of six of predicted targets were in fact true positives while 8/10 negatives were true negatives (Figures 3 and 4). Thus, the computational strategy described herein proved to be comparable in success rate to the experimental identification of HBSs by means of ChIP–chip, and it is hence an attractive alternative until these high-throughput techniques become more cost efficient.

In our strategy, the identification of relevant HBSs relies on the similitude of the potential HBS sequences to a PSSM that includes positions other than the core RCGTG. This PSSM was obtained by our analysis of a set of 46 sequences derived from well-characterized HREs (Supplementary Table S4). In contrast to our result, the analysis of genomic fragments bound by HIF failed to identify extended sequence preferences beyond the core RCGTG (28,29). Thus, it could be argued that the extended motif revealed by our analysis is consequence of the (relatively) reduced number of sequences in the alignment (46 sequences). However, using a PSSM based on this extended motif, we found that the HBSs identified by ChIP–chip had an associated score significantly higher than background sites (Figure 3B) and that, within a given locus, the functional HBS coincides with the highest scoring one (Figure 3C). These results strongly argue in favor of the PSSM-based score as a reliable parameter to discriminate functional HBSs and justify its inclusion in our computational strategy. The information content of the extra conserved positions is low as compared to that of the core HRE (Supplementary Table S4), suggesting that probably each individual position plays a minor role on HIF sequence preference. However, its combined effect could explain the preferential binding of HIF to these sequences. Further work is required to prove the relevance of these conserved positions outside the core RCGTG and, in the event of them being relevant, to determine whether they are part of the HIFα (or HIFβ) binding site, form the binding site of an unrelated transcription factor or just confer a favorable structure. Another premise in our strategy is that HREs are restricted to genomic sequences conserved during evolution. We imposed this restriction to our model knowing that not all *cis*-regulatory motifs are conserved (44). Nevertheless, this restriction was required to minimize the number of false positives while allowing a good sensitivity (~80%). In fact, this high sensitivity suggests that most real HBSs do in fact lie within conserved regions. In agreement, we found that 79% (254 out of 320) of the genomic fragments reported to bind HIF (29) contained one or more PhastCons elements. Thus, evolutionary conservation constraints are useful in reaching an optimum trade-off between sensitivity and specificity. In addition, our results imply that most (80%) of the experimentally identified HBSs are associated with CNSs. A further potential limitation of our strategy is imposed by the meta-analysis of gene expression profiling experiments. In our meta-analysis, genes showing a tissue-specific regulation by hypoxia, such as erthyropoietin (EPO), fail to be identified as hypoxia regulated genes. In order to mitigate this effect, in our strategy, the contribution of the meta-analysis to the classification of a gene as an HIF target is weighted by the consistency of its regulation by hypoxia (*P*-value) across the panel of microarray experiments (see 'Materials and Methods' section for details). In fact, this correction led us to the identification of RCOR2 as a HIF target in spite of it being induced by hypoxia in a limited number of experiments. Our analysis correctly identified the HRE driving EPO expression, but this gene was not selected as an HIF target because its associated P_T/P_B value was 5.17, right below the threshold of 6.5. In fact, EPO ranked in position 360 out of 11 672 analyzed genes. The design of our strategy tries to minimize the false positives to give a highly reliable list of HIF-targets but, because of the restrictions imposed, several HIF targets are missed as is the case of EPO. Hence, the list of 217 HIF-targets reported herein is clearly an underestimation of the whole complement of genes regulated by HIF.

An unexpected conclusion from our results is the lack of HBS enrichment in hypoxia-downregulated genes (Table 1). In agreement with our statistical approach, a recent ChIP–chip study (28) also failed to find association between HIF binding and transcriptional downregulation. In fact, although a direct effect on gene downregulation has been documented for some genes, such as CAD (45), they are rare exceptions, being upregulation of targets the predominant effect upon HIF binding. Thus, it is tempting to speculate that, in contrast to gene induction, most of

the transcriptional downregulation triggered by hypoxia is either HIF independent or mediated by a secondary factor downstream of HIF (indirect effect). The lack of genes consistently downregulated by hypoxia in gene expression data sets (Figures 3 and 4) supports that HIF does not play a direct role in gene downregulation. Interestingly, among the HIF targets identified in our study, there are several factors involved in transcriptional repression, including the novel HIF target RCOR2 described herein (Figure 4). Thus, it is plausible that HIF indirectly promotes the transcriptional repression of specific genes by controlling the expression of co-repressors. However, other mechanisms could be envisioned to explain an indirect effect of HIF on gene downregulation. For example, it has been recently described that hypoxia/HIF leads to the induction of microRNAs (46) that, in turn, could lead to downregulation of specific groups of genes. Thus, much work is required to understand the molecular mechanisms responsible for hypoxia-induced gene repression.

An important feature of our strategy is that it is not restricted to a particular HIF isoform. Most of the GEO data sets used for the meta-analysis (Supplementary Table S2) derive from experiments that used hypoxia or the EGLN inhibitor DMOG as stimuli and thus activated all HIFa subunits present in the cells. Only in two tables (GSE2020), a specific isoform was activated by overexpression. On the other hand, it is assumed that HIF1α and HIF2α bind to the same motif (RCGTG) and that their differential target preference stems from isoform-specific cooperation with other transcription factors (47,48). In fact, the binding of both isoforms to a common motif was recently confirmed by comparison of the genomic sequences immunoprecipitated with HIF1α and HIF2α (28). Thus, the approaches used in our strategy are not biased toward the preferential identification of isoform-specific targets. In fact, the list of candidates (Supplementary Table S6) includes genes reported as HIF1α [BNIP3, (48)] and HIF2α specific [CITED2, (47)].

The precise identification of HBSs did not only lead to the identification of direct HIF targets but also allowed us to predict polymorphisms that could affect gene regulation by hypoxia. In this work, we identified several SNPs mapping to predicted HBSs and demonstrated, in the case of MIF promoter, that specific allelic variants result in a severely impaired response to hypoxia. Thus, individuals presenting this variant probably fail to properly upregulate MIF in response to hypoxia. To our knowledge, these results constitute the first demonstration that the response to hypoxia could vary slightly between individuals of a population. It is difficult to predict the physiological consequences of the lack of hypoxic induction of MIF and further work is necessary to address this question. However, it is likely that the abrogation of hypoxic gene induction had dramatic consequences. In agreement, elimination of the HRE driving the hypoxic upregulation of VEGFA leads to motor neuron degeneration (49) and abnormal retinal neovascularization (50). Given the number of pathologies that course with hypoxia, our results point to a potential source of variability in the clinical course and/or response to treatments among different individuals. In addition, these results support the hypothesis that mutations in regulatory regions, rather than in coding sequences, are important to explain inter-individual variation. With the completion of ongoing sequencing projects aimed at the identification of novel SNPs, we foresee that the number of variants affecting HBSs will increase.

In conclusion, the data presented herein demonstrate that integration of gene expression profiling and *in silico* identification of TFBSs is a successful approach for the identification of direct target genes. In agreement, during the writing of our manuscript, a report was published (51) that employs a similar strategy to identify HIF targets. Interestingly, although both works identify a list of about 200 HIF targets, there is little overlap in the identity of the individual target genes (only 37 genes were coincident, see Supplementary Table S8), reflecting important differences in the approaches followed in each work. The application of our strategy led to the identification of a set of novel (potential) HIF targets and our experimental validation demonstrated the reliability of these predictions. Moreover, we have found that, at least one of the predicted targets, RCOR2, is an HIF target gene regulated by hypoxia. In addition, elsewhere we demonstrate that two additional novel targets, GYS1 and UGP2, are also regulated by hypoxia in an HIF-dependent manner (manuscript submitted for publication). Finally, we identified polymorphisms mapping to our predicted HBSs and demonstrated that specific alleles have a profound impact on the regulation of transcription by hypoxia. Altogether, these results expand our understanding of the adaptive responses to hypoxia and suggest, for the first time, that this response can vary among individuals.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Gema Moreno (Department of Biochemistry. Universidad Autónoma de Madrid. Madrid, Spain) for critical reading of the manuscript and Benjamin A.T. Rodríguez (Human Cancer Genetics Camprehensive Cancer Center. The Ohio State University. Columbus, OH, USA) for his valuable suggestions regarding validation of ChIP by qPCR. We also thank Dr. Manuel O. Landazuri and the researchers in his team for their kind support and help in many aspects of this work.

FUNDING

Ministerio de Ciencia y Tecnología/Ministerio de Ciencia e Innovación (SAF2005-00180 and SAF2008-03147), Comunidad Autónoma de Madrid (S-SAL-0311_2006); and the METOXIA, project ref. HEALTH-F2-2009-222741, under the 7th Research Framework Programme of the European Union. Funding for open access charge: SAF2008-03147.

Conflict of interest statement. None declared.

REFERENCES

1. Wenger,R.H., Stiehl,D.P. and Camenisch,G. (2005) Integration of oxygen signaling at the consensus HRE. *Sci. STKE*, **2005**, re12.
2. Semenza,G.L. and Wang,G.L. (1992) A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. *Mol. Cell. Biol.*, **12**, 5447–5454.
3. Salceda,S., Beck,I. and Caro,J. (1996) Absolute requirement of aryl hydrocarbon receptor nuclear translocator protein for gene activation by hypoxia. *Arch. Biochem. Biophys.*, **334**, 389–394.
4. Salceda,S. and Caro,J. (1997) Hypoxia-inducible factor 1alpha (HIF-1alpha) protein is rapidly degraded by the ubiquitin-proteasome system under normoxic conditions. Its stabilization by hypoxia depends on redox-induced changes. *J. Biol. Chem.*, **272**, 22642–22647.
5. Jiang,B.H., Zheng,J.Z., Leung,S.W., Roe,R. and Semenza,G.L. (1997) Transactivation and inhibitory domains of hypoxia-inducible factor 1alpha. Modulation of transcriptional activity by oxygen tension. *J. Biol. Chem.*, **272**, 19253–19260.
6. Ivan,M., Kondo,K., Yang,H., Kim,W., Valiando,J., Ohh,M., Salic,A., Asara,J.M., Lane,W.S. and Kaelin,W.G. Jr (2001) HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. *Science*, **292**, 464–468.
7. Jaakkola,P., Mole,D.R., Tian,Y.M., Wilson,M.I., Gielbert,J., Gaskell,S.J., Kriegsheim,A., Hebestreit,H.F., Mukherji,M., Schofield,C.J. *et al.* (2001) Targeting of HIF-alpha to the von Hippel–Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. *Science*, **292**, 468–472.
8. Bruick,R.K. and McKnight,S.L. (2001) A conserved family of prolyl-4-hydroxylases that modify HIF. *Science*, **294**, 1337–1340.
9. Epstein,A.C., Gleadle,J.M., McNeill,L.A., Hewitson,K.S., O'Rourke,J., Mole,D.R., Mukherji,M., Metzen,E., Wilson,M.I., Dhanda,A. *et al.* (2001) C. elegans EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell*, **107**, 43–54.
10. Maxwell,P.H., Wiesener,M.S., Chang,G.W., Clifford,S.C., Vaux,E.C., Cockman,M.E., Wykoff,C.C., Pugh,C.W., Maher,E.R. and Ratcliffe,P.J. (1999) The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature*, **399**, 271–275.
11. Hewitson,K.S., McNeill,L.A., Riordan,M.V., Tian,Y.M., Bullock,A.N., Welford,R.W., Elkins,J.M., Oldham,N.J., Shoumo,B., Gleadle,J.M. *et al.* (2002) Hypoxia-inducible factor (HIF) asparagine hydroxylase is identical to factor inhibiting HIF (FIH) and is related to the cupin structural family. *J. Biol. Chem.*, **277**, 26351–26355.
12. Lando,D., Peet,D.J., Gorman,J.J., Whelan,D.A., White,M.F. and Bruick,R.K. (2002) FIH-1 is an asparaginyl hydroxylase enzyme that regulates the transcriptional activity of hypoxia-inducible factor. *Genes Dev.*, **16**, 1466–1471.
13. Lando,D., Peet,D.J., Whelan,D.A., Gorman,J.J. and Whitelaw,M.L. (2002) Asparagine hydroxylation of the HIF transactivation domain a hypoxic switch. *Science*, **295**, 858–861.
14. Hu,C.-J., Wang,L.-Y., Chodosh,L.A., Keith,B. and Simon,M.C. (2003) Differential roles of hypoxia-inducible factor 1α (HIF-1α) and HIF-2α in hypoxic gene regulation. *Mol. Cell. Biol.*, **23**, 9361–9374.
15. Raval,R.R., Lau,K.W., Tran,M.G.B., Sowter,H.M., Mandriota,S.J., Li,J.-L., Pugh,C.W., Maxwell,P.H., Harris,A.L. and Ratcliffe,P.J. (2005) Contrasting properties of hypoxia-inducible factor 1 (HIF-1) and HIF-2 in von Hippel–Lindau-associated renal cell carcinoma. *Mol. Cell. Biol.*, **25**, 5675–5686.
16. Papandreou,I., Cairns,R.A., Fontana,L., Lim,A.L. and Denko,N.C. (2006) HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metab.*, **3**, 187–197.
17. Kim,J.-W., Tchernyshyov,I., Semenza,G.L. and Dang,C.V. (2006) HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.*, **3**, 177–185.
18. Bostrom,P., Magnusson,B., Svensson,P.-A., Wiklund,O., Boren,J., Carlsson,L.M.S., Stahlman,M., Olofsson,S.-O. and Hulten,L.M. (2006) Hypoxia converts human macrophages into triglyceride-loaded foam cells. *Arterioscler. Thromb. Vasc. Biol.*, **26**, 1871–1876.
19. Elvidge,G.P., Glenny,L., Appelhoff,R.J., Ratcliffe,P.J., Ragoussis,J. and Gleadle,J.M. (2006) Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition. *J. Biol. Chem.*, **281**, 15215–15226.
20. Kasper,L.H., Boussouar,F., Boyd,K., Xu,W., Biesen,M., Rehg,J., Baudino,T.A., Cleveland,J.L. and Brindle,P.K. (2005) Two transactivation mechanisms cooperate for the bulk of HIF-1-responsive gene expression. *EMBO J.*, **24**, 3846–3858.
21. Allen,J.W., Khetani,S., Johnson,R. and Bhatia,S. (2006) In vitro liver tissue model established from transgenic mice: role of HIF-1alpha on hypoxic gene expression. *Tissue Engg.*, **12**, 3135–3147.
22. Mense,S.M., Sengupta,A., Zhou,M., Lan,C., Bentsman,G., Volsky,D.J. and Zhang,L. (2006) Gene expression profiling reveals the profound upregulation of hypoxia-responsive genes in primary human astrocytes. *Physiol. Genomics*, **25**, 435–449.
23. Ray,J.B., Arab,S., Deng,Y., Liu,P., Penn,L., Courtman,D.W. and Ward,M.E. (2008) Oxygen regulation of arterial smooth muscle cell proliferation and survival. *Am. J. Physiol. Heart Circ. Physiol.*, **294**, H839–H852.
24. Wang,V., Davis,D.A., Haque,M., Huang,L.E. and Yarchoan,R. (2005) Differential gene up-regulation by hypoxia-inducible factor-1α and hypoxia-inducible factor-2α in HEK293T cells. *Cancer Res.*, **65**, 3299–3306.
25. Guimbellot,J., Erickson,S., Mehta,T., Wen,H., Page,G., Sorscher,E. and Hong,J. (2009) Correlation of microRNA levels during hypoxia with predicted target mRNAs through genome-wide microarray analysis. *BMC Med. Genomics*, **2**, 15.
26. Irigoyen,M., Ansó,E., Martínez,E., Garayoa,M., Martínez-Irujo,J.J. and Rouzaut,A. (2007) Hypoxia alters the adhesive properties of lymphatic endothelial cells. A transcriptional and functional study. *Biochim. Biophys. Acta – Mol. Cell. Res.*, **1773**, 880–890.
27. Cahan,P., Rovegno,F., Mooney,D., Newman,J.C., St. Laurent,G. III and McCaffrey,T.A. (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, **401**, 12–18.
28. Mole,D.R., Blancher,C., Copley,R.R., Pollard,P.J., Gleadle,J.M., Ragoussis,J. and Ratcliffe,P.J. (2009) Genome-wide association of hypoxia-inducible factor (HIF)-1α and HIF-2α DNA binding with expression profiling of hypoxia-inducible transcripts. *J. Biol. Chem.*, **284**, 16767–16775.
29. Xia,X., Lemieux,M.E., Li,W., Carroll,J.S., Brown,M., Liu,X.S. and Kung,A.L. (2009) Integrative analysis of HIF binding and transactivation reveals its role in maintaining histone methylation homeostasis. *Proc. Natl Acad. Sci. USA*, **106**, 4260–4265.
30. Elnitski,L., Jin,V.X., Farnham,P.J. and Jones,S.J.M. (2006) Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res.*, **16**, 1455–1464.
31. Pescador,N., Cuevas,Y., Naranjo,S., Alcaide,M., Villar,D., Landázuri,M.O. and del Peso,L. (2005) Regulation of the egl nine homologue 3 (egln3/phd3) gene: Identification of a functional hypoxia-responsive element. *Biochem. J.*, **390**, 189–197.
32. Kimura,H., Weisz,A., Ogura,T., Hitomi,Y., Kurashima,Y., Hashimoto,K., D'Acquisto,F., Makuuchi,M. and Esumi,H. (2001) Identification of hypoxia-inducible factor-1 (HIF-1) ancillary sequence and its function in vascular endothelial growth factor gene induction by hypoxia and nitric oxide. *J. Biol. Chem.*, **276**, 2292–2298.
33. Gilligan,P., Brenner,S. and Venkatesh,B. (2002) Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene*, **294**, 35–44.
34. Nobrega,M.A. and Pennacchio,L.A. (2004) Comparative genomic analysis as a tool for biological discovery. *J. Physiol.*, **554**, 31–39.

35. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
36. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–174.
37. Hoek,K.S., Schlegel,N.C., Eichhoff,O.M., Widmer,D.S., Praetorius,C., Einarsson,S.O., Valgeirsdottir,S., Bergsteinsdottir,K., Schepsky,A., Dummer,R. *et al.* (2008) Novel MITF targets identified using a two-step DNA microarray strategy. *Pigment Cell Melanoma Res.*, **21**, 665–676.
38. Jeffery,I.B., Madden,S.F., McGettigan,P.A., Perriere,G., Culhane,A.C. and Higgins,D.G. (2007) Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, **23**, 298–305.
39. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
40. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
41. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
42. Wood,S.M., Gleadle,J.M., Pugh,C.W., Hankinson,O. and Ratcliffe,P.J. (1996) The role of the aryl hydrocarbon receptor nuclear translocator (ARNT) in hypoxic induction of gene expression. Studies in ARNT-deficient cells. *J. Biol. Chem.*, **271**, 15117–15123.
43. Baugh,J.A., Gantier,M., Li,L., Byrne,A., Buckley,A. and Donnelly,S.C. (2006) Dual regulation of macrophage migration inhibitory factor (MIF) expression in hypoxia by CREB and HIF-1. *Biochem. Biophys. Res. Comm.*, **347**, 895–903.
44. Alonso,E., Pernaute,B., Crespo,M., Gómez-Skarmeta,J.L. and Manzanares,M. (2008) Understanding the regulatory genome. *Int. J. Dev. Biol.*, **53**, 1367–1378.
45. Chen,K.-F., Lai,Y.-Y., Sun,H.S. and Tsai,S.-J. (2005) Transcriptional repression of human cad gene by hypoxia inducible factor-1α. *Nucleic Acids Res.*, **33**, 5190–5198.
46. Kulshreshtha,R., Ferracin,M., Wojcik,S., Garzon,R., Alder,H., Agosto-Perez,F., Davuluri,R., Liu,C., Croce,C. and Negrini,M. (2007) A microRNA signature of hypoxia. *Mol. Cell. Biol.*, **27**, 1859–1867.
47. Hu,C.-J., Iyer,S., Sataur,A., Covello,K.L., Chodosh,L.A. and Simon,M.C. (2006) Differential regulation of the transcriptional activities of hypoxia-inducible factor 1 alpha (HIF-1α) and HIF-2α in stem cells. *Mol. Cell. Biol.*, **26**, 3514–3526.
48. Aprelikova,O., Wood,M., Tackett,S., Chandramouli,G.V.R. and Barrett,J.C. (2006) Role of ETS transcription factors in the hypoxia-inducible factor-2 target gene selection. *Cancer Res.*, **66**, 5641–5647.
49. Oosthuyse,B., Moons,L., Storkebaum,E., Beck,H., Nuyens,D., Brusselmans,K., Dorpe,J.V., Hellings,P., Gorselink,M., Heymans,S. *et al.* (2001) Deletion of the hypoxia-response element in the vascular endothelial growth factor promoter causes motor neuron degeneration. *Nat. Genet.*, **28**, 131–138.
50. Vinores,S.A., Xiao,W.-H., Aslam,S., Shen,J., Oshima,Y., Nambu,H., Liu,H., Carmeliet,P. and Campochiaro,P.A. (2006) Implication of the hypoxia response element of the VEGF promoter in mouse models of retinal and choroidal neovascularization, but not retinal vascular development. *J. Cell. Physiol.*, **206**, 749–758.
51. Benita,Y., Kikuchi,H., Smith,A.D., Zhang,M.Q., Chung,D.C. and Xavier,R.J. (2009) An integrative genomics approach identifies hypoxia inducible factor-1 (HIF-1)-target genes that form the core response to hypoxia. *Nucleic Acids Res.*, **37**, 4587–4602.

Adaptive selection of an incretin gene in Eurasian populations.

Chang CL, Cai JJ, Lo C, Amigo J, Park JI, Hsu SY

Genome Research. 10/2010; 21(1):21-32.

Las diversidades en la fisiología humana han sido parcialmente moduladas por la adaptación a los entornos naturales y culturas cambiantes. Recientes análisis genómicos han revelado polimorfismos de un solo nucleótido (SNPs) que están asociados con las adaptaciones en las respuestas inmunes, los cambios evidentes en las formas del cuerpo humano, o adaptaciones a climas extremos en poblaciones humanas concretas. A continuación, informamos que el *locus* humano GIP fue seleccionado diferencialmente entre las poblaciones humanas basándonos en el análisis de un SNP no sinónimo (rs2291725). Los análisis comparativos y funcionales mostraron que el gen GIP humano codifica una isoforma de polipéptido críptico glucosa-dependiente insulinotrópico (GIP) (GIP55S o GIP55G) que abarca el SNP y es resistente a la degradación en suero con respecto al péptido GIP maduro conocido. Notablemente, descubrimos que GIP55G, que está codificado por el alelo derivado, exhibe una bioactividad mayor en comparación con GIP55S, que se deriva a partir del alelo ancestral. El análisis del estructura haplotípica sugiere que el alelo derivado en rs2291725 se convirtió en dominante en los asiáticos orientales ~8100 años atrás, debido a la selección positiva. Los resultados combinados sugieren que rs2291725 representa una mutación funcional y puede contribuir al estudio de la genética de poblaciones. Dado que la señalización de GIP juega un papel crítico en la regulación de la homeostasis en los ejes digestivos insulínico y adiposo, nuestro estudio pone de relieve la importancia de la comprensión de las adaptaciones en la regulación del balance de energía ante las emergentes epidemias de diabetes y obesidad.

Research

Adaptive selection of an incretin gene in Eurasian populations

Chia Lin Chang,[1] James J. Cai,[2,3] Chiening Lo,[4] Jorge Amigo,[5] Jae-Il Park,[6] and Sheau Yu Teddy Hsu[7,8]

[1]*Department of Obstetrics and Gynecology, Chang Gung Memorial Hospital Linkou Medical Center, Chang Gung University, Kweishan, Taoyuan 333, Taiwan;* [2]*Department of Biology, Stanford University, Stanford, California 94305, USA;* [3]*Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas 77840, USA;* [4]*Department of Clinical and Experimental Epilepsy, Institute of Neurology, University College London, London WC1N 3BG, United Kingdom;* [5]*Genomic Medicine Group, University of Santiago de Compostela, CIBERER, Santiago de Compostela 15706, Spain;* [6]*Hormone Research Center and School of Biological Sciences and Technology, Chonnam National University, Kwangju 500-712, Republic of Korea;* [7]*Reproductive Biology and Stem Cell Research Program, Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, California 94305-5317, USA*

Diversities in human physiology have been partially shaped by adaptation to natural environments and changing cultures. Recent genomic analyses have revealed single nucleotide polymorphisms (SNPs) that are associated with adaptations in immune responses, obvious changes in human body forms, or adaptations to extreme climates in select human populations. Here, we report that the human *GIP* locus was differentially selected among human populations based on the analysis of a nonsynonymous SNP (rs2291725). Comparative and functional analyses showed that the human *GIP* gene encodes a cryptic glucose-dependent insulinotropic polypeptide (GIP) isoform (GIP55S or GIP55G) that encompasses the SNP and is resistant to serum degradation relative to the known mature GIP peptide. Importantly, we found that GIP55G, which is encoded by the derived allele, exhibits a higher bioactivity compared with GIP55S, which is derived from the ancestral allele. Haplotype structure analysis suggests that the derived allele at rs2291725 arose to dominance in East Asians ~8100 yr ago due to positive selection. The combined results suggested that rs2291725 represents a functional mutation and may contribute to the population genetics observation. Given that GIP signaling plays a critical role in homeostasis regulation at both the enteroinsular and enteroadipocyte axes, our study highlights the importance of understanding adaptations in energy-balance regulation in the face of the emerging diabetes and obesity epidemics.

[Supplemental material is available online at http://www.genome.org.]

Recent studies have revealed that genetic variation underlies a variety of diversities in human physiology and pathology (Sabeti et al. 2005; Voight et al. 2006, 2010; Sulem et al. 2007; Tishkoff et al. 2007; Genovese et al. 2010; Leslie 2010; Simonson et al. 2010; Yi et al. 2010). Among the sundry forms of genetic variation, single nucleotide polymorphisms (SNPs) with high population differentiation are regarded as candidates of adaptation to recent changes in human environment and culture, and have been shown to play an important role in acquiring distinct physiological traits and susceptibility to different diseases among the populations (Barreiro and Quintana-Murci 2010; Chen et al. 2010; Gibbons 2010; Ingelsson et al. 2010; Laland et al. 2010; Luca et al. 2010; Richerson et al. 2010). Thus, the identification of causal SNPs with signatures of positive selection and underlying functional changes is crucial to a better understanding of the relationship between genomic variation and human health as well as gene–environmental interactions (Nielsen et al. 2007; Laland et al. 2010; Nei et al. 2010). To systematically analyze the contributions of genetic variation in intercellular signaling molecules to physiological diversities in humans, we studied SNPs in the coding region of 839 human polypeptide hormones and their cognate receptors for evidence of selection using the data from the International HapMap project phases I and II (International HapMap Project 2003; International HapMap Consortium 2007). We focused on these polypeptide ligands and receptors because they represent half of the targets of modern medicine (Drews 2000) and because the genes associated with intercellular communication or the responses to environmental factors (e.g., pathogens and food sources) have been implicated in the evolution of a variety of common traits and pathologies in humans and other vertebrates (Seminara et al. 2003; Sabeti et al. 2005; Hoekstra et al. 2006; Lalueza-Fox et al. 2007; Sulem et al. 2007; Chambers et al. 2008; Prokopenko et al. 2008; Shiao et al. 2008; Shimomura et al. 2008; Anderson et al. 2009; Topaloglu et al. 2009). Here, based on genomic and biochemical analyses, we show that variants in an incretin hormone gene, *GIP*, were differentially selected in human populations, and a nonsynonymous SNP, rs2291725, represents a functional mutation. Because changes in the food source represented one of the most important selection pressures during the transitions of human culture and because incretin hormones play critical roles in homeostasis maintenance at the enteroinsular and enteroadipocyte axes, future studies of potential genotype–phenotype relationships for the selected *GIP* variants could provide a better understanding of which and how these variants contribute to phenotypic variation in energy-balance regulation among individuals or human populations.

[8]**Corresponding author.**
E-mail teddyhsu@stanford.edu; fax (650) 725-7102.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.110593.110.

Results

A nonsynonymous SNP (rs2291725) in the human glucose-dependent insulinotropic polypeptide gene exhibits high population differentiation

To investigate whether variation in the polypeptide intercellular signaling molecules contribute to physiological diversities in humans, we curated 457 human G-protein–coupled receptors (GPCRs) and their cognate ligand genes as well as 382 human non-GPCR receptor and ligand genes (Supplemental Table 1; Ben-Shlomo et al. 2003; Semyonov et al. 2008). We started by using the population differentiation statistic F_{ST} to identify the leads for functional characterization (Lewontin and Krakauer 1973; Akey et al. 2002; Li et al. 2008).

We computed the F_{ST} for the coding SNPs of GPCR and ligand genes and compared the result with those for coding SNPs of all other human genes. F_{ST} was computed between all possible pairs of HapMap II populations (YRI [African, Yoruba from Ibadan], CEU [European, United States residents with northern and western European ancestry], and ASN [East Asian, pooled samples of Chinese from Beijing {CHB} and Japanese from Tokyo {JPT}]). Distributions of F_{ST} for either the synonymous or nonsynonymous SNPs between any two HapMap II populations showed no difference (all $P > 0.01$, Kolmogorov–Smirnov test) (Supplemental Fig. 1). Likewise, studies of the F_{ST} of coding SNPs in 382 non-GPCR receptor and ligand genes have shown similar cumulative distribution function (CDF) plots, suggesting that there is no difference in F_{ST} distribution between the GPCR and the non-GPCR groups ($P > 0.01$, Kolmogorov–Smirnov test) (Supplemental Fig. 2). These results suggest that, when analyzed as a whole set, the coding SNPs of human polypeptide receptor and ligand genes do not have a significantly elevated F_{ST}.

Nevertheless, at the individual-gene level, dozens of coding SNPs in these receptor and ligand genes have a high F_{ST} (>0.5) between select pairs of populations (Supplemental Tables 2, 3). Among GPCRs and their cognate ligands, *DRD5*, *DARC*, *CELSR1*, *CCL23*, *GIP*, *MC1R*, *EMR1*, *GRM1*, *CALCR*, *CXCR6*, *GPR39*, and *DRD3* were found to have nonsynonymous SNPs with a high F_{ST}, which suggests that these SNPs are likely to be targets of selection. On the other hand, nonsynonymous SNPs in *EDAR*, which has been repeatedly shown to be under positive selection (Bryk et al. 2008; Fujimoto et al. 2008), and several immune response–related genes (e.g., *IL4R*, *TNFRSF10A*, *TRAF3*, *IL29*, *IL20RA*, *IL1RL1*, *TNFRSF6B*, and *PTPRA*) in the non-GPCR group were found to have high F_{ST} scores in select pair(s) of the populations.

