BMC Genomics

BioMed Central

Research article

Open Access

# Simple sequence repeats and compositional bias in the bipartite *Ralstonia solanacearum* GMI1000 genome

Tom Coenye* and Peter Vandamme

Address: Laboratorium voor Microbiologie, Ghent University,K.L. Ledeganckstraat 35, B-9000 Gent, Belgium

Email: Tom Coenye* - Tom.Coenye@rug.ac.be; Peter Vandamme - Peter.Vandamme@rug.ac.be

* Corresponding author

Published: 17 March 2003

*BMC Genomics* 2003, **4**:10

Received: 23 December 2002

Accepted: 17 March 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/10



## Abstract

**Background:** *Ralstonia solanacearum* is an important plant pathogen. The genome of *R. solananearum* GMI1000 is organised into two replicons (a 3.7-Mb chromosome and a 2.1-Mb megaplasmid) and this bipartite genome structure is characteristic for most *R. solanacearum* strains. To determine whether the megaplasmid was acquired via recent horizontal gene transfer or is part of an ancestral single chromosome, we compared the abundance, distribution and compositon of simple sequence repeats (SSRs) between both replicons and also compared the respective compositional biases.

**Results:** Our data show that both replicons are very similar in respect to distribution and composition of SSRs and presence of compositional biases. Minor variations in SSR and compositional biases observed may be attributable to minor differences in gene expression and regulation of gene expression or can be attributed to the small sample numbers observed.

**Conclusions:** The observed similarities indicate that both replicons have shared a similar evolutionary history and thus suggest that the megaplasmid was not recently acquired from other organisms by lateral gene transfer but is a part of an ancestral *R. solanacearum* chromosome.

## Background

The paradigm that bacterial genomes consist of a single circular chromosome is no longer valid. Linear chromosomes have been identified in *Borrellia burgdorferi* [1], various *Streptomyces* species [2,3], *Agrobacterium tumefaciens* [4] and various other species. In addition, it is now appreciated that genomes of several bacterial taxa consist of multiple replicons. Most organisms with a multi- or bipartite genome structure belong to the α-*Proteobacteria* (including *Rhodobacter sphaeroides* [5,6] and various *Rhizobium* [7,8], *Agrobacterium* [4,8], *Brucella* [9,10] and *Azospirillum* [11] species) or the β-*Proteobacteria*. Most isolates from species belonging to the β-proteobacterial genera *Burkholderia* and *Ralstonia* harbour multiple replicons, including members of the *Burkholderia cepacia* complex [12–16], *Burkholderia gladioli* [15], *Burkholderia pseudomallei* [17], *Burkholderia glumae* [13], *Burkholderia glathei* [13], *Burkholderia* sp. LB400 [18], *Ralstonia pickettii* [13], *Ralstonia eutropha* [13] and *Ralstonia metallidurans* [18]. Multiple replicons may have arisen from the need to achieve higher overall replication rates [19]. The origin of these multiple replicons is at present unclear but it has been suggested that they could have their origin in gene duplication followed by divergence; in this case intrachromosomal recombinational events within a duplicated region could give rise to the formation of two stable replicons [8]. In the genus *Brucella* these rearrangements have occurred in the region containing the ribosomal RNA genes [10] but in theory the rearrangements can occur at any repeated sequence [20]. An additional

explanation is that the presence of multiple replicons within an organism involved horizontal DNA transfer [19,21,22]. This hypothesis was used to explain the presence of two chromosomes in *Vibrio cholerae*: the small chromosome was suggested to be derived from a megaplasmid captured by an ancestral *Vibrio* [23,24]. This megaplasmid probably acquired genes from diverse bacterial species before its capture by the ancestral *Vibrio*; subsequent relocation of essential genes from the chromosome to the megaplasmid completed its stable structure.

*Ralstonia solanacearum* is a soil-borne phytopathogen with an unusually broad host-range, causing bacterial wilt on a wide range of crops, including economically important crops like potato, tomato, ginger and banana [25]. Recently the genome sequence of *R. solanacearum* strain GMI1000 was determined [26]. It was shown that the 5.8-Mb genome is organised into two replicons, a 3.7-Mb chromosome and a 2.1-Mb megaplasmid. This bipartite genome structure is characteristic for most *R. solanacearum* strains [27] and derivatives of strain GMI1000 without the megaplasmid have not been obtained [26]. The larger replicon contains all the basic genes required for survival of the bacterium; the smaller replicon carries several metabollically essential genes also present on the chromosome (including a rDNA locus, a gene coding for the α-subunit of DNA polymerase III and the gene for protein elongation factor G) but also contains several genes coding for enzymes involved in primary metabolism (including amino acid and cofactor biosynthesis) not present on the chromosome. The smaller replicon also contains all the *hrp* genes (required to cause disease in plants) and it has been suggested that it has a significant function in overall fitness and adaptation of the organism to various environmental conditions [26]. The origin of the bipartite genome structure of *R. solanacearum* is not clear. To determine whether the megaplasmid was formed through intrachromosomal recombinational events within a duplicated region or was recently acquired from other organisms we compared the abundance, distribution and composition of simple sequence repeats between the chromosome and the megaplasmid of *R. solanacearum* GMI1000. We also compared the compositional bias of di- and tetranucleotides between both replicons.

