Disentangling the population structure and evolution of the clam pathogen *Vibrio tapetis*.

Sabela Balboa, Asmine Bastardo, Jesús L. Romalde*

Departamento de Microbiología y Parasitología. CIBUS-Facultad de Biología. Universidad de Santiago de Compostela. 15782, Santiago de Compostela. Spain.

Running title: Population structure and evolution of *V. tapetis*.

Key words: *Vibrio tapetis*, MLST, population, evolution.

Submitted to: **FEMS Microbiology Ecology**, February 2013.

* Jesús L. Romalde. Departamento de Microbiología y Parasitología. CIBUS-Facultad de Biología. Campus vida. Universidad de Santiago de Compostela. C/ Lope Gómez de Marzoa s/n. 15782, Santiago de Compostela. Spain. Phone: +34 881816908. Fax: +34 881896938. E-mail: jesus.romalde@usc.es

SUMMARY

Vibrio tapetis is a fastidious slow-growing microorganism that causes the Brown Ring Disease (BRD) in clams. Recently, two subspecies for this bacterial pathogen have been proposed. We have developed a MSLT scheme and performed evolutionary studies of *V. tapetis* population using the great majority of isolates of *V. tapetis* obtained worldwide until now (30 isolates). *V. tapetis* constitutes a high polymorphic population, showing low diversity indexes and some genetic discontinuity among the isolates. Mutation events are more frequent than recombination, although both are approximately equally important for genetic diversification. In fact, the divergence between subspecies occurred exclusively by mutation but the diversity observed among isolates of the same subspecies appeared to be generated mostly by recombination. Between the subspecies, genetic distance is very high and almost no recurrent gene flow exists. This pathogen displays a non-clonal population structure with an ancient spatial segregation population and some degree of geographical isolament, followed by a population expansion, at least for *V. tapetis* subsp. *tapetis*. A database from this study was created and hosted on publmlst.org, being freely available (http://pubmlst.org/vtapetis/).

INTRODUCTION

A population is defined as an assemblage of individuals of the same species in a more or less geographically restricted area (Hanage *et al*., 2006; Sikorski, 2008). The evolutionary fate of a population is defined by mutation, recombination, natural selection and genetic drift. Combinations of these interactions can be schematically assigned to different evolutionary models (Cohan & Perry, 2007). Recombination and/or mutation play a major role in determining genetic variability and establishing different ways of species evolution (i.e., diversifying, directional, purifying). Therefore, their study is a fundamental prerequisite for better understanding the epidemiology and the disease cycle of a pathogen. On the other hand, dispersal, geographic isolation, past range restrictions and expansions, drift processes, founder events and selection leave their signature in the lineage composition of contemporary populations (Avise, 2000).

Multilocus Sequence Typing (MLST) is a powerful technique for inferring the genetic relationships among bacteria, at both interspecific and intraspecific levels, by using core housekeeping genes (Maiden, 2006). This technique was successfully applied for many pathogenic and commensal bacterial species, and is currently regarded as the gold-standard in molecular typing being able to even replace Pulsed Field Gel Electrophoresis (PFGE) and Multilocus Enzime Electrophoeresis (MLEE) (Maiden, 2006). Moreover, when coupled with the appropriate statistical tests and algorithms, it can help to unravel evolutionary dynamics (i.e., population evolution).

Vibrio tapetis is the causative agent of Brown Ring Disease (BRD) in adult clams. The disease has been detected for years in Europe and, more recently, in Asia. Some typing studies have been carried out for this species. At genetic level, differences were first detected in the plasmid content and ribotypes of the strains (Paillard, 2004; Borrego *et al*., 2011). Subsequently, on the basis of studies with PCR-based typing techniques, three major groups associated with the host origin were established (Rodríguez *et al*., 2006). More recently, our group has carried out a MLSA study and described a new subspecies for this pathogen, *V. tapetis* subsp. *britannicus* (Balboa & Romalde, 2013).

In this work, we have developed a MLST scheme and performed evolutionary studies on the population of *V. tapetis*, including both subspecies, to achieve a better knowledge of the distribution and the evolution of this pathogen.

MATERIALS AND METHODS

Bacterial Strains

A total of 30 strains of *V. tapetis* were used in this study (Table 1). These isolates were obtained from different hosts and were temporal and geographically diverse. To our knowledge, they constitute the great majority of *V. tapetis* strains isolated until now worldwide. The collection includes strains belonging to the two subspecies and representatives for the three genetic groups previously described for the species (Rodríguez *et al*., 2006; Balboa & Romalde, 2013). Bacteria were routinely grown on marine agar (MA, Pronadisa) and incubated at 15ºC for 72h.

DNA Purification and Gene Amplification and Sequencing

Chromosomic DNA was extracted from pure bacterial cultures using the Insta-Gene matrix (Bio-Rad, Madrid, Spain) as previously described by Romalde *et al.* (1999). Ten gene loci were selected for MLST analysis, including (*atp*A (α subunit of ATPase), *fst*Z (cell division protein), *gap*A (glyceraldehydes-3-phosphodehydrogenase), hsp60, (60-KDs heat shock protein), *pyr*H (uridyl monophosphate kinase), *rct*B (Replication origin-binding protein), *rec*A (recombinase A), *rpo*A (α subunit of RNA polymerase) and *rpo*D (RNA polymerase sigma factor), *top*A (topoisomerase I). PCR conditions and primer sequence are described in http://pubmlst.org/vtapetis/.

DNA sequences were performed as described Balboa & Romalde (2013). Briefly, amplified fragments were sequenced in both directions using GenomeLab DTCS-Quick Start Kit (Beckman Coulter, Ireland) in a Beckman coulter sequencer (Beckman Coulter, Ireland). *fst*Z and *rpo*D sequences were performed at Stabvida Lcd. (Portugal). Sequence data analysis was performed with DNAstar Seqman program (Lasergene, USA) and the DNA sequences were aligned and translated into amino acid sequences using the MEGA5 software (Tamura *et al*., 2011).

Genetic Analysis of the Population

According to MLST standard (Urwin & Maiden, 2003), a distinct ST number was attributed to each distinct profile of alleles types (ATs; particular alleles at particular loci).

A descriptive analysis of nucleotide and allele diversity was performed for the individual genes and the concatenated sequence (5826 bp in length). Thus, the average

gene diversity was calculated using Nei's index of diversity (H) (Nei, 1978) and the type of selection was determined by the ratio of mean non-synonymous substitutions per non-synonymous site/mean synonymous substitution per synonymous site (d_N/d_S ratio) according to Nei and Gojobori (Nei & Gojobori, 1986) method with the Jukes-Cantor correction.

Number of polymorphic sites and the index of association (I_A) was performed using the START2 program (http://pubmlst.org/software/analysis/start2/). The "standardized" index of association (I^S_A) from the allelic profile dataset of the whole collection as well as for the STs defined in the study were calculated (Maynard-Smith *et al*., 1993).

Population Structure

The clonal structure of the population was determined using eBURST v3.0 program http://eburst.mlst.net. On a different approach, a dot graph was constructed using a 50% majority rule consensus tree was done using ClonalFrame version 1.1 (Didelot & Falush, 2007) using 10 independent runs of MCMC (Markov Chain Monte Carlo). On the other hand, a phylogenetic reconstruction was performed for the whole collection of isolates by Neighbour joining method, using Kimura-2-parameters method and bootstrap of 1,000 replicates, with MEGA5 program (Tamura *et al*., 2001).

The relationships among haplotypes were displayed in a network constructed by Median Joining method using the programs DNA Alignment 1.3 and NETWORK 4.1 (Fluxus-engineering) (Bandelt *et al*., 1999).

Recombination

Putative recombination events were detected using three different approaches. The minimal number of recombination events R_{min} was computed on biallelic site only by using DnaSP5 software (Librado & Rozas, 2009). The identification of recombination sequences and breakpoints were identified using RDP4 Beta (Martin *et al*., 2010). Finally, SplitsTree versión 4.0 (Huson & Bryant, 2006) was employed to calculate the phi test for recombination for the whole strain collection, and to generate a Split-tree for each individual locus as well as for the concatenated sequence.

Clonal Genealogy

The clonal genealogy of the whole population was inferred with ClonalFrame (Didelot & Falush, 2007) using 100000 MCMC. Mixing and convergence properties were found

to be satisfactory based on comparisons of independent runs. For presentation purposes, the estimated genealogies were represented in a 50% majority rule consensus tree.

Demographic History and Gene Flow

The average number of nucleotides per site (π) and the number of segregating sites (θ) were computed using both DnaSP5 (Librado & Rozas, 2009) and Arlequin v3.5 (Excoffier & Lischer, 2010) programs that were also employed to calculate the values of selection tests based on polymorphism distribution. The tests of Tajima (Tajima, 1989) and Fu and Li (Fu & Li, 1993) F* and D* were used to confirm the hypothesis that all mutations are selectively neutral (Kimura, 1983).

