Comparative analysis of non-local events in strain evolution within prokaryotic species

Smirenina L.K.

Moscow State University, Leninskie Gori 1, build.40, Moscow119992, Russia,

lks@belozersky.msu.ru

Alexeevski A.V.

Moscow State University, Leninskie Gori 1, build.40, Moscow119992, Russia,
Scientific Research Institute for System Studies, the Russian Academy of Science (NIISI RAS), Moscow 117281, Russia

aba@belozersky.msu.ru

Non-local events, long deletions and insertions, inversions, translocations, duplications, mobile genetic elements expansions and foreign DNA uptakes, are frequent in evolution of bacteria and archaea. They play significant biological role in prokaryotes adaptation and fitness [1]. Due to growing number of completely (up to chromosomes) assembled prokaryotic genomes, it is of interest to compare abundance of non-local events within species of various taxa.

For this purpose, we identified all prokaryotic species with at least 10 completely assembled genomes of strains, according PATRIC DB (https://www.patricbrc.org/).   About one hundred species were selected. We downloaded genomes from PATRIC. Nucleotide pangenomes were constructed by NPG-explorer program with default parameters for 80 species.

Nucleotide pangenome (NPG) is a set of alignments (blocks) simply covering all input sequences. Major and minor blocks are distinguished. By definition, percent of identical positions in a major block is over an identity threshold, 90% by default, and the length of a major block is over a length threshold, 100 positions by default. A major block may contain any number of sequence fragments. Particularly blocks with several fragments from one genome are allowed as well as ‘blocks’ of one fragment. A minor block consists of fragments connecting the same major blocks. All its fragments are shorter than the length threshold.

Major blocks are classified into four categories.

Stable blocks (s-blocks) consist of exactly one fragment from each genome. Presumably, all fragments of an s-block are descendants of one fragment of the genome of last common ancestor (LCA). Thus, joined alignment of all s-blocks can be considered 'nucleotide core' of genomes. The above conclusion is correct only if identity of all (or almost all) alignments of orthologous fragments exceeds this threshold. This is why we restricted comparison of NPGs to 44 species that have average identity of s-blocks over 92.5%. Percent of nucleotides in all s-blocks with respect to total nucleotides on input we used as a characteristic of genome stability.

Stability of species' genomes may depend on the number of strains analyzed. However, among 44 species NPGs correlation between the number of genomes and nucleotide core percent was not found (r = 0.15).

To some extent, percent of identical positions in joined alignment of s-blocks reflects distance to LCA. Predictably, core size positively correlate with identity of joined s-blocks alignment, see Fig. 1. However, there is a number of outliers. We added to our data the NPG of two strains of *Orientia tsutsugamushi* species, which is susceptible to multiple large scale evolutionary events [2]. In fact, its nucleotide core size (45%) is minimal in our data although identity of nucleotide core is rather high, about 97% .

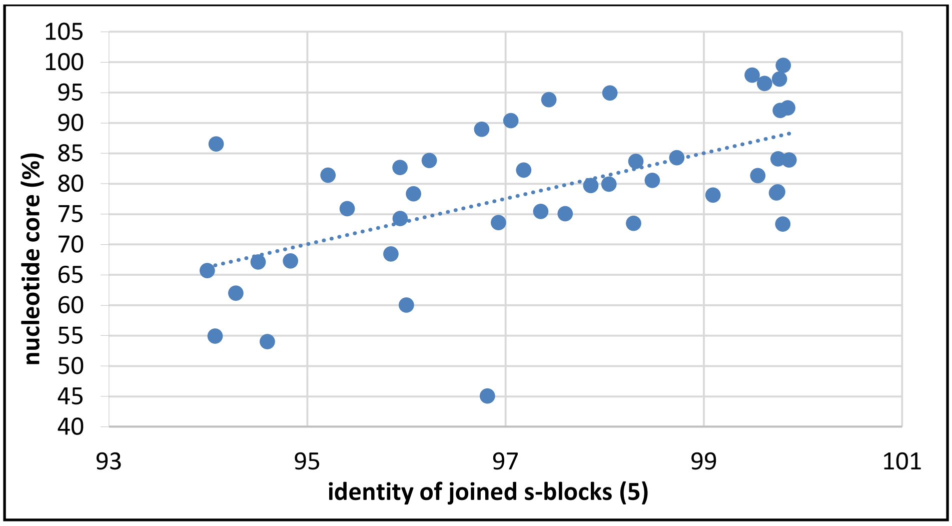


Fig.1. Nucleotide core size dependence of nucleotide core identity for 45 prokaryotic species.