Repeated DNA consists of homopolymeric tracts of a single nucleotide or of small or large numbers of multimeric classes of repeats. These multimeric repeats can be homogenous (i.e. built from identical units), heterogeneous (i.e. built from mixed units) or are built from degenerate repeat sequence motifs [28]. A special category of repeats are tandem repeats which are made up of periodically repeated monomeric sequences of varying length, arranged in a 'head-to-tail' configuration [29]. Several mechanisms have been proposed for the creation of tandem repeats, including 'slipped strand mispairing' in which illegitimate base-pairing during replication gives rise to addition of repeat units [30,31]. There is growing evidence that small tandem repeats (also called simple sequence repeats or SSRs) affect gene expression. A first effect of SSRs is the mediation of phase variation through the loss or gain of one or more repeats [29]. Phase variation is the process by which many bacterial species undergo reversible phenotypic changes resulting from genetic alterations in certain loci [32,33]. SSRs can also be involved in gene regulation by affecting spacing between flanking regions [34] or spacing between the -35 and -10 promotor regions [35]. Variation in abundance, distribution and composition of SSRs has been described [28] and it has been proposed that variation in SSR results in variation in gene expression and key phenotypes and hence provides an important target for natural selection and evolution [28,36].

The comparison of genome-wide compositional biases as a tool to study bacterial evolution has been introduced by Karlin and co-workers [37–39]. It is thought that dinucleotide relative abundance values are constant within a genome because the factors that work on them are constant throughout the genome; and it has been postulated that the set of dinucleotide relative abundance values constitute a genomic signature that reflects the pressures of these factors [38]. Differences in genome signature between different organisms can be attributed to differences in context-dependent mutation rates generated by the replication-repair system and differences in efficiency of the replication machinery on different sequences. In addition, many DNA structural properties (including curvature, flexibility and helix stability), which may play an important role in biological processes like replication, are determined by dinucleotide arrangements [38,40]. Tetranucleotide relative abundances are also characteristic for a given genome [39]. It has been postulated that frequent tetranucleotides may include parts of repetitive structural, regulatory and transposable elements, while low values for some palindromic tetranucleotides have been attributed to restriction avoidance [39].

## Results

### *Distribution and composition of SSRs in the* R. solanacearum *genome*

A total of 221729 SSRs with a motif length between 1 and 10 bp and minimum three repeats were found in the entire R. solanacearum genome. Of those, 139993 (63.14%) were located on the chromosome (Table 1) and 81736 (36.86%) were located on the megaplasmid (Table 2). This corresponds wel with the size distribution between both replicons (63.96% of all bases are in the chromosome, 36.04% are in the megaplasmid). The SSRs were evenly distributed both over the chromosome as over the megaplasmid (Fig. 1). The total number of repeats is

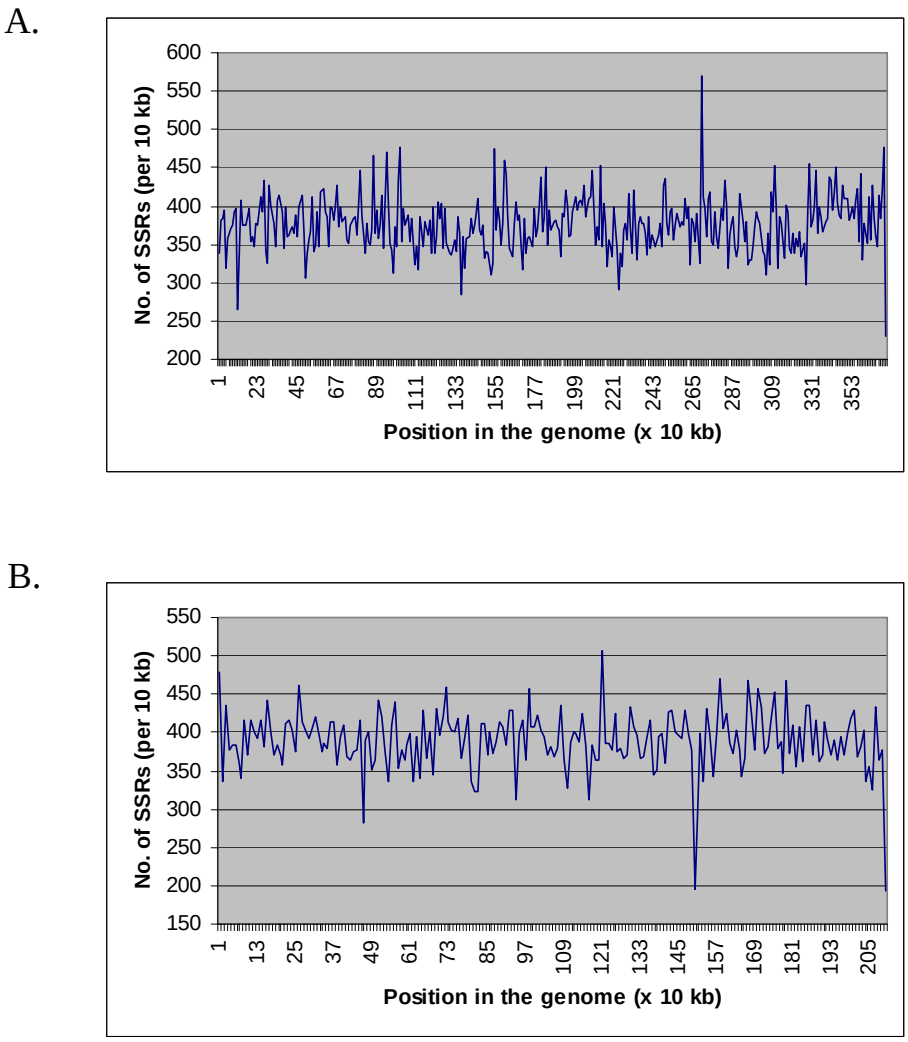


**Figure 1**
Distribution of simple sequence repeats in the *R. solanacearum* chromosome (panel A) and megaplasmid (panel B).

