Tremblaya phenacola PPER: an evolutionary beta-gamma-proteobacterium collage

Bougainvillea mealybug *Phenacoccus peruvianus* (PPER) harbors single betaproteobacterial symbiont *Tremblaya phenacola*. The genome of *Tremblaya phenacola* PPER is highly rearranged, in contrast to the high genomic stability of all previously sequenced *Tremblaya* lineages, with an almost absolute synteny conservation among *Tremblaya phenacola* strains (McCutcheon and von Dohlen, 2011; Husnik and McCutcheon, 2016), and a single inversion in *Tremblaya phenacola* in *Phenacoccus avenae* (PAVE) (Husnik and McCutcheon, 2016). Chromosome rearrangements cause perturbations in GC skew, which have a deleterious impact upon the replication system (Rocha, 2004). Therefore, although bacterial chromosomes can undergo many rearrangements at the beginning of an endosymbiotic relationship (see the marked example of *Candidatus* Sodalis pierantonius SOPE; Oakeson et al., 2014), long-term endosymbionts tend to present a typical GC skew, an indication that it is recovered with evolutionary time. Contrary to the PAVE genome, with a typical GC-skew pattern (Rocha, 2008), PPER genome presents a non-polarized and highly disrupted GC skew, except in the most syntenic region between both genomes, containing most ribosomal protein genes, suggesting that the chimeric genomic architecture is not stabilized.

The *Tremblaya phenacola* PPER genome contains 192 different CDSs, 188 with an assigned function. There are only four duplicated genes inside repeats (*rpsU*, *hisG*, *prmC* and $\mathrm{TPPER}_0$0169/220), *andtwohavetwohomologs*(*infAandrlmE*).*Itpossessesasingleribosomaloperonandacompletesetof*

In annotated CDSs, 102 CDSs appear to be of betaproteobacterial origin, but another 80 appear to belong to a gammaproteobacterium. Furthermore, there is a relationship between the taxonomic affiliation of each identified CDS and their G+C content. Generally, genes with gammaproteobacterial assignation have lower G+C values than betaproteobacterial assignation (Agashe and Shankar, 2014), which is consistent with genes in *Tremblaya phenacola* PPER. *Tremblaya phenacola* PPER genes not assigned to any category have a wide range in G+C content, and most of them have very short length. There is also differences in codon usage depending on the beta or gammaproteobacterial assignation in genes of *Tremblaya phenacola* PPER genes. The distribution of gamma or beta genes along the *Tremblaya phenacola* PPER genome is not random: most contigs contain only genes of one taxonomic origin, some

others change the gene affiliation in the middle, and only one contig is completely intermixed.

The functional distribution of *Tremblaya phenacola* PPER genes is not random either. The transcriptional machinery and the ribosomes are of betaproteobacterial origin, while aminoacyl-tRNA synthetases (not the complete set, as in other mealybugs) appear to be of gammaproteobacterial origin. The only exception is serS (a pseudogene in several *Tremblaya princeps* strains; Husnik and McCutcheon, 2016), which gave no clear affiliation. Except for iscSUA (involved in (FeS) cluster assembly), genes devoted to tRNA maturation are also of gammaproteobacterial origin. This pattern is similar to the nested endosymbiotic consortia from pseudococcinae mealybugs, *Tremblaya* has retained most of its own transcriptional and translational machinery except for aminoacyl-tRNA synthetases, which must be provided by the gammaendosymbiont. Furthermore, all maintained subunits of the DNA polymerase (also preserved in other *Tremblaya princeps*) are of beta origin. However, the other proteins involved in DNA replication (helicase and ligase) are of gamma origin; the first one has been preserved in all other *Tremblaya* genomes sequenced, while the second is absent in all of them. Genes involved in translation initiation (*infA*, *infB* and *infC*) and elongation (*fusA* and *tufA*) are of beta origin, although there is an additional gammaproteobacterial *infA*. Genes involved in translation termination (*prfA*, *prfB* and *prmC*), ribosome recycling (*frr*) and degradation of proteins stalled during translation (*smpB*), as well as N-formyltransferase (*fmt*) and peptide deformylase (*def*) are of gamma origin.