Importantly, we found that the nonsynonymous SNP (rs2291725) in exon 4 of the glucose-dependent insulinotropic polypeptide gene (or gastric inhibitory peptide, *GIP*; referred to as the $GIP^{103T/C}$ mutation in the following text) on chromosome 17 is highly linked with a cluster of neighboring SNPs within a 250-kb region (starting from rs8079874 to rs2291726) and has an F_{ST} value in the top 0.5% of all nonsynonymous SNPs in comparisons between the ASN and YRI populations (Fig. 1A; Supplemental Table 2). In contrast, F_{ST} estimates in comparisons between CEU and YRI or between CEU and ASN were not different from the

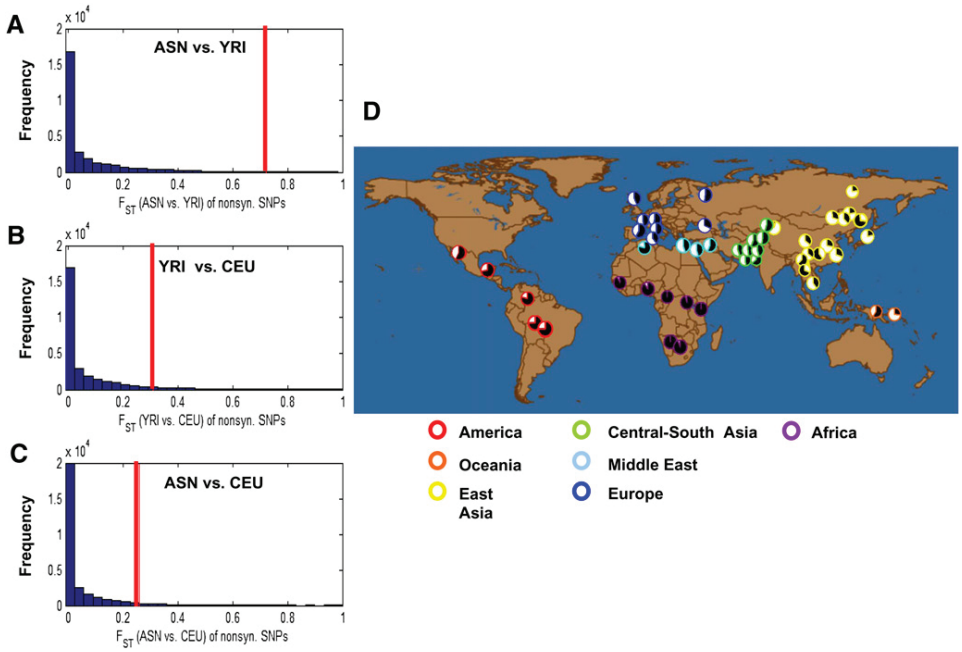


Figure 1. Differential distribution of alleles at rs2291725 in human populations. (*A*–*C*) The distribution of the F_{ST} for nonsynonymous SNPs across the human genome and the F_{ST} for rs2291725. F_{ST} estimations between ASN and YRI (*A*), YRI and CEU (*B*), and ASN and CEU (*C*) are shown in the *x*-axis. The red vertical bar indicates the corresponding values of F_{ST} for rs2291725. The *y*-axis represents the frequency of SNPs with a given F_{ST} estimate. Statistical significance for comparisons in *A*, *B*, and *C* is indicated with empirical $P = 0.0039$, 0.058, and 0.047, respectively. The number of coding SNPs that were analyzed in the *A*, *B*, and *C* histograms was 48,526, 48,674, and 48,075, respectively. Among these SNPs, the number of nonsynonymous SNPs in the *A*, *B*, and *C* histograms was 29,033, 29,026, and 28,774, respectively. (*D*) Distribution of rs2291725 in the HGDP–CEPH (944 unrelated samples) populations. Pie charts represent the proportion of each genotype by geographic region. The ancestral GIP^{103T} allele (black pie) occurs at a higher frequency in African and American populations, whereas the majority of Eurasian populations have a higher frequency of the derived GIP^{103C} allele (white pie).

genome average (Fig. 1B,C). Consistently, data from the HGDP-CEPH project (Center d'Etude du Polymorphisme Humain-Human Genome Diversity Panel; 944 unrelated individuals from 52 populations) (Cann et al. 2002; Rosenberg 2006; Li et al. 2008) and the Human Genome Center at the University of Tokyo (752 Japanese individuals, GIP^{103C}/GIP^{103T} = 0.732/0.268) showed that the derived GIP^{103C} allele frequency at rs2291725 is much higher (>60%) in the majority of East Asian populations and varies widely among other populations: ranging from 0.0%–9.5% in sub-Saharan Africans and increasing to >40.0% in European and Middle Eastern populations (Fig. 1D; Table 1; Supplemental Table 5).

GIP encodes one of the two incretin hormones (glucagon-like peptide-1[GLP-1] and GIP) in humans and plays a critical role in normal carbohydrate and lipid metabolism (Kim and Egan 2008). After ingestion of nutrients, GIP secreted from duodenal and jejunal K cells acts on pancreatic β cells to stimulate the release of insulin, which thereby ensures the prompt uptake of glucose and lipids into the tissues. Abnormal regulation of GIP signaling leads to altered carbohydrate metabolism and lipid accumulation at the enteroinsular and enteroadipocyte axis, respectively (Miyawaki et al. 2002; Fulurija et al. 2008; Isken et al. 2008; Kim and Egan 2008). In mice, the deletion of the GIP receptor (*Gipr*) led to impaired first-phase glucose-stimulated insulin release (Miyawaki et al. 2002), whereas exogenous GIP was found to worsen postprandial hyperglycemia in patients with type 2 diabetes (Chia et al. 2009). In addition to effects on glucose homeostasis, GIP has been shown to promote obesity in mice that were fed a high-fat diet (McClean et al. 2007; Gniuli et al. 2010). Because the incretin effect induced by the oral glucose intake leads to a higher insulin response compared with that from a matched intravenous glucose stimulation, the regulation of glucose levels by GIP represents a critical endocrine circuit to monitor exogenous energy intake and regulate subsequent storage (Kim and Egan 2008). In addition, the critical role of GIP signaling in energy-balance regulation has been highlighted by recent studies that showed that variants at the *GIPR* locus are associated with glucose levels 2 h after an oral glucose challenge test used in the diagnosis of type 2 diabetes (Ingelsson et al. 2010; Saxena et al. 2010). On the other hand, epidemiological studies have shown that the prevalence of different forms of diabetes and obesity as well as the regulation of glucose metabolism vary widely among ethnic groups (Buchanan and Xiang 2005; Kim and Egan 2008). Thus, the observed high population differentiation in *GIP* polymorphisms could be associated with the adaptations of incretin physiology to recent changes in diets and cultures in select human populations. To test this hypothesis, we fine mapped the *GIP* locus and tested the functions of GIP variants in vitro and in vivo.

Variants at the *GIP* locus were partially selected in Eurasian populations

Analyses of linkage disequilibrium (LD) in the *GIP* region for the three HapMap II populations showed that extended LD blocks are present in the CEU and ASN chromosomes, and these blocks encompass *GIP* and the neighboring *UBE2Z*, *SNF8*, and *ATP5G1* genes (Supplemental Fig. 3A,B, red square, D′ = 1, likelihood of odds [LOD] scores >2). In contrast, the majority of SNP pairs in the *GIP* region of the YRI chromosomes exhibited a low LOD and a low r^2 (Supplemental Fig. 3C, blue square, D′ = 1, LOD scores <2). Consistent with the LD analysis, plots depicting the haplotype map showed that most of the chromosomes with the derived GIP^{103C} allele in the ASN population have haplotypes extending >200 kb and are significantly longer compared with chromosomes with the ancestral GIP^{103T} allele (Fig. 2A, middle panel, $P < 0.001$). In contrast, most chromosomes in YRI did not exhibit an extended haplotype surrounding the GIP^{103} allele (Fig. 2A, bottom panel). These results were reflected in the plots of extended haplotype homozygosity (EHH) decay curves (Fig. 2B). In ASN and CEU, the EHH curve for chromosomes carrying the derived GIP^{103C} allele extends much further than that for the ancestral allele. To test whether the area under the EHH curve is greater for a selected allele than for a neutral allele, we calculated the integrated haplotype score (iHS) for rs2291725 as a core marker (Voight et al. 2006). By using a coalescent model that generated 10,000 data sets from the same length of genome region (750 kb) with the same sample size (180 chromosomes), we found that the observed iHS for the derived allele in ASN (iHS = −0.753)—but not CEU (iHS = −0.131)—deviated significantly from the neutral distribution ($P < 1 \times 10^{-4}$) (Supplemental Fig. 4A). We also performed coalescent simulations that took into account the effects of demography (e.g., population bottleneck) (Supplemental Fig. 4B) and recombination hotspot, respectively (Hellenthal and Stephens 2007; Gutenkunst et al. 2009; Supplemental Methods). Consistently, we found that no iHS for the simulated haplotype sets is more negative than the observed iHS for rs2291725 under the bottleneck scenario (empirical $P < 1 \times 10^{-4}$), and only a few are in the presence of a recombination hotspot (empirical $P < 8 \times 10^{-4}$) (Supplemental Fig. 4A). These results indicated that the observed iHS has deviated significantly from the neutral distribution even after adjusting for the bias that may have been introduced by a bottleneck in demography or heterogeneous recombination events. Thus, it is unlikely that neutral evolution alone explains the observed long haplotypes that carry the derived GIP^{103C} allele in the ASN population.

Further examination of the extended haplotype blocks in ASN showed that rs2291725 is highly linked with another 43 SNPs

Table 1. Human *GIP* SNP rs2291725 exhibits a high population differentiation characteristic in the International HapMap II Project data set

		rs2291725 allele frequency		rs2291725 genotypes			F_{ST} vs.		
Population	Chromosome no.	GIP^{103T}	GIP^{103C}	T/T	T/C	C/C	YRI	CEU	ASN
YRI	120	0.950	0.050	0.900	0.100	0.000	—	0.24	0.48[a]
CEU	120	0.517	0.483	0.267	0.500	0.233	0.24	—	0.08
ASN	178	0.245	0.755	0.090	0.314	0.595	0.48[a]	0.08	—
All	418	0.524	0.476						

Genotypes were analyzed as described using the SPSmart v3 and Haplotter (Voight et al. 2006; Amigo et al. 2008). The average frequency of two alleles—derived allele GIP^{103C} and ancestral chimpanzee allele GIP^{103T}—at rs2291725 is approximately even in the overall HapMap population: Only 5.0% of YRI chromosomes contain GIP^{103C} compared with 48.3% and 75.5% of CEU and ASN chromosomes that carry the derived allele, respectively.
[a] $P < 0.05$.

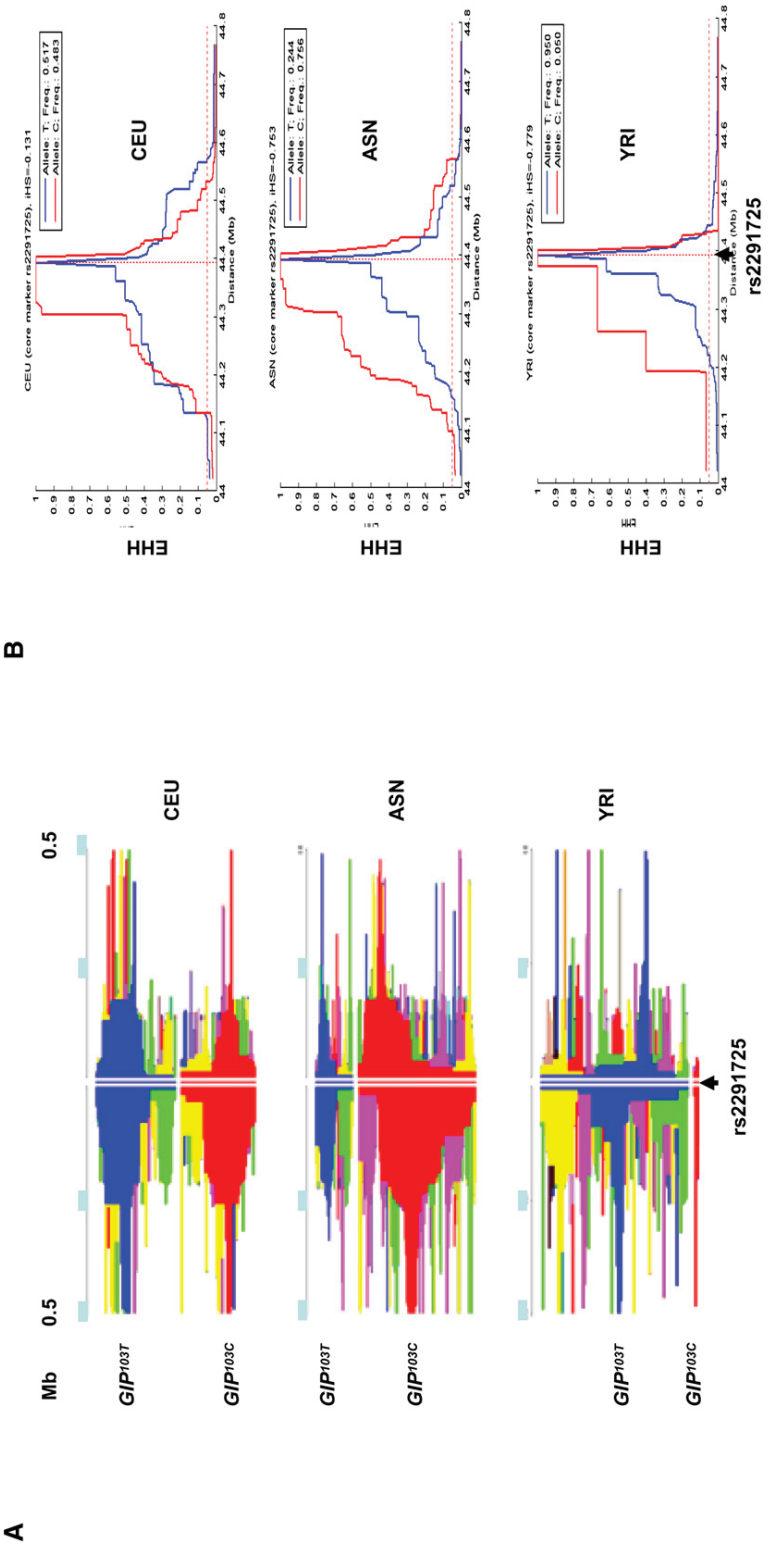


Figure 2. (Continued on next page)

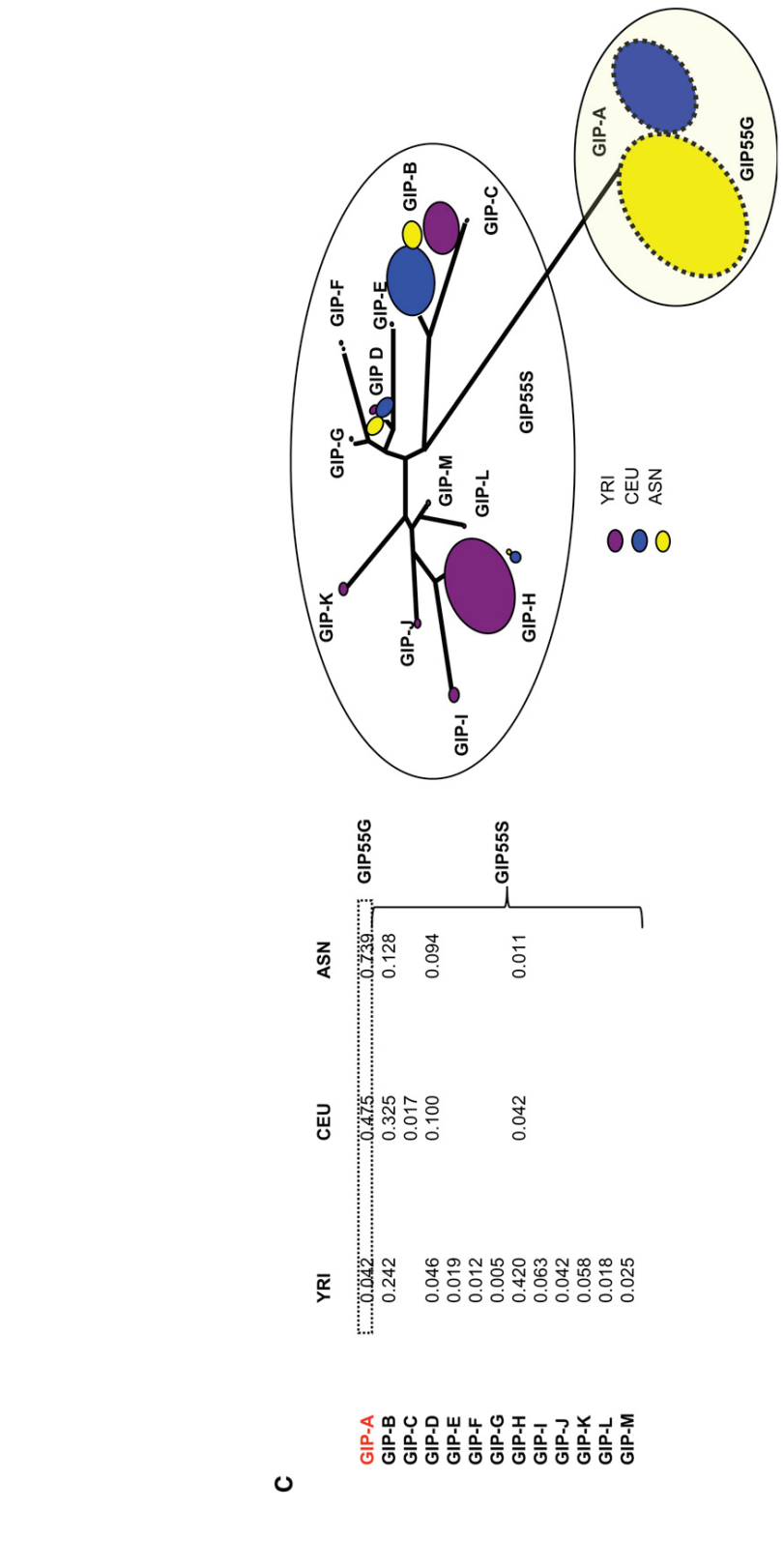


Figure 2. Evolution of haplotypes encompassing the *GIP* locus. (*A*) Plots of the extent of haplotype homozygosity in the 1.0-Mb region surrounding rs2291725 in CEU, ASN, and YRI as assessed by Haplotter (Voight et al. 2006). These plots are divided into two parts. The *upper* portion shows haplotypes with the ancestral GIP^{103T} allele in blue, and the *lower* portion shows haplotypes with the derived GIP^{103C} allele in red. Adjacent haplotypes with the same color carry identical genotypes spanning the region between a select SNP and rs2291725. (*B*) Plots of the breakdown of EHH over the distance between rs2291725 and neighboring SNPs at increasing distances. EHH decays much slower at the derived allele (red) compared with the ancestral allele (blue) in ASN and CEU. The position of rs2291725 at the *center* of the plots is indicated by a vertical dotted line. (*C*) Evolution of haplotypes within a 70-kb core haploblock in the three HapMap II populations. A total of 13 haplotypes was found in the core region from rs196241 to rs2291726 (37 SNPs at chr 17, 44325258–44394253). The frequencies of these haplotypes (GIP-A~M) in YRI, CEU, and ASN are shown in the *lower left* panel. There are 12 haplotypes in YRI, whereas CEU and ASN are represented by five and four haplotypes, respectively. GIP-A is the only haplotype containing the derived GIP^{103C} allele and is indicated by a dotted box. An unrooted tree analysis of these haplotypes showed that the GIP-A haplotype diverged from others early in human evolution (GeneBee). The frequency of each haplotype in a select population is indicated by the size of the pie at the *tip* of each branch.

within the adjacent 250-kb region, and in a 70-kb core region (rs1962412 to rs2291726; Genome build 36.3, chr 17, 44325258–44394253), the derived GIP^{103C} allele is associated with a single haplotype (Fig. 2C, haplotype GIP-A; Supplemental Fig. 3D; Supplemental Table 4). In contrast, the ancestral GIP^{103T} allele found in the majority of the YRI chromosomes is represented by 12 different haplotypes (Fig. 2C, haplotypes GIP-A, GIP-B, and GIP-D–M; Supplemental Fig. 3D). An unrooted tree analysis of these haplotypes confirmed that the evolutionary trajectory of the GIP^{103C}-associated haplotype is distinct from other haplotypes (Fig. 2C). Given the presence of several characteristic patterns, including highly differentiated alleles, high frequency–derived alleles, and relatively long derived haplotypes, these data suggested that pre-existing polymorphisms at the *GIP* locus were partially selected in ASN and possibly in CEU at times post-dating the separation of the YRI and Eurasian populations (Smith and Haigh 1974).

GIP encodes a GIP peptide containing the variable residue (Ser103 or Gly103) at rs2291725

Whereas the selection at the *GIP* locus could be attributed to a single variant or a combination of SNPs, the nonsynonymous rs2291725 provided a tangible target for functional analyses of the causal mutation. To explore whether rs2291725 represented a causal variant and provided a benefit to its carriers, we investigated the function of the GIP peptides containing the variable residue (Ser103 for GIP^{103T} or Gly103 for GIP^{103C}). GIP was originally characterized as a 42-amino-acid peptide derived from proteolytic processing at the monobasic cleavage sites at residues 51 and 94 of the GIP open reading frame (Moody et al. 1984; Takeda et al. 1987; Kim and Egan 2008). Although residue 103 is located outside the conventional mature GIP sequence (residues 52–93) (Fig. 3A, upper panel), we noticed that a conserved dibasic cleavage site (Arg-Lys, residues 106–107) is located 13 amino acids downstream from the conventional cleavage site in primates, dogs, cats, cows, pigs, and horses (Fig. 3A, upper panel). Thus, post-translational cleavage at this alternative processing site could generate an extended GIP isoform 13 amino acids longer (Fig. 3A, lower panel; the ancestral GIP55S and the derived GIP55G). To

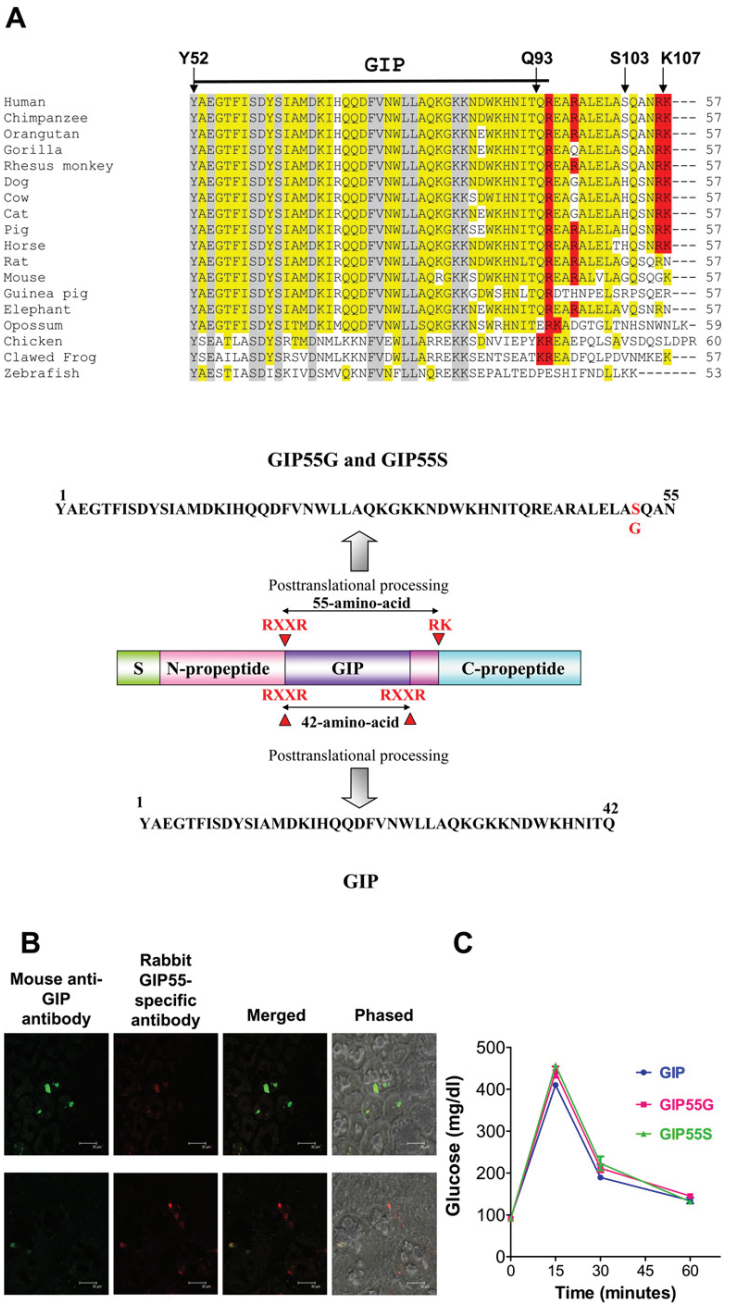


Figure 3. An extended GIP peptide is expressed in human gut cells. (*A*) Alignment of human GIP (residues Y52 to K107) with corresponding residues from 17 other vertebrates showed that the GIP open reading frame contains alternative basic cleavage sites for the generation of multiple GIP isoforms in primates, dogs, cats, cows, pigs, and horses (*upper* panel). The mature peptide region is indicated by a dark horizontal bar *above* the alignment. Residues that are conserved from teleosts to humans are indicated by a gray background. Putative basic cleavage sites are indicated by a red background. The position of the variable residue 103 is indicated by an arrow. Alternative post-translational processing of proGIP could lead to the generation of a 42-amino-acid mature GIP and an extended 55-amino-acid isoform (GIP55G or GIP55S) that differ at position 52 (*lower* panel). (*B*) GIP and GIP55 peptides are colocalized in the duodenum cells. Immunoreactive GIP and GIP55 were detected in select duodenum cells by immunofluorescent staining. The *right* panels showed the dark field and phased contrast images of merged immunofluorescent signals (800×). The white horizontal bar in each panel represents a distance of 30 μm. (*C*) GIP, GIP55G, and GIP55S suppressed exogenous glucose in fasting rats in vivo. Each of the three GIP peptides reduced glucose contents in the blood to basal levels at 1 h after injection of the peptide and glucose. Each data point represents the mean ± SEM of triplicate samples. Similar results were observed in five separate experiments.

investigate this possibility, we analyzed whether the GIP55 peptide is expressed in gut cells of the proximal small intestine. In support of our hypothesis, the immunohistochemical analysis of human duodenum sections with a rabbit GIP55-specific antibody and a mouse anti-GIP antibody showed that immunoreactive GIP55 and GIP are colocalized in select duodenum cells, suggesting that GIP55 is present in gut cells that normally express GIP (Fig. 3B). To study whether GIP55 is secreted into general circulation, we then analyzed the presence of GIP55 in serums of individuals 30 min after a regular breakfast meal using a sandwich ELISA assay that detects the extended C-terminal sequences specific to GIP55. Consistently, we found that GIP55 is present in human serum and constitutes ~1%–3% of the total GIP after a meal (GIP55, 2.91 ± 0.41 pmol/L; total GIP, 95.1 ± 9.41 pmol/L, N = 6). Thus, depending on the genotype, humans could contain two ($GIP^{103T/T}$: GIP+GIP55S; $GIP^{103C/C}$: GIP+GIP55G) or three GIP isoforms ($GIP103^{T/C}$: GIP+GIP55S+GIP55G) (Fig. 3A).

Because the receptor-activation domain of GIP is located at the N terminus of the peptide, we reasoned that alternative processing at the C terminus of GIP55 peptides is unlikely to decimate their bioactivity. Indeed, functional testing of the synthetic GIP isoforms (GIP, GIP55G, and GIP55S) in vivo showed that similar to conventional GIP, the extended GIP55G and GIP55S suppress hyperglycemia to similar extents in a time-dependent manner in fasting rats (Fig. 3C).

The ancestral GIP55S and the derived GIP55G peptides exhibit distinct bioactivity profiles in vitro

To compare the bioactivity of GIP isoforms and variants, we measured their receptor-activation activities in vitro using HEK293T cells expressing a recombinant human GIP receptor. As expected, treatments of GIP led to dose-dependent increases of cAMP production in transfected cells (Fig. 4A, top left panel). Unlike conventional GIP, which exhibits an EC_{50} of ~0.9 ± 0.21 nM, the extended GIP isoforms have approximately threefold lower potencies (GIP55G, 3.2 ± 0.21 nM; GIP55S, 2.6 ± 0.23 nM) (Supplemental Table 6). Importantly, we found that the derived GIP55G consistently increases cAMP production to significantly higher levels compared with the ancestral GIP55S (Fig. 4A, top left panel). Because GIP is known to be susceptible to serum degradation in vivo, we also studied the stability of GIP isoforms in human serum in vitro. Surprisingly, we found that GIP55G and GIP55S are more resistant to degradation by either pooled normal human serum (Fig. 4A) or pooled complement-preserved human serum (Supplemental Fig. 5). The ranking of potency on receptor activation shifted from GIP > GIP55G > GIP55S at 0 h to GIP = GIP55G = GIP55S and GIP55G > GIP55S ≥ GIP, respectively, after a 6-h or a 12-h preincubation with either normal serum (Fig. 4A), or complement-preserved serum (Supplemental Fig. 5). Plots of EC_{50} data in relation to the length of incubation showed that the slope of changes in the bioactivity for GIP is significantly steeper than that of GIP55G (Fig. 4B, P = 0.0023). On the other hand, coincubation with a recombinant dipeptidyl peptidase IV (DPP IV) led to similar extents of degradation of these peptides (Supplemental Table 7). Thus, the resistance to serum degradation by GIP55G cannot be attributed to a resistance of DPP IV, which represents the major processing enzyme that degrades GIP and GLP-1 in vivo by cleaving these peptides at position 2 of the N terminus (Kim and Egan 2008). These data suggested that the rise in the frequency of GIP^{103C} in Eurasian populations could be associated with the quantitative increase in the overall potency of GIP55G. Although the nature of the selective

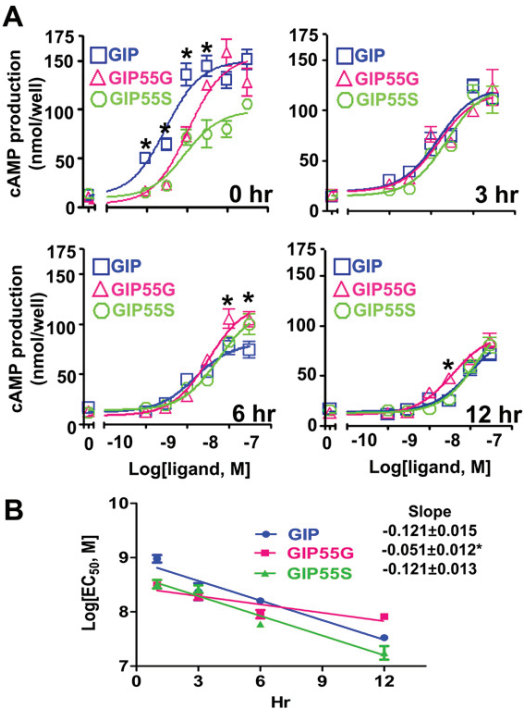


Figure 4. The variation at rs2291725 affects the bioactivity of translated products. (*A*) GIP55G peptide is resistant to serum degradation. Treatments of GIP receptor-expressing HEK293T cells with GIP, GIP55G or GIP55S led to dose-dependent increases of cAMP production (*top left* panel). Receptor-activation activities of peptides were also analyzed following incubation with pooled normal human serum for 3, 6, or 12 h. Cells were treated with synthetic peptides for 12 h; the signaling is reported as total cAMP contents in cell lysates. Error bars, SEM of triplicate samples. Significant differences in cAMP production between GIP and GIP55G treatments at a given peptide concentration are indicated by asterisks (P < 0.01). In the control group, cells were treated with an aliquot of human serum without a synthetic peptide. Similar results were observed in three separate experiments. (*B*) Comparison of the slopes of EC_{50} trend lines for GIP, GIP55G, and GIP55S after treatments with pooled human serum for the indicated time-spans. The slope of the GIP55G group is significantly different from that of the GIP group (*P = 0.0023).

advantage conferred by the GIP55 peptides is not clear, we speculate that GIP^{103C} could represent a risk allele in ancestors of YRI populations but provide a selective advantage in Eurasians.