lower than expected by chance; especially the number of mononucleotide repeats is significantly lower than expected (Tables 1 and 2). Trinucleotide repeats occur more than expected by chance alone, both in the chromosome and the megaplasmid (Tables 1 and 2). Mononucleotide repeats of length = 3 bp and dinucleotide repeats are distributed over coding and non-coding regions as expected, both in the chromosome and the megaplasmid. As mononucleotide repeats become larger, there is more and more deviation from the expected distribution; these larger mononucleotide repeats are almost exclusively located in non-coding regions. Our data also show that trinucle-

**Table 1: Number of simple sequence repeats of given structure in the chromosome of *R. solanacearum* GMI1000**

| | Motif length | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of repeats** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| 3 | 98068$^{-}$ | 12307$^{+}$ | 3712$^{+}$ | 90 | 21 | 11 | 1 | - | 3 | 1 | 114214$^{-}$ |
| 4 | 17943$^{-}$ | 2122$^{+}$ | 213$^{+}$ | - | - | - | - | - | - | - | 20278$^{-}$ |
| 5 | 4053$^{-}$ | 277 | 17 | - | - | - | - | - | 2 | - | 4349$^{-}$ |
| 6 | 859$^{-}$ | 18 | - | - | 1 | - | - | 1 | 1 | - | 880$^{-}$ |
| 7 | 199$^{-}$ | 3 | - | - | 1 | 1 | - | - | 1 | - | 205$^{-}$ |
| 8 | 45$^{-}$ | - | - | - | - | - | - | - | - | - | 45$^{-}$ |
| 9 | 14$^{-}$ | - | - | - | - | - | - | - | - | - | 14$^{-}$ |
| 10 | 2 | - | - | - | - | - | - | - | - | - | 2 |
| 11 | 4 | - | - | - | - | - | - | - | - | - | 4 |
| 12 | - | - | - | - | - | - | - | - | - | - | - |
| 13 | 2 | - | - | - | - | - | - | - | - | - | 2 |
| Total | 121189$^{-}$ | 14727$^{+}$ | 3942$^{+}$ | 90 | 23 | 12 | 1 | 1 | 7 | 1 | 139993$^{+}$ |

$^{+}$ significantly overrepresented compared to mean frequencies in computer-generated randomised genomes ($P < 0.001$) $^{-}$ significantly underrepresented compared to mean frequencies in computer-generated randomised genomes ($P < 0.001$)

**Table 2: Number of simple sequence repeats of given structure in the megaplasmid of *R. solanacearum* GMI1000**

| | Motif length | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **No. of repeats** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
| 3 | 57345- | 6440 | 1986+ | 49 | 10 | 8 | - | 3 | - | - | 65841$^{-}$ |
| 4 | 11216$^{-}$ | 1096 | 108$^{+}$ | 2 | 1 | 4 | 2 | 2 | 2 | - | 12433$^{-}$ |
| 5 | 2560- | 146 | 10 | - | - | 2 | - | 3 | - | - | 2721$^{-}$ |
| 6 | 529$^{-}$ | 9 | 1 | - | - | 2 | - | 1 | - | - | 542$^{-}$ |
| 7 | 141- | 1 | - | - | - | - | - | 1 | - | - | 143 |
| 8 | 41$^{-}$ | 1 | - | - | - | - | - | - | - | - | 42$^{-}$ |
| 9 | 11 | - | - | - | - | - | - | - | - | - | 11 |
| 10 | 2 | - | - | - | - | - | - | - | - | - | 2 |
| 11 | 1 | - | - | - | - | - | - | - | - | - | 1 |
| Total | 71846_ | 7693 | 2105$^{+}$ | 51 | 11 | 16 | 2 | 4 | 8 | 0 | 81736 |

$^{+}$ significantly overrepresented compared to mean frequencies in computer-generated randomised genomes ($P < 0.001$) $^{-}$ significantly underrepresented compared to mean frequencies in computer-generated randomised genomes ($P < 0.001$)

otides are overrepresented in protein-coding regions of both replicons (Table 3). The nucleotide composition of the SSR tracts in the R. solanacearum chromosome and megaplasmid are shown in Tables 4 and 5, respectively. Our data show that (i) the G+C composition of mononucleotide repeats in both replicons is significantly lower than the overall composition, but this difference can exclusively be attributed to non-coding regions; (ii) G and C mononucleotide repeats are underrepresented in coding and non-coding regions of both replicons and (iii) CG and GC dinucleotide repeats are vastly overrepresented both in coding and non-coding regions of both replicons, while other dinucleotide repeats are underrepresented.