Like all other mealybug endosymbionts, *Tremblaya phenacola* PPER mediates essential amino acid synthesis. As in most studied pseudococcinae mealybugs, all genes retained for the biosynthesis of methionine, threonine, isoleucine, leucine and valine, and the production of phenylalanine from chorismate are of betaproteobacterial origin, while the pathways for the production of chorismate and lysine retain the same patchwork pattern. Histidine biosynthesis is an exception, as PPER has only retained genes of gammaproteobacterial origin. The cysteine biosynthetic pathway is more complete in PPER, with all genes of gamma origin. Regarding tryptophan biosynthesis, dominated by gammaproteobacterial genes in previously analyzed mealybugs endosymbiotic consortia, in *Tremblaya phenacola* PPER the first step is performed by beta proteins, while the rest of the genes are of gamma origin, a similar pattern to that found in other insects endosymbiotic consortia (that is, *Serratia/Buchnera* in lachninae aphids and some *Carsonella*/secondary systems in psyllids; Lamelas *et al.*, 2011; Sloan and Moran, 2012; Manzano-Marn *et al.*, 2016).

Tremblaya phenacola PPER genome goes beyond what could be considered a standard horizontal gene transfer event, and rather resembles the complete fusion of two genomes to form a new chimeric organism. Independent phylogenomic analyses of two concatenations of the *Tremblaya phenacola* PPER genes assigned as beta or gammaproteobacterial placed beta origin genes in *Tremblaya phenacola* clade,

while the gammaproteobacterial genes were placed into the *Sodalis*-allied clade (Husnik and McCutcheon, 2016) as a sister species of *Candidatus* Mikella endobia, nested gamma-endosymbiont of the pseudoccinae mealybug *Paracoccus marginatus*.

How could the genomic fusion have ocurred? Although HGT is uncommon in modern endosymbionts, it is an extended phenomenon in flowering-plant mitochondria (Sanchez-Puerta, 2014), derived from an ancestral -proteobacterial endosymbiont (Andersson *et al.*, 1998). The most notable case corresponds to *Amborella trichopoda*, whose mitochondrial DNA has incorporated the complete mitochondrial genomes of three green algae and one moss, plus two mitochondrial genome equivalents from other angiosperms (Rice *et al.*, 2013). Such a high frequency of HGT has been explained by mitochondrial fusion and subsequent genomes fusion and rearrangements, mediated by homologous recombination systems (Marchal and Brisson, 2010). Something similar might have occurred in *Tremblaya phenacola* PPER. On the basis of current evidences, the ancestor of all *Tremblaya* probably had a reduced genome (Husnik and McCutcheon, 2016). In the lineage driving to *Tremblaya phenacola* PPER, a gammaproteobacterium must have entered the consortium and, instead of replacing *Tremblaya phenacola* (as in the tribe Rhizoecini and genus *Rastrococcus*; Gruwell *et al.*, 2010), or establishing a nested endosymbiosis (as in the *Tremblaya phenacola* clade; reviewed by Husnik and McCutcheon, 2016), a cellular fusion event must have occurred, followed by genomic fusion. It cannot be discarded that a nested endosymbiosis preceded the cellular and genomic fusions. Because this phenomenon implies the existence of a DNA recombination machinery, the most plausible hypothesis is that such genes were present in the genome of the gammaproteobacterial donor, similarly to what has been described in citrus mealybug (Lpez-Madrigal *et al.*, 2013). In fact, most mealybugs gamma-endosymbionts that have been completely sequenced (McCutcheon and von Dohlen, 2011; Lpez-Madrigal *et al.*, 2013; Husnik and McCutcheon, 2016) or screened for homologous recombination genes (Lpez-Madrigal *et al.*, 2015) present a more or less complete recombination machinery. Transposable elements might also facilitate a fusion process. Some authors suggest that in arthropod intracellular environments, the possibility of two bacteria co-infecting the same cell generates an intracellular arena where distantly related bacterial lineages can exchange mobile elements (Duron, 2013). However, although insertion sequences are frequent in early endosymbiotic stages (Latorre and Manzano-Marn, 2016), they have not been identified in any sequenced mealybugs gamma-endosymbiont, and no indication of their former presence in *Tremblaya phenacola* PPER. After the fusion, the chimeric genome must have undergone massive gene loss, getting rid of almost all redundant and non-essential genes. The initial presence of homologs might have accelerated gene losses through recombination until DNA recombination genes disappeared. The remnant repeats involved in intrachromosomal recombination might have been maintained due to the loss of such genes, leading to the current, complex genome

organization.