Fu's F_S (Fu, 1997) and Strobeck's S (Strobeck, 1987) statistics were calculated to assess the haplotype structure based on the haplotype frequency distribution. The later is also considered as an index of admixture, characteristic of populations originated from multiple sources. Onsis R_2-test is based on the difference between the number of singleton mutations in a population and the average number of nucleotide differences in that population, with lower values of the statistics expected after a recent episode of severe population growth (Ramos-Onsins & Rozas, 2002).

The gene flow between the two subspecies was assessed by the fixation index (F_{ST}), which measures the differentiation among individuals within population (Wright, 1951). Nei's $G_{ST,}$ to analyze the population differentiation based on differences in allele frequencies, and $N_{M,}$ to estimate gene flow on the basis of G_{ST} (Nei, 1978), were calculated with the Arlequin v3.5 software. To estimate the correlation between genetic (determined using F_{ST}) and geographical distances, a Isolation By Distance (IBD) test was performed using the program IBDWS (Jensen *et al*., 2005). Non-parametric Mantel test was carried out with the full dataset to analyse non-random associations between genetic and geographical distances. Genetic distances were computed using Slatkin's (Slatkin, 1993) similarity measure: M= ([1/ Fst] 21)/4. The significance of the Mantel test was determined by a permutation test of n=1000. Isolates were grouped in four groups according to their country of origin. Geographical distances were measured using geographical coordinates.

The software packages DnaSP5 and Arlequin v3.5 were also used to calculate frequency distributions of number of mismatches between pairwaise sequences, and to model the expected distributions under the demographic scenarios of constant and decline growth of population size, population expansion and spatial expansion. Model fitness was

examined by raggedness *r* statistic, which indicates the smoothness of the mismatch distribution (Rogers and Harpending, 1992).

RESULTS

Nucleotide Diversity

A total of 5,826 bp at 10 loci were sequenced from 30 strains of *V. tapetis* (Table 1). The in-frame DNA fragments ranged in size from 516 bp (*pyr*H) to 609 bp (*fst*Z) (http://pubmlst.org/vtapetis/). A descriptive analysis of nucleotide and allele diversity is shown in Table 2. Four hundred and fifty polymorphic sites were recorded, being only 8 of them (located in genes *fst*Z, *gap*A, hsp60, *rct*B and *rec*A) triallelic. The locus with more polymorphic sites was *rct*B (93 biallelic and 2 triallelic sites), in contras with *rpo*A in where only 7 sites were found polymorphic, being all of them biallelic. The number of alleles observed for each locus varied from 3 (*rpo*A) to 9 (*rec*A and *top*A), being identified a total of 72 different alleles, 34 of which were synapomorphic (i.e. shared and derived). The most frequently found allele was allele 1 in all loci analyzed, with percentages ranging from 56.7% for *rct*B, *rec*A, *rpo*D and *top*A and 80% for *fst*Z. The most polymorphic loci were *rec*A and *top*A that showed 9 different alleles, although those genes, together with *atp*A, *gap*A, *pyr*H and r*po*D retained more synapomorphic alleles. When the 72 alleles were combined for the 30 isolates of *V. tapetis*, a total of 10 sequence types (STs) could be identified (Table 1), indicating high diversity in the population.

The number of polymorphic sites was overall low and found to be 7.7% for the concatenated sequences. The average number of nucleotides per site (π) and the number of segregating sites (θ) were 0.015 and 0.020, respectively. Values of π and θ for each individual gene in the whole population and in each subspecies are shown in Tables 2 and S1.

Values of test for selection (d_N/d_S ratio) calculated for the concatenated sequence and for each individual locus were <1 (in particular *gap*A gave a $d_N/d_S = 0$), indicating that all these loci were subjected to positive selection, i.e., under stabilizing or purifying selection.

Linkage disequilibrium was estimated with the I_A parameter from allelic profiles with randomized data sets, rendering values of 6.300 (P = 0.000), for whole strain collection, and 1.584 (P = 0.000) when only the STs were analyzed. Standardized I^S_A values,

calculated to avoid dependence on the number of loci, were lower, 0.700 (P = 0.000)(whole strain collection) and 0.176 (P = 0.000) (STs), but still significantly different from zero. Similar results were obtained *V. tapetis* subsp. *tapetis*, I^S_A = 0.649 (P = 0.000) analysing all the isolates, and I^S_A = 0.111 (P = 0.000) obtained in the STs analysis. Hence, a non-random distribution of alleles in the *V. tapetis* population was observed, although some recombination may also occur.

Population Structure

Seven STs were represented by a single isolate, while the other three included 2, 4 and 17 isolates. It was no possible to identify any clonal complexes by eBURST with the most stringent criteria (9/10 shared alleles). Thus, all STs were considered as singletons with no apparent clonal relationship to each other, which probably reflects some genetic discontinuity among the strains (Fig. S1A). When the strains were visualized in a dot graph constructed using a 50% majority rule consensus tree (Fig. S1B), some missing intermediates showed up and the existence of some putative clonal complexes was evidenced. Therefore, *V. tapetis* population displays a level of genetic structure that cannot be detected using conventional MLST. The existence of these intermediates was also evident in the network constructed by Median Joining algorithm where these putative variants constituted the linkage among the established STs (Fig. 1).

To explore the phylogenetic relationships among 30 *V. tapetis* isolates, a phylogenetic tree was performed using the Neighbor joining (NJ) algorithm. This tree shows two clearly differentiated clusters (Fig S2), each corresponding with one subspecies.

Recombination

The phi test (Pairwise Homoplasy Test) did not indicated significant statistical evidences for recombination in any gene or the concatenated sequences. However, some putative recombination events were detected with the minimal number of recombination events R_{min}. Specifically, one event of recombination was found in *atp*A, *pyr*H and *rct*B, four events were found in *rec*A and five in *rpo*D. Using RDP4 software, five events of recombination were found in *atp*A and one in *fst*Z.

The split graph based on the concatenated sequences displayed an interconnecting network structure rather than a single bifurcation tree (Fig. 2). This structure indicates the presence of homoplasies (repeated mutations at the same nucleotide sites at independent links of a phylogenetic tree) that are most probably a result from intragenic

recombination. The network structures of those genes showing recombination can be also seen in the splits graphs constructed for each individual gene (Fig S3).

ClonalFrame analyses estimated that mutation was considerably more frequent than recombination within this dataset (ρ/θ = 0.084, 95% CI: 0.045-0.183) but they were approximately equally important for genetic diversification (r/m = 1.470, 95% CI: 0.842-2.724). The clonal genealogy confirmed the mutational character of the divergence between subspecies (Fig. 3). Sequence variations on the oldest branch of the tree (node A) seems to arise exclusively by mutational events while in the younger branches recombination imports are common although mutational events are also present. It is noteworthy that intermediate-evolutive branches (nodes B and H) are the ones that showed more recombinatorial events, suggesting that the diversity within each subspecies was mostly generated by recombination.

Gene Flow

Isolates from the two subspecies differed between 6.8 and 7.1% in the concatenated sequence. These nucleotide differences were accompanied by high population specific fixation index values (F_{ST} = 0.952, P = 0). Such a value, together with a G_{ST} = 0.105 illustrates in full view the magnitude of the separation of those groups of isolates. Moreover, a N_M of 0.07 indicates that the genetic exchange between both groups is almost non-existent.

Demographic Trends

All the statistics and site frequent spectra analysed were consistent with a deviation from the neutral mode, although some incongruences among the different indexes were observed (Table 3). Tajima´s D showed negative although non-significant value, reflecting an excess of rare polymorphisms in the population, which is consistent with either positive selection or an increase in population size. On the contrary, Fu and Li F* and D* showed positive, and most of them, non-significant values reflecting an excess of intermediate-frequency alleles in the population that can be the result of either balancing selection or population bottleneck (i.e. population subdivision). The positive values of Fu´s F_S constituted an evidence for a deficiency of alleles, as would be expected from a recent population bottleneck or from overdominant selection. Onsis R_2 values close to zero supported the population expansion pattern, while Strobeck´s S´ statistic indicated the non-admixture of the population.

Regarding to the subpopulations, the indexes were calculated only for *V. tapetis* subsp. *tapetis* that showed evidences for positive selection or an increase in population size regardless the index employed, as well as evidences for admixture in the population (Table S2).

It is noteworthy that some genes, namely *atp*A and *fst*Z, showed a deviation for the general trend of the population. The same was observed for *atp*A, *rct*B and *rpo*D genes in the subspecies *tapetis*.

The mismatch distribution was not significantly different from those expected under either spatial expansion or population size expansion models (Fig. 4). The observed mismatch distribution was multimodal and showed four distinct peaks. The distribution fitted with the expected model for spatial expansion, although Harpending´s raggedness index (r = 0.176) was only significant under a population expansion model (P = 0.01). When the pairwise mismatch distributions were calculated for *V. tapetis* subsp. *tapetis* evidence for population expansion was apparent (r = 0.086, P = 0.5).

