Insights into evolution history of Burkholderia spp.

E.V. Moroz (Lopatina)
Institute for Information Transmission Problems, Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia
e lena.v.lopatina@gmail.com

M.D. Kazanov
Institute for Information Transmission Problems, Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia
m.kazanov@gmail.com

A.S. Kalinina
Institute for Information Transmission Problems, Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia
a.s.kalinina@gmail.com

O.O. Bochkaryova
Institute for Information Transmission Problems, Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia
olga.bochkaryova@gmail.com

Abstract. Bacteria with multiple chromosomes belong to Actinobacteria, Chloroflexi, Deinococcus-Thermus, Firmibacteres, Firmicutes, Proteobacteria and Spirochaetes. In our research we consider genus Burkholderia belonging to Betaproteobacteria. Evolution of these bacteria is of great interest because of their multichromosomal genome organization, their species consists of two or three chromosome. We reconstructed translocations between chromosomes. Also we made a reconstruction of events of gain/loss. It was done by two methods for orthologs and synteny blocks. Another important force shaping the genomic evolution is homologous recombination. We identified homologous recombinations in mallei/pseudomallei group.

1 Introduction

Burkholderia are Gram-negative, motile, aerobic rod-shaped bacteria. They occupy wide range of ecological niches. This genus includes human, animal and plant pathogens [1, 2], polychlorinated biphenyl-degrader [3], using in biodegradation studies of pollutants and plant-beneficial endophyte [4], that is exploited for plant growth promotion purposes.

Besides of their diversity, these bacteria are interesting because of their multichromosomal genome organization. Species’ genome consists of two or three circular chromosomes. One of chromosomes is larger and harbors essential genes associated with metabolism and cell growth. Other chromosomes are smaller and contain more niche-specific genes [5]. Studies on Vibrio cholera which contains two chromosomes, demonstrated that chromosomes in bacteria have different mechanisms for replication and segregation and try to avoid competition [6].

Research of Burkholderia evolution is important for gain further insights in their diversity and genome organization.

2 Pan-genome analysis

We provide pan-genome analysis for 28 Burkholderia strains. Fig.1 shows how core genome size depends on the number of analyzed strains. This graphic demonstrated that while the size of a single genome varies from 2,000 to 8,000 orthologs, the number of universal genes present in all strains saturates at 1,400.

By definition [7] the pan-genome is the set of genes (orthologs) that belong to at least one strain. The pan-genome size for all strains is 30,000 orthologs and it shows no signs of saturation (Fig. 2a). Hence, the gene diversity of Burkholderia species has not been captured yet. However, this is caused mainly by orphan genes observed in a single genome. Indeed, an analogous plot that ignores single-strain genes reaches plateau (Fig.2b).

3 Reconstruction of translocations of orthologs between chromosomes

To analyze interchromosomal translocations we considered single-copy universal gene. We determine their distribution among chromosomes (Tab.1) and constructed the phylogenetic tree using these orthologs (Fig.3a). Distribution of orthologs is consistent with tree as closely related strains have similar composition of universal
orthologs on chromosomes. While number of orthologs on first and second chromosomes is comparable, number of universal orthologs on the first chromosome is by an order of magnitude greater than on second chromosome. This fact confirms different roles of primary and secondary chromosomes. Above mentioned translocation between first and third chromosome in Burkholderia cenocepacia AU 1054 is reflected in table as this strain simultaneously lacks on first chromosome and obtains on third chromosome exactly 208 universal genes relative to other cenocepacia.

Another tree (Fig.3b) was produced, using pairwise distance matrix 

\[
D_{ij} = 1 - \frac{|\text{Strain}_i \cap \text{Strain}_j|}{|\text{Strain}_i \cup \text{Strain}_j|},
\]