The derived GIP^{103C} allele arose to dominance in East Asians ~8100 yr ago

Because energy-balance regulation–related loci are subject to selection pressures that fluctuate over time in response to environmental and culture changes, genetic responses to changes in the human diet are more likely associated with incomplete signatures of selection or even signatures of balancing selection (Charlesworth 2006; Pritchard et al. 2010). Based on this understanding, we speculated that the GIP^{103C} and GIP^{103T} alleles likely confer distinct advantages depending on the history of culture changes (Allison 1956; Turner et al. 1979) and that the selection of GIP^{103C} could occur at a time when humans experienced major shifts in subsistence culture. To investigate this possibility, we estimated the age of the GIP^{103C}-associated haplotype. Under the neutrality, the average age of a polymorphism with the frequency

p is estimated to be $-4N_e[p(\log p)/(1-p)]$ (Kimura and Ota 1973; Slatkin and Rannala 2000). With the assumption of $N_e = 5000$ for each population, this yielded 77,500, 350,000, and 425,000 yr for the derived allele to arise to its current frequencies in the YRI, CEU, and ASN populations, respectively. These estimates are obviously incompatible with the archaeological evidence showing that modern humans originated ~195 kyr ago in Sub-Saharan Africa and that a first wave of migration to the Arabic peninsula occurred ~60 to 55 kyr ago, which was followed by migration toward Northern Eurasia ~40 kya (McDougall et al. 2005; Klein 2009). Consequently, we estimated the age of the GIP^{103C}-associated haplotype on the basis of the decay of haplotypes (Reich 1998; Stephens et al. 1998). Based on a recombination rate derived from estimates of LD of the HapMap data set (McVean et al. 2004), the analysis showed that ASN contains a dominant ancestral haplotype and that the GIP^{103C}-associated haplotypes arose to dominance ~8100 yr ago in East Asians. Because this dating approach relies on the linkage map derived from estimates of LD, and because discrepancies between LD maps and the pedigree-based recombination maps are significant in many genomic regions (Clark et al. 2010), we also performed the analysis using alternative estimates of recombination rate, which range from 0.5–3.03 cM/Mb in three studies of pedigree-based recombination maps (i.e., deCODE, Marshfield, and Genethon) (Dib et al. 1996; Broman et al. 1998; Kong et al. 2002). Using these recombination rates, we obtained alternative estimates of the age to be between 11,800 and 2000 yr. On the other hand, dating is unattainable for CEU because the ancestral GIP^{103C} haplotype in this population cannot be identified unambiguously, perhaps because that mutation in the region is effectively "saturated" (Supplemental Fig. 6). Given that GIP signaling plays a critical role in the regulation of glucose and lipid metabolism, it is plausible that the increased prevalence of the GIP^{103C} allele in Eurasians is a consequence of changes in foraging skills or in population dynamics associated with the emergence of agricultural societies in the Eurasian continent 12,000 to 7000 yr ago (Balter 2007; Fuller et al. 2009; Jones and Liu 2009; Richards and Trinkaus 2009; Gibbons 2010).

Discussion

Based on the gene age estimation and biochemical analyses, our study revealed a functional mutation that is associated with the selection of the *GIP* locus in East Asian populations ~8100 yr ago and the presence of a cryptic GIP isoform. Specifically, we showed that the inventory of human GIP peptides has recently diverged and that individuals could express three different combinations of GIP isoforms (GIP, GIP55S, and GIP55G) with distinct bioactivity profiles. Future study of how this phenotypic variation affects glucose and lipid homeostasis in response to different diets and of which physiological variations in humans can be attributed to prior gene–environmental interactions at the *GIP* locus is crucial to a better understanding of human adaptations in energy-balance regulation.

Taking advantage of the availability of genome information from diverse human populations, recent studies have revealed a number of loci and variants that describe phenotypic variation in appearance, physiological parameters, and pathological responses to diseases (Sabeti et al. 2005; Voight et al. 2006, 2010; Sulem et al. 2007; Tishkoff et al. 2007; Genovese et al. 2010; Gibbons 2010; Leslie 2010; Ng et al. 2010; Simonson et al. 2010). In addition, it has been shown that human genomes contain hundreds of loci that exhibit varying degrees of positive selection (Kelley et al. 2006; Voight et al. 2006; Sabeti et al. 2007; Barreiro et al. 2008; Akey 2009; Cai et al. 2009; Pickrell et al. 2009). These recent in silico findings on gene selection, perhaps due to differential gene–environmental interactions among populations, have opened the doors to human history and a better understanding of the prevalence of adaptations among human populations. However, few causal variants have been confirmed or linked to a phenotype (Akey 2009; Gibbons 2010; Nei et al. 2010). Thus, our finding that the high-frequency GIP^{103C}-associated haplotypes in Eurasian populations are functionally relevant has provided a rare opportunity to better understand the environmental impact on human physiology at the enteroinsular and enteroadipocyte axes. It is generally accepted that a food source represents a potent force that influences selection and adaptive radiation in nature (Darwin 1859; Freeman and Herron 2003). For example, the external differences in beak morphology and speciation of Darwin's finches are believed to be results of adaptations that exploit particular types of seeds, insects, and cactus flowers on the Galápagos Islands (Schluter 2000). Likewise, changes in food source represented one of the most important selection pressures during transitions of human culture and posed a wide spectrum of challenges to the digestive and endocrine systems of our ancestors (Piperno et al. 2004; Balter 2007; Fuller et al. 2009; Jones and Liu 2009; Richards and Trinkaus 2009; Gibbons 2010). For instance, the ability to digest lactose in milk (lactase persistence) and the adoption of a pastoral culture has been linked to the selection of a variety of variants at the lactase locus in multiple ethnic groups at various time points during the last 10 millenniums (Tishkoff et al. 2007; Itan et al. 2009). Similarly, individuals from populations with high-starch diets tend to have more copies of amylases than do those from populations with low-starch diets (Perry et al. 2007). It has long been hypothesized that our ancestors ate a low-carbohydrate, high-protein diet and that the adaptive response was manifested as insulin resistance, perhaps for coping with low dietary glucose (Miller and Colagiuri 1994). Because normal GIP signaling is crucial to normal glucose and lipid metabolism, we speculate that the selection of the GIP^{103C} haplotypes could be pertinent to shifts in long-term energy-balance regulation after the emergence of agriculture, which provided a stable supply of high-starch staples and a reduced need for metabolic efficiency as opposed to the traditional hunter-gatherer societies (Miller and Colagiuri 1994; Wang et al. 2006; Balter 2007; Fuller et al. 2009; Jones and Liu 2009; Richards and Trinkaus 2009; Gibbons 2010; Luca et al. 2010; Richerson et al. 2010).

It was hypothesized by Neel almost 50 yr ago that mismatches between prior physiological adaptations and contemporary environments can lead to health risks because the ancestral variants that have been selected for the organism's fitness or reproductive success may not be optimal for the individual's health in the new environment (Neel 1962). In support of this thrifty genotype hypothesis, a number of genes in humans and house mice have been implied to have coevolved with the emergence of agricultural societies (Prentice et al. 2005; Vander Molen et al. 2005; Shiao et al. 2008), and a rapid shift in diets is associated with the detrimental effects on human survival in a number of human populations (Anonymous 1989). Conceptually, the serum-resistant GIP55G carried by the GIP^{103C} haplotype may have been beneficial for individuals who have unconstrained access to the food supply in many agricultural societies by preventing severe hyperglycemia. As selection pressure changed in these societies, the ancient GIP^{103T} haplotype could have become a liability and conferred a loss of fitness in the new environment. In addition, we speculate that the selection of *GIP* in East Asians may contribute to the heterogeneity

in the risk of diabetes among major ethnic groups at the present time (Retnakaran et al. 2006; Nystrom et al. 2008; Ma and Chan 2009). Importantly, regardless of what the selection pressure might have been, our study indicated that the *GIP* locus was susceptible to recent changes in human environment and that the physiological variation stemming from this selection could bear important implications for understanding the phenotypic variation of metabolic syndromes such as diabetes and obesity. Nonetheless, it is important to note that the selection of the derived GIP^{103C} haplotypes could be a consequence of the coordinated actions of multiple SNPs at the *GIP* locus. Future analysis of the functionality of rs2291725 and linked SNPs is needed to elucidate the exact nature of the selection of GIP^{103C} haplotypes in the last 10 millenniums.

Taken together, our study has highlighted an unexpected complexity in the regulation of sugar and lipid metabolism among human populations; it has also illuminated the importance of understanding adaptations in genes associated with energy-balance regulation in the face of ongoing changes in our diets toward refined and high-energy convenience food as well as the emerging pandemics of diabetes and obesity in modern society.

Methods

Genotype analysis

F_{ST} estimation and LD plot

The selected loci that have rapidly increased in frequency due to local positive selection are likely to show high levels of genetic differentiation, which can be quantified with the F_{ST} statistic (Lewontin and Krakauer 1973; Weir 1996). The principle of a number of estimators of F_{ST} is $F_{ST} = (H_T - H_S)/H_T$, where H_T is the heterozygosity of the total population and H_S is the average heterozygosity across subpopulations. The empirical distribution of F_{ST} or the average of F_{ST} over multilocus windows for the genome-wide polymorphism data has been used to detect the genomic regions under positive selection (Akey et al. 2002). We computed the F_{ST} for all nonsynonymous SNPs in the HapMap Project phases I and II, and the function snp_fst.m in PGEToolbox (Cai 2008) was used to calculate and present an unbiased estimate of F_{ST}.

LD plots for SNPs within the *CALCOCO2*, *ATP5G1*, *UBE2Z*, *SNF8*, *GIP*, and *IGF2BP1* gene regions from rs1422645 to rs8069452, spanning 200 kb, were generated using HaploView 4.1 (Barrett et al. 2005). The LD blocks were determined using an r^2 threshold of 0.8.

Analysis of EHH and iHS

The EHH plots were generated as described (Voight et al. 2006). We also computed the EHH statistic for the core SNP (rs2291725) as previously described (Sabeti et al. 2002). The EHH curve measures the decay of identity of haplotypes that carry the core SNP as a function of distance. When an allele rises rapidly in frequency due to strong selection, it tends to exhibit high levels of haplotype homozygosity, extending further than expected under a neutral model. For the analysis of EHH decay, haplotypes in the genomic region that encompasses a 750-kb region centered on the *GIP* gene and contains 388 phased SNPs were downloaded from the HapMap project website.

The iHS statistic detects whether the area under the EHH curve is greater for a selected allele than for a neutral allele (Voight et al. 2006). Large negative values indicate unusually long haplotypes carrying the derived allele. To test whether the observed iHS for rs2291725 is significantly deviated from the expected values derived from neutral evolution, we used a coalescent model and generated 10,000 replicated haplotype sets using the coalescent simulator *ms* (Hudson 2002). We adopted parameters that are compatible with the sample size (180 chromosomes) of the ASN population and the length of genomic region (750 kb) that was analyzed. The simulation was conditioned based on recombination rate = 9.7×10^{-7} cM/bp, N_0 = 10,000, and number of segregating sites = 388. We also required that each replicate contain at least one segregating site in which the derived allele frequency reaches ~0.75 and that the site is located in the center of the haplotype (that is, between a 370- and 380-kb area of the 750-kb region in total). With these conditions, we found no iHS value for a simulated data set to be more negative than the observed iHS (empirical $P < 1 \times 10^{-4}$) (Supplemental Fig. 4A).

Estimation of the age of GIP^{103C} haplotype

Under positive selection, a derived, or novel, allele can quickly rise to a high frequency. As the allele ages, recombination causes the breakdown of the LD within the existing linked alleles, and mutation leads to the accumulation of new linked variations. The allele's age can be estimated by the decay of the haplotype carrying the allele (Stephens et al. 1998). The length of the haplotype retained between different copies of the allele shortens over time due to recombination and mutation, which is modeled using a Poisson process by Reich and Goldstein (1999). The probability that a haplotype remains ancestral (i.e., that the haplotype that carries the derived allele retains the status of high frequency right after being selected) is $P = e^{-G(r+u)}$, where G is generations, r is the recombination rate, and u is the mutation rate. For a given allele in the derived haplotype, the haplotype-decay approach estimates the number of generations G in terms of P (the probability that a given haplotype does not change from its ancestor) (Stephens et al. 1998; Reich and Goldstein 1999).

In ASN, the *GIP* LD block is 91 kb long and contains eight distinct haplotypes (1–8 from top to bottom in Supplemental Fig. 6). Haplotypes 1–4 are GIP^{103C} associated and are derived. Among these derived haplotypes, haplotype 1 is apparently the ancestral haplotype from which haplotypes 2–4 were derived. The frequency of haplotype 1 in GIP^{103C}-associated haplotypes is P' = 0.803. We obtained the recombination rate derived from estimates of LD (McVean et al. 2004) from the website of the HapMap project (http://hapmap.ncbi.nlm.nih.gov/). Regression analysis with data points of different physical lengths versus genetic map lengths gave the regression function: $y = 0.93x - 0.018$, where x is the physical distance (Mb) and y is the genetic distance (cM). The genetic distance of the 91-kb region was estimated to be 0.0666 cM (i.e., 0.0666% chance of crossing over in a single generation), which gives a rate of recombination $r = 6.66 \times 10^{-4}$ per generation. We took the mutation rate of the region $u = 1.0 \times 10^{-5}$, based on an estimation of the haploid mutation rate of $\sim 1.1 \times 10^{-8}$ per base per generation (Roach et al. 2010). We assumed that the most common haplotype is the ancestral haplotype and used P' obtained above as an approximation of P. Using $G = -\ln(P')/(r+u)$, we obtained G = 325 (or 8100 yr).

Phenotype tests

Reagents

Synthetic GIP peptides were obtained from Genescript and the American Peptide Company. The extended GIP55G and GIP55S isoforms were synthesized by the Stanford University PAN facility and GL Biochem. Pooled normal human serum, pooled complement-preserved human serum, and serum from individuals were obtained from Innovative Research and ProteoGenex. Stocks of different hormones were prepared in phosphate-buffered saline

and diluted in a serum-free culture medium. In addition to routine chemistry and mass-spectrometry assessments, we verified the quantity of different GIP isoforms using a human GIP enzyme-linked immunosorbent assay (ELISA; Linco Research).

Immunohistochemical analyses and ELISA

Total GIP levels in human serums were measured using a sandwich human GIP ELISA kit (Linco Research), and assays were performed with a programmable ELISA plate washer. The minimum detection limit of the assay was 8.1 pg/mL. For the measurement of GIP55, human serums were partially purified with C18 chromatography, and the secondary antibody of the human GIP ELISA kit was substituted with a rabbit polyclonal antibody specific for the last 13 amino acids of GIP55 (REARALELASQAN). The GIP55-specific antibodies were generated using a KLH-conjugated CREARALELASQAN peptide (Covance Research Products and Genescript).

For immunohistochemical analyses, human duodenum sections (BioChain Institute) were dewaxed with xylene and maintained for 10 min at 95°C–99°C in 10 mM sodium citrate buffer (pH 6.0), followed by cooling on the bench top. Sections were immunostained with a mouse monoclonal anti-GIP antibody (Abbiotec) and a rabbit GIP55-specific antibody overnight at 4°C, followed by incubation with fluorochrome-conjugated secondary antibodies (Invitrogen) for 1 h at room temperature in the dark. Signals for GIP and GIP55 were detected using an Alexa 488 donkey anti-mouse antibody (488 nm, 1:2500 dilution) and a Texas Red-conjugated goat anti-rabbit antibody (594 nm, 1:500 dilution), respectively, with a Leica SP2 single- and multi-photon confocal microscope.

The in vivo glucose suppression assay

Eight-week-old Sprague-Dawley rats (Charles River Laboratories, Inc., Wilmington, MA) were fasted overnight for 20 h. To measure the GIP isoforms' ability to reduce blood glucose levels in vivo, fasting rats were injected with a select GIP peptide (100 nmol/kg) dissolved in 0.8 mL of PBS together with glucose (3.8 g/kg body weight). Blood samples were obtained via the tail vein at select time points after the intraperitoneal injection of glucose and the peptide. Glucose levels in the blood were measured with a One-Touch Ultra Blood Glucose Monitoring System and OneTouch Ultra Test Strips (Johnson and Johnson).

Receptor–activation assay

The expression vector for the human GIP receptor (Origene) was transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen). Receptor-activation activities were assayed based on cAMP production in transfected cells and were performed as described (Park et al. 2008; Chang et al. 2010). To quantify resistance to serum degradation by GIP isoforms, aliquots of GIP peptides were incubated in microfuges at 37°C with human serum in a final concentration of 10 μM for indicated time spans. To analyze the mechanism underlying the serum-degradation–resistant property of GIP55, peptides (10 μM) were preincubated with a recombinant human DPP IV (1 mU/mL reaction in PBS; Enzo Life Science) for 3, 6, or 12 h and were frozen at −80°C before being tested for receptor-activation activity. The receptor-activation results were analyzed using the GraphPad Prism 5 package (GraphPad Software).

Acknowledgments

We thank Yi Wei and Augustin Sanchez (Stanford University Pan Facility) for technical assistance. We thank Drs. Aaron J.W. Hsueh and Renee A. Reijo Pera (Department of OB/GYN, Stanford University) for the critical review of the manuscript, and C.L.C. further thanks Dr. Yung-Kuei Soong (Department of OB/GYN, Chang Gung Memorial Hospital, Taiwan). J.J.C. thanks Dr. Dmitri Petrov (Dept. of Biology, Stanford University) for his invaluable advice and long-lasting support. We acknowledge the support of NIH (DK70652, SYTH), Avon Foundation (02-2009-054, SYTH), and Chang Gung Memorial Hospital (CMRPG34002, CLC).

Author contributions: S.Y.T.H. conceived and supervised the study. C.L.C., J.P., and S.Y.T.H. were involved in functional testing and data analyses. J.J.C., C.L., and J.A. processed and performed genome data analyses. The paper was written primarily by C.L.C., J.J.C., and S.Y.T.H.

References

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19:** 711–722.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12:** 1805–1814.

Allison AC. 1956. The sickle-cell and haemoglobin C genes in some African populations. *Ann Hum Genet* **21:** 67–89.

Amigo J, Salas A, Phillips C, Carracedo A. 2008. SPSmart: Adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics* **9:** 428.

Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, et al. 2009. Molecular and evolutionary history of melanism in North American gray wolves. *Science* **323:** 1339–1343.

Anonymous. 1989. Thrifty genotype rendered detrimental by progress? *Lancet* **2:** 839–840.

Balter M. 2007. Plant science. Seeking agriculture's ancient roots. *Science* **316:** 1830–1835.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nat Rev Genet* **11:** 17–30.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* **40:** 340–345.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Ben-Shlomo I, Yu Hsu S, Rauch R, Kowalski HW, Hsueh AJ. 2003. Signaling receptome: A genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci STKE* **203:** RE9. doi: 10.1126/stke.2003.187.re9.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am J Hum Genet* **63:** 861–869.

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S. 2008. Positive selection in East Asians for an EDAR allele that enhances NF-κB activation. *PLoS ONE* **3:** e2209. doi: 10.1371/journal.pone.0002209.

Buchanan TA, Xiang AH. 2005. Gestational diabetes mellitus. *J Clin Invest* **115:** 485–491.

Cai JJ. 2008. PGEToolbox: A Matlab toolbox for population genetics and evolution. *J Hered* **99:** 438–440.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet* **5:** e1000336. doi: 10.1371/journal.pgen.1000336.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, et al. 2002. A human genome diversity cell line panel. *Science* **296:** 261–262.

Chambers JC, Elliott P, Zabaneh D, Zhang W, Li Y, Froguel P, Balding D, Scott J, Kooner JS. 2008. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* **40:** 716–718.

Chang CL, Park J-I, Hsu SYT. 2010. Activation of calcitonin receptor and calcitonin receptor-like receptor by membrane-anchored ligands. *J Biol Chem* **285:** 1075–1080.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2:** e64. doi: 10.1371/journal.pgen.0020064.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res* **20:** 393–402.

Chia CW, Carlson OD, Kim W, Shin YK, Charles CP, Kim HS, Melvin DL, Egan JM. 2009. Exogenous glucose-dependent insulinotropic polypeptide worsens post prandial hyperglycemia in type 2 diabetes. *Diabetes* **58:** 1342–1349.

Clark AG, Wang X, Matise T. 2010. Contrasting methods of quantifying fine structure of human recombination. *Annu Rev Genomics Hum Genet* **11:** 45–64.

Darwin C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380:** 152–154.

Drews J. 2000. Drug discovery: A historical perspective. *Science* **287:** 1960–1964.

Freeman S, Herron JC. 2003. *Evolutionary analysis*, 3rd ed. Prentice-Hall, Upper Saddle River, NJ.

Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, Mustofa MS, Samakkarn U, Settheetham-Ishida W, Ishida T, et al. 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet* **17:** 835–843.

Fuller DQ, Qin L, Zheng Y, Zhao Z, Chen X, Hosoya LA, Sun GP. 2009. The domestication process and domestication rate in rice: Spikelet bases from the Lower Yangtze. *Science* **323:** 1607–1610.

Fulurija A, Lutz TA, Sladko K, Osto M, Wielinga PY, Bachmann MF, Saudan P. 2008. Vaccination against GIP for the treatment of obesity. *PLoS ONE* **3:** e3163. doi: 10.1371/journal.pone.0003163.

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, et al. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329:** 841–845.

Gibbons A. 2010. Human evolution. Tracing evolution's recent fingerprints. *Science* **329:** 740–742.

Gniuli D, Calcagno A, Dalla Libera L, Calvani R, Leccesi L, Caristo ME, Vettor R, Castagneto M, Ghirlanda G, Mingrone G. 2010. High-fat feeding stimulates endocrine, glucose-dependent insulinotropic polypeptide (GIP)-expressing cell hyperplasia in the duodenum of Wistar rats. *Diabetologia* **53:** 2233–2240.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5:** e1000695. doi: 10.1371/journal.pgen.1000695.

Hellenthal G, Stephens M. 2007. msHOT: Modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatics* **23:** 520–521.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313:** 101–104.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337–338.

Ingelsson E, Langenberg C, Hivert MF, Prokopenko I, Lyssenko V, Dupuis J, Magi R, Sharp S, Jackson AU, Assimes TL, et al. 2010. Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes* **59:** 1266–1275.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

International HapMap Project. 2003. The International HapMap Project. *Nature* **426:** 789–796.

Isken F, Pfeiffer AF, Nogueiras R, Osterhoff MA, Ristow M, Thorens B, Tschop MH, Weickert MO. 2008. Deficiency of glucose-dependent insulinotropic polypeptide receptor prevents ovariectomy-induced obesity in mice. *Am J Physiol Endocrinol Metab* **295:** E350–E355.

Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG. 2009. The origins of lactase persistence in Europe. *PLoS Comput Biol* **5:** e1000491. doi: 10.1371/journal.pcbi.1000491.

Jones MK, Liu X. 2009. Archaeology. Origins of agriculture in East Asia. *Science* **324:** 730–731.

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16:** 980–989.

Kim W, Egan JM. 2008. The role of incretins in glucose homeostasis and diabetes treatment. *Pharmacol Rev* **60:** 470–512.

Kimura M, Ota T. 1973. The age of a neutral mutant persisting in a finite population. *Genetics* **75:** 199–212.

Klein RG. 2009. Darwin and the recent African origin of modern humans. *Proc Natl Acad Sci* **106:** 16007–16009.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241–247.

Laland KN, Odling-Smee J, Myles S. 2010. How culture shaped the human genome: Bringing genetics and the human sciences together. *Nat Rev Genet* **11:** 137–148.

Lalueza-Fox C, Rompler H, Caramelli D, Staubert C, Catalano G, Hughes D, Rohland N, Pilli E, Longo L, Condemi S, et al. 2007. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science* **318:** 1453–1455.

Leslie M. 2010. Genetics. Kidney disease is parasite-slaying protein's downside. *Science* **329:** 263.

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74:** 175–195.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Luca F, Perry GH, di Rienzo A. 2010. Evolutionary adaptations to dietary changes. *Annu Rev Nutr* **30:** 291–314.

Ma RC, Chan JC. 2009. Pregnancy and diabetes scenario around the world: China. *Int J Gynaecol Obstet* **104:** S42–S45.

McClean PL, Irwin N, Cassidy RS, Holst JJ, Gault VA, Flatt PR. 2007. GIP receptor antagonism reverses obesity, insulin resistance, and associated metabolic disturbances induced in mice by prolonged consumption of high-fat diet. *Am J Physiol Endocrinol Metab* **293:** E1746–E1755.

McDougall I, Brown FH, Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433:** 733–736.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.

Miller JC, Colagiuri S. 1994. The carnivore connection: Dietary carbohydrate in the evolution of NIDDM. *Diabetologia* **37:** 1280–1286.

Miyawaki K, Yamada Y, Ban N, Ihara Y, Tsukiyama K, Zhou H, Fujimoto S, Oku A, Tsuda K, Toyokuni S, et al. 2002. Inhibition of gastric inhibitory polypeptide signaling prevents obesity. *Nat Med* **8:** 738–742.

Moody AJ, Thim L, Valverde I. 1984. The isolation and sequencing of human gastric inhibitory peptide (GIP). *FEBS Lett* **172:** 142–148.

Neel JV. 1962. Diabetes mellitus: A "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet* **14:** 353–362.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* **11:** 265–289.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42:** 790–793.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* **8:** 857–868.

Nystrom MJ, Caughey AB, Lyell DJ, Druzin ML, El-Sayed YY. 2008. Perinatal outcomes among Asian-white interracial couples. *Am J Obstet Gynecol* **199:** 385. doi: 10.1016/j.ajog.2008.06.065.

Park JI, Semyonov J, Chang CL, Yi W, Warren W, Hsu SY. 2008. Origin of INSL3-mediated testicular descent in therian mammals. *Genome Res* **18:** 974–985.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39:** 1256–1260.

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* **19:** 826–837.

Piperno DR, Weiss E, Holst I, Nadel D. 2004. Processing of wild cereal grains in the Upper Palaeolithic revealed by starch grain analysis. *Nature* **430:** 670–673.

Prentice AM, Rayco-Solon P, Moore SE. 2005. Insights from the developing world: Thrifty genotypes and thrifty phenotypes. *Proc Nutr Soc* **64:** 153–161.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: Hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* **20:** R208–R215.

Prokopenko I, McCarthy MI, Lindgren CM. 2008. Type 2 diabetes: New genes, new understanding. *Trends Genet* **24:** 613–621.

Reich DE. 1998. Estimating the age of mutations using variation at linked markers. In *Microsatellites: Evolution and applications*, pp. 129–138. Oxford University Press, Oxford.

Reich DE, Goldstein DB. 1999. Estimating the age of mutations using variation at linked markers. In *Microsatellites: Evolution and applications*, (ed. DB Goldstein and C Schlotterer), pp. 129–138. Oxford University Press, Oxford.

Retnakaran R, Hanley AJ, Connelly PW, Sermer M, Zinman B. 2006. Ethnicity modifies the effect of obesity on insulin resistance in pregnancy: A comparison of Asian, South Asian, and Caucasian women. *J Clin Endocrinol Metab* **91:** 93–97.

Richards MP, Trinkaus E. 2009. Out of Africa: Modern human origins special feature: Isotopic evidence for the diets of European Neanderthals and early modern humans. *Proc Natl Acad Sci* **106:** 16034–16039.

Richerson PJ, Boyd R, Henrich J. 2010. Colloquium paper: Gene-culture coevolution in the age of genomics. *Proc Natl Acad Sci* **107:** 8985–8992.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328:** 636–639.
Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* **70:** 841–847.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.
Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N, et al. 2005. The case for selection at CCR5-Delta32. *PLoS Biol* **3:** e378. doi: 10.1371/journal.pbio.0030378.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449:** 913–918.
Saxena R, Hivert MF, Langenberg C, Tanaka T, Pankow JS, Vollenweider P, Lyssenko V, Bouatia-Naji N, Dupuis J, Jackson AU, et al. 2010. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42:** 142–148.
Schluter D. 2000. *The ecology of adaptive radiation*. Oxford University Press, Oxford.
Seminara SB, Messager S, Chatzidaki EE, Thresher RR, Acierno JS Jr, Shagoury JK, Bo-Abbas Y, Kuohung W, Schwinof KM, Hendrick AG, et al. 2003. The GPR54 gene as a regulator of puberty. *N Engl J Med* **349:** 1614–1627.
Semyonov J, Park JI, Chang CL, Hsu SY. 2008. GPCR genes are preferentially retained after whole genome duplication. *PLoS ONE* **3:** e1903. doi: 10.1371/journal.pone.0001903.
Shiao MS, Liao BY, Long M, Yu HT. 2008. Adaptive evolution of the insulin two-gene system in mouse. *Genetics* **178:** 1683–1691.
Shimomura Y, Wajid M, Ishii Y, Shapiro L, Petukhova L, Gordon D, Christiano AM. 2008. Disruption of P2RY5, an orphan G protein-coupled receptor, underlies autosomal recessive woolly hair. *Nat Genet* **40:** 335–339.
Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB, et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329:** 72–75.
Slatkin M, Rannala B. 2000. Estimating allele age. *Annu Rev Genomics Hum Genet* **1:** 225–249.
Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23:** 23–35.
Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, et al. 1998. Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. *Am J Hum Genet* **62:** 1507–1515.
Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, et al. 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39:** 1443–1452.
Takeda J, Seino Y, Tanaka K, Fukumoto H, Kayano T, Takahashi H, Mitani T, Kurono M, Suzuki T, Tobe T, et al. 1987. Sequence of an intestinal cDNA encoding human gastric inhibitory polypeptide precursor. *Proc Natl Acad Sci* **84:** 7005–7008.
Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39:** 31–40.
Topaloglu AK, Reimann F, Guclu M, Yalin AS, Kotan LD, Porter KM, Serin A, Mungan NO, Cook JR, Ozbek MN, et al. 2009. TAC3 and TACR3 mutations in familial hypogonadotropic hypogonadism reveal a key role for Neurokinin B in the central control of reproduction. *Nat Genet* **41:** 354–358.
Turner JR, Johnson MS, Eanes WF. 1979. Contrasted modes of evolution in the same genome: Allozymes and adaptive change in Heliconius. *Proc Natl Acad Sci* **76:** 1924–1928.
Vander Molen J, Frisse LM, Fullerton SM, Qian Y, Del Bosque-Plata L, Hudson RR, Di Rienzo A. 2005. Population genetics of CAPN10 and GPR35: Implications for the evolution of type 2 diabetes variants. *Am J Hum Genet* **76:** 548–560.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72. doi: 10.1371/journal.pbio.0040072.
Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42:** 579–589.
Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci* **103:** 135–140.
Weir BS. 1996. *Genetic data analysis II: Methods for discrete population genetic data*. Sinauer Associates, Sunderland.
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329:** 75–78.

Received May 17, 2010; accepted in revised form October 22, 2010.

ENGINES: exploring single nucleotide variation in entire human genomes.

Amigo J, Salas A, Phillips C

BMC Bioinformatics. 01/2011; 12:105.

