# NPGe, a new tool for closely related genomes alignment and analysis

Boris Nagaev[1,2], Maxim Nikolaev[2], Andrei Alexeevski[1,2,3]

[1]*Moscow State University, A.N. Belozersky Institute, Leninskye gory 1-40, Moscow 119992, Russia*

[2]*Moscow State University, Faculty of Bioengineering and Bioinformatics, Leninskye gory 1-73, Moscow 119234, Russia*

[3]*Scientific Research Institute for System Studies, the Russian Academy of Science (NIISI RAS), Moscow 117281, Russia.*

*aba@belozersky.msu.ru*

Alignment of highly similar genomic sequences is required for several purposes. First, alignment of genomes of closely related prokaryotes allows reconstructing evolutionary events and improving gene annotations. Second, alignment of highly similar genomic sequences is useful for comparison of genome assemblies of the same organism or closely related ones. The alignment allows improving assembly in a number of cases. Particular interest is in comparison of assemblies of genomes with divergent (10 – 15% of SNPs) haplotypes.

As a result of high sequence similarity orthologous fragments of genomic sequences can be almost unambiguously determined just by sequence identity percent above certain threshold (e.g. 90%). Here we consider this property the definition of closely related genomes.

Ideally all differences between highly similar genomes might be detected up to a couple dozens of nucleotides. Practically, available multiple genome aligners (progressiveMAUVE, VISTA, etc,) are universal and do not explore high sequence similarity on a full scale. This is why we developed a new tool designed primarily for multiple alignments of genomes of closely related organisms. The goals of our work was to develop tools for detailed analysis of evolutionary events and for comparison of genome annotations.

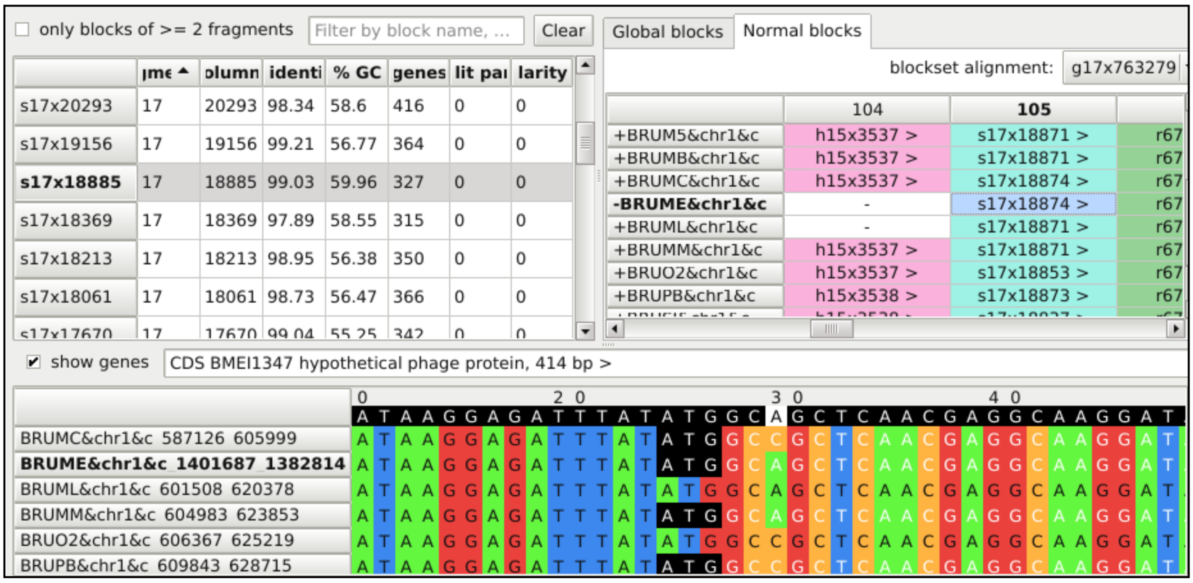
To achieve the goals as perfect genome alignments as possible are needed. We specify the definition of multiple genome alignment [1] for the case of closely related genomes and call it ‘a nucleotide pangenome’.

**Definition.** Nucleotide pangenome (NPG) of an input set of genomes is a set of aligned blocks, each block composed of orthologous fragments. A block may contain fragments either from all genomes or from a subset of them, fragments of the same genome are allowed. Blocks must meet the criteria: (i) conservation within a block must be over a threshold (e.g. > 90% of conserved positions) along alignment, including its ends; (ii) block length must be over a threshold (e.g. >100 positions); fragments having no orthologs are considered dummy blocks of one fragment; (iii) each nucleotide from input genomes belongs to exactly one fragment of one block; (iv) no similar sequences of required similarity and length can be detected between any pair of blocks.