where \(|\text{Strain}_i|\) is number of genes belonging to given strain.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Chromosome 1 Universal</th>
<th>Chromosome 1 Sum</th>
<th>Chromosome 2 Universal</th>
<th>Chromosome 2 Sum</th>
<th>Chromosome 3 Universal</th>
<th>Chromosome 3 Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burkholderia sp. 383</td>
<td>1065</td>
<td>3334</td>
<td>75</td>
<td>3174</td>
<td>1</td>
<td>1209</td>
</tr>
<tr>
<td>ambifaria AMMD</td>
<td>1065</td>
<td>3206</td>
<td>75</td>
<td>2346</td>
<td>1</td>
<td>1013</td>
</tr>
<tr>
<td>vietnamiensis G4</td>
<td>1065</td>
<td>3274</td>
<td>75</td>
<td>2066</td>
<td>1</td>
<td>1114</td>
</tr>
<tr>
<td>multivorans ATCC 17616 IG</td>
<td>1065</td>
<td>3146</td>
<td>75</td>
<td>2152</td>
<td>1</td>
<td>853</td>
</tr>
<tr>
<td>multivorans ATCC 17616</td>
<td>1065</td>
<td>3104</td>
<td>75</td>
<td>2130</td>
<td>1</td>
<td>766</td>
</tr>
<tr>
<td>thailandensis E264</td>
<td>1054</td>
<td>3276</td>
<td>87</td>
<td>2358</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>pseudomallei 688</td>
<td>1054</td>
<td>3515</td>
<td>87</td>
<td>3279</td>
<td>0</td>
<td>1335</td>
</tr>
<tr>
<td>pseudomallei KS6243</td>
<td>1054</td>
<td>3599</td>
<td>87</td>
<td>2329</td>
<td>0</td>
<td>1080</td>
</tr>
<tr>
<td>pseudomallei 1710b</td>
<td>1054</td>
<td>3611</td>
<td>87</td>
<td>3736</td>
<td>0</td>
<td>260</td>
</tr>
<tr>
<td>pseudomallei 1301a</td>
<td>1054</td>
<td>4015</td>
<td>87</td>
<td>3159</td>
<td>0</td>
<td>267</td>
</tr>
<tr>
<td>mallei ATCC 23344</td>
<td>1027</td>
<td>2995</td>
<td>114</td>
<td>2029</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>mallei NCTC 10229</td>
<td>1027</td>
<td>3333</td>
<td>114</td>
<td>2217</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>mallei NCTC 10247</td>
<td>1027</td>
<td>3476</td>
<td>114</td>
<td>2176</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>glutinar RGR1</td>
<td>1065</td>
<td>3287</td>
<td>76</td>
<td>2079</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>glutol RGR1</td>
<td>1065</td>
<td>3152</td>
<td>76</td>
<td>2459</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>phytoxytrom SmR15</td>
<td>1065</td>
<td>3072</td>
<td>76</td>
<td>2349</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>sp. CGGE2001</td>
<td>1065</td>
<td>3116</td>
<td>76</td>
<td>2258</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>pseudomallei 18400</td>
<td>1071</td>
<td>4480</td>
<td>70</td>
<td>2960</td>
<td>0</td>
<td>1312</td>
</tr>
<tr>
<td>phytoformans P41N</td>
<td>1071</td>
<td>3922</td>
<td>70</td>
<td>2352</td>
<td>0</td>
<td>1312</td>
</tr>
<tr>
<td>sp. CGGE2001</td>
<td>1071</td>
<td>3280</td>
<td>70</td>
<td>2236</td>
<td>0</td>
<td>1312</td>
</tr>
<tr>
<td>sp. CGGE2005</td>
<td>1071</td>
<td>5270</td>
<td>70</td>
<td>2350</td>
<td>0</td>
<td>1312</td>
</tr>
<tr>
<td>mallei AHO 454</td>
<td>1084</td>
<td>1921</td>
<td>57</td>
<td>384</td>
<td>0</td>
<td>1312</td>
</tr>
</tbody>
</table>

Table 1. Distribution of universal orthologs among chromosomes.
Reconstructed translocations of universal genes are represented on the figure 4.

We reconstructed 338 events. Due to results common ancestor of *Burkholderia* has 1068 universal orthologs on the first chromosome and 73 on second. The interesting thing that some orthologs translocated at the same time on distant branches, for example 12 orthologs translocated on branch with *thailandensis, pseudomallei*, *mallei* and on branch with *cenocpecia* AU 1054. These orthologs belong to one translocated segment of the genome in *cenocpecia, thailandensis* and *pseudomallei* and to two segments of genome in *mallei*. There are 18 such universal orthologs, one of them took a part in translocations on three branches. So this tree gives us a comprehensive view of translocations between chromosomes since existence of common ancestor till appearance of each contemporary strain.
4 Gene Gain/Loss Events

We assessed gene gain/loss on the phylogenetic tree using stochastic mapping approach implemented in GLOOME (1). Results are reflected on Figure 5.