Hemi-stable blocks (h-blocks) consist of exactly one fragment from a subset of genomes. The following scenarios may be manifested in h-blocks. First is long deletion occurred in last common ancestor of a clade in phylogenetic tree of strains. Second is lateral transfer to last common ancestor of a clade. More complicated scenarios are possible. For example, in NPG of 55 strains of *Brucella* genus we observed a number of long deletions of the same fragments in parallel branches of the phylogenetic tree. This effect is known [3].

Unique sequences (u-blocks) are ‘blocks’ of one fragment. According requirements on NPG, it implies that no homologous fragments of prescribed length and identity exists in input genomes. Unique sequences as well as h-blocks appeared according second scenario (see above) are good candidates for lateral transfer study. We have examined such possibility in NPGs of a few species.

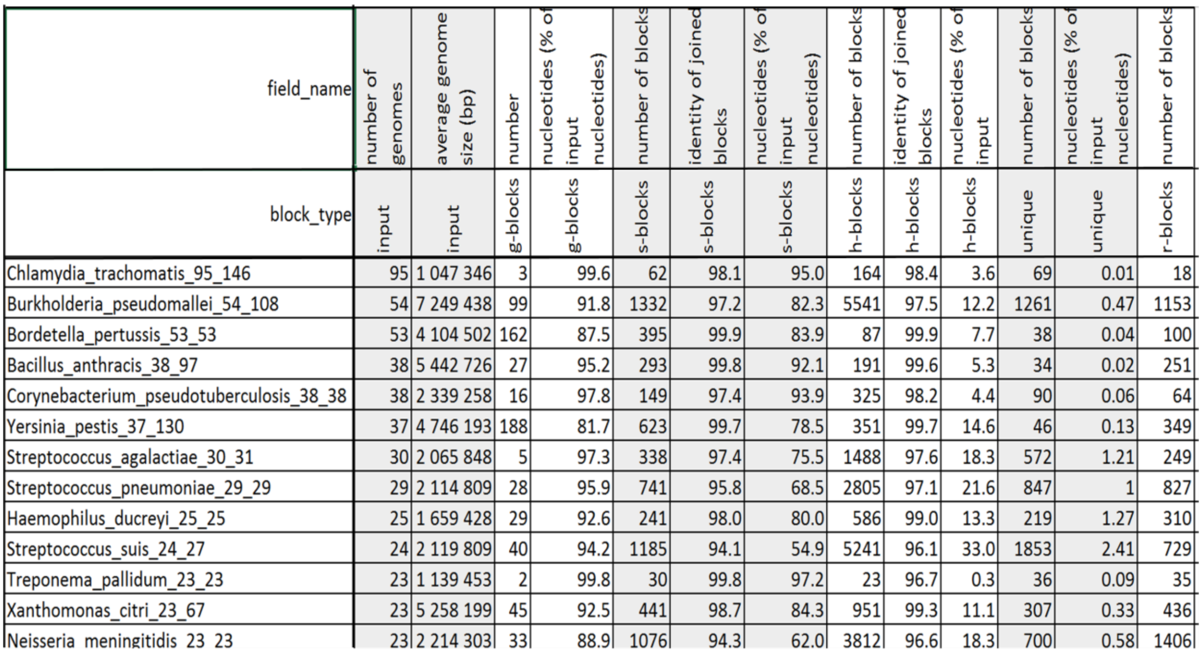
Blocks with repeats (r-blocks) include two or more fragments of at least one genome. Most frequent r-blocks contain single duplications in one genome. Another source of r-blocks is duplications in the LCA genome. Demonstrative examples are ribosomal clusters in genomes. Additional sources of r-blocks are active mobile elements. Different kinds of r-blocks may be preliminary distinguished using data on distributions of r-block fragments among input genomes. These data are included NPG-explorer output.