**NPG block types.** Blocks are classified as follows. Stable blocks (s-blocks) contain exactly one fragment from each input genomic sequence. Hemi-stable blocks (h-blocks) include no more than one fragment of a genome and lack a fragment of at least one of genomes. Unique sequences (u-blocks of exactly one fragment) are appropriately long sequences that meet no long enough similar fragments in input data. Blocks with repeats (r-blocks) include more than one fragment from one or several genomes. Minor blocks (m-blocks) are blocks shorter than a threshold and typically with similarity less than a threshold. These blocks represent zone of uncertainty in nucleotide pangenome. Additionally we define global blocks (g-blocks), which are joined collinear s-blocks and intermediate blocks between them. Global blocks are analogs of synteny regions usually determined by conserved sequences of orthologous genes. Intermediate blocks (i-blocks) are union of all block fragments between two g-blocks.

**Algorithm.** Two types of data are used. Blockset is any set of blocks. Pre-pangenome is a blockset meeting all the criteria except (iv). The algorithm iterates main procedure, which takes on input a pre-pangenome and a blockset and returns new pre-pangenome. At first iteration pre-pangenome is trivial, i.e. all input sequences are considered one fragment blocks. Initial blockset is created by anchors detection, as sets of identical words of fixed length (e.g. 20 bp) [1]. For acceleration we used Bloom filter [2]. Anchors are used for constructing a pre-pangenome by expanding anchors to blocks of appropriate quality, resolving blocks intersections, joining collinear blocks. At further iterations consensuses of all pre-pangenome blocks are calculated. Pair blocks between consensuses are found by BLAST ‘all against all’. These blocks are transformed into blocks of genomes’ fragments. The same operations as described above are applied. Iterations repeated until all conditions of nucleotide pangenome are satisfied.

**Program.** The algorithm is implemented in MakePangenome program within NPG-explorer package. NPG-explorer includes also (i) *Prepare* utility facilitating downloading genomic sequences and their annotations from GeneBank, ENA or other databanks and reformatting them appropriately; (ii) *PostProcessing* module which computes a number of analytical data; (iii) NPG visualization module *qnpge* (Fig. 1). The programs are written in Lua and C++ languages. Current version of NPGexplorer is freely available at http://mouse.belozersky.msu.ru/tools/npge.html.



□ only blocks of >= 2 fragments | Filter by block name, ... | Clear

| | ıme ▲ | olumn | identi | % GC | genes | lit pai | larity |
|---|---|---|---|---|---|---|---|
| s17x20293 | 17 | 20293 | 98.34 | 58.6 | 416 | 0 | 0 |
| s17x19156 | 17 | 19156 | 99.21 | 56.77 | 364 | 0 | 0 |
| **s17x18885** | 17 | 18885 | 99.03 | 59.96 | 327 | 0 | 0 |
| s17x18369 | 17 | 18369 | 97.89 | 58.55 | 315 | 0 | 0 |
| s17x18213 | 17 | 18213 | 98.95 | 56.38 | 350 | 0 | 0 |
| s17x18061 | 17 | 18061 | 98.73 | 56.47 | 366 | 0 | 0 |
| s17x17670 | 17 | 17670 | 99.04 | 55.25 | 342 | 0 | 0 |

Global blocks | Normal blocks

blockset alignment: g17x763279

| | 104 | **105** | |
|---|---|---|---|
| +BRUM5&chr1&c | h15x3537 > | s17x18871 > | r67 |
| +BRUMB&chr1&c | h15x3537 > | s17x18871 > | r67 |
| +BRUMC&chr1&c | h15x3537 > | s17x18874 > | r67 |
| **-BRUME&chr1&c** | - | s17x18874 > | r67 |
| +BRUML&chr1&c | - | s17x18871 > | r67 |
| +BRUMM&chr1&c | h15x3537 > | s17x18871 > | r67 |
| +BRUO2&chr1&c | h15x3537 > | s17x18853 > | r67 |
| +BRUPB&chr1&c | h15x3538 > | s17x18873 > | r67 |