The spatial structure of allele frequencies was assessed using Moran´s I statistic, a coefficient of spatial dependence for pairs of samples separated into distance classes according to their country of origin (4 distance classes). The resulting correlogram of spatial dependence (Fig. 5) revealed that the spatial correlation decreased with increasing pairwise distances, being in addition lower than it would be expected in a random dataset at two of the larger distance classes. This finding indicates that allele frequencies in neighbouring sites tend to be more similar when they are geographically closer. Nonetheless, the Mantel test (M), for correlation between genetic similarity and geographic distance using the shortest distance over land, showed non-significant negative correlation for the full dataset (Z = 4463.971, r = 0.180□, P = 0.657).

DISCUSSION

The MLST approach of this study was motivated by a previous phylogenetic work performed by our group using, to our knowledge, the great majority of isolates of *V. tapetis* obtained worldwide until now (Balboa & Romalde, 2013). This study led to the description of two subspecies for *V. tapetis* associated to the geographical origin of the isolates. Thus, the subspecies *britannicus* is restricted to the British Isles, while subspecies *tapetis* comprises isolates from all other locations. The variability within each subspecies is associated to host origin, in agreement with the results previously obtained by Rodríguez *et al.* (2006). In this work, we tried to shed light on the

emergence and evolution of *V. tapetis* since its genetic structure is very intriguing, showing high heterogeneity with genetic discontinuity.

The first obvious achievement was the polymorphism found among the isolates. Thus, and having in mind the scarce number of isolates, a total of 450 polymorphic sites were described among the 5826 bp. Despite this fact, the nucleotide diversity per site (π) and the average number of nucleotide differences per site (θ) indicated a relative low degree of sequence diversity. This fact has been partially attributed in other aquatic bacterial pathogens to the selection of particular strains in the collection (Yan *et al*., 2011; Bastardo *et al*., 2012), as occur in this case where ST1 is formed by 17 strains while most of STs are constituted by a single isolate. On the other hand, the average of π in each subspecies is ten-fold lower than in the species as a whole, suggesting that barriers to genetic exchange between subspecies could be formed in a DNA sequence homology-dependent manner (Tanabe & Watanabe, 2011). Only 51 out of the 450 polymorphic sites corresponded to non-synonymous changes, consistent with a strong selection against amino acid changes, as typically observed for housekeeping genes (Pérez-Losada *et al*., 2006).

The lack of clonal complexes reflected some genetic discontinuity among the strains, but it has been demonstrated that the eBURST can only identify recent divergence of clones on the basis of shared allele at multiple loci. Therefore, it can produce incongruent results if the role of recombination has increased in recent times (Feil *et al*., 2004). The dot graph generated by ClonalFrame confirmed the existence of missing intermediates in the population, suggesting that the two subspecies diverged in a long-term evolutionary consequence rather than a recent diversification. On the other hand, it is important to emphasize at this point that the small amount of isolates belonging to this species may be distorting the structure of the population. The lack of isolates can be probably related to the fact that *V. tapetis* is a slow-growing difficult-to-isolate pathogen (Balboa *et al*., 2012) that most of times, even when is detected by indirect procedures such as PCR, it cannot be recovered on culture media (Drummond *et al*., 2007; Paillard, 2004; Borrego *et al*., 2011).

The differentiation between subspecies together with the lack of clonal relationships exhibited by the STs could lead to think that they are two different species. However, the NJ tree supported a high phylogenetic relationship, showing all branches more than 87% bootstrap. In addition, our previous study of MLSA demonstrated a clear monophyletic nature of *V. tapetis* as bacterial species and further DDH experiments

confirmed that the three isolates of *V. tapetis* subsp. *britannicus* belong indeed to *V. tapetis* species (Balboa & Romalde, 2013).

Mutation and recombination are equally important for genetic diversification of *V. tapetis*. As it was demonstrated with the clonal genealogy, the separation of the two subspecies occurred exclusively by mutation, and these mutational events were quantified in 370 with the construction of the median joining network. Such amount of mutational events together with the differences between the I^S_A values for the whole population and for STs, indicate that although recombination occurs at a considerably frequency, it is not enough for allele random association (de las Rivas *et al*., 2006). On the other hand, the amount of recombination showed by Clonal genealogy could not be demonstrated using other methods such as phi test, which can be probably due to the fact that the genes used in this study are under purifying selection, purging any sequence variation required for the detection of intragenic recombination by most algorithms (Feil *et al*., 2004)

Despite the limited sequence diversity evidenced by the low π and θ values, the majority of STs occurred as singletons, which suggests that the sequence diversity was magnified by extensive intraspecies recombination, generating genotypes which can be readily distinguished by sequence analysis (Brisse *et al*., 2009).

The existence of the high genetic distance between the two subspecies is confirmed by F_{ST} and G_{ST} indexes and little or no recurrent gene flow (Nm<1), explaining the lack of recombinational events among strains of the two different subspecies. Such a genetic distance may be the result of vicariance or obstruction by natural barriers but, although the inference of range expansion was well supported by Moran´s index, the isolation by distance (IBD) was not significant.

Despite the high I^S_A showed by the entire population, the lack of IBD together with the low nucleotide diversity and the frequent recombination among strains within each subspecies evidence a non-clonal population structure for *V. tapetis*. Moreover, the appearance of the dominant ST1 that, together with ST2, are the only STs of demonstrated pathogenicity (Paillard, 2004) reminds to the emergence of highly adaptive bacterial clones in epidemic populations, where a small number of successful clones dramatically increase their population size with respect to the other genotypes (Smith *et al*., 1993).

The frequency distribution of pairwise mismatches of neutral-evolving loci showed multimodal distribution with high degree of raggedness, as expected for a population in

constant expansion (Slatkin & Hudson, 1991; Rogers & Harpending, 1992; Ray *et al*., 2003). Tajima's D* and Onsis R_2 (Tajima, 1989; Ramos-Onsis & Rozas, 2002) values confirmed the evidences for population expansion while Fu´s F_S (Fu, 1997) indicates that the population suffered a bottleneck, for example a subdivision of the population. Although it is well known that F_S is a more sensitive indicator for population expansion and genetic hitchhiking than D*, it has also been described that the former statistic is highly affected by recombinational events (Fu, 1997). However, the powerful Onsis R_2 test, which is particularly suited for small size samples with recombination, also supports the population expansion detected by the mismatch distribution. This apparent contradiction may be explained by the lower statistical power of these tests compared to mismatch distribution to detect spatial expansions when local demes exchange only few migrants (Ray *et al*., 2003; Städler *et al*., 2009). F* and D* values are positive, indicating a deficit of mutation in the external branches. This statistic reflects an unusually amount of ancient alleles probably maintained by balancing selection. Nonetheless, it is impossible to distinguish between selection and expansion as the cause of negative F_S, D* and F* values. It is noteworthy that some genes showed a deviation from the general trend of the population, being indicative of selective pressures unique to one population (i.e., local adaptation), different demographic histories, spurious results, or most likely some complex combination of all of these factors. The evidences for admixture found in *V. tapetis* subsp. *tapetis* population are probably a result of the introduction of new genetic lineages into the population.

These statistics in combination with mismatch distribution indicate an ancient spatial segregation of the population with some degree of geographical isolation, that probably caused the apparition of the two subspecies, followed by a population expansion, at least for *V. tapetis* subsp. *tapetis*.

A database from this study was created and hosted on publmlst.org (Jolley & Maiden, 2010) and is freely available in Internet (http://pubmlst.org/vtapetis/). This database will be useful in future works studying evolutionary and epidemiological relationships between strains and for unambiguous comparison of data generate from laboratories around the world.

Acknowledgements

This work was supported in part by grant AGL2010-18438 from the Ministerio de Ciencia e Innovación (Spain). S.B. acknowledges the Ministerio de Ciencia e Innovación (Spain) for a research fellowship.

The authors have no conflict of interest to declare.

REFERENCES

Avise JC (2000) Phylogeography: the history and formation of species. Harvard University Press, Harvard.

Balboa S & Romalde JL (2013) Multilocus sequence analysis of *Vibrio tapetis*, the causative agent of Brown Ring Disesase. Description of *Vibrio tapetis* subsp. *britannicus* subsp. nov. *Syst Appl Microbiol* doi: 10.1016/j.syapm.2012.12.004.

Balboa S, Dieguez, AL, Doce A, Barja JL & Romalde JL (2012) Evaluation of different culture media for the isolation and growth of the fastidious *Vibrio tapetis*, the causative agent of brown ring disease. *J Invertebr Pathol* 111: 74-81.

Bandelt HJ, Fortser P & Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16: 37–48.

Bastardo A, Ravelo C & Romalde JL (2012) Multilocus sequence typing reveals high genetic diversity and epidemic population structure for the fish pathogen *Yersinia ruckeri. Environ Microbiol* 14: 1888-97.

Borrego JJ, Romalde JL & Castro D (2011) La enfermedad del anillo marrón. In *Enfermedades de moluscos bivalvos de interés en acuicultura.* (Figueras A & Novoa B, eds) pp 149-179. Fundación OESA.

Brisse F, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L & Grimont P (2009) Virulent clones of *Klebsiella pneumoniae*: identification and evolutionary scenario based on genomic and phenotypic characterization. *PLoS One* 4: e4982.