### *Compositional biases in the R. solanacearum genome*

Dinucleotide relative abundances are shown in Table 6. The dinucleotides TA and AT are strongly underrepresented in both replicons while GC is moderately overrepresented in both replicons. CC and GG are moderately underrepresented in the chromosome. The average absolute dinucleotide relative abundance difference ($\delta^*$) between both replicons is 9.78. To assess the variability of dinucleotide relative abundances within a replicon, both replicons were divided into 12 and 7 (for the chromosome and the megaplasmid, respectively) equally-sized, nonoverlapping fragments and $\rho^*_{XY}$ values were calculated for each fragment. $\delta^*(f,g)$ values within repli-

**Table 3: Distribution of simple sequence repeats among protein coding and non-coding regions of the *R. solanacearum* genome**

| | Chromosome | | | | | Megaplasmid | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total no. | Coding regions | | Non-coding regions | | Total no. | Coding regions | | Non-coding regions | |
| | | No. | % | No. | % | | No. | % | No. | % |
| **Mononucleotides** | | | | | | | | | | |
| 3 bp | 98068 | 81805 | 83.4 | 16263 | 16.6 | 57345 | 47087 | 82.1 | 10258 | 17.9 |
| 4 bp | 17943 | 13192 | 73.5 | 4751 | 26.5 | 11216 | 8509 | 75.9 | 2707 | 24.1 |
| 5 bp | 4053 | 2555 | 63.0 | 1498 | 37.0 | 2560 | 1752 | 68.4 | 808 | 31.6 |
| 6 bp | 859 | 373 | 43.4 | 486 | 56.6 | 529 | 282 | 53.3 | 247 | 46.7 |
| 7 bp | 199 | 62 | 31.2 | 137 | 68.8 | 141 | 49 | 34.8 | 92 | 65.2 |
| 8 bp | 45 | 7 | 15.6 | 38 | 84.4 | 41 | 16 | 39.0 | 25 | 61.0 |
| 9 bp | 14 | 1 | 7.1 | 13 | 92.9 | 11 | 3 | 27.3 | 8 | 72.7 |
| 10 bp | 2 | 1 | 50.0 | 1 | 50.0 | 2 | 0 | 0 | 2 | 100 |
| 11 bp | 4 | - | 0 | 4 | 100 | 1 | 0 | 0 | 2 | 1001 |
| 13 bp | 2 | - | 0 | 4 | 100 | - | - | - | - | - |
| **Dinucleotides ≥ 6 bp** | 14727 | 13090 | 88.9 | 1637 | 11.1 | 7693 | 6699 | 87.1 | 994 | 12.9 |
| **Trinucleotides ≥ 9 bp** | 3942 | 3672 | 93.2 | 270 | 6.8 | 2105 | 1933 | 91.8 | 172 | 8.2 |
| **Tetranucleotides** | 90 | 77 | 85.6 | 13 | 14.4 | 51 | 37 | 72.5 | 14 | 27.5 |
| **Genome partition** | | | 87.8 | | 12.2 | | | 86.5 | | 13.5 |

**Table 4: Nucleotide composition of simple sequence repeats in the *R. solanacearum* chromosome**

| | Total | | Coding | | Non-coding | |
|---|---|---|---|---|---|---|
| | No. | % | No. | % | No. | % |
| **Genome composition** | | | | | | |
| A | 608615 | 16.37 | 543922 | 16.62 | 64693 | 14.53 |
| C | 1238438 | 33.32 | 1112352 | 34.00 | 126086 | 28.31 |
| G | 1252933 | 33.71 | 1099390 | 33.60 | 153543 | 34.47 |
| T | 616427 | 16.58 | 515362 | 15.75 | 101065 | 22.70 |
| **Mononucleotide SSRs ≥ 6 bp** | | | | | | |
| Total | 1125 | | 444 | | 681 | |
| A | 299 | 26.58 | 109 | 24.55 | 190 | 27.90 |
| C | 220 | 19.56 | 106 | 23.87 | 114 | 16.74 |
| G | 261 | 23.20 | 153 | 34.46 | 108 | 15.86 |
| T | 345 | 30.67 | 76 | 17.12 | 269 | 39.50 |
| G+C | 481 | 42.76 | 259 | 58.33 | 222 | 32.60 |
| A+T | 644 | 57.24 | 185 | 41.67 | 459 | 67.40 |
| **Dinucleotide SSRs ≥ 6 bp** | | | | | | |
| Total | 14764 | | 13121 | | 1643 | |
| AC/CA | 278 | 1.88 | 205 | 0.15 | 73 | 4.44 |
| AG/GA | 205 | 1.39 | 80 | 0.61 | 125 | 7.61 |
| AT/TA | 47 | 0.32 | 11 | 0.08 | 36 | 2.19 |
| CG/GC | 13710 | 92.86 | 12521 | 95.43 | 1189 | 72.37 |
| CT/TC | 210 | 1.42 | 130 | 0.99 | 80 | 4.87 |
| GT/TG | 323 | 2.19 | 174 | 1.33 | 149 | 9.07 |

cons ranged from 6.63 to 31.77 (mean ± standard deviation: 14.49 ± 5.35) (for the chromosome) and from 4.83 to 20.63 (13.11 ± 4.55) (for the megaplasmid). These differences are not significantly smaller than the between-replicon differences (data not shown). Significantly over- or underrepresented tetranucleotides are shown in Table 7. CTAG, AATT, CATG, GATA and TATA are underrepresented in both replicons. GTAG and TTAA are overrepresented in both replicons.

## Discussion

To study the origin of the bipartite genome structure of *R. solanacearum* GMI1000 we compared the abundance, distribution and composition of simple sequence repeats and differences in compositional biases between the chromosome and the megaplasmid of *R. solanacearum* GMI1000.