Las tecnologías de ultra-secuenciación de nueva generación a producir cantidades extensas de datos de secuencias de genoma o exoma humano, y por lo tanto se necesitan nuevos programas para presentar y analizar esta gran cantidad de información. El Proyecto 1000 Genomes ha publicado recientemente los datos crudos de 629 genomas completos que representan varias poblaciones humanas a través del análisis interno de la Fase I y, aunque hay ciertas herramientas públicas disponibles que permiten la exploración de estos genomas, hasta la fecha no existe una herramienta que permita un análisis poblacional amplio de la variación catalogada por estos datos. Hemos desarrollado un explorador de variantes genéticas capaz de obtener datos de variaciones de nucleótido único (SNVs), población por población, a partir de genomas enteros sin comprometer la capacidad de ampliación futura y su agilidad. ENGINES (interfaz de genoma completo para la exploración de SNVs) utiliza datos de 1000 Genomes Fase I para demostrar su capacidad para manejar grandes cantidades de variación genética (>7,3 mil millones de genotipos y 28 millones de SNVs), así como de derivar estadísticas resumen de interés para aplicaciones médicas y de genética de poblaciones. El conjunto de datos es pre-procesado y se resume en un repositorio estático de datos accesible a través de una interfaz web. El sistema de consultas permite la combinación y la comparación de cada muestra de población disponible, buscando por lista de códigos *rs*, región cromosómica, o genes de interés. Filtros de frecuencia y F_{ST} están disponibles para refinar las consultas, mientras que los resultados pueden ser comparados visualmente con otros grandes repositorios de polimorfismos de nucleótido único (SNPs) como HapMap o Perlegen. ENGINES es capaz de acceder a datos a gran escala de repositorios de variación de una manera rápida y comprensible. Permite rápida inspección de la variación del genoma, y suministra información estadística para cada posición variante como la frecuencia alélica, heterocigosidad o valores F_{ST} para la diferenciación genética.

DATABASE Open Access

ENGINES: exploring single nucleotide variation in entire human genomes

Jorge Amigo[1,2*], Antonio Salas[2] and Christopher Phillips[2]

Abstract

Background: Next generation ultra-sequencing technologies are starting to produce extensive quantities of data from entire human genome or exome sequences, and therefore new software is needed to present and analyse this vast amount of information. The 1000 Genomes project has recently released raw data for 629 complete genomes representing several human populations through their Phase I interim analysis and, although there are certain public tools available that allow exploration of these genomes, to date there is no tool that permits comprehensive population analysis of the variation catalogued by such data.

Description: We have developed a genetic variant site explorer able to retrieve data for Single Nucleotide Variation (SNVs), population by population, from entire genomes without compromising future scalability and agility. ENGINES (ENtire Genome INterface for Exploring SNVs) uses data from the 1000 Genomes Phase I to demonstrate its capacity to handle large amounts of genetic variation (>7.3 billion genotypes and 28 million SNVs), as well as deriving summary statistics of interest for medical and population genetics applications. The whole dataset is pre-processed and summarized into a data mart accessible through a web interface. The query system allows the combination and comparison of each available population sample, while searching by rs-number list, chromosome region, or genes of interest. Frequency and F_{ST} filters are available to further refine queries, while results can be visually compared with other large-scale Single Nucleotide Polymorphism (SNP) repositories such as HapMap or Perlegen.

Conclusions: ENGINES is capable of accessing large-scale variation data repositories in a fast and comprehensive manner. It allows quick browsing of whole genome variation, while providing statistical information for each variant site such as allele frequency, heterozygosity or F_{ST} values for genetic differentiation. Access to the data mart generating scripts and to the web interface is granted from http://spsmart.cesga.es/engines.php

Background

The appearance of large-scale online compilations of human variation has profoundly changed the population genetics field in the last decade. Private companies such as Perlegen Sciences [1], global collaborations such as HapMap [2] and high density Single Nucleotide Polymorphism (SNP) genotyping of the CEPH human genome diversity panel by groups from the Universities of Stanford [3] and Michigan, have provided extensive variation catalogues for geneticists to examine differences amongst a wide range of human populations. But although most genome studies have released their raw data to the public there has been a lack of web interfaces that allow population genetics based interpretation of the data. Indeed, in the current era of rapidly expanding numbers of publicly released complete human sequences there is an evident need to develop online data browsers that can collate and represent portions of the data relevant for particular fields of research.

The 1000 Genomes project http://www.1000genomes.org/ is a public initiative that aims to collect a very large proportion of variation detectable by next generation sequencing techniques of human genomes from several worldwide populations. The first pilot study (Pilot 1) assessed the strategy of sharing data across samples on whole genome sequencing results with relatively low coverage (2-4x). It presented 179 genomes from the four different population panels previously

* Correspondence: jorge.amigo@usc.es
[1]Grupo de Medicina Xenómica, CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain
Full list of author information is available at the end of the article

BioMed Central

© 2011 Amigo et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

characterised by HapMap (CEU, CHB, JPT and YRI) describing ~14 million variants. The recent release of an interim analysis of the project's Phase I has considerably enriched the data available: 629 entire genomes from 12 different populations, describing ~28 million variants. These populations are: individuals of African ancestry in Southwest USA (ASW), Utah residents with N & W European ancestry from the CEPH collection (CEU), Han Chinese in Beijing, China (CHB), Han Chinese South (CHS), Finnish in Finland (FIN), British in England and Scotland (GBR), Japanese in Tokyo, Japan (JPT), Luhya in Webuye, Kenya (LWK), individuals of Mexican ancestry in Los Angeles, California (MXL), Puerto Ricans in Puerto Rico (PUR), Tuscans in Italy (TSI), and Yoruba in Ibadan, Nigeria (YRI).

Although the 1000 Genomes project has already started to release results there are few publicly available bioinformatics tools that allow thorough exploration of such data. The Integrative Genomics Viewer http://www.broadinstitute.org/igv/home is a Java-based desktop application that permits visual browsing of the 1000 Genomes Pilot 1, 2, and 3 calls (among other tracks). Alternatively the 1000 Genomes Browser http://browser.1000genomes.org/ is a web tool that permits visualization of the variant sites against the reference sequence, and dynamic loading of tracks of interest (functional consequence, conservation, etc.). The latter provides a very simple and intuitive way to browse the 1000 Genomes results, but it does not provide basic variation statistics for population studies such as allele frequency or genetic differentiation of the genomes included in the project. More importantly, the 1000 Genomes Browser reviews the sequence surrounding just a single query at a time whether variant site, gene or chromosome segment. Furthermore, the 1000 Genomes browser is currently confined to the six Pilot 2 sequences.

Table 1 Data mart facts

1000 Genomes Phase I

Populations	*N* genomes	Variant Sites*[1]	Variant Genotypes
ASW	24	14,037,711	336,905,064
CEU	90	10,983,038	988,473,420
CHB	68	9,490,259	645,337,612
CHS	25	7,588,537	189,713,425
FIN	36	8,680,985	312,515,460
GBR	43	9,376,836	403,203,948
JPT	84	10,071,464	846,002,976
LWK	67	17,279,531	1,157,728,577
MXL	17	8,513,411	144,727,987
PUR	5	6,354,128	31,770,640
TSI	92	11,368,655	1,045,916,260
YRI	78	16,567,193	1,292,241,054
TOTAL	629	28,210,483	7,394,536,423

HapMap release 28

Populations	*N* samples	Variant Sites	Variant Genotypes
ASW	53	1,543,440	81,802,320
CEU	121	2,816,160	340,755,360
CHB	139	2,635,473	366,330,747
CHD	109	1,312,139	143,023,151
GIH	101	1,409,285	142,337,785
JPT	116	2,561,639	297,150,124
LWK	110	1,527,108	167,981,880
MEX	58	1,453,424	84,298,592
MKK	156	1,532,287	239,036,772
TSI	102	1,420,285	144,869,070
YRI	153	3,151,427	482,168,331
TOTAL	1218	4,170,392	2,489,754,132

The comparison of all the variability information present on the 1000 Genomes Phase I with HapMap release 28 indicates that although HapMap doubles the sample size, 1000 Genomes triples the number of genotypes due to the superior density of variants (this is particularly interesting in the YRI population which is now even more completely described than before). The number of non-monomorphic sites is reported as "variant sites".

*[1]Variant sites refer to the number of bi-allelic markers observed in each population group. Note that the Phase I does not contain information on tri- or tetra-allelic variants while in Pilot 1 there are more than 16,000 tri-allelic SNVs plus 12 tetra-allelic SNVs (data not shown).

Construction and content

We have developed a human genome variant site browser: ENGINES dedicated, in the first instance, to the flexible and thorough analysis of the Single Nucleotide Variation (SNV) catalogue generated from the 1000 Genomes Phase I interim analysis, although it will subsequently integrate new whole genome sequence data from other sources as this becomes publicly available.

Design and capabilities

As shown in Table 1 the volume of data is already very large, and with the goal to aggregate all available new whole genome data, summarizing approaches are essential to allow easy data management and to perform quick non-batched queries [4]. The whole dataset is pre-processed using a pipeline of customized PERL scripts and collated into a seven gigabyte MySQL data mart, containing only the summarized statistics, arranged by population, including allele frequencies, heterozygosity or minor allele frequency (MAF). This data mart is then queried through a PHP web interface with the main aim of permitting multiple SNV queries of entire genomes with a single step, dictated by user-defined nucleotide range, HGNC gene symbol list or rs-number list applied to the user's selection from different global population panels (Figure 1).

The statistics tab displays a table describing each variation result in columns: variation code, chromosome, chromosome position, gene, reference allele (from the

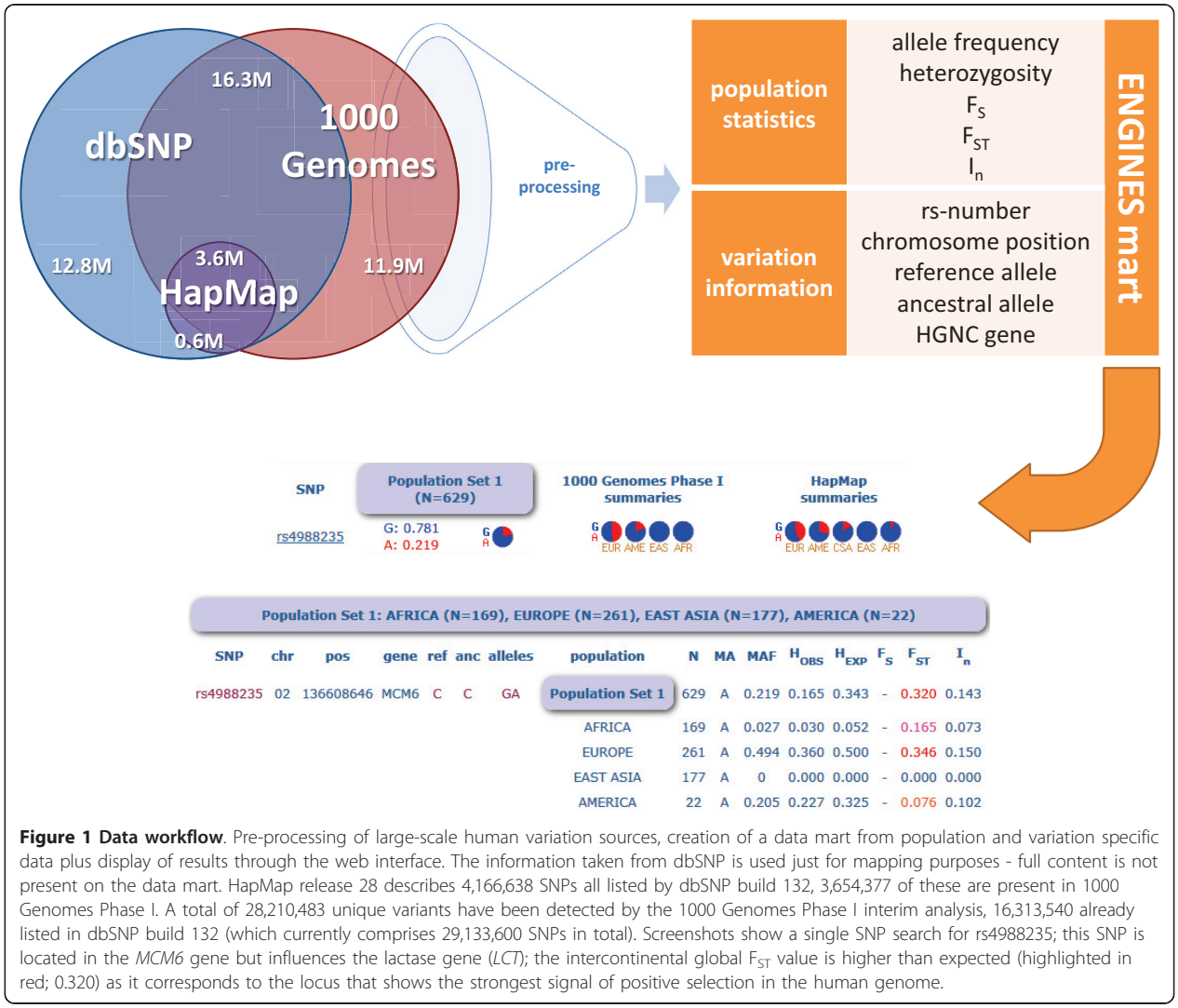


Figure 1 Data workflow. Pre-processing of large-scale human variation sources, creation of a data mart from population and variation specific data plus display of results through the web interface. The information taken from dbSNP is used just for mapping purposes - full content is not present on the data mart. HapMap release 28 describes 4,166,638 SNPs all listed by dbSNP build 132, 3,654,377 of these are present in 1000 Genomes Phase I. A total of 28,210,483 unique variants have been detected by the 1000 Genomes Phase I interim analysis, 16,313,540 already listed in dbSNP build 132 (which currently comprises 29,133,600 SNPs in total). Screenshots show a single SNP search for rs4988235; this SNP is located in the *MCM6* gene but influences the lactase gene (*LCT*); the intercontinental global F_{ST} value is higher than expected (highlighted in red; 0.320) as it corresponds to the locus that shows the strongest signal of positive selection in the human genome.

current human reference genome GRCh37), ancestral allele (from the Chimpanzee genome), alleles found in all present genotypes, populations queried, number of samples (N), the minor allele (MA) and its frequency (MAF), observed and expected heterozygosities (H_{OBS} and H_{EXP}), local inbreeding (F_S), genetic differentiation (F_{ST}, which is presented on different colours depending on meaning steps: under 0.05, 0.15, 0.25 and above 0.25) and informativeness of population group assignment (I_n). In ENGINES the emphasis is on multiple queries as a flexible, and in terms of genome portions that can be queried, broader alternative to the single marker queries offered by e.g. the 1000 Genomes browser.

Rapid responses to queries of dense genomic data have been engineered into the browser by use of pre-calculated SNV allele frequencies based on population groupings, an approach already successfully implemented in the population-based SNP frequency browser: SPSmart [5]. ENGINES therefore exploits one of the major assets of the 1000 Genomes Pilot 1 data, the improved detection and characterization of low frequency nucleotide variation, whether defined by population, genome position or overall MAF with linked references made at the same time to existing data in dbSNP or HapMap. For example, ENGINES can be used to search in batch mode for:

1) SNVs in specific genes or gene families;
2) SNVs at varying frequencies in different global population panels;
3) Novel variants or SNVs at very low MAF, which are now adequately catalogued and validated; For any selected SNV set, ENGINES can also calculate a range of statistical indices of interest for human population genetics studies.

Maintaining the data mart

The update frequency of the databases currently accessed by ENGINES varies considerably. Thus, while dbSNP is expected to release updates on a yearly basis, having been updated once or twice a year since 2004, Phase I is a static resource, and the project's final data releasing policy has not been publicly stated. The data mart will be updated with the 1000 Genomes final variant data upon release, in addition relevant whole genome sequencing data in the public domain from other initiatives will also be collated and included.

Originally, ENGINES used 1000 Genomes' Pilot 1 as an appropriate testing dataset. It was mapped to the old NCBI36/hg18 human genome reference, and for that reason we were forced to use dbSNP build 130 as the most up to date standard for describing all variants when possible. When the 1000 Genomes project released this Phase I interim analysis we decided to update our tool to a more appropriate testing dataset, implying adapting the data parsing scripts and upgrading the mapping reference to the new GRCh37/hg19. This later fact allowed ENGINES to update the variants description reference to dbSNP build 132, and considering that human reference versions tend to be fixed for a long time this should allow the internal data marts to be easily updated when new data is released, either from the number of genotypes side (new projects or existing projects update) or either from the variants description point of view (dbSNP updates, which occur approximately once a year).

The most common population genetics statistical indices have been implemented and summarized in the ENGINES data mart, but other metrics of interest could be easily implemented with just the raw data pre-processing script requiring updates: equivalent to two computing days due to the flexibility of the pipeline developed. In fact, and although it took ENGINES 1 month to be adapted to the new 1000 Genomes Phase I interim analysis data release policy, updating the data mart with the whole project's final data would take only 1 week even considering that the number of genomes is expected to be multiplied by 5.

Utility and discussion

Since several alternative means are available for researchers to access 1000 Genomes SNP data it is important to outline the advantages offered by the ENGINES browser in comparison to other approaches, which we see as complementary in their output, rather than competing to provide the same type of data. ENGINES is primarily designed to serve population genetics studies and therefore has several key features built in:

1. A straightforward system to download the individual genotypes for the SNPs, genes and populations queried. This permits direct input into population analysis algorithms such as *Structure* [6] or *Arlequin* [7].
2. Each database, population and SNV can be visually compared side by side, and the relevant data for SNVs and populations can be downloaded in one session from each database query.
3. F_{ST} values, amongst other metrics, can be collated for the entire genome-wide or exome SNV catalogue.
4. Lists of SNPs or genes are easily handled offering a more rapid and straightforward system than the SNP by SNP queries of the 1000 Genomes browser.
5. Genotyping coverage can be assessed at a glance by reviewing which SNPs and databases show incomplete genotyping.
6. Different filters are available that allow the selective listing of sets of variants according to different thresholds defined by the user (e.g. F_{ST} , MAF, etc).

ENGINES processed more than 7.3 billion genotypes and ~28 million unique variants in the Phase I interim analysis of the 1000 Genomes project (Table 1), of which 11.9 million were not previously described in dbSNP 132 (Figure 1). To illustrate the ease with which the ENGINES browser can add extra data to existing genome-wide analyses, of relevance for population genetics studies, we collated the total variant number by population group (Table 1). As expected from the demographic history of human populations, ENGINES clearly indicates the two sub-Saharan samples (LWK and YRI) contain more variants than any other population or set of populations, followed by the African-American sample (ASW). The data in this population break-down is different to the one provided by the 1000 Genomes analysis [8] because the latter targeted low coverage analysis of only the CEU, YRI, CHB, and JPT (Pilot 1) or exon regions (Pilot 3). Our data reveals interesting differences of SNP density that could contribute to the study of global patterns of natural selection (Table 1).

F_{ST} is a metric of genetic differentiation [9] between populations. It is also well known that the action of natural selection can locally cause systematic deviation in F_{ST} values for a selected gene and nearby markers. Thus, when compared with the action of a neutral evolving gene, high F_{ST} values might signal the action of local directional selection, while a decrease of F_{ST} values would be suggestive of balancing selection. Analysis of F_{ST} values on a genome-wide scale has already been demonstrated to be very useful for mapping genes

under selection [10]. The 1000 Genomes pilot project has allowed the calculation of F_{ST} values for the first time in the framework of a whole genome sequencing project [8], and has already revealed preliminary features relating to new regions that could have been subject to natural selection. In a step forward, ENGINES provides F_{ST} values for different population or continental combinations selected by the user and centred on the most current data release of 1000 Genomes. Access to this information is straightforward, and genotypes can be easily downloaded *ad hoc* for the regions of interest in order to carry out further analyses. By way of example, additional file 1 provides a snapshot of genome-wide F_{ST} values when considering a four-way inter-continental comparison (Africa, Europe, Asia, and America). Additional file 2 records the top F_{ST} values (>0.9) plotted in Figure S1, indicating that a large proportion of these values fall within known genes but notably a significant proportion are also located in uncharacterized genomic regions; therefore, providing new targets of considerable interest for further evolutionary and population genetic research. In addition, analysis of populations to a more extended intra-continental scale allows a refinement in the ability to search at greater population depth signals of localized adaptation.

Finally, an indirect assessment of the quality of ENGINES can be undertaken by the user by comparing SNP frequencies in Phase I with those of HapMap for the overlapping SNPs and populations (CEU, CHB, JPT, and YRI). Minor differences or discrepancies are possible but can be attributed to missing data or potential genotyping errors (due e.g. to Phase I SNV detection based on ultra-sequencing at low coverage). We have indeed observed genotyping discrepancies between genotypes reported in HapMap and those reported in Phase I for the same samples (data not shown).

Conclusions

ENGINES is capable of accessing large variation data repositories in a fast and comprehensive manner. We have shown that 1000 Genomes variant data, which represents the largest current whole human genome variation repository, is easily summarized and queried by ENGINES with a straightforward yet thorough approach for handling multiple sites across multiple genomes. ENGINES allows fast and easy browsing of whole genome variation by using a simple and intuitive web interface that performs queries in seconds and displays results in an efficient manner, while providing statistical information of each variation site such as frequency, heterozygosity or genetic differentiation among populations that are already pre-calculated and presented on demand.

Availability

The data mart generating scripts are a set of Perl files that are freely available on the software section of ENGINES. Access to these scripts and to the main web interface is granted from http://spsmart.cesga.es/engines.php

Additional material

Additional file 1: Figure S1 - Genome-wide F_{ST} values. Chromosome position in Mb is given in the X-axis, and F_{ST} values are plotted on the Y-axis. F_{ST} values are shown in black or red (red shows values that are exceptionally high: corresponding to the upper 2.5% of the empirical distribution of F_{ST} values). The yellow line shows the average of F_{ST} values for non-overlapping genomic windows of 1 Mb. Gaps correspond to heterochromatic staining regions near centromeres.

Additional file 2: Table S1 - Top F_{ST} values. List of SNVs showing the top F_{ST} values (above 0.9) for the four main continental group and their pairwise combinations (AFR = Africa; EAS = East Asia; EUR = Europe, and AME = America). Genes and rs-numbers are provided when available.

Acknowledgements and funding

This work was supported by grants from Ministerio de Ciencia e Innovación (SAF2008-02971), and Fundación de Investigación Médica Mutua Madrileña (2008/CL444) given to AS, and from Xunta de Galicia PGIDJT06PXIB228195PR given to CP. We would like to acknowledge CESGA (Supercomputing Centre of Galicia, Santiago de Compostela, Spain) for its supercomputing availability, web hosting and support. We would also like to thank Paul Flicek and Laura Clarke of the EBI (European Bioinformatics Institute, Hinxton, United Kingdom) for their extensive help to enable a full understanding of the 1000 Genomes data.

Author details

[1]Grupo de Medicina Xenómica, CIBERER, Universidade de Santiago de Compostela, Santiago de Compostela, Galicia, Spain. [2]Unidade de Xenética Forense, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela. Santiago de Compostela, Galicia, Spain.

Authors' contributions

JA carried out the design, programming and implementation of the software, and drafted the manuscript. AS and CP participated in the design of the software, and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 20 September 2010 Accepted: 19 April 2011
Published: 19 April 2011

References

1. Peacock E, Whiteley P: **Perlegen sciences, inc.** *Pharmacogenomics* 2005, **6(4)**:439-442.
2. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437(7063)**:1299-1320.
3. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science (New York, NY* 2008, **319(5866)**:1100-1104.
4. Amigo J, Phillips C, Salas A, Carracedo A: **Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes.** *BMC Bioinformatics* 2009, **10(Suppl 3)**:S5.
5. Amigo J, Salas A, Phillips C, Carracedo A: **SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access.** *BMC Bioinformatics* 2008, **9**:428.
6. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155(2)**:945-959.

7. Excoffier L, Laval G, Schneider S: **Arlequin ver. 3.0: An integrated software package for population genetics data analysis.** *Evolutionary Bioinformatics Online* 2005, **1**:47-50.
8. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467(7319)**:1061-1073.
9. Lewontin RC, Krakauer J: **Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms.** *Genetics* 1973, **74(1)**:175-195.
10. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res* 2002, **12(12)**:1805-1814.

doi:10.1186/1471-2105-12-105
Cite this article as: Amigo *et al.*: ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics* 2011 **12**:105.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Call for participation in the neurogenetics consortium within the Human Variome Project.

Haworth A, Bertram L, Carrera P, Elson JL, Braastad CD, Cox DW, Cruts M, den Dunnen JT, Farrer MJ, Fink JK, Hamed SA, Houlden H, Johnson DR, Nuytemans K, Palau F, Rayan DL, Robinson PN, Salas A, Schüle B, Sweeney MG, Woods MO, Amigo J, Cotton RG, Sobrido MJ

Neurogenetics. 06/2011; 12(3):169-73.

La tasa de descubrimiento de variantes de ADN ha acelerado la necesidad de recopilar, almacenar e interpretar los datos de manera coherente y estandarizada, y se está convirtiendo en un paso crítico para maximizar el impacto de los descubrimientos en la comprensión y tratamiento de las enfermedades humanas. Esto se aplica particularmente al campo de la neurología ya que la función neurológica se altera en muchos trastornos humanos. Además, el campo de neurogenética ha probado que muestra relaciones genotipo-a-fenotipo notablemente complejas. Para facilitar el acopio de variación de secuencias de ADN correspondientes a los trastornos neurogenéticos, hemos iniciado el "Consorcio de Neurogenética" bajo el amparo del Proyecto Varioma Humano. El grupo fundador del Consorcio consistió en investigadores básicos, clínicos, informáticos y los creadores de bases de datos. Este informe resume los objetivos estratégicos establecidos en las reuniones preliminares del Consorcio de Neurogenética y pide la participación de la comunidad neurogenética en general para facilitar el desarrollo de este importante recurso

NEUROGENETICS NEWS

Call for participation in the neurogenetics consortium within the Human Variome Project

Andrea Haworth · Lars Bertram · Paola Carrera · Joanna L. Elson · Corey D. Braastad · Diane W. Cox · Marc Cruts · Johann T. den Dunnen · Matthew J. Farrer · John K. Fink · Sherifa A. Hamed · Henry Houlden · Dennis R. Johnson · Karen Nuytemans · Francesc Palau · Dipa L. Raja Rayan · Peter N. Robinson · Antonio Salas · Birgitt Schüle · Mary G. Sweeney · Michael O. Woods · Jorge Amigo · Richard G. H. Cotton · Maria-Jesus Sobrido

Received: 20 March 2011 / Accepted: 10 May 2011 / Published online: 1 June 2011
© Springer-Verlag 2011

Abstract The rate of DNA variation discovery has accelerated the need to collate, store and interpret the data in a standardised coherent way and is becoming a critical step in maximising the impact of discovery on the understanding and treatment of human disease. This particularly applies to the field of neurology as neurological function is impaired in many human disorders. Furthermore, the field of neurogenetics has been proven to show remarkably complex genotype-to-phenotype relationships. To facilitate the collection of DNA sequence variation pertaining to neurogenetic disorders, we have initiated the "Neurogenetics Consortium" under the umbrella of the Human Variome Project. The Consortium's founding group consisted of basic researchers, clinicians, informaticians and database creators. This report outlines the strategic aims

A. Haworth · M. G. Sweeney
Neurogenetics Unit, Department of Molecular Neurosciences,
National Hospital of Neurology and Neurosurgery,
Queen Square,
London, UK

L. Bertram
Neuropsychiatric Genetic Vertebrate Genomics,
Max-Planck Institute for Molecular Genetics,
Berlin, Germany

P. Carrera
San Raffaele Scientific Institute, Center for Translational Genomics and Bioinformatics, Unit of Genomics for Human Disease Diagnosis and Laboraf,
Milan, Italy

J. L. Elson
Mitochondrial Research Group, Institute for Ageing and Health,
Newcastle University,
Newcastle upon Tyne, UK

C. D. Braastad
Athena Diagnostics,
Worcester, MA, USA

D. W. Cox
Department of Medical Genetics, University of Alberta,
Edmonton, AB, Canada

M. Cruts · K. Nuytemans
Neurodegenerative Brain Diseases Group, Department of Molecular Genetics, VIB, University of Antwerp,
Antwerp, Belgium

J. T. den Dunnen
Human and Clinical Genetics, Leiden University Medical Center,
Leiden, the Netherlands

M. J. Farrer
Centre for Applied Neurogenetics, Brain Research Centre,
University of British Columbia,
Vancouver, BC, Canada

J. K. Fink
Department of Neurology and Geriatric Research Education and Clinical Center, Ann Arbor Veterans Affairs Medical Center,
University of Michigan,
Ann Arbor, MI, USA

S. A. Hamed
Department of Neurology and Psychiatry,
Assiut University Hospital,
Assiut, Egypt

H. Houlden · D. L. R. Rayan
MRC Centre for Neuromuscular Diseases,
UCL Institute of Neurology,
Queen Square,
London WC1N 3BG, UK

established at the preliminary meetings of the Neurogenetics Consortium and calls for the involvement of the wider neurogenetic community in enabling the development of this important resource.

Keywords Human Variome project · Neurogenetics consortium · Database · Genetic variation · Standardisation · Phenotype

Meeting report

The vision of the Human Variome Project (HVP) is to develop a global collaboration with the aim of building systems and strategies for the collection, storage, interpretation and sharing of human genetic variation and its implications for disease [1, 2]. To reach these objectives, the HVP is organised in working groups to produce consensus recommendations for areas such as variant nomenclature, clinical data collection, laboratory data collection, informatics data integration, ethics and other relevant issues. In addition to these problem-driven working groups, the HVP has two complementary approaches to ensure data collection: country-specific collections and disease-specific collections (see Roadmap at www.humanvariomeproject.org, adapted in Fig. 1) [3].