Cohan FM & Perry EB (2007) A systematics for discovering the fundamental units of bacterial diversity. *Curr. Biol* 17: R373-R386.

de Las Rivas B, Marcobal A & Muñoz R (2006) Development of a multilocus sequence typing method for analysis of *Lactobacillus plantarum* strains. *Microbiology* 152:85-93.

Didelot X & Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175: 1251-1266.

Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci USA* 88(16): 7160-7164.

Drummond LC, Balboa S, Beaz R, Mulcahy MF, Barja JL, Culloty SC & Romalde, J.L. (2007) The susceptibility of Irish-grown and Galician-grown Manila clams, *Ruditapes philippinarum*, to *Vibrio tapetis* and brown ring disease. *J Invertebr Pathol* 95: 1-8.

Excoffier L & Lischer JEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Res* 10: 564-567.

Feil EJ, Li, BC, Aanensen DM, Hanage WP & Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clústers of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* 186: 1518-1530.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915–925.

Fu Y-X & Li, W-H (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693-709.

Gordon DM, Bauer S & Johnson, JR (2002) The genetic structure of *Escherichia coli* populations in primary and secondary habitats. *Microbiology* 148: 1513–1522

Hanage WP, Spratt BG, Turner KM & Fraser C (2006) Modelling bacterial speciation. Philos Trans R Soc Lond B Biol Sci 361: 2039-2044.

Huson DH & Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254-267.

Jensen JL, Bohonak AJ & Kelley ST (2005) Isolation by distance, web service. *BMC Genet* 6: 13.

Jolley KA & Maiden MCJ (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11: 595.

Kimura M (1983) *The neutral theory of molecular evolution*. Cambridge University Press. Cambridge.

Librado P & Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

Maiden MC (2006) Multilocus sequence typing of bacteria. *Ann Rev Microbiol* 60: 561-588.

Martin DP, Lemey P, Lott M, Moulton V, Posada D & Lefeuvre, P. (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26: 2462–2463.

Maynard-Smith J, Smith NH, O'Rourke M & Spratt BG (1993) How clonal are bacteria? *Proc Natl Acad Sci USA* 90: 4384-4388.

Nei M (1978) Estimation of average heterozigosity and genetic distance from a small number of individuals. *Genetics* 89: 583-590.

Nei M & Gojobori T (1986) Simple methods for estimating the numbers of synonymous and non synonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.

Paillard C (2004) A short-review of brown ring disease, a vibriosis affecting clams, *Ruditapes philippinarum* and *Ruditapes decussatus*. *Aquat Living Resour* 17: 467-475.

Pérez-Losada M, Browne EB, Madsen A, Wirht T, Viscidi RP & Crandall KA (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* 6: 97-112.

Ramos-Onsins SE & Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Mol Biol Evol* 19: 2092–2100.

Ray N, Currat M & Excoffier L (2003) Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* 20: 76-86.

Rodríguez JM, López-Romalde S, Beaz R, Alonso MC, Castro D & Romalde JL (2006) Molecular fingerprinting of *Vibrio tapetis* strains using three PCR-based methods: ERIC-PCR, REP-PCR, and RAPD. *Dis Aquat Org* 69: 175-183.

Rogers AR & Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9: 552-569.

Romalde JL, Magariños B, Villar C, Barja JL & Toranzo AE (1999) Genetic analysis of turbot pathogenic *Streptococcus parauberis* strains by ribotyping and random amplified polymorphic DNA. *FEMS Microbiol Lett* 179: 297-304.

Sikorski J (2008) Populations under microevolutionary scrutiny: what will we gain? *Arch Micriobiol* 189: 1-5.

Slatkin M (1993) Isolation by distance in equilibrium and non equilibrium populations. *Evolution* 47: 264–279.

Slatkin M & Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555-62.

Smith JM, Smith NH, O´Rourke M & Spratt BG (1993) How clonal are bacteria? *Proc Natl Acad Sci USA* 90:4384-4388.

Städler T, Haubold B, Merino C, Stephan W & & Pfaffelhuber P (2009) The impact of sampling schemes on the site frequency spectrum in subdivided populations. *Genetics* 182: 205-216.

Strobeck C (1987) Average number of nucleotide differences in a simple from a single subpopulation: a test for population subdivisión. *Genetics* 117: 149-153.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. *Genetics* 123: 585-595.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using Maximum Likelihood, evolutionary distance, and Maximum Parsimony methods. *Mol Biol Evol* 28: 2731-2739.

Tanabe Y & Watanabe MM (2011) Local Expansion of a Panmictic Lineage of Water Bloom-Forming Cyanobacterium *Microcystis aeruginosa PLoS One* 6(2): e17085.

Urwin R & Maiden MC (2003) Multilocus sequence typing: a tool for global epidemiology. *Trends Microbiol* 11: 479-87.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics* 15: 323-354.

Yan Y, Cui Y, Han H, Xiao X, Won HC, Tan Y, Guo Z, Liu X, Yang R & Zhou D (2011) Extended MLST-based population genetics and phylogeny *of Vibrio parahaemolyticus* with high levels of recombination. *Int J Food Microbiol* 145:106-12.

Figure legends

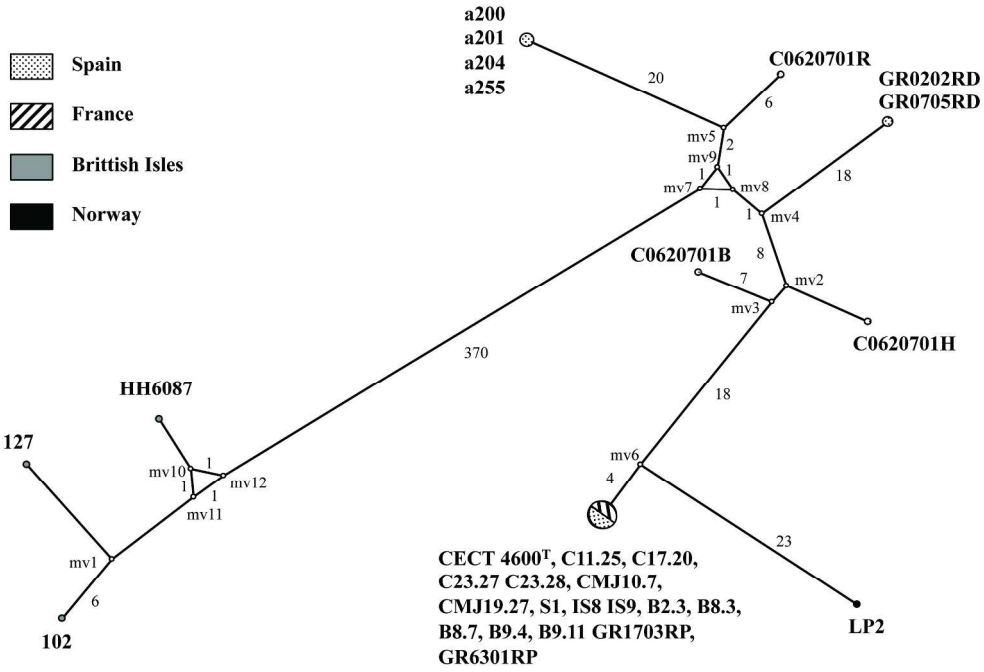
Fig. 1. Median Joining network based on the concatenated sequence of the 10 housekeeping genes. Size of the circle are proportional to the size of each ST. Next to each circle are shown the strains belonging to each ST. Color of the circles indicates the country origin of the isolates. Open circles represent putative intermediated sequences. Numbers indicates nucleotidic substitutions.

Fig. 2. Neighbor-Net graph based on the concatenated sequence of the 10 housekeeping genes of the 30 isolates of *V. tapetis* showing a bushy network structure indicative of pervasive recombination.

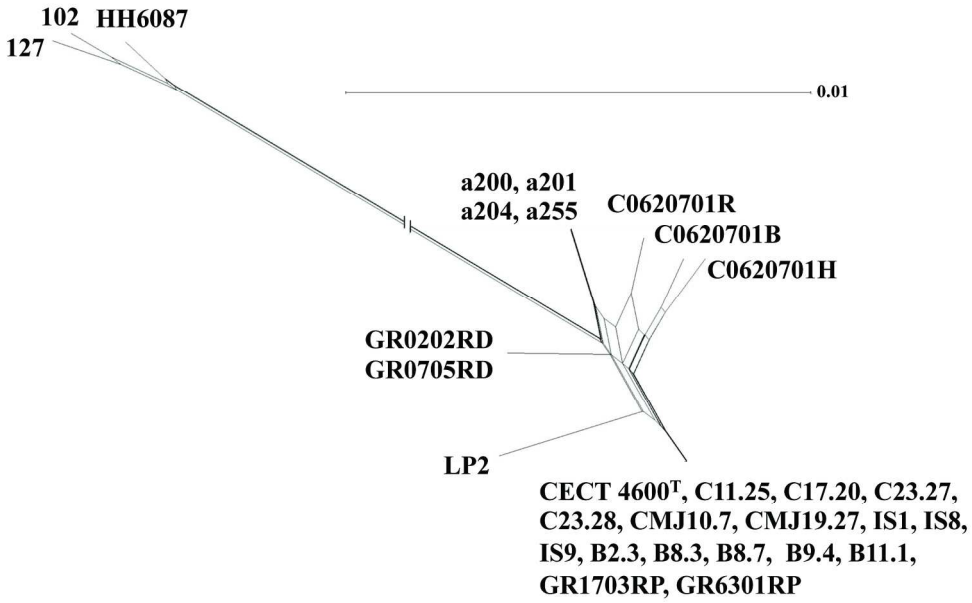
Fig. 3. Clonal genealogy (A) and evolutionary events on each node (B) reconstructed following the majority rule from the concatenated sequence. The designations for the nodes (A-L) in part (A) are congruent with the designations of rows in part (B). Each column in part B delineates the extent of each gene constituting the concatenated. The height of the red line indicates the probability of recombination on a scale from 0 (row bottom) to 1 (row top). Each nucleotide substitution is represented by a black cross.