### *Occurrence of simple sequence repeats*

Our data clearly show that the R. solanacearum genome contains numerous SSRs with a motif length between 1

and 10 bp, although not as many as expected by chance alone. Mutations in SSRs are thought to be the result of slipped strand mispairing during DNA replication; slipped strand mispairing can occur because the tertiary structure of SSRs allows mismatching and repeats can be inserted or excised during DNA duplication [41–43]. The observation of upper limits for SSR length in Escherichia coli suggested that the tendency for repeat length to arise via mutation is counteracted by selection [36]. We observed similar upper limits: the upper limit for total length of mononucleotide SSRs is 13 bp and 11 bp for the chromosome and megaplasmid, respectively and, in addition, very few other SSRs with a total length >15 bp (for the chromosome) or >18 bp (for the megaplasmid) are observed. Both strand separation and slippage are more likely for mononculeotide SSRs, explaining why mononucleotide SSRs are more likely to undergo slipped strand mispairing; longer SSRs with a lower repeat number have less opportunity to undergo slipped strand mispairing and there will be less mutability in their repeat number [36]. This may explain why larger mononucleotide SSRs are overrepresented in non-coding regions of the R. solanacearum genome as selection has ample opportunity to operate against these larger repeats that cause frameshift and nonsense mutations in coding regions. This hypothesis is supported by the fact that poly(A) and poly(T) SSRs are overrepresented, especially in the non-coding regions, in both replicons (Tables 4 and 5): strand separation for these poly(A) and poly(T) tracts is considerably easier than for poly(G) or poly(C) tracts, increasing the possibility of slipped strand mispairing.

### *Compositional biases*

The dinucleotide TA is underrepresented in both replicons. TA is underrepresented in almost all prokaryotic genomes; this could be due to the fact that (i) TA forms the thermodynamically least stable DNA (allowing unwinding of the helix), (ii) RNases preferentially degrade UA dinucleotides in mRNA, and/or (iii) TA is part of many regulatory sequences [38]. AT is significantly underrepresented in the *R. solanacearum* genome but is overrepresented in the genome of most α-*Proteobacteria* and in the genomes of the β-proteobacterial species *R. eutropha* and *Bordetella pertussis* [39]. CC and GG are slightly underrepresented in the chromosome but not in the megaplasmid, although the differences in relative abundances are small (Table 6). The dincucleotide GC is overrepresented in both replicons; this is also the case in most other β-*Proteobacteria* and γ-*Proteobacteria* [39]. In general, within species $\delta^*(f,g)$-differences among nonoverlapping 50 kb contigs of bacteria are in the range 18–43 [39] and genome signatures of chromosomes and plasmids from the same host are at least weakly similar to each other [$\delta^*(f,g) < 115$] [44,45]. $\delta^*(f,g)$ values reported for the multiple chromosomes of *A. tumefaciens*, *Deinococcus radiodurans*, *V. cholerae* and *B. melitensis* were between 27.0 and 30.8 [45]. A comparison of both *R. solanacearum* replicons based on dinucleotide relative abundances indicates that they are very similar with $\delta^*(f,g) = 9.78$. A comparison of $\delta^*(f,g)$ values within and between replicons revealed that the variability in $\delta^*(f,g)$ within a replicon is not significantly smaller than the difference in $\delta^*(f,g)$ between both replicons. CTAG is significantly underrepresented in the *R. solanacearum* genome as it is in most proteobacterial organisms. Possible reasons for the underrepresentation of this tetranucleotide include structural defects or special functional roles associated with CTAG [38]. AATT, CATG, GATA and TATA are underrepresented in both replicons while GTAG and TTAA are overrepresented. ATTG, CATC and TTGG occur slightly less than expected in the megaplasmid but their relative abundance in the chromosome is in the normal range. The general mechanisms underlying tetranucleotide extremes are unclear but besides the above-mentioned structural defects or functional roles associated with specific tetranucleotides, it has been suggested that restriction avoidance may play an important role in the maintenance of tetranucleotide extremes [39]).

## Conclusions

It can be concluded that both replicons that constitute the *R. solanacearum* genome are very similar in respect to distribution and composition of SSRs and presence of compositional biases, although minor differences between both replicons are present. The megaplasmid carries the *hrp* genes required to cause disease in plants, genes coding for constituents of the flagellum and genes involved in exopolysaccharide production; it also contains 315 genes of unknown function [26]. The minor variations in SSR and compositional biases observed between both replicons may therefore be attributable to minor differences in gene expression and regulation of gene expression between both replicons. Alternatively, it is not unlikely that some of the observed differences are the result of the small sample numbers observed (for example the minor differences in tetranucleotide SSR distribution over coding and non-coding regions in both replicons [Table 3]). At present no completely sequenced and fully annotated genomes of other β-*Proteobacteria* with multiple replicons are available for comparison and therefore it is difficult to place the observed differences in a broader perspective. Nevertheless, the observed similarities in SSRs and compositional biases indicate that both replicons have shared a similar evolutionary history and suggest that the megaplasmid was not recently acquired from other organisms by lateral gene transfer but is a part of an ancestral *R. solanacearum* chromosome. Alernatively, the hypothesis of an ancient acquisition by lateral gene transfer followed by a long co-evolution with the chromosome cannot be completely ruled out.