Neurological diseases are a particularly complex range of disorders. Neurological dysfunction is often insidious and progressive, with age-associated penetrance and with variable expressivity. Genetic and allelic heterogeneities are the norm, with many different disorders sharing phenotypic features in addition to biological mechanisms. This leads to a continuum of both phenotype and genotype in this group of diseases, hampering the construction of neurogenetic locus-specific databases (LSDBs) [4]. A neurologic diagnosis is best informed through longitudinal clinical observation, brain imaging and postmortem pathology, to which molecular genetics analysis of leucocyte DNA can make a major contribution. The latter may inform the development of novel targeted therapeutics and appropriate selection of patients for phase II clinical trials (for efficacy and to avoid adverse drug responses, based on an individual's molecular aetiology). A number of databases related to neurological diseases have already been developed, most of them thanks to individual efforts [5–7]. The routine use of these databases by both the research and clinic communities in their daily work reflects the fundamental need for these repositories [8]. However, databases maintained by resources of individual groups or centres are always at risk of no longer receiving the support they need. Furthermore, as the number of genetic variants associated with neurologic traits rapidly increases, the necessity for efficient organisation of the genotype-to-phenotype information becomes even more crucial. This is especially true if the opportunities of whole exome/genome sequencing are to be fully realised. Thus, with the objective of making a global and coordinated effort to develop this key resource for clinicians and researchers focused on neurogenetic disorders, a Neurogenetics Consortium (NGC)

D. R. Johnson
Evidence Based Healthcare Consulting, LCC,
Cheshire, CT, USA

F. Palau
Institute of Biomedicine of Valencia, CSIC and Center for Biomedical Network Research on Rare Diseases (CIBERER), Institute of Health Carlos III,
Valencia, Spain

P. N. Robinson
Institute for Medical Genetics and Human Genetics,
Charité Universitätsmedizin Berlin,
Berlin, Germany

A. Salas
Genetics Unit, Department of Pathology and Forensic Sciences, Institute of Legal Medicine, School of Medicine,
University of Santiago de Compostela,
Santiago de Compostela, Galicia, Spain

B. Schüle
The Parkinson's Institute and Clinical Center,
Sunnyvale, CA, USA

M. O. Woods
Discipline of Genetics, Memorial University of Newfoundland,
St. John's, NL, Canada

J. Amigo
Genomic Medicine Group, University of Santiago de Compostela, Center for Biomedical Network Research on Rare Diseases (CIBERER), Institute of Health Carlos III,
Santiago de Compostela, Spain

R. G. H. Cotton
Genomic Disorders Research Centre and Department of Medicine (St. Vincent's), University of Melbourne,
Melbourne, Victoria, Australia

M.-J. Sobrido (✉)
Fundación Pública Galega de Medicina Xenómica,
Travesía da Choupana s/n,
15706, Santiago de Compostela, Spain
e-mail: ssobrido@telefonica.net

M.-J. Sobrido
Center for Biomedical Network Research on Rare Diseases (CIBERER), Institute of Health Carlos III,
Santiago de Compostela, Spain

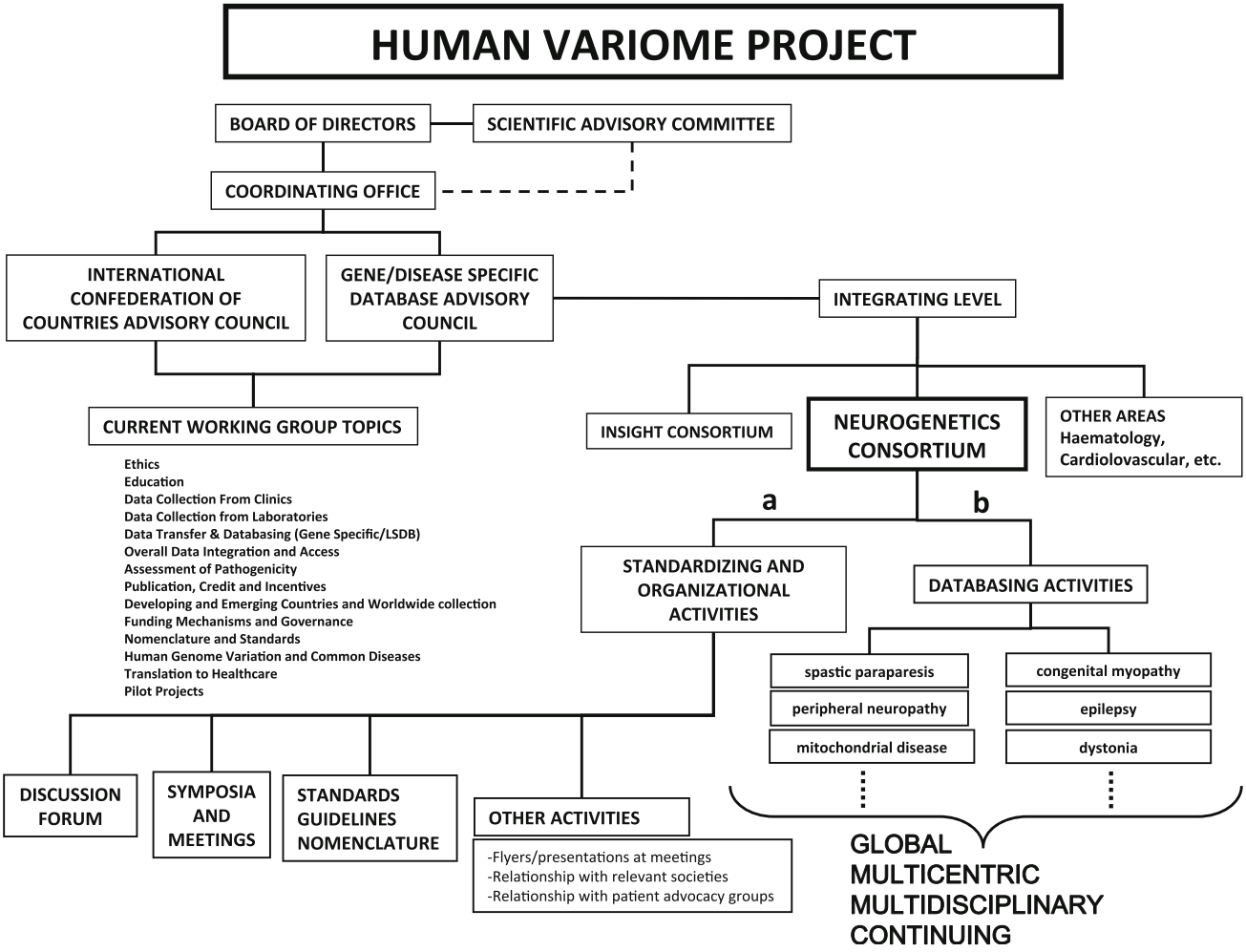


Fig. 1 Schematic representation of the potential organisation structure of the Neurogenetics Consortium and its relationship to other activities within the Human Variome Project. Two sets of actions will be needed: (*a*) those directed to meet standardisation and organisation goals and (*b*) activities dedicated to database development and curation by multidisciplinary, disease-centred, expert working groups

was initiated within the HVP. The aim of this consortium is to discuss the most important challenges and actions towards the construction of coordinated neurological LSDB resources.

The first meeting of the NGC was held in Honolulu, 19th October 2009, as a satellite of the American Society of Human Genetics meeting. It was a full-day meeting with over 70 attendees including expert speakers on neurogenetic disorders, researchers, informaticians and database curators. Among the main problems identified and discussed regarding neurogenetic LSDBs were:

1. The need for a coordinated and standardised system to collect and curate human variants and their associated clinical manifestations in LSDBs, this vision of the HVP being especially important for the field of neurogenetics.
2. Among the main reasons making the current neurological LSDBs inefficient are: (a) diverse structure and nomenclature of the existing databases; (b) lack of coverage for many genes/disorders; (c) duplication of efforts for some genes/disorders across independent databases, sometimes with overlapping or contradictory information; and (d) insufficient multidisciplinary knowledge in the curating teams. Much effort is lost when databases are constructed but there is no legacy plan developed to maintain the database after key individual(s) move on. Furthermore, "fossil" databases that are left behind uncurated could be subject to misuse by non-experts who may be unaware that the data are no longer current.
3. The need for standards and strategies for: (a) phenotype coding; (b) phenotype registration in the databases; (c) assessment of pathogenicity; (d) collection of specific types of variants relevant in neurogenetics (mitochondrial variations, repeat expansions, copy number variations); and (e) ethical and legal aspects, both general and specific to neurodegenerative disorders.

4. The need to capture and interpret the impact of genetic variants on complex disorders (e.g. as assessed by association studies), modifying genetic factors and epigenetics.

Other important issues regarding neurogenetic mutation databases brought up at the meeting were the lack of sufficient quality control of submissions, the frequently inadequate interpretation of mitochondrial variants, and the need for increased engagement from the clinical community and basic neuroscientists to achieve success in this endeavour.

The second meeting of the Neurogenetics consortium (UNESCO headquarters, Paris, 10th May 2010) was a follow-up of the inaugural meeting. The main points raised and discussed at the meeting, as well as proposed strategies and directions, were:

1. It was emphasized that it would be advantageous to have neurogenetic databases tailored towards specific diseases or syndromes (Fig. 1), allowing database queries for clinical terms as well as by gene. This would enable usage in both clinical and research settings, ultimately improving the health of patients with inherited neurological conditions. During the course of the meeting, several existing databases, both in-house and publicly available, were demonstrated, ranging from approaches based largely on collecting data from the literature to those that were used in clinical settings to collect longitudinal data. Commitments were undertaken to explore database construction/amalgamation for motor neuron diseases, Charcot–Marie–Tooth disease (CMT)/hereditary sensory neuropathies, mitochondrial diseases and disorders related to mutations in genes for neuromuscular ion channels. A working meeting for the hereditary spastic paraparesis/motor neuron disease mutation database occurred in September 2010.
2. Country-driven initiatives may be a very useful complement to international, disease-centred working groups. Country nodes should facilitate local collection of genetic variants under specific cultural, ethical and legal frameworks. In this respect, two pilot country-centred initiatives were described from Italy (for hereditary spastic paraplegias) and Spain (for CMT) with the ultimate aims of establishing an international network of experts to aid database curation and establish a global registry. One possible option for the envisioned NGC is to build networks of such country-wide disease-specific databases all linked at an operational level. This approach would help ensure maximum integration and minimum overlap.
3. To allow interoperability between such networks, efforts are being made to address uncontrolled vocabulary use in phenotype description, e.g. as in the Human Phenotype Ontology described at the meeting by Peter Robinson [9]. Standardisation of platforms was discussed in depth, as well as database structure and the integration of networks of databases. Demonstration of the Leiden Open Variation Database by Johan den Dunnen showed how this particular platform could be used to create either a disease or gene “front-end”. Another suggestion was to implement a “Wiki-like” format and environment.
4. The quality and interpretation of data was the focus of several talks. Accurate assignment of pathogenicity is of utmost importance, and even though there is no “gold standard”, it is reassuringly apparent that independent groups have developed similar systems/pipelines to analyse variants in both nuclear and mitochondrial genomes. Where the development of these databases will excel is the coherent, accurate inclusion of all relevant data related to genotype, phenotype, family history, healthy controls and functional studies to facilitate a more accurate interpretation for clinicians and their patients.
5. Patient access and input were discussed, e.g. the possibility that patients could enter their own longitudinal phenotypic data. The advantages and disadvantages of this approach were discussed at length. Concerns included the lack of standardised phenotype descriptors and the ethics involved in such a venture. Most participants agreed, however, that patient involvement would be positive and useful. It was also agreed that the consortia should participate in and be overseen by the HVP ethics working group.
6. The importance of involving a wide range of experts, such as population geneticists in database curation, was highlighted by two presentations on mitochondrial genetics, a field in which the lack of phylogenetic knowledge has caused a number of errors in the literature [10]. This is of particular importance as it is estimated that up to 1 in 200 people carry a pathological mtDNA mutation.

In summary, the field of neurogenetics is extensive and the confluence of many experts will be needed in systematic attempts to collect and annotate as much of the relevant data as possible in order to build high-quality databases of genetic variation and its significance. The issues related to information and data collection for different neurogenetic disorders should be worked out in a coordinated manner if the best possible level of integration is sought. Funding support will be a key issue to provide the necessary continuity and to guarantee the delivery of open-access, high-quality and up-to-date information. Involvement of the wider neurogenetics community as a whole, including patient advocates, will be crucial and is highly encouraged. To this end, we invite all members of the neurogenetics community worldwide to join this effort, which we believe will be for a widespread benefit

of patients suffering from neurological disorders. If you are interested in joining this effort, you can write to the corresponding author of this paper, María-Jesús Sobrido, and/or join as a Human Variome Project Consortium member free of charge on the Human Variome Project web site www.humanvariomeproject.org

Acknowledgements The authors are grateful to Rania Horaitis, Heather Howard and Lauren Martin for their assistance with the meeting organisation. Both meetings were supported by funds from the network REGENPSI from the Consellería de Educación, Xunta de Galicia (2009/019). Further support for the meeting came from the National Health and Medical Research Council, Australia. MJS received funding support from the Institute of Health Carlos III and from the European Funds for Regional Development (FEDER). The AlzGene database is supported by a grant from the Cure Alzheimer Fund (to L.B.). JKF is supported by the NIH (R01NS069700) and the Department of Veterans Affairs (Merit Review Award). FP is supported by the Spanish Ministry of Science and Innovation (SAF2009-07653), the European Comission, DG Research (7th Framework Programme, HEALTH-2009-2.4.4-1/242193), the Generalitat Valenciana (Prometeo 2009/051), Fundació Marató TV3 and the CIBERER, Instituto de Salud Carlos III. AH, HH, DLRR and MGS work was undertaken at UCLH/UCL who received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme. HH has grant support from MRC UK. DRR is supported by an MRC Clinical Research Training Fellowship (G1000347) and by The Consortium for Clinical Investigation of Neurologic Channelopathies (CINCH) funded by the National Institute of Health. MC and KN were in part supported by the Special Research Fund of the University of Antwerp, The Research Foundation–Flanders (FWO); the Foundation for Alzheimer Research (SAO-FRMA); and the Interuniversity Attraction Poles (IAP) programme P6/43 of the Belgian Federal Science Policy Office, Belgium. MJF is supported by a Canada Excellence Research Chair.

References

1. Cotton RG, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, Cantu JM, Cassiman JJ, Claustres M, Concannon P, Cotton RG, den Dunnen JT, Flicek P, Gibbs R, Hall J, Hasler J, Katz M, Kwok PY, Laradi S, Lindblom A, Maglott D, Marsh S, Masimirembwa CM, Minoshima S, de Ramirez HZ, Scriver CR, Sherry S, Shimizu N, Shimizu N, Stein L, Tadmouri GO, Taylor G, Watson M (2007) Recommendations of the 2006 Human Variome Project Meeting. Nat Genet 39:433–436
2. Kaput J, Cotton RG, Hardman L, Watson M, Al Aqeel AI, Al-Aama JY, Al-Mulla F, Alonso S, Aretz S, Auerbach AD, Bapat B, Bernstein IT, Bhak J, Bleoo SL, Blöcker H, Brenner SE, Burn J, Bustamante M, Calzone R, Cambon-Thomsen A, Cargill M, Carrera P, Cavedon L, Cho YS, Chung YJ, Claustres M, Cutting G, Dalgleish R, den Dunnen JT, Díaz C, Dobrowolski S, dos Santos MR, Ekong R, Flanagan SB, Flicek P, Furukawa Y, Genuardi M, Ghang H, Golubenko MV, Greenblatt MS, Hamosh A, Hancock JM, Hardison R, Harrison TM, Hoffmann R, Horaitis R, Howard HJ, Barash CI, Izagirre N, Jung J, Kojima T, Laradi S, Lee YS, Lee JY, Gil-da-Silva-Lopes VL, Macrae FA, Maglott D, Marafie MJ, Marsh SG, Matsubara Y, Messiaen LM, Möslein G, Netea MG, Norton ML, Oefner PJ, Oetting WS, O'Leary JC, de Ramirez AM, Paalman MH, Parboosingh J, Patrinos GP, Perozzi G, Phillips IR, Povey S, Prasad S, Qi M, Quin DJ, Ramesar RS, Richards CS, Savige J, Scheible DG, Scott RJ, Seminara D, Shephard EA, Sijmons RH, Smith TD, Sobrido MJ, Tanaka T, Tavtigian SV, Taylor GR, Teague J, Töpel T, Ullman-Cullere M, Utsunomiya J, van Kranen HJ, Vihinen M, Webb E, Weber TK, Yeager M, Yeom YI, Yim SH, Yoo HS, Contributors to the Human Variome Project Planning Meeting (2009) Planning the Human Variome Project: the Spain Report. Hum Mutat 30:496–510
3. Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Macrae FA, Marafie MJ, Paalman MH, Patrinos GP, Qi M, Ramesar RS, Scott RJ, Sijmons RH, Sobrido MJ, Vihinen M, members of the Human Variome Project (2010) Data Collection from Clinics, Data Collection from Laboratories and Publication, Credit and Incentives Working Groups. Capturing all disease causing mutations for clinical and research use: towards an effortless system for the Human Variome Project. Genet Med 11:843–849
4. Cotton RG, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting G, den Dunnen JT, Flicek P, Freimer N, Greenblatt MS, Howard HJ, Katz M, Macrae FA, Maglott D, Möslein G, Povey S, Ramesar RS, Richards CS, Seminara D, Smith TD, Sobrido MJ, Solbakk JH, Tanzi RE, Tavtigian SV, Taylor GR, Utsunomiya J, Watson M (2008) GENETICS. The Human Variome Project. Science 322:861–862
5. Nuytemans K, Theuns J, Cruts M, Van Broeckhoven C (2010) Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: a mutation update. Hum Mutat 31:763–780
6. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 39:17–23
7. Lill CM, McQueen MB, Roehr JT, Schjeide BMM, Zauft U, Bagade S, Zipp F, Bertram L. Schjeide BMM, McQueen MB, Bertram L (2010) The MSGene Database Systematic Meta-Analyses and Field Synopsis of Genetic Association Studies in Multiple Sclerosis Slide presentation at the 62nd Annual Meeting of the American Academy of Neurology, Toronto, Canada [S30.003]
8. Tuffery-Giraud S, Béroud C, Leturcq F, Yaou RB, Hamroun D, Michel-Calemard L, Moizard MP, Bernard R, Cossée M, Boisseau P, Blayau M, Creveaux I, Guiochon-Mantel A, de Martinville B, Philippe C, Monnier N, Bieth E, Khau Van Kien P, Desmet FO, Humbertclaude V, Kaplan JC, Chelly J, Claustres M (2009) Genotype–phenotype analysis in 2,405 patients with a dystrophinopathy using the UMD-DMD database: a model of nationwide knowledgebase. Hum Mutat 30:934–945
9. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S (2008) TheHuman Phenotype Ontology: a tool for annotating and analyzing humanhereditary disease. Am J Hum Genet 83:610–615
10. Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem Biophys Res Commun 335:891–899

The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome.

Catelli ML, Alvarez-Iglesias V, Gomez-Carballa A, Mosquera-Miguel A, Romanini C, Borosky A, Amigo J, Carracedo A, Vullo C, Salas A

BMC Genetics. 08/2011; 12:77.

Los antecedentes genéticos de los argentinos es un mosaico de diferentes ascendencias continentales. De tiempos coloniales hasta la actualidad, la contribución genética de los europeos y los africanos al sur del Sahara ha superpuesto o reemplazado el "estrato" genético indígena. Se recogió una muestra de 384 personas que representan a diferentes provincias argentinas y se genotipó para la primera y la segunda región de ADN mitocondrial (mtDNA) regiones hipervariables, y selectivamente se genotiparon SNPs de mtDNA. Estos datos fueron analizados en conjunto con otros 440 perfiles de las zonas rurales y poblaciones urbanas, más de 304 argentinos nativos americanos, todos ellos disponibles en la literatura. Se utilizó una base de datos mundial para las inferencias filogeográficas, comparaciones entre poblaciones, y el análisis de mezcla poblacional. Se secuenciaron muestras identificadas como pertenecientes a hg (hg) H2a5 para todo el genoma del mtDNA.

Los análisis filogenéticos y de mezcla poblacional indican que sólo la mitad del componente nativo americano en argentinos urbanos se puede atribuir al legado de los extintos argentinos ancestrales y que la contribución genética española es ligeramente superior a la italiana. Genomas enteros H2a5 vinculan estos mtDNAs argentinos al País Vasco y ha mejorado la filogenia de esta rama autóctona vasca. La huella de los esclavos africanos en mtDNAs de zonas urbanas de Argentina es baja y se puede atribuir filogeográficamente principalmente a África occidental. El componente europeo es mucho más frecuente en la provincia de Buenos Aires, la puerta principal de entrada para inmigración atlántica en Argentina, mientras que el componente nativo americano es más grande en el Norte y Sur de Argentina. AMOVA, Análisis de Componentes Principales y patrones *hgs*/haplotipo en Argentina revelaron un importante nivel de la sub-estructura genética en el país.

Los estudios destinados a comparar frecuencias de perfiles de ADN mitocondrial de diferentes regiones geográficas argentinas (es decir, los forenses y los estudios caso-control) deben tener en cuenta la heterogeneidad genética importante del país a fin de evitar falsas afirmaciones positivas de la asociación en estudios de enfermedades o la evaluación inadecuada de las pruebas forenses.

RESEARCH ARTICLE Open Access

The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome

María Laura Catelli[1†], Vanesa Álvarez-Iglesias[2†], Alberto Gómez-Carballa[2], Ana Mosquera-Miguel[2], Carola Romanini[1], Alicia Borosky[3], Jorge Amigo[2], Ángel Carracedo[2], Carlos Vullo[1,3] and Antonio Salas[2*†]

Abstract

Background: The genetic background of Argentineans is a mosaic of different continental ancestries. From colonial to present times, the genetic contribution of Europeans and sub-Saharan Africans has superposed to or replaced the indigenous genetic 'stratum'. A sample of 384 individuals representing different Argentinean provinces was collected and genotyped for the first and the second mitochondrial DNA (mtDNA) hypervariable regions, and selectively genotyped for mtDNA SNPs. This data was analyzed together with additional 440 profiles from rural and urban populations plus 304 from Native American Argentineans, all available from the literature. A worldwide database was used for phylogeographic inferences, inter-population comparisons, and admixture analysis. Samples identified as belonging to hg (hg) H2a5 were sequenced for the entire mtDNA genome.

Results: Phylogenetic and admixture analyses indicate that only half of the Native American component in urban Argentineans might be attributed to the legacy of extinct ancestral Argentineans and that the Spanish genetic contribution is slightly higher than the Italian one. Entire H2a5 genomes linked these Argentinean mtDNAs to the Basque Country and improved the phylogeny of this Basque autochthonous clade. The fingerprint of African slaves in urban Argentinean mtDNAs was low and it can be phylogeographically attributed predominantly to western African. The European component is significantly more prevalent in the Buenos Aires province, the main gate of entrance for Atlantic immigration to Argentina, while the Native American component is larger in North and South Argentina. AMOVA, Principal Component Analysis and hgs/haplotype patterns in Argentina revealed an important level of genetic sub-structure in the country.

Conclusions: Studies aimed to compare mtDNA frequency profiles from different Argentinean geographical regions (e.g., forensic and case-control studies) should take into account the important genetic heterogeneity of the country in order to prevent false positive claims of association in disease studies or inadequate evaluation of forensic evidence.

Background

The inhabitation of the Americas took place with the passage of people from northeast Asia to North America, who then rapidly moved southwards along the continent [1-4]. The first human settlements in Argentina were found in the Patagonia and dated to ~13,000 years ago (y.a.) [5]. The colonial period (roughly 1550-1810) began with the arrival of Spanish conquerors, and their domination lasted until the independence wars. During the colonial era, the Spaniards entered Argentina from Peru and Bolivia mainly through the northern 'Camino Real' and by the Río de la Plata, and they established a permanent colony on the site of what would later become Buenos Aires. Río de la Plata was also one of the main gates of entrance for other trans-Atlantic immigrants, such as African slaves. Indigenous people were under the domination of Spanish colonizers and many of these groups were exterminated or progressively admixed with the colonizers. Only natives

* Correspondence: antonio.salas@usc.es
† Contributed equally
[2]Unidade de Xenética, Instituto de Medicina Legal and Departamento de Anatomía Patolóxica e Ciencias Forenses, Calle San Francisco sn, Facultade de Medicina, Universidade de Santiago de Compostela, CIBERER, Santiago de Compostela, 15782, Galicia, Spain
Full list of author information is available at the end of the article

BioMed Central

© 2011 Catelli et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

inhabiting the mountainous north-western and southern Argentina survived the repression. At the end of the 19th century, the Native populations were exterminated in the central region and upper Patagonia. The Argentinean National Constitution of 1853 promoted immigration from Europe, and the country received large waves of European immigrants, predominantly Italians (e.g., from South Italy) and Spanish (e.g., Galicia in northwest Spain [6]). In about 100 years, the census of Argentina increased by one order of magnitude to about 20 million people in 1960. Internal demographic movements were also important in Argentina during the industrialization period (1930-1950). Thus, waves of Native Americans moved from northern Native Argentinean enclaves to the largest cities of the country. In the seventies, massive numbers of immigrants would also arrive to the main cities coming from bordering countries (Bolivia, Paraguay, Uruguay, Chile, and Peru) [7-9].

Argentina is a melting pot of people with different continental ancestries but a majority of the citizens are descendents of colonial-era settlers and of the late19th and early20th century European immigrants. The official census in Argentina, INDEC (Instituto Nacional de Estadística y Censos; http://www.indec.gov.ar/), indicates that the country is populated by more than 40 million people, of which only about 600,000 (~1.7%) considered themselves as belonging to or descending from indigenous groups. About 30 officially recognized indigenous populations survived the colonial and post-colonial period up to the present and nowadays there are more than 25 Native speaking live languages [10]. The most important ones in terms of population size are the Mapuches in the South, and the Collas (also spelled Kollas), Tobas, Wichí and Guaraní in the North.

It is difficult to determine the real impact of the different demographic changes occurred in Argentina during the last few centuries. From a genetic point of view, one could indirectly predict the impact of the different contributors by looking at the census; however, the census can be somehow misleading for several reasons. Thus, the INDEC indicates that the proportion of Italians arriving to Argentina in the 1980 and 1991 was ~47% and ~51% involving about 236,467 and 167,977 individuals, respectively; while the Spaniards were 41% and 38% involving about 202,523 and 124,667 individuals, respectively. However, historical sources [7-9] indicate that Spain contributed more significantly to the Argentinean pool in several periods of the last 150 years (Table 1). On the other hand, the 'masculinity index' (as the amount of male immigrants each 100 female immigrants [8]) was larger for Italians than for Spaniards [8], which would contribute e.g. to inflating the signal left by Spaniards on the mitochondrial DNA (mtDNA) of contemporary Argentineans.

The study of mtDNA data has been demonstrated to be very useful in unraveling the patterns of human worldwide migrations, in particular, those occurred in America [1-3,11-15]. Several studies have been devoted to the analysis of mtDNA in Argentinean populations. Ginther et al. [16] analyzed the first hypervariable region (HVS-I) in a sample of indigenous Mapuches (South); the study revealed the predominant Native American nature of this population. Cabana et al. [17] analyzed the HVS-I of individuals belonging to different ethnic groups from Gran Chaco (North), and focused on the historical events occurring in this northern Argentinean region. Álvarez-Iglesias et al. [18] showed a SNP-based methodological approach to allocate Native American mtDNAs into hgs [18]. A sample from Córdoba (Argentina) was also analyzed by Salas et al. [19]; a high proportion of the Native American component was observed in the mtDNA lineages (~41%) but not on the Y-chromosome (~2%). Martínez-Marignac et al. [20] analyzed a sample from the city of La Plata (Central Argentina); the results corroborated the hg distribution observed in previous studies. In a sample from Argentina, the results of Bobillo et al. [21] showed that Amerindian hgs were most frequent in North and South (60%) and decreased to less than 50% in Central. García and Demarchi [22] reported hg frequencies in nine villages from central Argentina, indicating that ~80% of the lineages belonged to native American hgs. In a congress report, Catelli et al. [23] presented broad hg frequencies of a subset of the sample used in the present study. Mitochondrial DNA sequences were also investigated in six Mbyá-Guaraní villages (northeastern) [24], being A2 and D1 the ones exhibiting the highest frequencies (~41% and ~36%, respectively). Most recently, Corach et al. [25] investigated the genetic admixture of unrelated male individuals from eight different provinces using different sets of markers; the results showed that different ancestry components were detectable in contemporary Argentineans, the amounts depending on the genetic system applied, exhibiting large inter-individual heterogeneity.

The present study has been motivated by the following reasons: (i) although several Argentinean populations have been analyzed to date, Argentina has not been analyzed from a global perspective (with the exception of the [25] study which however focus on a different sampling strategy, different methodology and aims), and several regions still remain uncharacterized, (ii) there is a need to explore the levels of population stratification within the Argentinean country since this could have important consequences in different biomedical studies, (iii) a comprehensive and comparative analysis of the mtDNA patterns observed in Native communities *versus* rural and urban population is still lacking, and

Table 1 Distribution of immigrants to Argentina coming from Italy, Spain and neighboring countries (modified from [8])

Inter-census period	Population (millions)	Immigrants (%)	Italians (%)	Spaniards (%)	M_{IT}/M_{SP}*[1]	Neighboring countries (%)
1869-1895	1.8-4.0	12.1	50.7	20.2	0.97	10.5
1895-1914	4.0-7.9	25.4	35.7	41.2	1.07	7.5
1914-1947	7.9-15.8	29.9	25	26.2	1.6	17.2
1947-1960	15.8-20.0	15.3	35.8	17.2	1.5	28.9
1960-1970	20.0	13	5.4	8.0	-	76.1

The percentages in the census indicate the approximate contribution regarding to the total immigrants. *[1] Lattes and Sautu [8] defines the 'masculinity index' (here denote as M) as the amount of man immigrants per 100 woman immigrants. Here, we define M_{IT}/M_{SP} as the quotient of the M value in Italians divided by the M value in Spaniards.

(iv) while Native American lineages in Argentina have been analyzed with certain resolution, the provenience of the trans-Atlantic immigration has been poorly inferred from control region sequences.

Methods

DNA samples

A total of 384 blood samples were collected from unrelated donors by the Equipo Argentino de Antropología Forense, and the Laboratorio de Inmunogenética y Diagnóstico Molecular de Córdoba representing different regions in Argentina (Figure 1). All the participants have permanent residence in Argentina. An undetermined proportion of them could descent from non-Argentinean parents or great-parents but this information was not recruited. One of the aims of the present study was to evaluate the proportion of Native American component that is autochthonous *versus* non-autochthonous in people that have permanent residence in the country. The analysis provides therefore a rough estimate of the amount of autochthonous lineages that are among present-day Argentineans. On the other

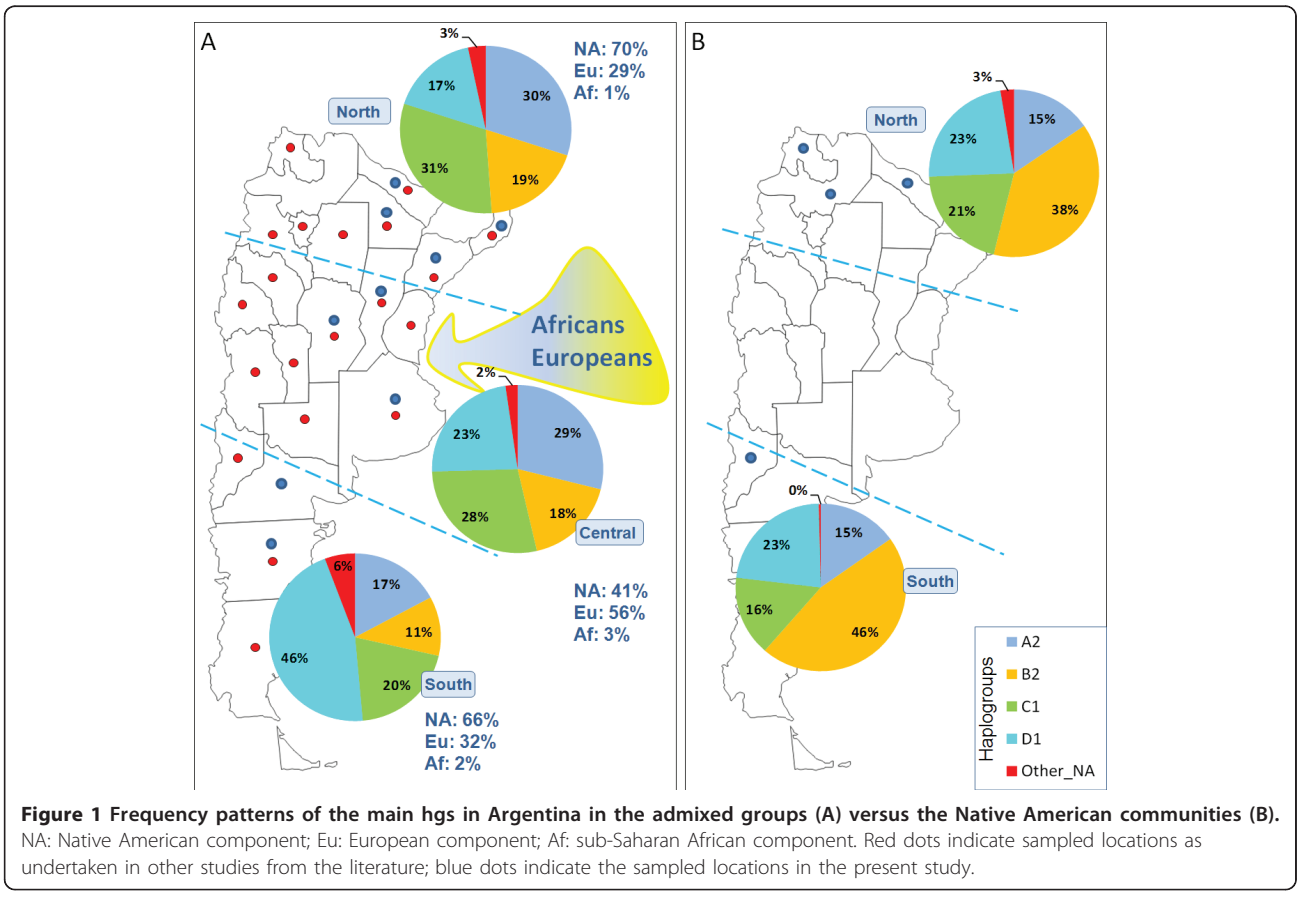


Figure 1 Frequency patterns of the main hgs in Argentina in the admixed groups (A) versus the Native American communities (B). NA: Native American component; Eu: European component; Af: sub-Saharan African component. Red dots indicate sampled locations as undertaken in other studies from the literature; blue dots indicate the sampled locations in the present study.

hand, since we have carried out a meta-analysis of Argentinean mtDNA profiles adding to our set of lineages those collected from the literature, uncertainty exists concerning the characterization of many donors (see discussion below).