☑ show genes | CDS BMEI1347 hypothetical phage protein, 414 bp >

```
                              0                   2 0                 3 0                 4 0
                              A T A A G G A G A T T T A T A T G G C A G C T C A A C G A G G C A A G G A T
BRUMC&chr1&c  587126  605999  A T A A G G A G A T T T A T A T G G C C G C T C A A C G A G G C A A G G A T
BRUME&chr1&c_1401687_1382814  A T A A G G A G A T T T A T A T G G C A G C T C A A C G A G G C A A G G A T
BRUML&chr1&c  601508  620378  A T A A G G A G A T T T A T A T G G C A G C T C A A C G A G G C A A G G A T
BRUMM&chr1&c  604983  623853  A T A A G G A G A T T T A T A T G G C A G C T C A A C G A G G C A A G G A T
BRUO2&chr1&c  606367  625219  A T A A G G A G A T T T A T A T G G C C G C T C A A C G A G G C A A G G A T
BRUPB&chr1&c  609843  628715  A T A A G G A G A T T T A T A T G G C C G C T C A A C G A G G C A A G G A T
```

*Fig.1. NPG of 17 Blucella genomes visualized with qnpge. The top left panel presents the list of all blockst; block parameters are shown. Part of the alignment of the highlighted stable block s17x18885 is shown in bottom panel. Genes are shown with white letters, start (and stop) codons are highlighted, name of selected gene is shown in the middle one line panel. It is seen that in two genomes gene annotations are missed. The right top panel presents the alignment of block IDs within g-block g17x763279. Column 104 reveals the deletion of 3537 bp in two genomes. Signs ">" and "<" indicate direction of a fragment in chromosome sequence*

**Results.** Using NPG-explorer we constructed nucleotide pangenomes of several taxa of prokaryotes. Statistical data is presented in table 1. NPG-explorer was successfully used also for comparison assemblies of genome of two biosamples *Amoeboaphelidium protococcarum* and for comparison mitochondrial genomes of 24 species of mosses.

NPG-explorer showed itself convenient for revealing all point mutations (listed in separate output file) as well as global evolutionary events – duplications (r-blocks), long deletions/insertions (h-blocks and u-blocks), rearrangements (appeared in alignment of blocks, Fig.1). Joined alignment of all stable blocks is used for construction phylogenetic tree of genomes, presented in one of output files.

Table 1. Nucleotide pangenomes of five groups of bacterial genomes.

| | Species Yersinia pestis | Genus Brucella | Species Rickettsia rickettsii | Species Burkholderia cenocepacia |
|---|---:|---:|---:|---:|
| Number of genomes | 12 | 17 | 8 | 5 |
| Number of chromosomes in a genome | 1 | 2 | 1 | 3 |
| Average genome size (b.p.) | 4584592 | 3306924 | 1266154 | 7693300 |
| Minimum block length threshold | 100 | 100 | 100 | 100 |
| Minimum identity threshold | 0.86 | 0.90 | 0.90 | 0.83 |
| Size of NPG (b.p.) | 4523299 | 3395644 | 1265809 | 10129300 |
| Number of g-blocks (global) | 129 | 26 | 6 | 69 |
| Sum of g-blocks' lengths (% of NPG) | 91.2 | 96.3 | 99.7 | 69.3 |
| Number of s-blocks (stem) | 532 | 193 | 100 | 1109 |
| Sum of s-blocks' lengths (% of NPG) | 87.6 | 90.4 | 98.4 | 56.7 |
| Identity of s-blocks | 99.5 | 98.4 | 99.2 | 92.6 |
| Number of h-blocks (hemi) | 119 | 62 | 10 | 1385 |
| Sum of h-blocks' lengths (% of NPG) | 9.0 | 7.0 | 1.0 | 21.6 |
| Number of u-blocks (unique) | 19 | 9 | 0 | 556 |
| Sum of u-blocks' lengths (% of NPG) | 0.6 | 1.4 | 0.0 | 16.8 |
| Number of r-blocks (repeats) | 228 | 90 | 51 | 925 |
| Sum of r-blocks' lengths (% of NPG) | 1.5 | 0.8 | 0.6 | 3.0 |
| Number of m-blocks (minor) | 789 | 281 | 24 | 1441 |
| Sum of m-blocks' lengths (% of NPG) | 1.3 | 0.3 | 0.0 | 1.9 |
| MakePangenome operation time (min) | 70 | 56 | 8 | 116 |
| Number of iterations | 21 | 15 | 13 | 21 |
| PostProcessing operation time (min) | 2 | 2 | 1 | 6 |

This work is partially supported by Russian Foundation for Basic Research grant 14-04-01693 (Nagaev, Nikolaev) and Russian Science Foundation grant 14-50-00029 (Alexeevski).

1. Dewey,C.N. (2012) Whole-genome alignment. *Methods Mol. Biol.*, **855:**237-257

2. Bloom,B.H. (1970) Space/Time Trade-offs in Hash Coding with Allowable Errors. *Communications of the ACM*, **13**:422–426.