Numbers that you can see on the figure are not integer because stochastic mapping approach considers all possible scenarios weighted by its probability of occurrence. As you can see the part of the tree with pseudomallei/mallei species is poor resolved. It was previously known that *Burkholderia mallei* strains are similar in their gene content and belong to one MLST [8]. To get better results for mallei/pseudomallei species we built nucleotide tree. And we carried out analysis of gene gain/loss processes for this part of tree separately (Fig. 6). Relatively big loss of genes occurred on the branch with *mallei*. Genome reduction among *mallei* strains is likely associated with loss of genes which are not required for living in a host [9].

Figure 5. Number of gain (red) and loss (blue) genes for whole tree.

Figure 6. Number of gain (red) and loss (blue) orthologs for mallei/pseudomallei.
5 Synteny blocks identification and genome rearrangements

We also reconstructed rearrangement history using synteny blocks. We produced trees that consider inversions, insertions and deletions. First tree for ten strains containing three chromosomes consider 14 inversions and 236 indel events.

Figure 7. Phylogenetic tree for *cenocepacia* branch that consider insertions and deletions.

Among them were found two parallel inversions that occurred in *B. cenocepacia* AU 1054 and *B. cenocepacia* J2315.

Two trees were produced for the branch with *thailandensis, mallei, pseudomallei* (Fig. 8, 9). We obtained 94 inversions and 169 indels for these strains. First tree takes into account only inversions. When we consider all events including insertions and deletions, this topology occurs non-optimal and it is required *mallei* and *pseudomallei* separate in two clusters. It is interesting that this topology is similar to topology of tree based on pairwise distance matrix (Fig. 3b).

Figure 8. Phylogenetic tree for *mallei* branch that consider inversions.

Figure 9. Phylogenetic tree for *mallei* branch that consider insertions and deletions.

The last tree was produced taking into account inversions in all strains. As strains in *cenocepacia* and *mallei* branches are closely related in comparison with other *Burkholderia* strains in our study, we chose only one representative strain from each of these branches (Fig. 10).
6 Homologous recombination in mallei/pseudomallei

Since genetic distances between all considered *Burkholderia mallei* are very small, and, in this sense, their genomes are almost identical, we use here one strain of *Burkholderia mallei*: ATCC 23344 (strain 0), and four strains of *Burkholderia pseudomallei*: 1106a, 1710b, 668 and K96243 (strains 1, 2, 3, 4 respectively). For each pairwise comparison of strains was built histogram of number of SNPs per 1000 nt segment (Fig. 11).

Figure 11. Histograms of number of SNPs per 1000 nt for pairwise comparisons between strains. For each plot considered strains and genetic distance between them are shown in the plot title.
7 Material and Methods

Genomes. Available complete genomes of twenty eight *Burkholderia* strains were selected for analysis. The genomes were taken from the NCBI Genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

Orthologs. Orthologs were constructed as follows. Bidirectional best hits (BBHs) were constructed for each pair of strains using BLASTP [10]. BLASTP results with identities of 50% or coverage of shorter sequence 67% were ignored. At the next step, if two paralogs were more similar to each other than to either BBH partner, both paralogs were added to the ortholog. Then, maximal connected components were constructed. This was done using ad hoc software based on the Relational Database Management System (RDBMS) ORACLE Express Edition.

Phylogenetic trees. For alignment of protein sequences was used Muscle 3.8 [11]. For tree building was used MEGA 5.1 [12].

Translocations. To reconstruct translocations between chromosomes we numbered orthologs from 1 to 1141. Then for each strain we made sequence in which number of character was equal to number of ortholog and character was number of chromosome that this ortholog belonged to in a given strain. For ancestral reconstruction was used PAML 4.6 [13] with REV(GTR) model.

Gene gain/loss events. To provide gain/loss analysis we used GLOOME [14]. We chose evolution model with variable gain/loss ratio. It means that this ratio could be different among different orthologs. And also our model allows root freq to differ from stationary ones. This condition adds a single free parameter to the model, doing it more flexible.

Synteny blocks. Synteny blocks for closely related strains were constructed using Sibelia algorithm with standard parameters. We filtered blocks that were found in some genome more than once. Synteny blocks for far strains were constructed using Drinn-Synteny algorithm based on sequences of universal genes. Rearrangements history for given topology and optimal phylogenetic tree based on rearrangements were obtained using MGRA 2.2 server.

Homologous recombinations. For homologous recombination analysis were used 3018 universal single-copy genes. These genes were aligned with ClustalW [15] and for each alignment were removed all gap-contained columns. Further, all alignments were concatenated in alignment of the total length 3082491 nt.

8 References