The geographic origin and sizes of the samples analyzed in the present study are summarized in Additional file 1: Table S1. Broad hg frequencies of a subset of these samples have been summarized in a previous congress report [23].

DNA was extracted using phenol-chloroform standard procedures. Written informed consent was obtained in Argentina from all the participants. In addition, an Institutional Ethical approval to carry out this study was obtained from the Equipo Argentino de Antropología Forense (EAAF) and the University of Santiago de Compostela.

PCR, sequencing and minisequencing analysis

Samples were PCR amplified and sequenced for HVS-I and HVS-II regions as described previously [23]. In addition, all the profiles were contrasted with the phylogeny in order to detect potential artifacts e.g. [26]. In order to increase the phylogenetic resolution, most of the samples were genotyped for sets of diagnostic SNPs mainly located in the coding region (mtSNPs). For the samples belonging to R0 (European ancestry), a set of 71 mtSNPs were genotyped as described previously [27] whereas samples belonging to Native American hgs were additionally genotyped for 31 mtSNPs as described in [18]. The full set of results for the control region sequences and the mtSNPs are shown in Additional file 2: Table S2.

Population database

A database of mtDNA profiles of rural and urban populations (referred to in this article as the admixed group/population) and indigenous Argentineans has been compiled from the literature. Together with the samples analyzed here, the Argentinean database contains 824 mtDNAs representing 24 different populations. The Native American groups were collected from (a) North Argentina (n = 265), and includes Coyas (n = 61) from the provinces of Jujuy and Salta [18], Pilagá (n = 38) and Toba (n = 24) from Gran Chaco (Formosa), Toba (n = 43) from Chaco (Formosa), and Wichí or Mataco (n = 99) from Gran Chaco [17], and (b) South Argentina, represented by 39 Mapuches [16]. The admixed populations were collected from: (a) North Argentina (n = 98), including Formosa (n = 19), Chaco (n = 5), Misiones (n = 48) and Corrientes (n = 26) [21]; (b) Central Argentina (n = 295) from Santa Fe (n = 6) and Buenos Aires (n = 187) [21] and Córdoba (n = 102) [19]; and (c) South Argentina (n = 47) from Río Negro (n = 46) and Chubut (n = 1) [21].

In addition, data from ancient DNA studies [28] or other studies aimed to target specific mtDNA lineages (such as [1]) were also used for database searching.

A database of European (Italian and Spanish) and other Argentinean neighboring populations (including Uruguay, Paraguay, Bolivia and Chile) were additionally used for the admixture analysis. Details on the samples used in this study are provided in Additional file 1: Table S1.

Admixture analysis

Here, we are interested in separately analyzing the origin of the European and the Native American component of urban Argentineans. It was known from the Argentinean census that Spain and Italy were the two main countries in supplying European immigrants to Argentina. In modern times, Argentina has been also the destination of thousands of immigrants coming from neighboring countries that have a predominant Native American component. A premise of admixture analysis is that the source populations considered in the model are genetically different. Figure 2 indicates this feature by way of exploring the number of sharing haplotypes between the population groups involved in the admixture analysis. Differences between Italy and Spain are small and cannot be detected when looking at statistical tests of population differentiation (yielding non-significant statistical differences; data not shown) or examining genetic distances (F_{ST} = 0.0022); an issue that could be improved in the future if adding more molecular information to the statistical model (e.g. entire genomes and larger sample sizes). Although F_{ST} is not informative at indicating differences between Spain and Italy, and given the fact that one of the admixed analysis carried out in the present study (see below) relies on haplotype sharing, we have carried out a simulation analysis in order to test if the two populations are sufficiently different in terms of haplotype sharing in order to support the results yielded by the admixture analysis. We performed a simulation that consists of (i) randomly distributing in 10,000 iterations the total number of individuals (from Spain and Italy jointly considered) in two groups (with samples sizes as in the original samples), (ii) compute the proportion of shared haplotypes each time, and (iii) reconstruct the distribution of this statistics under the null hypothesis of no differentiation. Clearly, the observed haplotype sharing is significantly smaller than 5^{th} percentile of this distribution (see Additional file 3: Figure S1). This allowed to conclude that haplotype sharing contains enough information to discriminate Spain and Italy and therefore to compute admixture proportions of Argentineans from Europe.

The first admixture model was undertaken as described by Salas et al. [29]; see also [30]. Since this

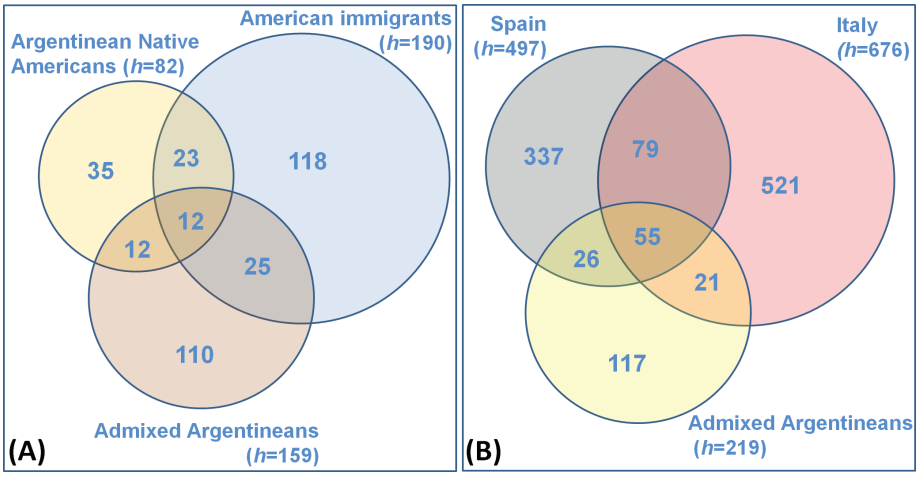


Figure 2 The share of identical haplotypes (H) between: (i) the Native American component of the admixed Argentinean populations (salmon-pink) versus the Argentinean Native Americans (yellow) and American immigrants (light blue) (Paraguay, Bolivia, Uruguay and Chile) (Figure 2A); and (ii) the European component of the admixed Argentinean populations (yellow) and a database of the Spanish (gray) and Italians (pink) (Figure 2B).

model is based on hg frequencies, it was only applied to infer the contribution of the European countries to the population of Argentina. This is because the Native American component was too homogeneous and the phylogenetic hg resolution was too low (at the control region level) to yield meaningful results.

The second admixed model was applied as described previously [14], but with an extension of the original model that is detailed below. The probability of origin of each of the sub-continental region can be computed as $po_s = \frac{1}{n}\sum_{i=1}^{n} k_i \frac{p_{is}}{p_{iC}}$, where n is the number of Argentinean sequences with matches (≥ 1) in the whole database; k_i, the number of times the sequence i is found in Argentina; p_{is}, the frequency of the sequence i in each regional datasets (e.g., Spain and Italy); and p_{ic}, the frequency of the sequence i in the whole database. The same analysis was carried out independently considering n to be the number of Argentinean sequences that have cero, one or two mutational differences from the sequences contained in the database. We will refer to P_0, P_1, and P_2, for the admixture components of sequences that match perfectly, differ by one mutational step, or two, respectively. In order to account for different sample sizes in the source populations, admixed components (and their 95% C.I.) were built by way of bootstrapping, taken 1000 re-samples of the source populations of size 300 each (other sample sizes yielded consistent results; data not shown).

Statistical analysis

DnaSP v.5 software [31] was used for the computation of haplotype (H) and nucleotide (π) diversities, and mean number of pairwise differences (M). AMOVA (Analysis of Molecular Variance) and the significance of the covariance components associated with different levels of genetic structure were tested on haplotypes and haplogroup frequencies applying a non-parametric permutation procedure. The latter analyses and population pairwise F_{ST} values, between/within population average nucleotide pairwise differences, and Nei's inter-population distances, were computed using Arlequin 3.5.1.2 [32]. Diversity indices, phylogeographic inferences and inter-population comparisons were carried out using the sequence range 16090 to 16365, since this is the common segment reported in the literature. Problematic variation located around 16189 usually associated to length heteroplasmy, e.g., 16182C or 16183C, was ignored. Principal Component Analysis (PCA) was carried out on population hg frequencies using R http://www.r-project.org/. AMOVA and PCA were performed on Argentinean samples of sample sizes ≥ 20 (see Additional file 1: Table S1).

Fisher's exact test and Pearson's chi-square test were undertaken using the R package http://www.r-project.org/, a significant value of $\alpha = 0.05$ was considered.

Finally, estimation of the time to the most recent common ancestor (TMRCA) and SDs of hg H2a5 were carried out according to Saillard et al. [33]and using an evolutionary rate estimate for the entire mtDNA molecule as reported by Soares et al. [34].

Results

Summary statistics in Argentinean mtDNAs

Summary statistics were computed for admixed Argentineans, Native Americans, and the whole Argentinean

Table 2 Diversity indices in the Argentinean population groups

	N [1]	*H* [1]	*H/n* [1]	*D* [1]	*Π* [1]	*M* [1]	*N* [2]	*H* [2]	*H/n* [2]	*D* [2]	*Π* [2]	*M* [2]	*N* [3]	*H* [3]	*H/n* [3]	*D* [3]	*Π* [3]	*M* [3]
Urban populations																		
North	37	28	0.76	0.980 (0.012)	0.0155 (0.0015)	4.9	90	46	0.51	0.958 (0.010)	0.0199 (0.0008)	6.3	129	76	0.59	0.978 (0.005)	0.0208 (0.0007)	6.6
Central	358	195	0.54	0.978 (0.004)	0.0131 (0.0005)	4.1	263	118	0.45	0.965 (0.005)	0.0196 (0.0004)	6.2	642	329	0.51	0.987 (0.002)	0.0187 (0.0004)	5.9
South	17	14	0.82	0.971 (0.001)	0.0134 (0.0024)	4.2	35	27	0.77	0.971 (0.018)	0.0187 (0.0013)	5.9	53	42	0.79	0.985 (0.009)	0.0192 (0.0010)	6.1
All	412	216	0.52	0.978 (0.004)	0.0133 (0.0004)	4.2	388	158	0.41	0.967 (0.004)	0.0197 (0.0003)	6.2	824	392	0.48	0.987 (0.001)	0.0192 (0.0003)	6.0
Native Americans																		
North	-	-	-	-	-	-	265	72	0.27	0.940 (0.008)	0.0181 (0.0005)	5.7	265	72	0.27	0.940 (0.008)	0.0181 (0.0005)	5.7
South	-	-	-	-	-	-	39	13	0.33	0.908 (0.020)	0.0171 (0.0006)	5.4	39	13	0.33	0.908 (0.020)	0.0171 (0.0006)	5.4
All	-	-	-	-	-	-	304	82	0.27	0.950 (0.006)	0.0181 (0.0004)	5.7	304	82	0.27	0.950 (0.006)	0.0181 (0.0004)	5.7
All Argentineans																		
All	-	-	-	-	-	-	1128	449	0.40	0.984 (0.001)	0.0192 (0.0003)	*id*	1128	449	0.40	0.984 (0.001)	0.0192 (0.0003)	6.0

Standard deviations are given in round brackets.

n = sample size, *H* = number of different haplotypes, *D* = sequence diversity, π = nucleotide diversity, *M* = average number of pairwise differences. Numbers in square brackets (above) indicate: [1]: haplotypes of European ancestry,[2]: haplotypes of Native American ancestry, [3]: all the haplotypes together

sample (Table 2). The analysis was also carried out separately for the the Native American and the European components (Table 2). The Native American component of the admixed populations has higher diversity values than the one of the indigenous groups (Table 2) for all the indices computed. Within the admixed groups, there is more sequence diversity in Central and South while nucleotide diversity is higher in Argentina.

The diversity of European lineages in the admixed group is higher in the North (Table 2). As expected, the European component is more diverse than the Native American one (Table 2) for the haplotype diversity, corresponding with their demographic histories, which is about four times older for the Europeans than for the Native Americans with the latter suffering strong bottlenecks at the time of entrance through the Bering Strait [1-4]; nucleotide diversity shows the opposite pattern which in this case most likely mirrors the low resolution of the HVS-I in a high proportion of European lineages (e.g. macro-hgs R0). Finally, admixed groups are genetically more diverse than the Native American ones (Table 2).

Phylogeography of mtDNA lineages in Argentina

The Native American component observed in the urban populations was 66%, 41%, and 70% in South, Central, and North, respectively (Figure 1) and it was virtually 100% in most Native American groups. The distribution of Native American hgs was substantially different in the main Argentinean regions especially when looking at urban populations (Figure 1A); for instance, hg A2 constitutes 30% in North admixed populations but only 17% in South admixed populations (Figure 1A) (Pearson's χ^2 test; un-adjusted *P*-value = 0.00561). Moreover, the percentages of the different Native American hgs significantly differ when comparing admixed with native populations (Figure 1A vs Figure 1B), even when comparing samples from the same geographical location; thus, for example, when considering only the Native American component of the urban populations, hg B2 is 19% in North admixed populations *versus* 38% in North Natives (Pearson's χ^2 test; un-adjusted *P*-value = 0.01808), or hgs B2 and D1 have frequencies of 11% *versus* 46% (hg B2; Pearson's χ^2 test; un-adjusted *P*-value < 0.0000) and 46% *versus* 23% (hg D1; Pearson's χ^2 test; un-adjusted *P*-value = 0.00334) in South admixed populations *versus* South Natives.

The lower prevalence of Native American hgs observed in Central Argentina coincides with the high proportion of European lineages in this region, mirroring the fact that this was the main European settlement area in the country; e.g. the European component is significantly more predominant in Central (56%) than in North (29%; Pearson's χ^2 test; un-adjusted *P*-value < 0.00901).

African slaves were brought to Argentina by Europeans during the period of the Atlantic slave trade [30,35,36] and they entered the country following the main entrance provided by the Río de la Plata, but the

impact of this process in the mtDNA pool of Argentina was much lower than in other American regions [11,14,37,38]. Sub-Saharan lineages represent only 1-3% of the total mtDNA component observed in Argentina.. The most prevalent sub-Saharan HVS-I mtDNAs in Argentina are: (i) the L2c2 profile C16223T C16264T C16278T T16311C, which also appears in Brazil [38,39] and other American locations [40]; exact matches of this mtDNA profile were found in Gabon [41], Cabinda [42], Mozambique [43] and some other South African locations; and (ii) the L3f1a mtDNA G16129A T16209C C16223T C16292T C16295T T16311C that also appeared in Brazil [39,44] and in US 'African Americans' [40,45]; this hg has a likely origin in East Africa [43] but probably arrived in America via West-Central Africa [41] or Southwest Africa [42]; see also [36]. Other typical North African profiles belonging to hg U6 reached Argentina via Portugal or Madeira (such as T16172C C16174T C16188T A16219G T16311C; hg U6b) [46,47], Canary islands (G16129A C16169T T16172C T16189C [48]) or directly from Morocco (T16172C A16183C T16189C A16219G C16239T C16278T T16362C [49]). Only two Argentinean mtDNAs belong to hg M1, the hg that is prevalent in the Middle East and East Africa and with a wide distribution in several African regions. For instance, matches for G16129A T16189C C16223T T16249C T16311C T16359C were observed in the Chad Basin [50], Ethiopia [51] and Egypt [52] while profile G16129A T16189C T16249C T16311C is present only in the Arabs in Chad [50] and outside Africa in Spain [53,54].

Finally, it is also interesting to note that haplotype frequencies vary substantially between populations (Additional file 4: Figure S2). For instance, Native American groups have several haplotypes at high frequencies (probably due to historical bottlenecks).

Admixture analysis and the mtDNA indigenous legacy in present-day Argentina

Admixture analysis, as carried out here, considers two potential source populations: (i) the Native American component of the available indigenous Argentinean populations, and (ii) the Native American component of neighboring countries as a proxy for the Native American component that has been introduced into Argentina through recent immigration. The model of admixture (Table 3) indicates that about half of the Native American component in the urban populations most likely comes from immigration arriving from neighboring countries, while the rest most likely corresponds with the indigenous inhabitants living in those regions before European colonization or arriving from rural Argentinean Native American enclaves. The data is roughly consistent when executing admixture analysis either looking at full HVS-I matches (P_0) or considering one or two mutational steps (P_1 *and* P_2).

Characterizing the most likely origin of the European component in present-day Argentina

The models employed here considers only the two main historical contributors to the European immigration in Argentina, namely Italy and Spain (representing > 80%

Table 3 Admixture proportions, P_0, P_1, P_2 (95% C.I. in brackets) of admixed Argentineans referring to their Native American (mainly hgs A2, B2, C1 and D1) and European components according to the main source populations

Urban Argentinean Populations (n = 800)	Argentinean Native Americans (n = 303; h = 82)	Argentinean Native Americans (n = 303; h = 82)	Argentinean Native Americans (n = 303; h = 82)
Native American Component (n = 388)	HS_0 = 25 (0.30)	HS_1 = 96 (1.17)	HS_2 = 141 (1.72)
Native American Component (n = 388)	P_0 = 0.50 (0.03)	P_1 = 0.46 (0.03)	P_2 = 0.42 (0.03)
	American immigrants (n = 488; h = 190)	**American immigrants (n = 488; h = 190)**	**American immigrants (n = 488; h = 190)**
Native American Component (n = 388)	HS_0 = 38 (0.20)	HS_1 = 108 (0.56)	HS_2 = 144 (0.76)
Native American Component (n = 388)	P_0 = 0.50 (0.03)	P_1 = 0.54 (0.03)	P_2 = 0.58 (0.03)
	Spain (n = 1467; h = 497)	**Spain (n = 1467; h = 497)**	**Spain (n = 1467; h = 497)**
European component (n = 412)	HS_0 = 81 (0.16)	HS_1 = 159 (0.32)	HS_2 = 174 (0.35)
European component (n = 412)	P_0 = 0.55 (0.5499-0.5564)	P_1 = 0.49 (0.4896-0.4961)	P_2 = 0.50 (0.5009-0.5074)
	Italy (n = 1667; h = 676)	**Italy (n = 1667; h = 676)**	**Italy (n = 1667; h = 676)**
European component (n = 412)	HS_0 = 76 (0.11)	HS_1 = 177 (0.26)	HS_2 = 185 (0.27)
European component (n = 412)	P_0 = 0.45 (0.4436-0.4501)	P_1 = 0.50 (0.5039-0.5104)	P_2 = 0.49 (0.4926-0.4991)

Shared haplotypes between groups are also indicated; HS_0, HS_1 and HS_2 refer to the number of shared haplotypes differing 0, 1 or 2 variants between the two main components of the urban mtDNAs (Native American and European) and the number of haplotypes (h) in the corresponding source populations. Note that these values can be > 1 because the same haplotype in the source population can count more than once for HS_1 and HS_2. The amount n indicates sample size. P_0, P_1, and P_2, are the admixture components referring to sequences that match perfectly, differ by one mutational step, or two, respectively, in the database; standard deviations are in round brackets.

of immigrants coming from Europe in the last 150 years; Table 1).

The mathematical admixed model based on hg frequencies indicates that Italy most likely contributed 33% (95% SD: 9.2) of the European mtDNA hgs to the Argentinean genome versus 67% (95% SD: 8.1) from Spain. The admixed model based on haplotype sharing yielded slightly different but quite consistent results, roughly indicating that Spain and Italy contributed almost similarly to the European component in Argentina (Table 3), although there are slight differences when considering perfect haplotype matches (P_0; indicating ~55% contribution from Spain) *versus* considering one or two mutational step differences between HVS-I profiles (P_1 and P_2, indicating about equal contribution from Spain and Italy). The haplotype shared between Argentina and Europe seems to favor the hypothesis that the Spanish legacy in Argentina is slightly larger than the one from Italy (Table 3), either when looking at perfect haplotype matches (HS_0) or one (HS_1) or two (HS_2) mutational step differences.

AMOVA analysis of Argentinean populations

When applying AMOVA on haplotypes, variance within populations accounts for ~84% of the total variance (Table 4) Grouping populations by geographic region or by Native American *versus* Admixed populations add little to the proportion of variance among groups (~1; Table 4); probably indicating that the HVS-I alone does not provide enough molecular information for the computation of F_{ST} based on molecular distances (pairwise differences). However, when applying AMOVA on haplogroup frequencies, among groups variance, by geography or by admixed vs Native groups, increases substantially to ~4 and ~6%, respectively. The figures are however not very high given that about half of the component of the admixed populations is Native American.

Table 4 AMOVA of Argentinean populations

	Within Populations	Among Groups	Among Pop./ Among Pop. within groups
Pairwise differences			
All populations	83.72	-	16.28
North/Central/ South	83.43	1.04	15.53
Native/Admixed	83.67	0.10	16.22
Haplogroup Freq.			
All populations	75.48	-	24.52
North/Central/ South	74.63	3.39	21.98
Native/Admixed	73.18	6.09	20.73

Values indicate the distribution of the variance components according to the different hierarchical population levels; all of them are statistically significant (Significant tests: 20,022 permutations; adjusted *P*-value < 0.0000)

Additional file 5: Figure S3A displays population pairwise F_{ST} values, indicating that the highest figures occur in comparisons involving Native American populations. Nei's genetic distances are in good agreement with pairwise F_{ST} matrix values (Additional file 5: Figure S3B). Population structure is also reveled when observing that values of the average number of nucleotide differences between are higher than those for within population comparisons (Additional file 5: Figure S3B).

Principal component analysis of Argentinean populations

PCA was carried out on hgs frequencies for Argentinean samples with sizes > 20 (Figure 3). PC1 accounts for 74% of the variation; it clearly separates Mapuches and Coyas to one side of the plot, from an amalgam of other population samples in the opposite side; Buenos Aires and Córdoba occupy an intermediate position. PC2 (13%) is clear at showing an important separation between the two admixed populations of Buenos Aires and Córdoba; the rest of the populations are located in between. The most important feature of PC3 (7%) is that it separates populations by geographic regions, with South being more distant from Central and North (Argentina). It is important to highlight that the merged groups of admixed and Native American populations are located very proximal in the plot (Figure 3) in agreement with AMOVA results.

Complete H2a5 genomes

Three entire genomes belonging to the recently described lineage, H2a5, have been completely sequenced. One of the entire Argentinean genomes belongs to the H2a5a1 branch (previously H2a5 [27]) defined by the transition T4592C (Figure 4). This clade has only been observed in the Basque country where it is supposed to be autochthonous [27]. The other two entire H2a5 genomes analyzed from Argentina are identical and belong to a new branch (defined by a synonymous transition at position T11233C), H2a5a2. The only known member belonging to this clade was observed by Achilli et al. [55]. The geographical location of its donor is unknown although his surnames (A. Achilli, personal communication) suggest a Galician origin (a region located in the westernmost corner of the Cantabrian region [6]); one of the main Spanish source populations to Argentina. The age of H2a5 is approximately 5.4 thousand years (kya) (95% C.I.: 0-12.9 kya) but the Basque autochthonous sub-clade H2a5a1 is much younger (~0.6 kys; 95% C.I.: 0.4-0.7 kya).

Finally, there is another entire genome sharing the same features as H2a5. It does not carry private mutations, lacks transition A1842G and was observed outside the Iberian Peninsula in the Czech Republic [56].

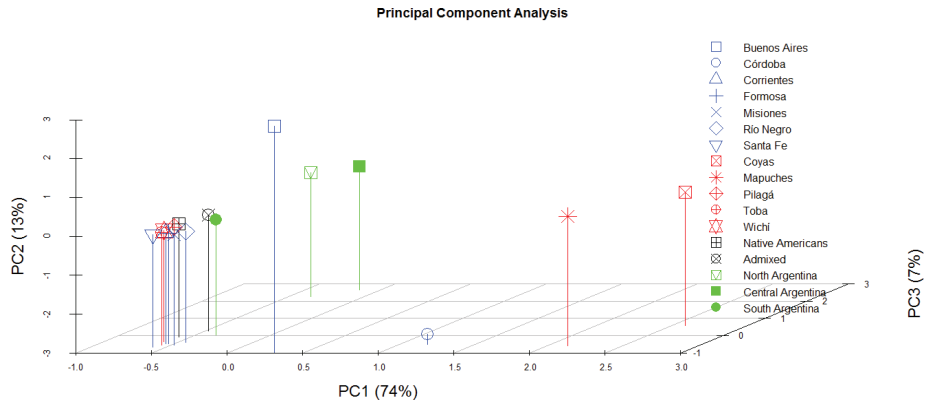


Figure 3 Principal Component Analysis of Argentinean populations. PC1, PC2 and PC3 stand for principal component one, two and three, respectively.

Discussion

Admixed Argentineans have an important Native American background. Admixture models indicated that about half of this Native American component could be non-autochthonous. The exact figures are only tentative given a main limitation of the present study, namely, we did not collect bio-geographic information for most of the donors of our samples, and this is information was not available for most of the data collected from the literature; therefore, some donors could be in reality Native American immigrants (or descents form parents) from neighboring countries. Given the results of admixture analysis, one could tentatively hypothesize an important demographic influence coming from neighboring countries that have a predominantly Native American background and where massive immigrations to Argentina have come from in recent times (such as Paraguay, Peru and Bolivia). There are several other pieces of evidence that would further support this hypothesis. Firstly, the Native American component in the urban admixed populations differs very significantly from the Native American component of the indigenous populations from North and South Argentina (Figure 1). A simple process of (recent) admixture of Europeans with indigenous peoples would tend to keep the same hg frequencies in admixed and indigenous people, which is not the case here. Secondly, several diversity indices are significantly higher in the Native American component of admixed Argentineans than in the indigenous groups (see above and Table 2). This could be easily explained if one assumes that the Native American component in the admixed populations has being

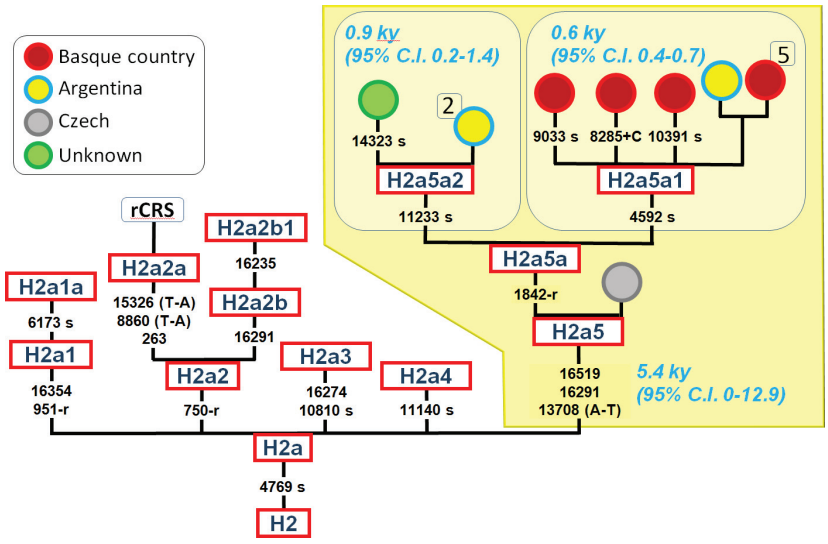


Figure 4 Phylogeny of hg H2a5 based on complete genome sequences.

continuously enriched with the arrival of a different Native American component coming from recent neighboring immigrants together with migrants arriving from rural Argentinean regions with large Native American components (from northeast and northwest Argentina). On the contrary the indigenous groups would tend to reduce its genetic diversity with time due to drift (smaller effective population size) and isolation from Europeans and other immigrants. The data therefore indicates that the Native American component observed in the urban groups only partially mirrors the populations that inhabited the regions in colonial times. An important proportion of the autochthonous Argentinean Native American component could have arrived to rural and urban cities in modern times. For instance, after the economical crisis suffered in the country in the 1930's, waves of people from rural areas with high Native American component moved to industrialized cities [7-9].

From the different analyses carried out, the contribution of Spain in the present Argentineans seems to be slightly higher than that of Italy, although the estimates vary significantly depending on the admixture model. The results agree quite well with the historical records (Table 1). Thus, until 1850 almost all immigrants came from Spain. From 1850 onwards, thousand of Spaniards and Italians left their countries with final destiny in Argentina; but Spaniards were generally more prevalent than Italians [7-9]. Moreover, 'Spanish ancestry' could have enriched the Argentinean European component through immigrants coming from neighboring countries, where Spaniards contributed significantly more than Italians (Uruguay, Chile, etc).

Some caveats should be added concerning admixture analysis. Computations are based on a meta-analysis by way of collecting samples that did not necessarily follow the same sampling criteria. Thus, for instance, samples from Argentina were collected in different forensic, anthropological or clinical laboratories, using different sampling criteria; a meta-analysis could contribute in the direction of balancing different sampling strategies or the opposite in case e.g. of some sample being much larger than others. Moreover, it is well-known from the census that some regions in some (European/American) countries contributed more than others to Argentina; but it is not possible to determine how the different regions should be represented in the source meta-populations; a reasonable solution seems to merge all the available data from each country without any *a priori* regarding sampling origin or institution involved (as done in the present study).

Admixture analysis as carried out in the present study only provides a view of the female historical and contemporary demography (as inferred from the mtDNA); there are however indications showing that the ancestral proportions inferred from other markers are different [19,25], indicating for instance a sex bias in the contribution coming from the different source populations (at least from Europe).

It is also interesting to note that the genetic diversity of European lineages in the admixed groups is higher in the North than in the other regions, independently to the fact that the proportion of European in higher in Central Argentina. This is consistent with the historical documentation indicating that the 'Camino Real' to Potosí (Bolivia) and Lima (Peru) was by far the most important trade route during colonial times. Thus, Río de la Plata was the main gate for European immigrants into Argentina in modern times, but contributing less mtDNA diversity than the northern 'Camino Real'.

The impact of the African slave trade on present day Argentineans seems minimal compared to other South American locations (e.g., Brazil and Colombia), and comes most likely from West-central Africa, but also from Angola and Mozambique (see [30]).