Fig. 4. Mismatch distribution of the whole population of *V. tapetis*. Bars indicates observed frequency of each pairwise difference. The line graphs correspond to the expected distribution of the mismatches following four different population expansion models.

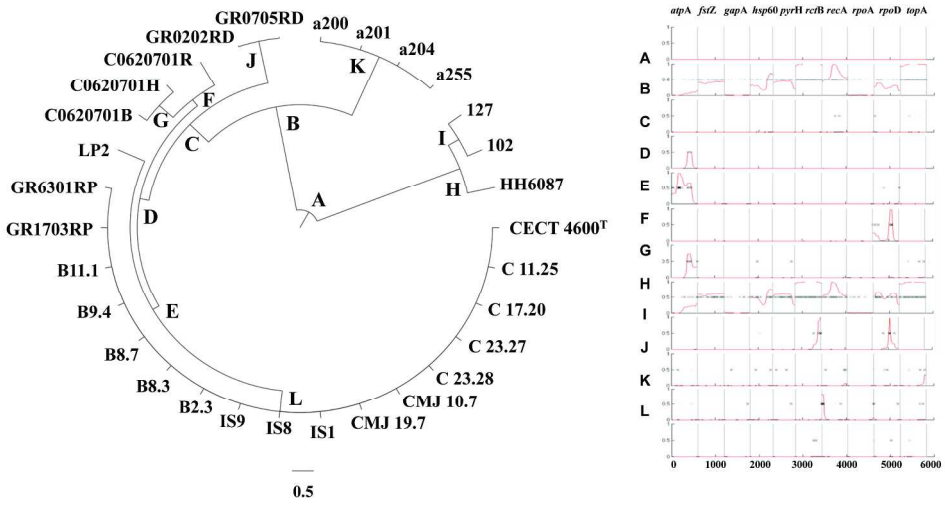
Fig. 5. Moran´s correlogram of individual allele frequencies. Moran´s *I*, and index of spatial autorrelation was plotted for individual allele frequencies across 4 qually sized distances.



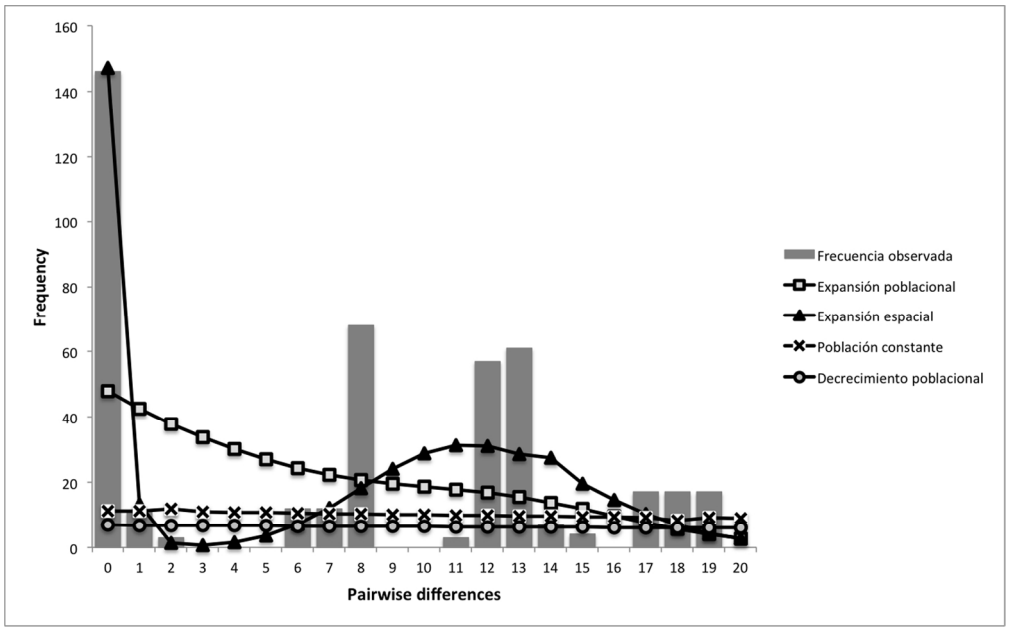
Median Joining network based on the concatenated sequence of the 10 housekeeping genes. Size of the circle are proportional to the size of each ST. Next to each circle are shown the strains belonging to each ST. Color of the circles indicates the country origin of the isolates. Open circles represent putative intermediated sequences. Numbers indicates nucleotidic substitutions.
239x163mm (300 x 300 DPI)



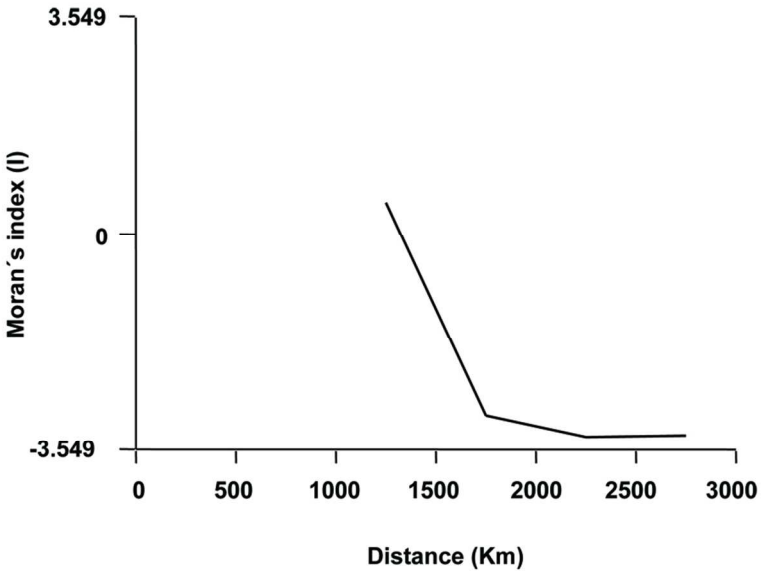
Neighbor-Net graph based on the concatenated sequence of the 10 housekeeping genes of the 30 isolates of V. tapetis showing a bushy network structure indicative of pervasive recombination.
189x116mm (300 x 300 DPI)



. Clonal genealogy (A) and evolutionary events on each node (B) reconstructed following the majority rule from the concatenated sequence. The designations for the nodes (A-L) in part (A) are congruent with the designations of rows in part (B). Each column in part B delineates the extent of each gene constituting the concatenated. The height of the red line indicates the probability of recombination on a scale from 0 (row bottom) to 1 (row top). Each nucleotide substitution is represented by a black cross.
235x125mm (300 x 300 DPI)



Mismatch distribution of the whole population of V. tapetis. Bars indicates observed frequency of each pairwise difference. The line graphs correspond to the expected distribution of the mismatches following four different population expansion models.
150x93mm (300 x 300 DPI)



Moran´s correlogram of individual allele frequencies. Moran´s I, and index of spatial autorrelation was plotted for individual allele frequencies across 4 qually sized distances.
99x82mm (300 x 300 DPI)

Disentangling the structure and evolution of the population of the clam pathogen *Vibrio tapetis*.