**Table 5: Nucleotide composition of simple sequence repeats in the *R. solanacearum* megaplasmid**

| | Total | | Coding | | Non-coding | |
|---|---|---|---|---|---|---|
| | **No.** | **%** | **No.** | **%** | **No.** | **%** |
| **Genome composition** | | | | | | |
| A | 347472 | 16.58 | 301622 | 16.62 | 45850 | 16.34 |
| C | 699983 | 33.41 | 613428 | 33.81 | 86555 | 30.83 |
| G | 700536 | 33.44 | 610348 | 33.64 | 90188 | 32.13 |
| T | 346518 | 16.54 | 288441 | 15.90 | 58077 | 20.69 |
| **Mononucleotide SSRs ≥ 6 bp** | | | | | | |
| Total | 722 | | 350 | | 372 | |
| A | 172 | 23.82 | 83 | 23.71 | 89 | 23.93 |
| C | 169 | 23.41 | 94 | 26.86 | 75 | 20.16 |
| G | 197 | 27.29 | 130 | 37.14 | 67 | 18.01 |
| T | 184 | 25.49 | 43 | 12.29 | 141 | 37.90 |
| G+C | 366 | 50.69 | 224 | 64.00 | 142 | 38.17 |
| A+T | 356 | 49.31 | 126 | 36.00 | 230 | 61.83 |
| **Dinucleotide SSRs ≥ 6 bp** | | | | | | |
| Total | 7708 | | 6727 | | 981 | |
| AC/CA | 190 | 2.47 | 143 | 2.13 | 47 | 4.79 |
| AG/GA | 135 | 1.75 | 71 | 1.06 | 64 | 6.52 |
| AT/TA | 38 | 0.49 | 15 | 0.22 | 23 | 2.35 |
| CG/GC | 7127 | 92.46 | 6297 | 93.61 | 830 | 84.61 |
| CT/TC | 129 | 1.67 | 82 | 1.22 | 47 | 4.79 |
| GT/TG | 166 | 2.15 | 119 | 1.77 | 47 | 4.79 |

**Table 6: Dinucleotide relative abundances in the *R. solanacearum* genome**

| **XY** | **Chromosome** | | **Megaplasmid** | |
|---|---|---|---|---|
| | $\rho^*_{XY}$ | **Over/under represented?** | $\rho^*_{XY}$ | **Over/under represented?** |
| AA | 1.0713 | | 1.0738 | |
| AC | 0.9014 | | 0.8928 | |
| AG | 0.8775 | | 0.8722 | |
| AT | 0.6787 | -- | 0.6957 | -- |
| CA | 1.1702 | | 1.1778 | |
| CC | 0.7750 | - | 0.7879 | |
| CG | 1.2029 | | 1.1880 | |
| CT | 0.8775 | | 0.8722 | |
| GA | 1.0601 | | 1.0446 | |
| GC | 1.2454 | + | 1.2440 | + |
| GG | 0.7750 | - | 0.7879 | |
| GT | 0.9014 | | 0.8928 | |
| TA | 0.4624 | --- | 0.4808 | --- |
| TC | 1.0601 | | 1.0446 | |
| TG | 1.1702 | | 1.1778 | |
| TT | 1.0713 | | 1.0738 | |

## Methods

### *DNA sequences*

The sequences of the chromosome (AL646052) and the megaplasmid (AL646053) of *R. solanacearum* strain GMI1000 were downloaded from the GenBank database.

### *Analysis of SSRs*

We used the software developed by Gur-Arie et al. [36] to screen the entire genome of *R. solanacearum* for SSRs withg a motif length between 1 and 10 bp and a minimal number of three repeats. This software can be downloaded from ftp://ftp.technion.ac.il/supported/biotech/ssr.exe and reports motif, motif length, repeat number and genomic location of all SSRs. To determine whether the ob-

**Table 7: Singificantly over- or underrepresented tetranucleotides in the *R. solanacearum* genome**

| XYZW | Chromosome | | Megaplasmid | |
|---|---|---|---|---|
| | $\tau^*_{XYZW}$ | Over/under represented? | $\tau^*_{XYZW}$ | Over/under represented? |
| AATT | 0.6965 | -- | 0.6560 | -- |
| ATTG | 1.0732 | | 0.7270 | - |
| CATC | 0.7878 | | 0.7738 | - |
| CATG | 0.7129 | - | 0.7167 | - |
| CTAG | 0.2747 | | 0.3442 | |
| GATA | 0.7165 | - | 0.7182 | - |
| GTAG | 1.2469 | + | 1.3802 | ++ |
| TATA | 0.7440 | - | 0.6723 | -- |
| TTAA | 1.3314 | ++ | 1.4219 | ++ |
| TTGG | 0.8569 | | 0.6476 | -- |

served SSR frequencies of a given motif length and repeat number occurred as expected by chance, they were compared with the mean frequencies observed in three randomly shuffled genomes. Randomised sequences were generated with *shuffleseq* (part of the EMBOSS package, http://www.hgmp.mrc.ac.uk/software/EMBOSS). Statistical significance was tested with two-tailed *t*-tests using SPSS 11.0.1 (SPSS). To determine the distribution of SSRs between coding and non-coding regions of the genome, all coding regions were extracted from the sequence using Artemis 4.0 [46] and parsed into a new sequence file using *seqret* (EMBOSS).