Conclusion

The issue of population stratification in Argentina has stimulated an intense debate concerning the use of autosomal markers in forensic casework and paternity tests. While some hold a position that stratification is an issue of little interest in forensic databases [57], others claim a more important role in both forensic and clinical genetics [19,58-61,19,21]. The present study certainly indicates the existence of a clear-cut sub-structure in the country; this is shown by the differences observed in hg distributions, AMOVA analysis, population differentiation tests, statistical hypothesis testing on hg frequencies, and PCA. Population stratification could have obvious implications in different biomedical applications in Argentina. This would not only be in forensic genetics (where inter-population haplotype differences can have important consequences for the weight of the evidence) but also in other population-based studies (e.g., case-control studies) dealing with the analysis of the potential role of mtDNA variants in common diseases, where false positives are unfortunately higher than desirable [62-64]. By extrapolation, and given the important ancestral components and regional differences observed in the mtDNA variation, stratification should also be a matter of interest when using autosomal SNPs. The forensic field should not ignore forensic stratification in their routine casework [60,65], especially if one takes into account that local databases do not exist in Argentina and that most of the forensic casework is carried out in the largest cities under the risk of using a single database for cases arriving from any province in the country.

Additional material

Additional file 1: Table S1. List of samples analyzed in the present study and collected from the literature (American neighboring countries and European ones) used for admixture analysis.

Additional file 2: Table S2. Haplotype and mtSNP profiles of the Argentinean samples analyzed in the present study.

Additional file 3: Figure S1. Simulation aimed to demonstrate that Italy and Spain are sufficiently different in terms of haplotype sharing, therefore, supporting the results of admixture analysis. First, two databases were considered jointly, the Spanish (n = 1467) and the Italian database (n = 1667) (see Additional file 1: Table S1, and text for more information on the databases). From this global database (n = 3134), two samples of sizes 1467 and 1667 each were taken at random without replacement 10,000 times. The distribution represents the number of identical shared haplotypes (horizontal axis) and their counts (vertical axis) between the 10,000 pairs of random samples. The red line indicates the observed number of haplotypes shared between the Italian and the Spanish database (n = 134; see also Figure 2).

Additional file 4: Figure S2. Patterns of haplotype frequencies in Argentinean population samples. Only those samples of sizes > 20 individuals were considered.

Additional file 5: Figure S3. Pairwise F_{ST} values (A), and average number of pairwise differences within and between populations and Nei's distances (B).

Acknowledgements
We would like to thank the donors for their participation in the present project. This project was supported by grants from "Fundación de Investigación Médica Mutua Madrileña" (2008/CL444) and "Ministerio de Ciencia e Innovación" (SAF2008-02971) given to AS. The project was also partially supported by "Argentinean Government, Agencia de Cooperación Española para el desarrollo and European Union". There are no conflicts of interest in this study. The complete genomes analyzed in the present study have been submitted to GenBank under accession numbers JF284816-JF284818.

Author details
[1]Equipo Argentino de Antropología Forense, Independencia 644 - 5C, Edif. EME1, Córdoba, Argentina. [2]Unidade de Xenética, Instituto de Medicina Legal and Departamento de Anatomía Patolóxica e Ciencias Forenses, Calle San Francisco sn, Facultade de Medicina, Universidade de Santiago de Compostela, CIBERER, Santiago de Compostela, 15782, Galicia, Spain. [3]Laboratorio de Inmunogenética y Diagnóstico Molecular, Independencia 644 - 4, Edif EME1, Córdoba, Argentina.

Authors' contributions
MLC, VAI, AGC, AMM, CR, AB and CV carried out the genotyping of the samples used in the present study. AS carried out the meta-analysis and statistical analysis, and drafted the manuscript, and JA performed the simulation analysis. CV, AC, and AS contributed materials and reagents. All authors approved the final version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 3 February 2011 Accepted: 30 August 2011
Published: 30 August 2011

References

1. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, Fedorova SA, Golubenko MV, Stepanov VA, Gubina MA, Zhadanov SI, Ossipova LP, Damba L, Voevoda MI, Dipierri JE, Villems R, Malhi RS: **Beringian standstill and spread of Native American founders.** *PLoS ONE* 2007, **2(9)**:e829.
2. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Kashani BH, Ritchie KH, Scozzari R, Kong Q-P, Myres NM, Salas A, Semino O, Bandelt H-J, Woodward SR, Torroni A: **Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups.** *Curr Biol* 2009, **19(1)**:1-8.
3. Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, Woodward SR, Salas A, Torroni A, Bandelt H-J: **The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies.** *PLoS ONE* 2008, **3(3)**:e1764.
4. Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Kashani BH, Carossa V, Ekins JE, Gomez-Carballa A, Huber G, Zimmermann B, Corach D, Babudri N, Panara F, Myres NM, Parson W, Semino O, Salas A, Woodward SR, Achilli A, Torroni A: **The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia.** *Genome Res* 2010, **20(9)**:1174-1179.
5. Mandrini RJ: **La Argentina aborigen. De los primeros pobladores a 1910.** Buenos Aires: Siglo XXI Editores Argentina S.A; 2008.
6. Salas A, Comas D, Lareu MV, Bertranpetit J, Carracedo Á: **mtDNA analysis of the Galician population: a genetic edge of European variation.** *Eur J Hum Genet* 1998, **6(4)**:365-375.
7. Lattes ZR, Lattes AE: **Migración internacional y dinámica demográfica en la Argentina durante la segunda mitad del siglo XX.** Buenos Aires: CEMLA; Estudios migratorios latinoamericanos; 2003**50.**
8. Lattes AE, Sautu R: **Immigration, Demographic Change and Industrial Development in Argentina.** Buenos Aires, Argentina: TAPINOS, GEORGES; 1974.
9. Lattes ZR, Lattes AE: **La población de Argentina.** Buenos Aires: Instituto Nacional de Estadistica y Censos; 1975.
10. Lewis MP: **Ethnologue. Languages of the world.** Dallas, Texas: SIR International;, 16 2009.
11. Salas A, Richards M, Lareu MV, Sobrino B, Silva S, Matamoros M, Macaulay V, Carracedo Á: **Shipwrecks and founder effects: Divergent demographic histories reflected in Caribbean mtDNA.** *Am J Phys Anthropol* 2005, **128**:855-860.
12. Gilbert MT, Kivisild T, Gronnow B, Andersen PK, Metspalu E, Reidla M, Tamm E, Axelsson E, Gotherstrom A, Campos PF, Rasmussen M, Metspalu M, Higham TF, Schwenninger JL, Nathan R, De Hoog CJ, Koch A, Moller LN, Andreasen C, Meldgaard M, Villems R, Bendixen C, Willerslev E: **Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland.** *Science* 2008, **320(5884)**:1787-1789.
13. Fagundes NJ, Kanitz R, Eckert R, Valls AC, Bogo MR, Salzano FM, Smith DG, Silva WA Jr, Zago MA, Ribeiro-dos-Santos AK, Santos SE, Petzl-Erler ML, Bonatto SL: **Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas.** *Am J Hum Genet* 2008, **82(3)**:583-592.
14. Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, Comas D: **Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba.** *BMC Evol Biol* 2008, **8**:213.
15. Sandoval K, Buentello-Malo L, Peñaloza-Espinosa R, Avelino H, Salas A, Calafell F, Comas D: **Linguistic and maternal genetic diversity are not correlated in Native Mexicans.** *Hum Genet* 2009, **126(4)**:521-531.
16. Ginther C, Corach D, Penacino GA, Rey JA, Carnese FR, Hutz MH, Anderson A, Just J, Salzano FM, King MC: **Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes.** *Exs* 1993, **67**:211-219.
17. Cabana GS, Merriwether DA, Hunley K, Demarchi DA: **Is the genetic structure of Gran Chaco populations unique? Interregional perspectives on native South American mitochondrial DNA variation.** *Am J Phys Anthropol* 2006, **131(1)**:108-119.
18. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A: **Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups.** *Forensic Sci Int Genet* 2007, **1**:44-55.
19. Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo Á: **Gender bias in the multi-ethnic genetic composition of Central Argentina.** *J Hum Genet* 2008, **53**:662-674.
20. Martínez-Marignac VL, Bravi CM, Lahitte HB, Bianchi NO: **Estudio del ADN mitocondrial de una muestra de la ciudad de la Plata.** *Revista Argentina de Antropología Biológica* 1999, **2(1)**:281-300.
21. Bobillo MC, Zimmermann B, Sala A, Huber G, Rock A, Bandelt H-J, Corach D, Parson W: **Amerindian mitochondrial DNA haplogroups predominate in the population of Argentina: towards a first nationwide forensic mitochondrial DNA sequence database.** *Int J Legal Med* 2009, **74(1)**:65-76.

22. Garcia A, Demarchi DA: **Incidence and distribution of Native American mtDNA haplogroups in central Argentina.** *Hum Biol* 2009, **81(1)**:59-69.
23. Catelli L, Romanini C, Borosky A, Salado-Puerto M, Prieto L, Vullo C: **Common mitochondrial DNA haplogroups observed in an Argentine population database sample.** *Forensic Sci Int Genet Supplement Series* 2010.
24. Sala A, Arguelles CF, Marino ME, Bobillo C, Fenocchio A, Corach D: **Genetic analysis of six communities of Mbya-Guarani inhabiting northeastern Argentina by means of nuclear and mitochondrial polymorphic markers.** *Hum Biol* 2010, **82(4)**:433-456.
25. Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, van Duijn K, Goedbloed M, Vallone PM, Parson W, de Knijff P, Kayser M: **Inferring continental ancestry of argentineans from Autosomal, Y-chromosomal and mitochondrial DNA.** *Ann Hum Genet* 2010, **74(1)**:65-76.
26. Salas A, Carracedo Á, Macaulay V, Richards M, Bandelt H-J: **A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics.** *Biochem Biophys Res Commun* 2005, **335(3)**:891-899.
27. Álvarez-Iglesias V, Mosquera-Miguel A, Cerezo M, Quintáns B, Zarrabeitia MT, Cuscó I, Lareu MV, García O, Pérez-Jurado L, Carracedo Á, Salas A: **New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0.** *PLoS One* 2009, **4(4)**:e5112.
28. García-Bour J, Pérez-Pérez A, Álvarez S, Fernández E, López-Parra AM, Arroyo-Pardo E, Turbón D: **Early population differentiation in extinct aborigines from Tierra del Fuego-Patagonia: ancient mtDNA sequences and Y-chromosome STR characterization.** *Am J Phys Anthropol* 2004, **123(4)**:361-370.
29. Salas A, Lovo-Gomez J, Alvarez-Iglesias V, Cerezo M, Lareu MV, Macaulay V, Richards MB, Carracedo A: **Mitochondrial echoes of first settlement and genetic continuity in El Salvador.** *PLoS One* 2009, **4(9)**:e6882.
30. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo Á: **The African diaspora: mitochondrial DNA and the Atlantic slave trade.** *Am J Hum Genet* 2004, **74(3)**:454-465.
31. Librado P, Rozas J: **DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.** *Bioinformatics* 2009, **25(11)**:1451-1452.
32. Excoffier L, Smouse PE, Quattro JM: **Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data.** *Genetics* 1992, **131(2)**:479-491.
33. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Norby S: **mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion.** *Am J Hum Genet* 2000, **67(3)**:718-726.
34. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB: **Correcting for purifying selection: an improved human mitochondrial molecular clock.** *Am J Hum Genet* 2009, **84(6)**:740-759.
35. Salas A, Carracedo Á, Richards M, Macaulay V: **Charting the Ancestry of African Americans.** *Am J Hum Genet* 2005, **77(4)**:676-680.
36. Salas A, Torroni A, Richards M, Quintana-Murci L, Hill C, Macaulay V, Carracedo Á: **The phylogeography of mitochondrial DNA haplogroup L3g in Africa and the Atlantic slave trade.** *Am J Hum Genet* 2004, **75**:524-526.
37. Salas A, Acosta A, Álvarez-Iglesias V, Cerezo M, Phillips C, Lareu MV, Carracedo Á: **The mtDNA ancestry of admixed Colombian populations.** *Am J Hum Biol* 2008, **20**:584-591.
38. Alves-Silva J, da Silva Santos M, Guimaraes PE, Ferreira AC, Bandelt H-J, Pena SD, Prado VF: **The ancestry of Brazilian mtDNA lineages.** *Am J Hum Genet* 2000, **67(2)**:444-461.
39. Carvalho BM, Bortolini MC, BdS SE, Ribeiro-dos-Santos ÂKC: **Mitochondrial DNA mapping of social-biological interactions in Brazilian Amazonian African-descendant populations.** *Genet Mol Biol* 2008, **31(1)**:12-22.
40. Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B: **The mtDNA Population Database: an integrated software and database resource for forensic comparison.** *Forensic Sci Commun* 2002, **4**:no 2.
41. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM: **Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers.** *Proc Natl Acad Sci USA* 2008, **105(5)**:1596-1601.
42. Beleza S, Gusmão L, Amorim A, Carracedo Á, Salas A: **The genetic legacy of western Bantu migrations.** *Hum Genet* 2005, **117(4)**:366-375.
43. Salas A, Richards M, De la Fé T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo Á: **The making of the African mtDNA landscape.** *Am J Hum Genet* 2002, **71(5)**:1082-1111.
44. Barbosa AB, da Silva LA, Azevedo DA, Balbino VQ, Mauricio-da-Silva L: **Mitochondrial DNA control region polymorphism in the population of Alagoas state, north-eastern Brazil.** *J Forensic Sci* 2008, **53(1)**:142-146.
45. Diegoli TM, Irwin JA, Just RS, Saunier JL, O'Callaghan JE, Parsons TJ: **Mitochondrial control region sequences from an African American population sample.** *Forensic Sci Int Genet* 2009, **4(1)**:e45-52.
46. Brehm A, Pereira L, Kivisild T, Amorim A: **Mitochondrial portraits of the Madeira and Acores archipelagos witness different genetic pools of its settlers.** *Hum Genet* 2003, **114(1)**:77-86.
47. Pereira L, Prata MJ, Amorim A: **Diversity of mtDNA lineages in Portugal: not a genetic edge of European variation.** *Ann Hum Genet* 2000, **64**:491-506.
48. Rando JC, Cabrera VM, Larruga JM, Hernández M, González AM, Pinto F, Bandelt H-J: **Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands.** *Ann Hum Genet* 1999, **63**:413-428.
49. Rhouda T, Martinez-Redondo D, Gomez-Duran A, Elmtili N, Idaomar M, Diez-Sanchez C, Montoya J, Lopez-Perez MJ, Ruiz-Pesini E: **Moroccan mitochondrial genetic background suggests prehistoric human migrations across the Gibraltar Strait.** *Mitochondrion* 2009, **9(6)**:402-407.
50. Černý V, Salas A, Hájek M, Žaloudková M, Brdička R: **A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome.** *Ann Hum Genet* 2007, **71(Pt 4)**:433-452.
51. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R: **Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears.** *Am J Hum Genet* 2004, **75(5)**:752-770.
52. Saunier JL, Irwin JA, Strouss KM, Ragab H, Sturk KA, Parsons TJ: **Mitochondrial control region sequences from an Egyptian population sample.** *Forensic Sci Int Genet* 2009, **3(3)**:e97-103.
53. Maca-Meyer N, Sánchez-Velasco P, Flores C, Larruga JM, González AM, Oterino A, Leyva-Cobian F: **Y chromosome and mitochondrial DNA characterization of Pasiegos, a human isolate from Cantabria (Spain).** *Ann Hum Genet* 2003, **67(Pt 4)**:329-339.
54. Crespillo M, Luque JA, Paredes M, Fernández R, Ramirez E, Valverde JL: **Mitochondrial DNA sequences for 118 individuals from northeastern Spain.** *Int J Legal Med* 2000, **114(1-2)**:130-132.
55. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogvali EL, Kivisild T, Bandelt H-J, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A: **The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool.** *Am J Hum Genet* 2004, **75(5)**:910-918.
56. Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ: **The role of selection in the evolution of human mitochondrial genomes.** *Genetics* 2006, **172(1)**:373-387.
57. Marino M, Sala A, Bobillo C, Corach D: **Inferring genetic sub-structure in the population of Argentina using fifteen microsatellite loci.** *Forensic Sci Int Genet* 2008, **1**:350-352.
58. Toscanini U, Berardi G, Amorim A, Carracedo Á, Salas A, Gusmão L, Raimondi E: **Forensic considerations on STR databases in Argentina.** *Int Congress Series* 2006, **1288**:337-339.
59. Toscanini U, Gusmao L, Berardi G, Amorim A, Carracedo A, Salas A, Raimondi E: **Testing for genetic structure in different urban Argentinian populations.** *Forensic Sci Int* 2007, **165(1)**:35-40.
60. Toscanini U, Gusmao L, Berardi G, Amorim A, Carracedo A, Salas A, Raimondi E: **Y chromosome microsatellite genetic variation in two Native American populations from Argentina: population stratification and mutation data.** *Forensic Sci Int Genet* 2008, **2(4)**:274-280.
61. Toscanini U, Salas A, Carracedo Á, Berardi G, Amorim A, Gusmão L, Raimondi E: **A simulation-based approach to evaluate population stratification in Argentina.** *Forensic Sci Int Genet SS* 2008, **1**:662-663.
62. Mosquera-Miguel A, Álvarez-Iglesias V, Vega A, Milne R, Cabrera de León A, Benitez J, Carracedo Á, Salas A: **Is mitochondrial DNA variation associated with sporadic breast cancer risk?** *Cancer Res* 2008, **68(2)**:623-625.
63. Salas A, Carracedo Á: **Studies of association in complex diseases: statistical problems related to the analysis of genetic polymorphisms.** *Rev Clin Esp* 2007, **207**:563-565.

64. Salas A, Fachal L, Marcos-Alonso S, Vega A, Martinón-Torres F, ESIGEM G: **Investigating the role of mitochondrial haplogroups in genetic predisposition to meningococcal disease.** *PLoS One* 2009, **4(12)**:e8347.
65. Toscanini U, Salas A, Garcia-Magarinos M, Gusmao L, Raimondi E: **Population stratification in Argentina strongly influences likelihood ratio estimates in paternity testing as revealed by a simulation-based approach.** *Int J Legal Med* 2010, **124(1)**:63-69.

doi:10.1186/1471-2156-12-77
Cite this article as: Catelli *et al.*: The impact of modern migrations on present-day multi-ethnic Argentina as recorded on the mitochondrial DNA genome. *BMC Genetics* 2011 **12**:77.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

GDF: Dealing with high-throughput genotyping multiplatform data for medical and population genetic applications.

Amigo J, Salas A, Costas J, Carracedo Á

Journal of Proteomics and Bioinformatics. 01/2012; 5(1):1-6.

Diferentes plataformas de genotipado de alto rendimiento han surgido recientemente. Estas plataformas generan grandes cantidades de datos genotípicos que son posteriormente procesados y almacenados en bases de datos públicas y/o privadas. Ambos, la variedad de plataformas empleadas por los diferentes laboratorios y la gran cantidad de datos que generan, conllevan graves problemas para la gestión de datos en la mayoría de los laboratorios. Algunos paquetes de software públicos o privados disponibles en la actualidad resuelven algunas necesidades importantes, pero lidian con los datos desde un punto de vista que el investigador puede no compartir y puede no realizarse una supervisión de los resultados (por ejemplo, las inconsistencias de genotipado o resúmenes de los datos de genotipado).

El objetivo principal del Filtro de Datos Genotípicos (GDF) es permitir al investigador administrar localmente una gran cantidad de genotipos generados por las plataformas de genotipado más estándar, generar estadísticas y resúmenes de los experimentos de genotipado, mientras se mantiene su privacidad. GDF también permite al usuario supervisar los datos para que el investigador pueda evaluar los parámetros importantes, como la proporción de datos faltantes en las muestras y los polimorfismos de un solo nucleótido (SNPs), equilibrio Hardy-Weinberg, etc. Además, GDF transforma los datos crudos en diferentes formatos de texto en archivos de entrada necesarios en paquetes de software populares utilizados con frecuencia en aplicaciones médicas y de genética de poblaciones.

GDF es un programa Perl que procesa de manera eficiente los datos de diferentes plataformas de genotipado, lo que permite a los investigadores inspeccionar fácilmente sus propios datos de genotipado y transformarlos para un amplio espectro de software de análisis especializado. Se ha preparado para ser ejecutado a través de una interfaz web sencilla en los casos más comunes, pero también se puede ejecutar como un script local en computadores personales, e incluso supercomputadores para proyectos de gran escala.

GDF: Dealing with High-throughput Genotyping Multiplatform Data for Medical and Population Genetic Applications

Jorge Amigo[1]*, Antonio Salas[2], Javier Costas[3] and Ángel Carracedo[1,2,3]

[1]*Grupo de Medicina Xenómica, CIBERER, Universidade de Santiago de Compostela, Galicia, Spain*
[2]*Unidade de Xenética Forense, Instituto de Medicina Legal, Facultade de Medicina, Universidade de Santiago de Compostela, Galicia, Spain*
[3]*Fundación Pública Galega de Medicina Xenómica (FPGMX), Hospital Clínico Universitario. 15706, Santiago de Compostela, Spain*

Abstract

Background: A number of different high throughput genotyping platforms have arisen recently. These platforms generate large amounts of genotyping data which is subsequently processed and stored in public and/or private databases. Both, the variety of platforms employed by the different laboratories and the large amount of data they generate, entail serious problems for data managing in most laboratories. Some public or private software packages available today solve some important needs, but they deal with the data from a point of view that the researcher may probably not share, and no supervision of the results (e.g. genotyping inconsistencies or summaries of the genotyping data) may be performed.

Results: The main goal of the Genotyping Data Filter (GDF) software is to allow the researcher to locally manage large numbers of genotypes generated by the most standard genotyping platforms, obtaining statistics and summaries of the genotyping experiments whilst maintaining their privacy. GDF also allows the user to supervise the data such that the researcher can easily evaluate important parameters, including the proportion of missing data in samples and single nucleotide polymorphisms (SNPs), Hardy-Weinberg equilibrium, etc. Additionally, GDF parses the raw data into different text formats needed as input files in popular software packages frequently used in medical and population genetic applications.

Conclusions: GDF is a Perl program that efficiently process data from various genotyping platforms, allowing researchers to easily inspect their own genotyping data and to parse it for a wide spectrum of well-known specialized analysis software. It has been prepared to be run through a user friendly web interface on the most common cases, but it can also be run as a local script on personal computers, or even supercomputers for very large-scale projects.

Background

A major interest of current genomics research is devoted to disease-gene association studies, that is, studies aimed to identify DNA variants presumably associated with susceptibility (or protection) to a common disease. Advances in genotyping and sequencing technologies, coupled with the development of sophisticated statistical methods, have afforded investigators novel opportunities to define the role of sequence variation in the development of common human diseases. [1,2]. Considering that during the last years the genotyping efficiency has heavily increased, research groups have now to cope with genomic large-scale and high density SNP association analysis through all the genome. It is expected that these genome-wide association studies may identify alleles related to complex disorders, and therefore finding the underlying causative relationships is currently a major challenge.

It is estimated that SNPs occur once per 100~300 bases in the human genome, which represents over 10 million SNPs in our whole genome [3]. Thus, in large-scale association studies, genotyping all SNPs in a candidate region for a large number of individuals is still costly and time-consuming. Sets of nearby SNPs on the same chromosome are inherited in blocks (this pattern of inherited SNP variants on a single block is a haplotype), and although blocks may contain a large number of SNPs and can be very variable in size, only few SNPs might be needed to uniquely tag and identify the haplotypes in a block (what is called a haplotype tagging SNP, or htSNP or tagSNP). This is due to the correlation between alleles at nearby variant sites, named linkage disequilibrium (LD), that exists because of the shared ancestry of contemporary chromosomes that is erode by mutation and recombination [4]. From the initial efforts to characterize the human genome by studying its common variability [5,6], the HapMap Project was born as a public effort to build a map of these haplotype blocks and their htSNPs. This map of blocks and htSNPs allows reducing significantly the number of SNPs required to interrogate the entire genome for association with a disease phenotype from more than 14 million SNPs that exist today to roughly 500,000 htSNPs. This will make genome scan approaches to find regions that affect diseases in a much more efficient and comprehensive way, since effort will not be wasted typing more SNPs than necessary and all regions of the genome can be included. Results from these whole genome scans promise to be successfully translated into useful applications in areas such as medical diagnosis [7] or pharmacogenomics [8]. HapMap is then a useful resource that allows selecting a group of SNPs to analyze a possible association between custom genomic regions with the studied pathology. HapMap, together with other ambitious genomic projects (e.g. Perlegen), has allowed changing the classical perspective of analyzing a single functional polymorphism on a single gene to the current analysis of multiple genes from the same pathway, or even the whole genome.

Several and very different genotyping techniques have arisen in

***Corresponding author:** Jorge Amigo, Grupo de Medicina Xenómica, CIBERER, Universidade de Santiago de Compostela, Galicia, Spain. E-mail: jorge.amigo@usc.es

Received October 22, 2011; **Accepted** December 27, 2011; **Published** January 02, 2012

Citation: Amigo J, Salas A, Costas J, Carracedo Á (2012) GDF: Dealing with High-throughput Genotyping Multiplatform Data for Medical and Population Genetic Applications. J Proteomics Bioinform 5: 001-006. doi:10.4172/jpb.1000206

Copyright: © 2012 Amigo J, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

less than a decade. Companies like Sequenom [9], Applied Biosystems [10,11], Illumina [12] or Affymetrix [13,14] have been developing their own exclusive techniques to identify SNPs with very diverse throughput capacities (which is constantly increasing due to continuous innovations, most of them in chemistry) but also developing different strategies in terms of hardware and software. What they all have in common is that they can be used for large-scale genotyping experiments, and so they all have to face the same issue: data management. The software packages that the companies usually provide with their genotyping platforms have been developed with the main aim in mind of making the generated data management as comfortable and powerful as possible. However, the lack of flexibility and serious limitations of these software packages encourages for local dedicated developments. In addition, it is often required to use multiple genotyping platforms to perform a single experiment, as the genotyping methods are very different and certain SNPs may be better detected with one or other genotyping technique. Therefore, corporate software does not usually allow dealing with all the experiment data as a whole. The complexity of these tools will well vary with the specific needs of each group: from a simple set of platform-specific Visual Basic macros like TIMS [15] to SNPator [16], an example of an *ad hoc* online package designed to cover the needs of the large-scale genotyping process of the Spanish National Genotyping Centre (http://www.cegen.org/) on multiple platforms.

The high throughput genotyping (HTG) capability does not only depend on the genotyping techniques, but also on the data handling approaches that had to manage all that new overwhelming amount of information [17]. The critical issues that arise on HTG projects always concern the data: inspecting it for possible errors, the whole management and the later analysis. As mentioned above, some useful free tools have appeared during the HTG expansion in order to cover the management part using databases and web interfaces [18,19], even with great visual aids [20], while they implement internal consistency checks and embed different algorithms for data analysis. Data managers such as SNPator [14], SNPP [17] or SNPLims [16] are also capable of exporting the data in appropriate formats for later deeper analysis, a basic request for any HTG project as it is very hard to implement all the algorithms that all users may need, but the limited format offer for some researchers may still force them to find a more appropriate tool such as GDF, that provides the input format for several different programs on the association and population studies field.

Although SNPator's data importing module implements many of the features considered in GDF, a major advantage of GDF is that it can work locally, and this feature may be of great help for researchers that have to deal with low to medium SNP genotyping projects, especially for those researchers that wish to preserve as much as possible the privacy of their research projects. Although web-based implementations are obviously useful, some researchers may not be fully comfortable with the idea of storing their data in servers where they do not have full control of it, and this may even lead to some bioethical concerns. GDF represents a much more flexible alternative that could even be embedded into a larger software package already developed, or into any local pipeline for specific research needs.

Implementation

The GDF has been designed to work as a flexible interface between the researcher and the raw genotyping data dumped by the platforms (Figure 1). The researcher may need to process the raw data in a particular way, and for that reason the proprietary software from the genotyping platforms may be not very flexible. The main idea is to allow merging complementary information with the raw data, allowing the researcher to obtain not only general summarizing reports, but also to perform a customizable quality control of the results and to have that raw data parsed into input files for specific analysis software packages (Table 1).

Data reading module

The reading process of the raw data exported directly from the genotyping platforms works line by line, as each line contains a single genotype. Most platforms share this characteristic in plain text tabulated files, and therefore, by using a different recognition pattern for each platform, it is possible to identify and dissect even any forthcoming technology.

Each platform has its own format, but all of them provide at least the information concerning the codes of the SNPs that were genotyped, the sample IDs and the genotyping calls. The reading module of GDF

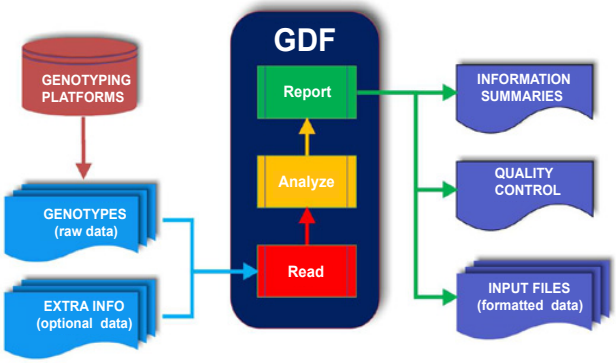


Figure 1: The GDF data workflow. The raw data coming from the genotyping platforms is directly read by GDF, while extra information may be added to characterize it. This combination is then processed and exported if requested into input files for specific software of analysis, along with a complete report of all the internal checks performed.

INPUT	Genotyping Platforms	Sequenom SNPlex Illumina Affymetrix*
	Generic formats	HapMap SNPs vs. Samples tables Samples vs. SNPs tables
OUTPUT	Association Studies	GeneHunter Haploview` PLINK MEGA-2 Arlequin Phase Hapblock EMLD Unphased MDR
	Population Studies	Structure Arlequin Haploview

Table 1: File formats handled by GDF. Being a flexible interface between raw data and specific analysis programs, GDF deals with several input formats coming directly from the genotyping platforms, from the HapMap project, or even from custom made tables. The Affymetrix format coverage is not full though (*), as only general reports may be obtained from it through GDF due to its usual data size, but for the rest of the input formats GDF is able to parse their genotypes and provide the input for several programs of analysis for association and population studies.

also allows to store additional information collected by the different platforms (plate information, well position on the plate, manual edition of the call or the score of the result for instance). All this information may be later analysed and used to enrich the output, but only these three fields are mandatory.

Once the line is recognized by the pattern recognition engine the reading function starts to process it, saving the platform detected and storing all the dissected data into an appropriate hash indexed by gene, sample ID and SNP position. The first and third indexes are automatically given by GDF unless they are indicated in a complementary input file, but they must be always consider to allow the user to group sets of SNPs and to sort them in case this is needed.

Complementary input files

Since the raw data from the genotyping platforms can possibly not contain all the data the researcher may need, GDF gives the option of providing extra information through complementary input files, and this information is then concatenated with the raw data. The kind of information that can be supplemented in these extra files could be related to the need of grouping information (e.g. by ethnic or population groups), sample characterization, translation for non-explicit platform alleles, or even signalling samples that should be excluded from some particular analysis.

Configuration file

This is a three column tabulated plain text file where the SNPs can be grouped in genes or any kind of grouping strategy. A number (e.g. chromosome location) can also be attached to each SNP. The first row must contain the fields' names, which must be "GENE", "SNP_ID" and "POSITION", and the rest of the lines must contain the data.

This file is often used to divide the output in different files, one for each gene, to sort the SNPs by their position to build the appropriate haplotypes or to filter SNPs to be processed as GDF will not process any SNP that is not present in this file.