Sabela Balboa, Asmine Bastardo, Jesús L. Romalde

Supplementary Material

Table S1: Average number of nucleotide differences per site for two randomly selected strains (π) and average number of nucleotide differences per site (θ) for each individual gene and the concatenated sequence in each subspecies

Gene	*V. tapetis* subsp. *tapetis*		*V. tapetis* subsp. *britannicus*	
	π	θ	π	θ
atpA	0.00752	0.00679	0.00089	0.00073
fstZ	0.00093	0.00123	0.00349	0.00286
gapA	0.00120	0.00171	0	0
hsp60	0.00150	0.00346	0.00203	0.00167
pyrH	0.00314	0.00396	0.00310	0.00255
rctB	0,00242	0,00259	0,00622	0,00511
recA	0.00646	0.00711	0.00265	0.00218
rpoA	0.00034	0.00049	0	0
rpoD	0.00930	0.00850	0.00486	0.00399
topA	0.00715	0.00735	0.00240	0.00197
MLSA	0.00523	0.00547	0.00200	0.00165

Table S2: Summary statistics of demographic history of the subspecies *V.tapetis* subsp. *tapetis*

Gene	D	D*	F*	F_S	S	R_2
atpA	0.608	0.772	0.810	1.462	0.506	0,203
fstZ	-1.132	-1.155	-1.195	-0.858	0.934	0,236
gapA	-1.358	-1.427	-1.522	-1.798	0.974	0,165
hsp60	-1.719	-1.435	-1.768	-1.377	0.927	0,091
pyrH	-1.024	-0.969	-1.066	-2-019	0.980	0,153
rctB	-0.195	-1.046	-0.925	0.241	0.688	0,123
recA	-0.504	-0.318	-0.396	-2.920	1.000	0,147
rpoA	-1.006	-1.049	-1.101	-0.095	0.890	0,350
rpoD	0.520	0.341	0.420	-2.208	1.000	0,350
topA	-0.151	-0.354	-0.339	-1.228	0.958	0,177
MLSA	-0.247	-0.258	-0.283	0.163	1.000	0,165

Note: D, Tajima´s D statistic (Tajima, 1989); D* and F*, Fu and Li statistics (Fu and Li, 1993); F_S, Fu´s statistic (Fu, 1997); S´, Strobecks´s statistic (Strobeck, 1987); R_2, Onsis R_2 statistic (Ramos-Onsins and Rozas, 2002)

Figure S1. Population structure of *V. tapetis*. (A) Population snapshot of all the sequence types (ST) of the MLST using eBurst program. The sizes of the circle are related to the number of strains within each ST. (B) Dot graph base don majority rule consensous tree. Ancestral nodes are showed in black and the location of isolates in red, with each red line indicating a single isolate. The empty circles indicate putative ancentral nodes

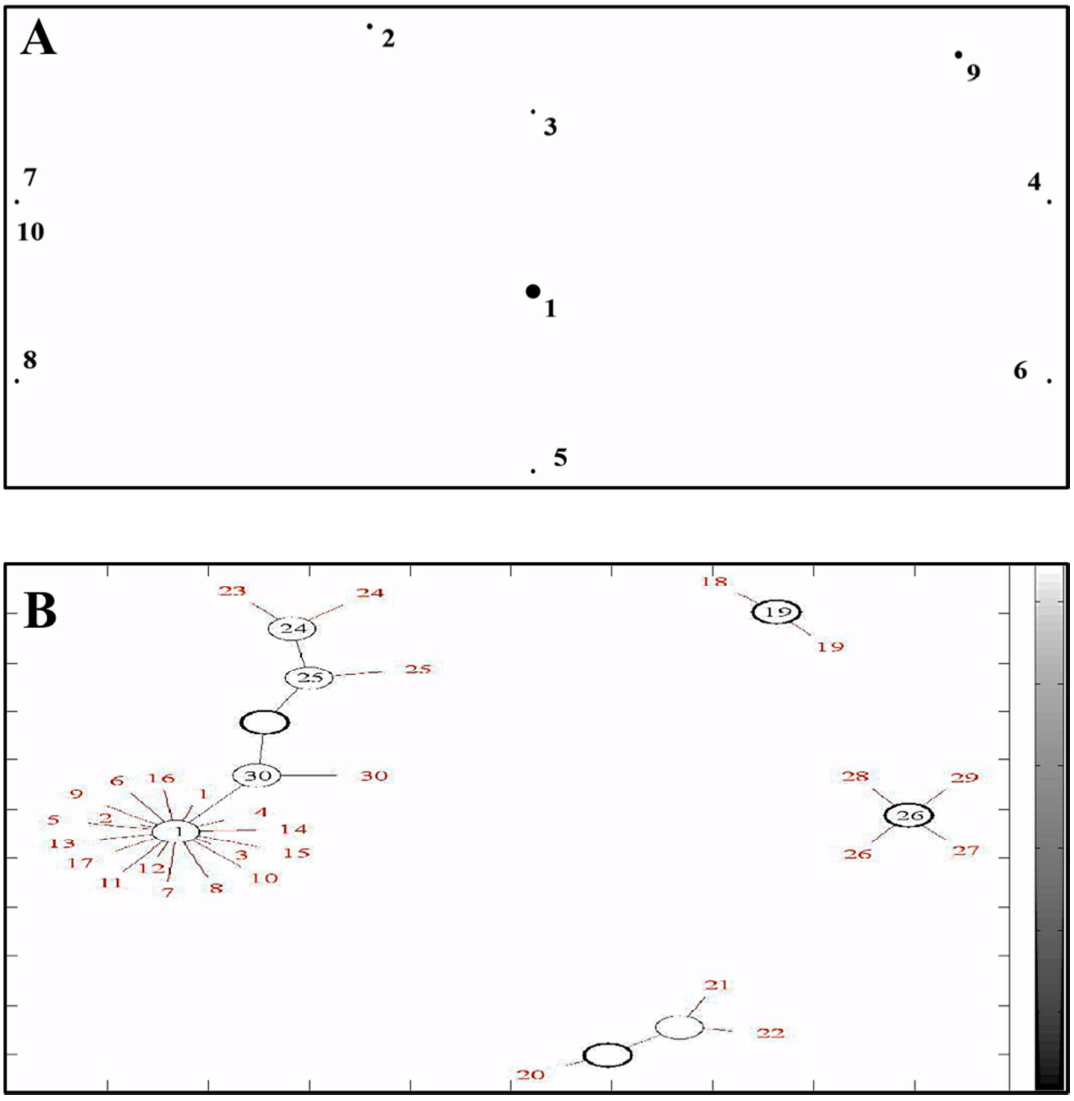


Figure S2. Phylogenetic reconstruction based on the concatenation of nucleotide sequences using the 30 isolates of *V. tapetis*. by Neighbor Joining algorithm. Bar, expected nucleotide substitutions per site. Only bootstrap values above 70% are shown (1000 resamplings) at each branch point.