### *Analysis of compositional bias*

We determined the compositional bias in di- and tetranucleotides in the chromosome and megaplasmid of *R. solanacearum* GMI1000. Both sequences were concatenated with their inverted complementary sequence using *revseq*, *yank* and *union* (EMBOSS). Mononucleotide frequencies were calculated using Artemis 4.0 [46], di-, tri- and tetra-nucleotide frequencies were calculated using *compseq* (EMBOSS). Dinucleotide relative abundances $\rho^*_{XY}$ were calculated using the equation $\rho^*_{XY} = f_{XY}/f_X f_Y$ where $f_{XY}$ denotes the frequency of dinucleotide XY and $f_X$ and $f_Y$ denote the frequencies of X and Y, respectively [38]. Similarly, the corresponding fourth-order oligonucleotide measures (which factor out all lower-order biases) is given by $\tau^*_{XYZW} = (f^*_{XYZW} f^*_{XY} f^*_{XNZ} f^*_{XN1N2W} f^*_{YZ} f^*_{YNW} f^*_{ZW})/(f^*_{XYZ} f^*_{XYNW} f^*_{YZW} f^*_{X} f^*_{Y} f^*_{Z} f^*_{W})$ were N is any nucleotide and X, Y, Z and W are each one of A, C, G and T [38]. Statistical theory and data from previous studies [38,39] indicate that the normal range of $\rho^*_{XY}$, is between 0.78 and 1.23. In this study we used the refined criteria of discrimination proposed by Karlin et al. [38]. Overrepresentation is indicated by + ($1.23 = \rho^* < 1.30$), ++ ($1.30 = \rho^* < 1.50$) and +++ ($\rho^* \geq 1.50$), while underrepresentation is indicated by - ($0.70 < \rho^* = 0.78$), -- ($0.50 < \rho^* = 0.70$) and --- ($\rho^* = 0.50$). The dissimilarities in relative abundance of dinucleotides between both sequences were calculated using the equation described by Karlin et al. [38]: $\delta^*(f,g) = 1/16\Sigma\, |\rho^*_{XY}(f) - \rho^*_{XY}(g)|$ (multiplied by 1000 for convenience), were the sum extends over all dinucleotides. To assess the variability of dinucleotide relative abundances within a replicon, both replicons were divided into 12 and 7 (for the chromosome and the megaplasmid, respectively) non-overlapping fragments and $\rho^*_{XY}$ values were calculated for each fragment. The average $\delta^*(f,g)$ within each replicon was also calculated.

## List of Abbreviations

SSR : simple sequence repeat

## Authors' Contribution

TC conceived the study and carried out the computational analyses. PV participated in experimental design. Both authors read and approved the final manuscript.

## Acknowledgements

T. C. and P. V. are indebted to the Fund for Scientific Research – Flanders (Belgium) for a position as postdoctoral fellow and research grants, respectively. T.C. also acknowledges the support from the Belgian Federal Government (Federal Office for Scientific, Technical and Cultural Affairs).

## References

1. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathirga R, White O, Ketchum KA, Dodson R and Hickey EK **Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*.** *Nature* 1997, **390:**580-586
2. Lin YS, Kieser HM, Hopwood DA and Chen CW **The chromosomal DNA of *Streptomyces lividans* 6 is linear.** *Mol Microbiol* 1993, **10:**923-933
3. Bentley SD, Chater KF, Cerdenno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H and Harper D **Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2).** *Nature* 2002, **417:**141-147
4. Allardet-Servent A, Michaux-Charachon S, Jumas-Bilak E, Karayan L and Ramuz M **Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome.** *J Bacteriol* 1993, **175:**7869-7874