Minor blocks (m-blocks) consist of short (less than the block length threshold) fragments connecting the same major blocks. They are grey zone in nucleotide pangenomes. Fortunately, m-blocks typically contain less than 1% of nucleotides.

NPG-explorer identifies also synteny regions (g-blocks). g-block consists of collinear s-blocks and blocks of other types between them. Presumably, g-blocks correspond to regions in LCA genome that underwent no rearrangements in evolution; long deletions, insertions and invasions of repetitive elements may have occurred. Thus, number of g-blocks characterize frequency of genome rearrangements, inversions and translocations. Alignments of g-blocks in genomes in comparison with phylogenetic tree of strains allows reconstructing rearrangements in evolution.

Table 1 demonstrates some data about NPGs provided by NPG-explorer.

Table 1. Information on all kinds of NPG blocks for 44 prokaryotic species (fragment).



field_name	number of genomes	average genome size (bp)	number	nucleotides (% of input nucleotides)	number of blocks	identity of joined blocks	nucleotides (% of input nucleotides)	number of blocks	identity of joined blocks	nucleotides (% of input	number of blocks	nucleotides (% of input nucleotides)	number of blocks
block_type	input	input	g-blocks	g-blocks	s-blocks	s-blocks	s-blocks	h-blocks	h-blocks	h-blocks	unique	unique	r-blocks
Chlamydia_trachomatis_95_146	95	1 047 346	3	99.6	62	98.1	95.0	164	98.4	3.6	69	0.01	18
Burkholderia_pseudomallei_54_108	54	7 249 438	99	91.8	1332	97.2	82.3	5541	97.5	12.2	1261	0.47	1153
Bordetella_pertussis_53_53	53	4 104 502	162	87.5	395	99.9	83.9	87	99.9	7.7	38	0.04	100
Bacillus_anthracis_38_97	38	5 442 726	27	95.2	293	99.8	92.1	191	99.6	5.3	34	0.02	251
Corynebacterium_pseudotuberculosis_38_38	38	2 339 258	16	97.8	149	97.4	93.9	325	98.2	4.4	90	0.06	64
Yersinia_pestis_37_130	37	4 746 193	188	81.7	623	99.7	78.5	351	99.7	14.6	46	0.13	349
Streptococcus_agalactiae_30_31	30	2 065 848	5	97.3	338	97.4	75.5	1488	97.6	18.3	572	1.21	249
Streptococcus_pneumoniae_29_29	29	2 114 809	28	95.9	741	95.8	68.5	2805	97.1	21.6	847	1	827
Haemophilus_ducreyi_25_25	25	1 659 428	29	92.6	241	98.0	80.0	586	99.0	13.3	219	1.27	310
Streptococcus_suis_24_27	24	2 119 809	40	94.2	1185	94.1	54.9	5241	96.1	33.0	1853	2.41	729
Treponema_pallidum_23_23	23	1 139 453	2	99.8	30	99.8	97.2	23	96.7	0.3	36	0.09	35
Xanthomonas_citri_23_67	23	5 258 199	45	92.5	441	98.7	84.3	951	99.3	11.1	307	0.33	436
Neisseria_meningitidis_23_23	23	2 214 303	33	88.9	1076	94.3	62.0	3812	96.6	18.3	700	0.58	1406

We conclude that suggested methods are useful for comparison of prokaryotic evolution on the level of strains in a large scale. In the talk examples and statistical data on many dozens of species will be presented.

The work was supported by RSF grant 16-14-10319.

1. Lobkovsky AE et al., (2015) Evolvability of an Optimal Recombination

Rate, *Genome Biol Evol.,* **8:**70-77

2. Nakayama K, et al., (2010), Genome comparison and phylogenetic analysis of *Orientia tsutsugamushi* strains. *DNA Res.* **17**:281-291

3. Raeside C, et al., (2014), Large chromosomal rearrangements during a long-term evolution experiment with Escherichia coli. *MBio,* **5**:e01377-14