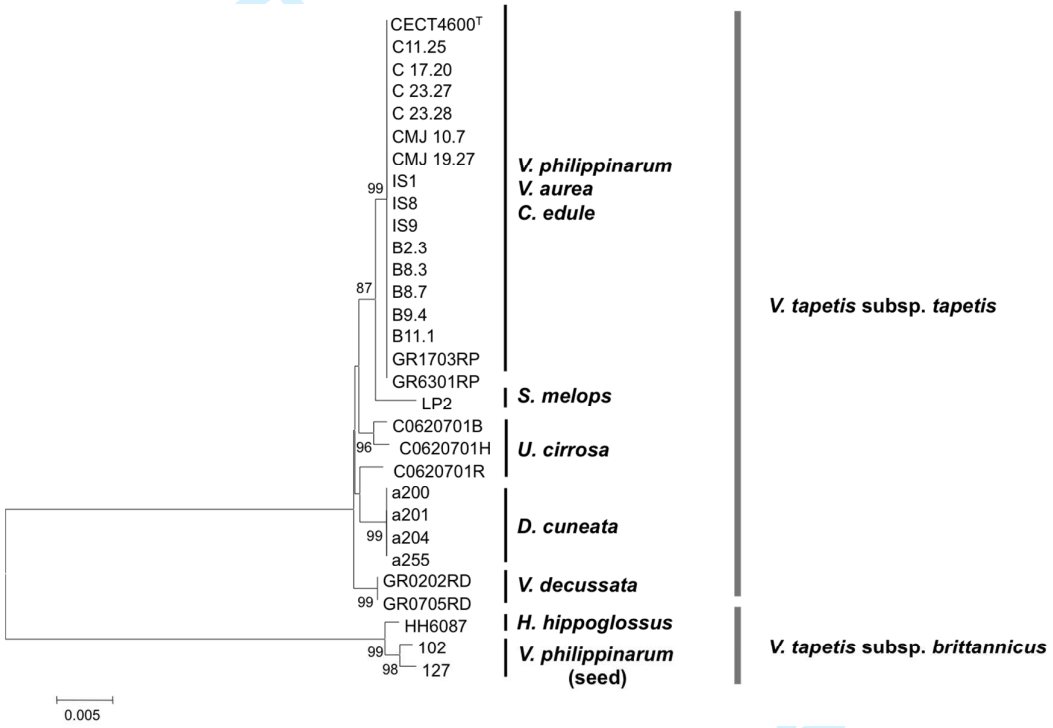
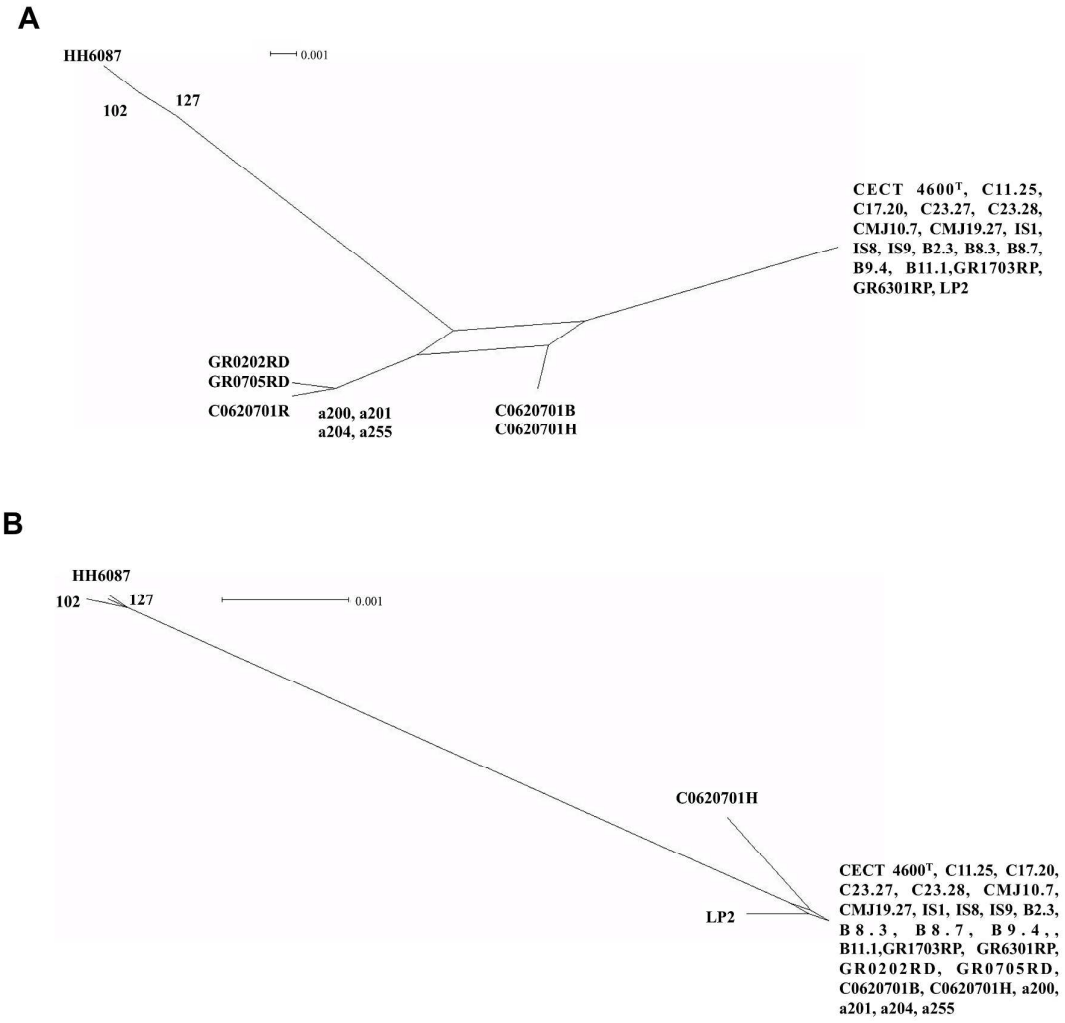
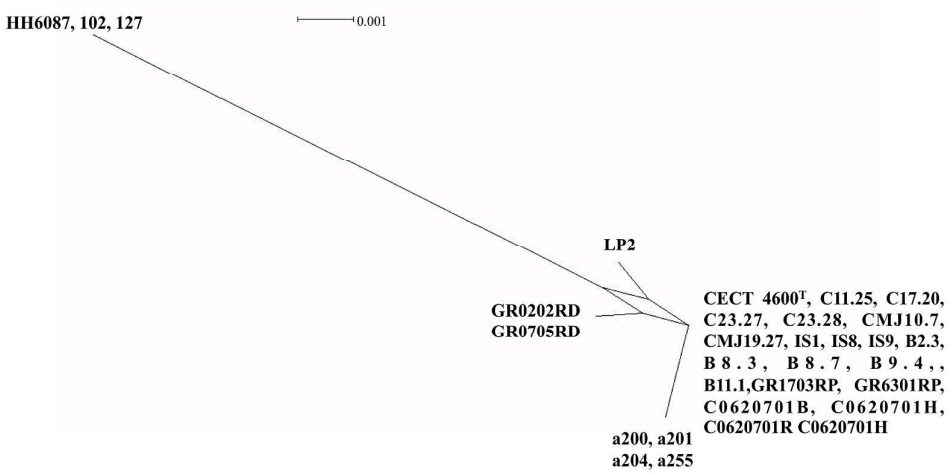


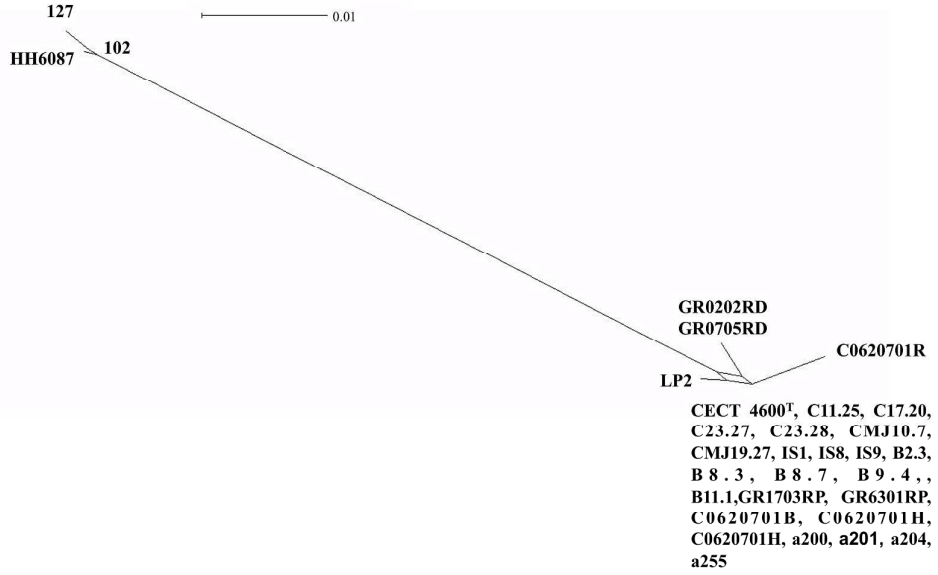
Figure S3. Neighbor-Net graph based on the individual housekeeping genes of the 30 isolates of *V. tapetis.* A, *atp*A; B, *fst*Z; C, *pyr*H; D, *rct*B; E, *rec*A; F, *rpo*D.



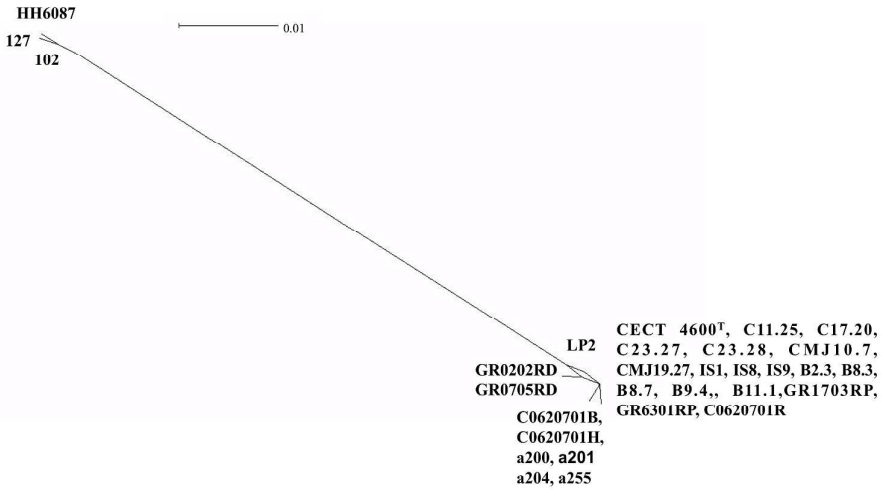
C


HH6087, 102, 127
0.001
LP2
GR0202RD
GR0705RD
CECT 4600T, C11.25, C17.20, C23.27, C23.28, CMJ10.7, CMJ19.27, IS1, IS8, IS9, B2.3, B 8 . 3 , B 8 . 7 , B 9 . 4 , , B11.1,GR1703RP, GR6301RP, C0620701B, C0620701H, C0620701R C0620701H
a200, a201
a204, a255

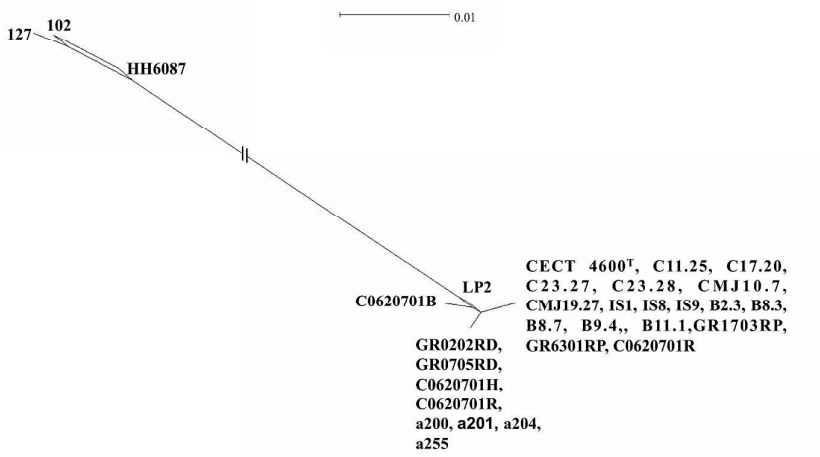
D


127
0.01
HH6087
102
GR0202RD
GR0705RD
C0620701R
LP2
CECT 4600T, C11.25, C17.20, C23.27, C23.28, CMJ10.7, CMJ19.27, IS1, IS8, IS9, B2.3, B 8 . 3 , B 8 . 7 , B 9 . 4 , , B11.1,GR1703RP, GR6301RP, C0620701B, C0620701H, C0620701H, a200, a201, a204, a255