5. Suwanto A and Kaplan S **Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome : presence of two unique circular chromosomes.** *J Bacteriol* 1989, **171:**5850-5859
6. Suwanto A and Kaplan S **Chromosome transfer in *Rhodobacter sphaeroides*: Hfr formation and genetic evidence for two unique circular chromosomes.** *J Bacteriol* 1992, **174:**1135-1145
7. Honeycutt RJ, McClelland M and Sobral BW **Physical map of the genome of *Rhizobium meliloti* 1021.** *J Bacteriol* 1993, **175:**6945-6952
8. Jumas-Bilak E, Michaux-Charachon S, Bourg G, Ramuz M and Allardet-Servent A **Unconventional genomic organisation in the alpha subgroup of the *Proteobacteria*.** *J Bacteriol* 1998, **180:**2749-2755
9. Michaux S, Paillisson J, Carles-Nurit MJ, Bourg G, Allardet-Servent A and Ramuz M **Presence of two independent chromosomes in the *Brucella melitensis* 16M genome.** *J Bacteriol* 1993, **175:**701-705
10. Jumas-Bilak E, Michaux-Charachon S, Bourg G, O'Callaghan D and Ramuz M **Differences in chromosome number and genome rearrangements in the genus *Brucella*.** *Mol Microbiol* 1998, **27:**99-106
11. Martin-Didonet CCG, Chubatsu LS, Souza EM, Kleina M, Rego FGM, Rigo LU, Yates MG and Pedrosa FO **Genome structure of the genus *Azospirillum*.** *J Bacteriol* 2000, **182:**4113-4116
12. Cheng HP and Lessie TG **Multiple replicons constituting the genome of *Pseudomonas cepacia* 17616.** *J Bacteriol* 1994, **176:**4034-4042
13. Rodley PD, Römling U and Tümmler B **A physical genome map of the *Burkholderia cepacia* type strain.** *Mol Microbiol* 1995, **17:**57-67
14. Lessie TG, Hendrickson W, Manning BD and Devereux R **Genomic complexity and plasticity of *Burkholderia cepacia*.** *FEMS Microbiol Lett* 1996, **144:**117-128
15. Wigley P and Burton NF **Multiple chromosomes in *Burkholderia cepacia* and *B. gladioli* and their distribution in clinical and environmental strains of *B. cepacia*.** *J Appl Microbiol* 2000, **88:**914-918
16. Parke JL and Gurian-Sherman D **Diversity of the *Burkholderia cepacia* complex and implications for risk assessment of biological control strains.** *Annu Rev Phytopathol* 2001, **39:**225-258
17. Songsivilai S and Dharakul T **Multiple replicons constitute the 6.5-megabase genome of *Burkholderia pseudomallei*.** *Acta Trop* 2000, **74:**169-179
18. **DOE Joint Genome Institute** [http://www.jgi.doe.gov/JGI_microbial/html/index.html]
19. Cole ST and Saint-Girons S **Bacterial genomes – all shapes and sizes.** *In: Organisation of the prokaryotic genome (Edited by: Charlebois RL) Washington DC, American Society for Microbiology* 1999, 35-62
20. Itaka M and Tanaka T **Experimental surgery to create subgenomes of *Bacillus subtilis* 168.** *Proc Natl Acad Sci USA* 1997, **94:**5378-5382
21. Ochman H, Lawrence JG and Groisman EA **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304
22. Kennedy SP, Ng WV, Salzberg SL, Hood L and DasSarma S **Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence.** *Gen Res* 2001, **11:**1641-1650
23. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD and Umayam L **DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*.** *Nature* 2000, **406:**477-483
24. Tagomori K, Iida T and Honda T **Comparison of genome structures of Vibrios, bacteria possessing two chromosomes.** *J Bacteriol* 2002, **184:**4351-4358
25. Hayward AC **Biology and epidemiology of bacterial wilt caused by *Pseudomonas solanacearum*.** *Annu Rev Phytopathol* 1991, **29:**65-87
26. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Biliaut A, Brottier P, Camus JC and Cattolico L **Genome sequence of the plant pathogen *Ralstonia solanacearum*.** *Nature* 2002, **415:**497-502
27. Rosenberg C, Casse-Delbart F, Dusha I, David M and Boucher C **Megaplasmids in the plant-associated bacteria *Rhizobium meliloti* and *Pseudomonas solanacearum*.** *J Bacteriol* 1982, **150:**402-406
28. van Belkum A, Scherer S, Van Alphen L and Verbrugh H **Short-sequence repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62:**275-293
29. Yeramian E and Buc H **Tandem repeats in complete bacterial genome sequences : sequence and structural analyses for comparative studies.** *Res Microbiol* 1999, **150:**745-754
30. van Belkum A, Van Leeuwen W, Scherer S and Verbrugh H **Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes.** *Res Microbiol* 1999, **150:**617-626
31. Bzymek M and Lovett ST **Instability of repetitive DNA sequences : the role of replication in multiple mechanisms.** *Proc Natl Acad Sci USA* 2001, **98:**8319-8325
32. Hallet B **Playing Dr Jekyll and Mr Hyde : combined mechanisms of phase variation in bacteria.** *Curr Opin Microbiol* 2001, **4:**570-581
33. Henderson IR, Owen P and Nataro JP **Molecular switches – the ON and OFF of bacterial phase variation.** *Mol Microbiol* 1999, **33:**919-932
34. Liu L, Panangala VS and Dybvig K **Trinucleotide GAA repeats dictate pMGA gene expression in *Mycoplasma gallisepticum* by affecting spacing between flanking regions.** *J Bacteriol* 2002, **184:**1335-1339
35. van der Ende A, Hopman CTP, Zaat S, Oude Essink BB, Berkhout B and Dankert J **Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the -10 and -35 regions of the promotor.** *J Bacteriol* 1995, **177:**2475-2480
36. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM and Kashi Y **Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition and polymorphism.** *Gen Res* 2000, **10:**62-71
37. Burge C, Campbell AM and Karlin SA **Over- and under-representation of short oligonucleotides in DNA sequences.** *Proc Natl Acad Sci USA* 1992, **89:**1358-1362
38. Karlin S, Mrazek J and Campbell AM **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179:**3899-3913
39. Karlin S, Campbell AM and Mrazek J **Comparative DNA analysis across diverse genomes.** *Annu Rev Genet* 1998, **32:**185-225
40. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH and Ussery DW **A DNA structural atalas for *Escherichia coli*.** *J Mol Biol* 2000, **299:**907-930
41. Strand M, Prolla TA, Liskay RM and Petes TD **Destabilisation of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair.** *Nature* 1993, **365:**274-276
42. Hauge XY and Litt M **A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR.** *Hum Mol Genet* 1993, **2:**411-415
43. Chiurazzi P, Kozak L and Neri G **Unstable triplets and their mutational mechanisms : size reduction of the CGG repeat vs. germline mosaicism in the fragile X syndrome.** *Am J Med Genet* 1994, **15:**517-521
44. Campbell A, Mrazek J and Karlin S **Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA.** *Proc Natl Acad Sci USA* 1999, **96:**9184-9189
45. Wong K, Finan TM and Golding GB **Dinculeotide compositional analysis of *Sinorhizobium meliloti* using the genome signature : distinguishing between chromosomes and plasmids.** *Funct Integr Genomics* 2002, **2:**274-281
46. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA and Barell B **Artemis : sequence visualisation and annotation.** *Bioinformatics* 2000, **16:**944-945