Pedigree and population file

Pedigree and population information may be assigned to each processed sample respectively in pedigree format and in a tabulated text headed file with three columns: samples, populations and population ids. This may be desired when expecting GDF to provide in its output the input formats for specific programs, such as Structure [21] (will not work without population information) or Phase [22] (the case-control running option will not be available if the appropriate column is not present in the input file) for instance. The GDF web interface detects which files have been uploaded, and allows the user to select only the available outputs for the data that is to be used.

Allele translation file

Raw data coming from platforms such as Taqman, Illumina or even older versions of the SNPlex format do not provide explicit calls for each genotype. A code is given to each result instead of the appropriate base (al1 or al2, A or B, A1 or A2 …), and the translation of that code is SNP specific and then stored in a configuration file. To deal with that code in the output data from any of the mentioned platforms GDF allows importing a tabulated text file that, without any headers, indicates in three columns the SNP name, the code and its translation. This information is then used to convert internally all the data to the proper formats and to give the desired outputs.

SNPs and samples not to be processed

As the raw genotyped data file may be difficult to edit, and being this not the best option, a filtering option was implemented through file input. Thus, if a list of SNPs or samples is provided (text files with all the desired SNPs or samples in a single column) GDF will not process the data associated with them. This may help, for instance, to remove from the statistics or from the specific output files samples or SNPs that have been wrongly genotyped due to experiment errors, or to treat separately data coming from different projects that have shared the experiment but that should not be analysed together.

Results

We have developed a program written in Perl, as it is one of the most popular reference programming language for fast and comprehensive text handling [23]. GDF performs a series of internal analysis such as quality control and consistency checks, and provides formatted data to be given as input for several different programs for association and population studies (Table 1).

Several projects present in the literature have used GDF as the data pre-processing tool when they had to deal directly with raw genotypes (e.g. [24,25]). As a practical example, GDF was used to deal with 137.015 genotypes generated by the Sequenom platform, creating the input files for additional analysis using Structure, Unphased, Haploview, PHASE, and MDR, in 31 seconds. This led to the replication of *DTNBP1* as a schizophrenia susceptibility locus [26].

Another common use of GDF is to parse tables of samples and SNPs that may have been manually edited, or data from the HapMap project extracted directly from its website or from intermediate repositories such as SPSmart [27,28]. For instance, the study of the *CYP21A2* gene reveals a low SNP density on HapMap, but it can be merged through GDF with the SNP information obtained by direct sequencing of 21-hydroxylase deficiency patients [29] in order to highlight haplotypes associated with this pathology.

Internal Data Analysis

Validation analyses tests

This group of analysis includes all the error and consistency checks performed by GDF. One of the primary aims of this sub-program is to let the user revise the raw data from the platform, and thus it allows the user to know if there are any incoherencies (e.g. a duplicated genotype does not match) present for a given genotype. If control samples were introduced in the experiment, or any sample is just genotyped more than once, a quality control will be carried out in order to detect inconsistencies.

Another analysis performed is the check for more than two alleles found for a single SNP. Considering that bi-allelic SNPs are the most common variation, highlighting these situations is needed as they would normally represent a flagrant error, possibly at the experimental design of the experiment or at the genotyping software assignation.

Informative analyses

All the rest of the tests performed by GDF are meant to describe the data analysed, in order to provide a broader understanding of the experiment. The most important ones would be the detection of data skews (inconsistent genotypes for the same SNP tested on the same sample), monomorphic SNPs, and the highlighting of SNPs or samples with no valid result in all the experiment, but the summary

and statistics performed with GDF given at the end of each run are also very informative: the total of the data input lines is presented, and compared with the total numbers of genotypes detected, valid results and failed genotypes. Table 2 gives a detailed description of these checks. In case repeated genotypes (such as quality controls or just replicas) are entered, GDF will display a quality control section at the output, detailing the numbers of repeats, how many did actually match and how many did not match.

Output Files

Files under demand

These files contain all the available input formats for the specific analysis programs. Currently, the linkage pre-makefile format is included, which is valid for popular association studies programs such as GeneHunter [30], Haploview [31] or PLINK [32], or for a meta-analysis software package named MEGA-2 [33] that is able to support 28 target programs by combining the linkage format with a pedigree and a mapping file. Additionally, GDF can also be used to obtain the input format for other popular association studies software such as Arlequin [34], EMLD [35], Hapblock [36], MDR [37], Phase [20] and Unphased [38,39]. Input formats for population studies software like Structure [21] may also be obtained. Except the input format for Arlequin, the rest of the programs will work directly with the files generated by GDF. Arlequin needs some minor manual edition in order to include the configuration headers.

Automatically generated files

This set of files contains valuable information obtained from the input files. For instance, the GDF generates a set of files that contain information on genotyping errors or undesired inconsistencies. There are also three useful files that are generated by default: i) a text table containing a matrix of SNPs versus samples which will state all the results in an efficient manner for visual inspection, ii) a statistic file for samples with information of the percentage of missing genotypes on each one, and iii) a statistics file containing information about the alleles observed, SNP heterozygosity values, minor allele frequencies, or the result of checking for Hardy-Weinberg equilibrium and its statistical significance using a simple chi-square test.

Performance in memory and time

As GDF is meant to work with large amounts of data coming from HTG experiments, its performance had to be measured in order to predict the running time and the computer resources that could be needed for the biggest experiments. For that reason we tested its performance on an ordinary PC with an Intel Pentium IV 3.4GHz processor and 1GB RAM. We then measured the computational resources demanded by GDF with respect to the number of genotypes that had to be processed. The summary of the benchmarking results of GDF are reflected on the top graph of Figure 2, where the memory and time linear tendencies can be observed, validating the adequacy of the internal GDF code which could otherwise depend exponentially on the amount of processed data. This fact is critical to allow future evolution of the program.

Unused genes	Genes present in the configuration file with all their SNPs untested
Unknown SNPs	SNPs present in the data file but not in the configuration file
Untested SNPs	SNPs present in the configuration file but not in the data file
Failed SNPs	SNPs that failed in all the genotyped samples
Failed samples	Samples with no successful genotype
No pedigree samples	Samples with no pedigree information present in the pedigree file if used
Unperformed tests samples	SNPs that were not genotyped on a sample but they were tested on the rest
Overlapping information samples	Samples that carry a third allele, assuming most genotyping techniques deal with bi-allelic SNPs

Table 2: Analyses performed by GDF. A description of the internal checks that GDF performs in order to improve the description of the data being analysed, which are given at the end of each run.

The lack of perfect linearity regarding the demand of memory and time needed by GDF as more SNPs and samples are processed seems to indicate that there are underlying factors affecting GDF's performance (see the bottom graphs of Figure 2), such as the percentage of repeated genotypes present on the experiment (quality control) that significantly reduces the amount of memory needed. The invested time per genotype strongly depends on the platform used, as there are platforms that provide much more information for each genotype that is also processed by GDF. Providing more information implies more complex line pattern recognition, and that is the major latency present on the program. Thus, platforms like SNPlex include in their outputs information concerning genotyping quality scores and manual edition flags.

Interacting with the program

GDF can be run locally in command line, executing its code through a locally installed Perl interpreter. This is the most versatile option, and as there are plenty of versions of Perl depending on the operating system, GDF has been designed in order to be also independent to the platform. In addition, as some researchers may not be comfortable with command line commands, several graphical user interfaces (GUIs) have also been developed to work around this issue: i) an online PHP interface to the most updated version of GDF, which runs it directly on the web server without having to install anything locally, and ii) a Visual Basic interface that runs an encapsulated executable version of GDF for Windows platforms only. In both cases the user gets a four steps interface: i) the data input, where all the files that are going to be used must be selected, ii) the options selection, where all the GDF's options may be chosen, iii) the formats request, where the programs to which the data should be formatted for should be highlighted, and iv) the final results. In this last step there will always be a screen output, accompanied by a link to all the files that were generated (one of them will be that screen output for later inspection).

Discussion

Performing of all of the automatic analyses described above provides more information about the raw data than the one provided by corporate software, independently from the genotyping platform used. The personalization of the analyses performed allows GDF users to find out data which is difficult to retrieve when using those corporate software packages, because of the appropriate options absence or the manual revising impossibility, such as platform errors or even pre-genotyping problems.

The genotyped data is never the final step of the full analysis process. The data must always be processed by deeper analysis and specialized programs that will go far beyond to find information such as associations (case-control, TDT,…), haplotypes, present population substructures, and so on. Some platforms may give as output the input for a certain analysis program, but the researcher may prefer to be able to transform any kind of data coming from any platform into any

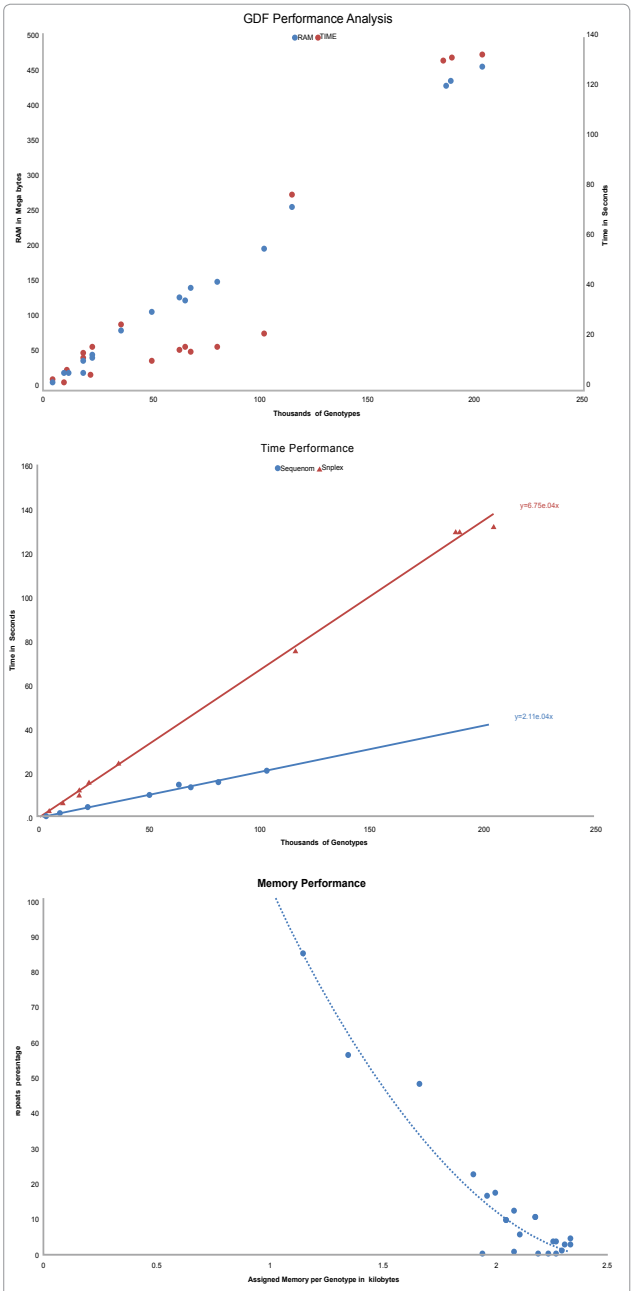


Figure 2: Performance analyses. Multiple experiments performed with Sequenom and SNPlex were grouped and summarized for this figure. Raw measurements of memory and time are presented on 2A to show the linear dependency of the resources needed, and a deeper analysis of those magnitudes is displayed on the two bottom graphs. The performance in time showed on 2B highly depends on the platform where the genotyping data comes from, due to the complexity of the pattern recognition applied to each line. The memory assigned to each genotype in each experiment is presented on 2C, depending on the amount of information that each genotype may be characterized with. This assigned memory does not really depend on the genotyping platform, but it drastically depends on the percentage of genotype repeats of the experiment.

desired format. GDF allows doing this through an appropriate internal variable structure and design that is prepared to easily deal with new upcoming input formats.

Providing a parallelizable version of the program is our next aim, as it would allow running it either on dedicated supercomputers or directly on personal computers with multiple-core machines.

Conclusion

GDF is a program to process HTG data specially produced by the biomedical community. Other fields of research are now benefiting from HTG such as those interested in quantitative characters' analysis in species of commercial interest, for instance. Any researcher may then workaround some of the corporate software packages' limitations embedding GDF in the genotyping routine. A set of improvements to this process have been implemented inside GDF, and their use is fairly straightforward. But the best advantage for the researcher is probably not to be forced to use a local database, which will need expertise on installation and maintenance, nor even a remote one that could compromise the data privacy. Previous similar work has been done using a SNP database to hold the data while processing it, but GDF allows dealing directly with the data.

Authors' Contributions

JA carried out the design, programming and implementation of the software, and drafted the manuscript. AS, JC and AC participated in the design of the software, suggesting and testing the implementation of new features, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Genoma España foundation through the National Genotyping Center (CeGen), and by grants from the Xunta de Galicia (Grupos Emerxentes; 2008/037), Ministerio de Ciencia e Innovación (SAF2008-02971), and Fundación de Investigación Médica Mutua Madrileña (2008/CL444) given to AS. Thanks to Beatriz Sobrino, María Torres and Inés Quintela for their feedback on their usage as genotyping platform managers, to Ceres Fernández and Laura Fachal for their analysis on additional usage possibilities, and to the many users that have provided us with the feedback about features to improve.

References

1. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366: 1121-1131.
2. Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56: 303-320.
3. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-796.
4. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, et al. (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
5. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, et al. (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29: 233-237.
6. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.
7. Ribas G, Gonzalez-Neira A, Salas A, Milne RL, Vega A, et al. (2006) Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes. Hum Genet 118: 669-679.
8. Deloukas P, Bentley D (2004) The HapMap project and its application to genetic studies of drug response. Pharmacogenomics J 4: 88-90.
9. Jurinke C, Oeth P, van den Boom D (2004) MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. Mol Biotechnol 26: 147-164.
10. De la Vega FM, Lazaruk KD, Rhodes MD, Wenz MH (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. Mutat Res 573,111-135.
11. Tobler AR, Short S, Andersen MR, Paner TM, Briggs JC,et al. (2005) The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. J Biomol Tech 16: 398-406.

12. Steemers FJ, Gunderson KL (2005) Illumina Inc Pharmacogenomics 6: 777-782.

13. Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, et al. (2005) Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Res 15: 269-275.

14. Moorhead M, Hardenbol P, Siddiqui F, Falkowski M, Bruckner C, et al. (2006) Optimal genotype determination in highly multiplexed SNP data. Eur J Hum Genet 14: 207-215.

15. Monnier S, Cox DG, Albion T, Canzian F (2005) T.I.M.S: TaqMan Information Management System, tools to organize data flow in a genotyping laboratory. BMC Bioinformatics 6: 246.

16. Morcillo-Suarez C, Alegre J, Sangros R, Gazave E, de Cid R, et al. (2008) SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. Bioinformatics 24: 1643-1644.

17. Li JL, Deng H, Lai DB, Xu F, Chen J, et al. (2001) Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. Genome Res 11: 1304-1314.

18. Orro A, Guffanti G, Salvi E, Macciardi F, Milanesi L (2008) SNPLims: a data management system for genome wide association studies. BMC Bioinformatics 9: S13.

19. Zhao LJ, Li MX, Guo YF, Xu FH, Li JL, et al. (2005) SNPP: automating large-scale SNP genotype data management. Bioinformatics 21: 266-268.

20. Tebbutt SJ, Opushnyev IV, Tripp BW, Kassamali AM, Alexander WL, et al. (2005) SNP Chart: an integrated platform for visualization and interpretation of microarray genotyping data. Bioinformatics 21: 124-127.

21. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155 :945-959.

22. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68: 978-989.

23. Stein LD (2001) Using Perl to facilitate biological analysis. Methods Biochem Anal 43: 413-449.

24. Salas A, Vega A, Milne R, García-Magariños M, Ruibal A, et al. (2008) The 'Pokemon' (ZBTB7) Gene: No Evidence of Association with Sporadic Breast Cancer. Clin Med Oncol 2:357-362.

25. Vega A, Salas A, Milne RL, Carracedo B, Ribas G, et al. (2009) Evaluating new candidate SNPs as low penetrance risk factors in sporadic breast cancer: a two-stage Spanish case-control study. Gynecol Oncol 112: 210-214.

26. Vilella E, Costas J, Sanjuan J, Guitart M, De Diego Y, et al. (2008) Association of schizophrenia with DTNBP1 but not with DAO, DAOA, NRG1 and RGS4 nor their genetic interaction. J Psychiatr Res 42: 278-288.

27. Amigo J, Phillips C, Salas A, Carracedo A (2009) Viability of in-house datamarting approaches for population genetics analysis of SNP genotypes. BMC Bioinformatics 10: S5.

28. Amigo J, Salas A, Phillips C, Carracedo A (2008) SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. BMC Bioinformatics 9: 428.

29. Loidi L, Quinteiro C, Parajes S, Barreiro J, Leston DG, et al. (2006) High variability in CYP21A2 mutated alleles in Spanish 21-hydroxylase deficiency patients, six novel mutations and a founder effect. Clin Endocrinol (Oxf) 64: 330-336.

30. Li H, Schaid DJ (1997) GENEHUNTER: application to analysis of bipolar pedigrees and some extensions. Genet Epidemiol 14: 659-663.

31. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

33. Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. Bioinformatics 21: 2556-2557.

34. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1: 47-50.

35. Huang Q (2005) EMLD.

36. Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, et al. (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics 21: 131-134.

37. Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19: 376-382.

38. Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25: 115-121.

39. Dudbridge F (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered 66: 87-98.

Submit your next manuscript and get advantages of OMICS Group submissions

Unique features:

- User friendly/feasible website-translation of your paper to 50 world's leading languages
- Audio Version of published paper
- Digital articles to share and explore

Special features:

- 200 Open Access Journals
- 15,000 editorial team
- 21 days rapid review process
- Quality and quick editorial, review and publication processing
- Indexing at PubMed (partial), Scopus, DOAJ, EBSCO, Index Copernicus and Google Scholar etc
- Sharing Option: Social Networking Enabled
- Authors, Reviewers and Editors rewarded with online Scientific Credits
- Better discount for your subsequent articles

Submit your manuscript at: http://www.editorialmanager.com/proteomics

IV. DISCUSIÓN

Desde el descubrimiento de la estructura del ADN a mediados del siglo XX se han ido sucediendo una serie de revoluciones tecnológicas que han ido cambiando la concepción de la biología molecular tal y como se entendía previamente. Y no se han tratado de evoluciones y optimizaciones de la técnica, que también las ha habido, sino de avances tan radicalmente sustanciales que acababan implantándose rápidamente como estándares en la rutina diaria, transformando la técnica anterior en obsoleta o relegándola a un ámbito de uso muy particular. Desde el descubrimiento de la reacción en cadena de la polimerasa, pasando por la secuenciación Sanger, hasta la aparición de los grandes repositorios en línea de información de referencia, por no mencionar las más recientes tecnologías de genotipado de alto rendimiento y ultra-secuenciación de la última década. El problema surge cuando la separación en el tiempo de estas revoluciones es muy corto, ya que resulta muy complicado primero entenderlas, luego planificar adecuadamente su uso y por último, pero no por ello menos importante, hacerse con ellas especialmente en los tiempos que corren.

La publicación del genoma humano en 2001 supone un punto de inflexión en la ya de por sí rápida evolución del campo, ya que motivó el desarrollo de nuevos proyectos y nuevas tecnologías orientadas al escrutinio profundo y sistemático de nuestra información genética. Así, a principios del siglo XXI surgen las primeras plataformas de genotipado, que eran capaces de interrogar un puñado de marcadores en un puñado de muestras, aunque ahora son capaces de trabajar con miles de muestras accediendo a la información de millones de marcadores. También hacia mediados de la primera década del siglo XXI comienzan a aparecer los primeros secuenciadores paralelos, que permiten leer el ADN de una forma masiva sin precedentes hasta la fecha. Lo

que hace poco más de 10 años necesitó varios años para su consecución, ahora somos capaces de obtenerlo en cuestión de unos pocos días o incluso horas. Las reglas del juego han cambiado totalmente. Ya no entendemos nuestra rutina diaria sin la presencia de ordenadores, sin la disposición de grandes equipos automáticos de procesamiento de las muestras, ni sin el acceso a información en línea. Ha crecido exponencialmente el número de variables a manejar, y nuestra capacidad de generación de resultados supera ya con creces nuestra capacidad de análisis, por lo que es necesario crear nuevos algoritmos y nuevas herramientas que permitan salvar esta limitación.

A pesar de los esfuerzos de las casas comerciales dedicadas al desarrollo de tecnología por proveer a sus usuarios de las herramientas de análisis necesarias para entender los datos generados, programas como GDF o SNPator ponen de manifiesto la necesidad de los investigadores de conseguir una rápida inspección, lo más cercana posible a las características de su investigación, de los resultados de las tecnologías de genotipado. Es más, a pesar de que son muchos los programas de análisis que trabajan con datos genotípicos, es difícil encontrar un formato consenso de resultados entre las tecnologías, y por tanto entre los formatos requeridos por los posteriores programas de análisis. La simple necesidad de conversión de formato se convierte pues en una necesidad al margen del propio análisis, y herramientas como las anteriormente mencionadas se encargan de salvar esta limitación.

Pero uno de los principales problemas de la investigación surge cuando se requiere difundir y compartir sus resultados. A veces es posible realizar una entrega de dichos resultados a un repositorio en línea de gran ámbito, manteniéndolos allí alojados en un marco informativo a un nivel superior. Otras veces, las más, uno se encuentra con la necesidad de compartir un conjunto de datos informativos en sí mismos, y trata de generar como resultado colateral una plataforma de difusión. Iniciativas como SNP*for*ID browser o popSTR son ejemplos de cómo los resultados de investigación no sólo se pueden listar o hacer accesibles vía descarga en masa, sino que a través de un pre-análisis de ciertos índices estadísticos de interés poblacional se puede proveer a la comunidad científica de una herramienta de consulta ágil y útil podrá servir de referencia para proyectos de investigación venideros, ya sean internos como externos.

Los artículos presentados en el capítulo de resultados que tratan sobre la selección adaptativa de un gen encargado del procesado de glucosa en poblaciones euroasiáticas o la identificación en todo el genoma de dianas genéticas inducidas por hipoxia son ejemplos de la exportación del conocimiento adquirido internamente y compartido hacia el exterior. Una vez adquirida la capacidad de procesamiento de grandes volúmenes de datos genotípicos como los contenidos en SPSmart, y una vez demostrada la agilidad en el acceso a estadísticas derivadas de interés poblacional a pesar de tener que manejar una alta densidad de marcadores a lo largo de todo el genoma humano como en el caso de ENGINES, la traslación de este conocimiento a líneas de investigación en genética de poblaciones no sólo es interesante a nivel de colaboraciones, sino también necesaria a fin de evitar la duplicidad de esfuerzos en un área donde no se puede malgastar el tiempo. Es necesario invertir un gran esfuerzo previo en conocer el estado del arte al comienzo y a lo largo de una investigación si no se quiere descubrir que se ha invertido tiempo y recursos de todo tipo en realizar un trabajo que, por lo tremendamente cambiante del campo, pudiera haberse publicado apenas unos pocos meses antes. Idealmente uno trata de preparar sus líneas de investigación lo mejor posible, pero dada la velocidad en la que este tipo de disciplinas evolucionan el investigador debe estar dispuesto a modificar ciertos objetivos previamente bien documentados haciendo uso de los recursos que vayan surgiendo. Los artículos previamente mencionados, relacionados ambos con el ámbito de la genética evolutiva, son el ejemplo de dos colaboraciones independientes que, ante la necesidad de investigar la variabilidad humana en ciertas regiones de interés, optaron por recurrir a herramientas desarrolladas apenas unos meses antes (SPSmart y ENGINES) en lugar de dedicar recursos al estudio interno de la diferenciación poblacional en dichas regiones de interés.

La necesidad de adaptación a nuevos retos de gestión de la información biológica queda patente con los trabajos presentados en el campo de la genética de poblaciones y en el de la neurogenética. En el primer caso se estudió la viabilidad real de manejar datos genotípicos a gran escala, midiendo los recursos necesarios a todos los niveles (computación, almacenamiento y consulta) y mostrando el uso de SPSmart como prueba de concepto. En el segundo caso se trató de aportar los conocimientos adquiridos del uso de bases de datos para variantes genéticas a una llamada a la formación de una comunidad internacional de neurogenetistas al amparo del Proyecto del

Varioma Humano (HVP), ya que éste tiene como fundamento la correlación de variantes genotípicas con la descripción lo más pormenorizada posible del fenotipo del donante de la variación. La particularidad de estas variantes en el campo de la neurogenética es que suelen estar relacionadas con enfermedades complejas, abarcando varios genes, y la descripción del fenotipo también es muy compleja, por lo que es fundamental una investigación detallada previa de las necesidades y de las posibles soluciones técnicas a desarrollar, así como la adopción y adaptación de herramientas ya disponibles para fines similares como pueden ser las bases de datos LOVD o MutaDataBase.

La aplicación de la bioinformática también resulta muy útil con volúmenes de datos mucho más pequeños que los hasta ahora descritos. La flexibilidad y potencia de los lenguajes de programación de alto nivel existentes hoy en día facilitan el desarrollo e implementación de algoritmos para ser empleados de manera automatizada. Los trabajos realizados en el campo del ADN mitocondrial reflejan la adecuación de los recursos de programación y computación disponibles a la obtención de conocimiento a partir de listados de variantes en crudo o de sus frecuencias alélicas en las poblaciones estudiadas. El artículo que describe la saturación de la diversidad haplotípica a partir de un número reducido de mtSNPs requirió no sólo de la elaboración de un programa en Perl similar en esencia a los ya desarrollados anteriormente para el procesado de datos genotípicos, sino también del uso del poder computacional del Centro de Supercomputación de Galicia (CESGA) para poder llevar a cabo un ingente número de simulaciones requeridas por el algoritmo propuesto. Por otra parte, el artículo que analiza el impacto de las migraciones modernas en Argentina desde un enfoque mitocondrial necesitó de la creación de un programa en R para el análisis estadístico de la compartición de haplotipos entre poblaciones, pudiendo así demostrar una suficiente diferenciación entre las muestras españolas y las italianas, lo cual confirmaba los resultados del análisis de mezcla poblacional realizados anteriormente.

V. CONCLUSIONES

1. La presencia de distintos formatos de datos hace necesario un esfuerzo por la estandarización y la normalización tanto en la elección de la tecnología a usar como en la generación de los resultados.

2. La rápida evolución de la genómica exige la creación de herramientas altamente flexibles para la realización de tareas muy concretas y que permitan manejar grandes volúmenes de datos.

3. Se requiere un acceso ordenado y racional a repositorios públicos de variabilidad humana en línea para la obtención de información útil.

4. La creación interna de repositorios estáticos de información pre-computada resulta una alternativa útil a la reiteración de análisis sobre almacenes masivos de variabilidad.

5. El tratamiento y almacenamiento de índices estadísticos de interés para la genética de poblaciones puede acometerse sobre todo el genoma como una iniciativa independiente requiriendo unos recursos moderados.

6. El uso de lenguajes de programación como Perl o R en el campo de la biología molecular permiten aportar nuevo conocimiento incapaz de ser generado anteriormente.

VI. REFERENCIAS

1. Freeman S, Herron JC: **Evolutionary analysis**, 2nd edn. Upper Saddle River, NJ: Prentice Hall; 2001.
2. Jobling MA, Hurles M, Tyler-Smith C: **Human evolutionary genetics : origins, peoples & disease**. New York: Garland Science; 2004.
3. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics**. *Genome Res* 2009, **19**(9):1639-1645.
4. Cummings MR: **Human heredity : principles and issues**, 5th edn. Pacific Grove [Calif.]: Brooks/Cole; 2000.
5. Cotton RG, Scriver CR: **Proof of "disease causing" mutation**. *Hum Mutat* 1998, **12**(1):1-3.
6. Condit CM, Achter PJ, Lauer I, Sefcovic E: **The changing meanings of "mutation:" A contextualized study of public discourse**. *Hum Mutat* 2002, **19**(1):69-75.
7. Butler JM: **Genetics and genomics of core short tandem repeat loci used in human identity testing**. *Journal of forensic sciences* 2006, **51**(2):253-265.
8. Edenberg HJ, Liu Y: **Laboratory methods for high-throughput genotyping**. *Cold Spring Harbor protocols* 2009, **2009**(11):pdb top62.
9. **Compgen Tool Suite** [http://csbio.unc.edu/CCstatus/index.py?run=Geneseek]
10. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature biotechnology* 2011, **29**(1):24-26.
11. Kim H, Dionne RA: **Lack of influence of GTP cyclohydrolase gene (GCH1) variations on pain sensitivity in humans**. *Molecular pain* 2007, **3**:6.
12. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics* 2005, **21**(2):263-265.
13. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE *et al*: **Estimating African American admixture proportions by use of population-specific alleles**. *Am J Hum Genet* 1998, **63**(6):1839-1851.
14. Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L *et al*: **Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina**. *Am J Phys Anthropol* 2001, **114**(1):18-29.

15. Burrell AS, Disotell TR: **Panmixia postponed: ancestry-related assortative mating in contemporary human populations**. *Genome Biol* 2009, **10**(11):245.
16. Wright S: **Evolution in Mendelian Populations**. *Genetics* 1931, **16**(2):97-159.
17. **Human Genome Project Information** [http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml]
18. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science (New York, NY* 2001, **291**(5507):1304-1351.
20. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.
21. Peacock E, Whiteley P: **Perlegen sciences, inc**. *Pharmacogenomics* 2005, **6**(4):439-442.
22. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A *et al*: **A human genome diversity cell line panel**. *Science (New York, NY* 2002, **296**(5566):261-262.
23. Rosenberg NA: **Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives**. *Ann Hum Genet* 2006, **70**(Pt 6):841-847.
24. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL *et al*: **Worldwide human relationships inferred from genome-wide patterns of variation**. *Science (New York, NY* 2008, **319**(5866):1100-1104.
25. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R *et al*: **Genotype, haplotype and copy-number variation in worldwide human populations**. *Nature* 2008, **451**(7181):998-1003.
26. Cardon LR, Abecasis GR: **Using haplotype blocks to map human complex trait loci**. *Trends in genetics : TIG* 2003, **19**(3):135-140.
27. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.
28. **International HapMap Project** [http://hapmap.ncbi.nlm.nih.gov/]
29. Siva N: **1000 Genomes project**. *Nature biotechnology* 2008, **26**(3):256.
30. **1000 Genomes** [http://www.1000genomes.org]

31. **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.
32. Mount DW: **Bioinformatics : sequence and genome analysis**. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press; 2001.
33. Attwood TK, Parry-Smith DJ: **Introduction to bioinformatics**. Harlow, England ; New York: Prentice Hall; 1999.
34. **NCBI-GenBank Flat File Release 191.0** [ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt]
35. Lesk AM: **Introduction to bioinformatics**, 2nd edn. Oxford ; New York: Oxford University Press; 2005.
36. **2Can Support Portal - Bioinformatics** [http://www.ebi.ac.uk/2can/bioinformatics/]
37. Galperin MY, Fernandez-Suarez XM: **The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection**. *Nucleic acids research* 2012, **40**(Database issue):D1-8.
38. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2012, **40**(Database issue):D13-25.
39. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic acids research* 2012, **40**(Database issue):D48-53.
40. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic acids research* 2001, **29**(1):308-311.
41. **dbSNP Summary** [http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi]
42. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2012**. *Nucleic acids research* 2012, **40**(Database issue):D84-90.
43. **Ensembl Species List** [http://www.ensembl.org/info/about/species.html]

Foto de portada: Museo do Pobo Galego – Escaleira (foto cedida por Turismo de Santiago)