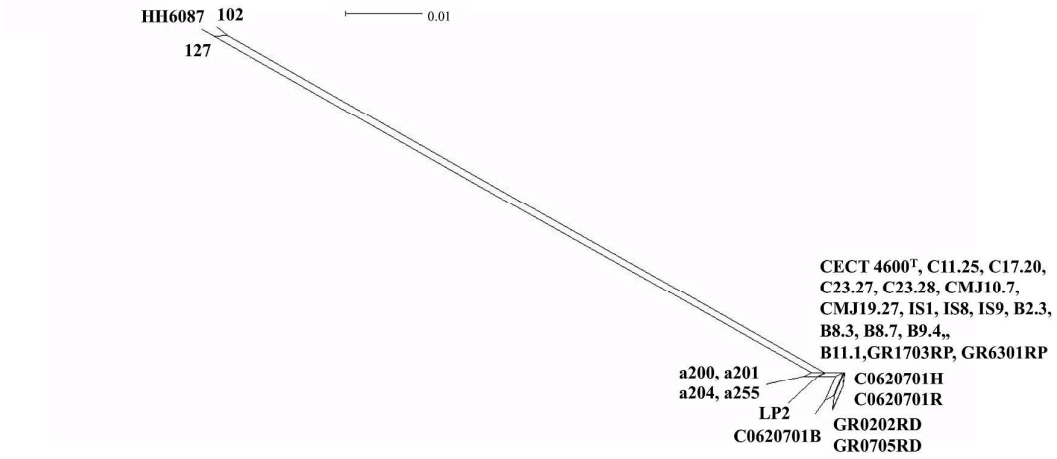
E


HH6087
127
102
0.01
LP2
GR0202RD
GR0705RD
C0620701B,
C0620701H,
a200, a201
a204, a255
CECT 4600T, C11.25, C17.20,
C23.27, C23.28, CMJ10.7,
CMJ19.27, IS1, IS8, IS9, B2.3, B8.3,
B8.7, B9.4,, B11.1,GR1703RP,
GR6301RP, C0620701R

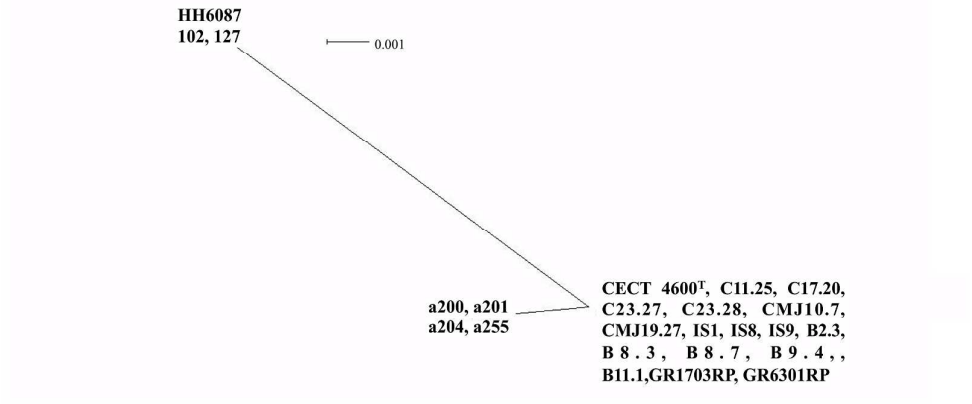
F


127
102
HH6087
0.01
C0620701B
LP2
GR0202RD,
GR0705RD,
C0620701H,
C0620701R,
a200, a201, a204,
a255
CECT 4600T, C11.25, C17.20,
C23.27, C23.28, CMJ10.7,
CMJ19.27, IS1, IS8, IS9, B2.3, B8.3,
B8.7, B9.4,, B11.1,GR1703RP,
GR6301RP, C0620701R

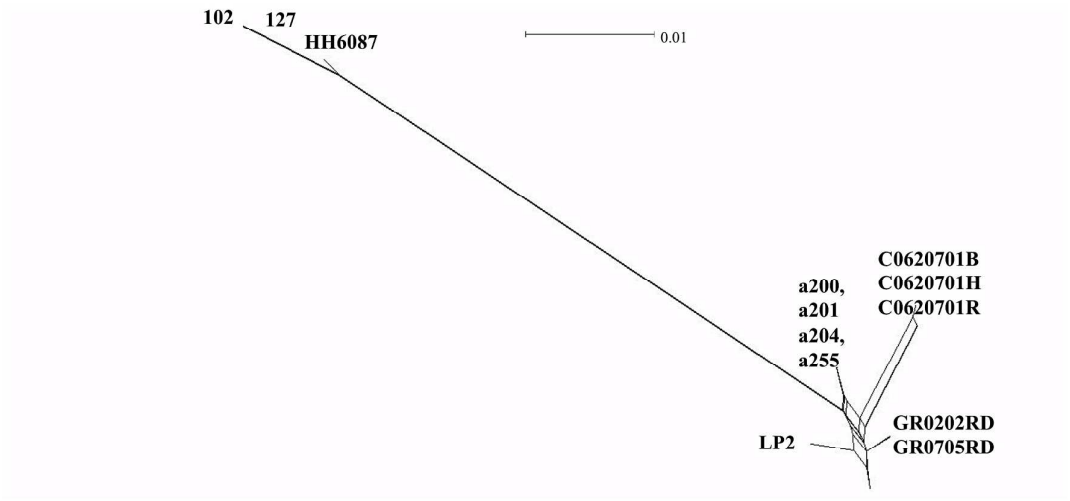
G


HH6087
102
127
0.01
CECT 4600T, C11.25, C17.20, C23.27, C23.28, CMJ10.7, CMJ19.27, IS1, IS8, IS9, B2.3, B8.3, B8.7, B9.4,, B11.1,GR1703RP, GR6301RP
a200, a201
a204, a255
C0620701H
C0620701R
LP2
C0620701B
GR0202RD
GR0705RD

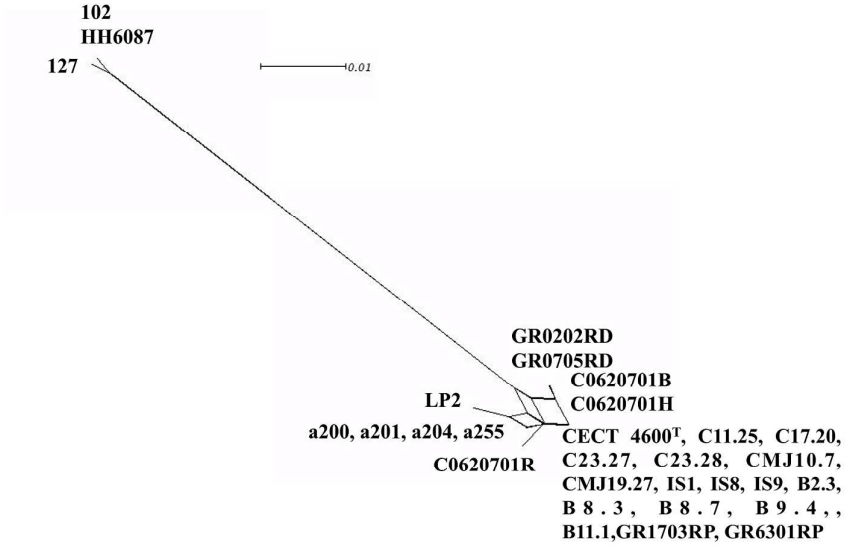
H


HH6087
102, 127
0.001
a200, a201
a204, a255
CECT 4600T, C11.25, C17.20, C23.27, C23.28, CMJ10.7, CMJ19.27, IS1, IS8, IS9, B2.3, B8.3, B8.7, B9.4,, B11.1,GR1703RP, GR6301RP

I


102
127
HH6087
0.01
C0620701B
C0620701H
C0620701R
a200,
a201
a204,
a255
GR0202RD
GR0705RD
LP2
CECT 4600T, C11.25, C17.20,
C23.27, C23.28, CMJ10.7,
CMJ19.27, IS1, IS8, IS9, B2.3,
B8.3, B8.7, B9.4,B11.1
GR1703RP, GR6301RP

J


102
HH6087
127
0.01
GR0202RD
GR0705RD
C0620701B
C0620701H
LP2
a200, a201, a204, a255
C0620701R
CECT 4600T, C11.25, C17.20,
C23.27, C23.28, CMJ10.7,
CMJ19.27, IS1, IS8, IS9, B2.3,
B8.3, B8.7, B9.4,,
B11.1,GR1703RP, GR6301RP