NPG-explorer: a new tool for nucleotide pangenome construction and closely related prokaryotic genomes analysis

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

[1]A.N.Belozersky Institute Of Physico-Chemical Biology and [2]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 119991, Russia [3]Scientific Research Institute of System Analysis (NIISI RAS), Moscow 117218, Russia

bnagaev@gmail.com

Genomes of closely related bacteria have highly similar sequences of orthologous fragments but usually undergo multiple rearrangements, longdeletions, insertions of mobile elements and occasionally horizontally transferred regions.

We developed a new tool, Nucleotide PanGenome explorer (NPG-explorer), designed for aligning and analysis of a number of input closely related genomes. NPG-explorer constructs nucleotide pangenome - a set of aligned blocks, each block consisting of orthologous fragments. Minimum length of block (default 100 bp) and minimum identity (default 90%) are algorithm parameters. NPG-explorer iterates block detection algorithm until the following criterion is satisfied: BLAST search all-against-all block consensuses detects no hits of appropriate size and identity.

Each nucleotide from input genomes belongs to exactly one block of NPG (it is a reason for NPG terminology). Blocks are classified into four categories. Stable blocks (named s-blocks) are composed of one fragment from each genome. Hemi-stable blocks (h-blocks) are presented by one fragment from a subset of genomes. Repeat containing blocks (r-blocks) contain more than one fragment from at least one genome. Unique sequence blocks (u-blocks) contain only one fragment of length greater than a threshold. Minor blocks (m-blocks) are blocks of fragments of length less than a threshold. Blockset of global and intermediate blocks. Global blocks consist of glued consequent collinear s-blocks and fragments of sequencesthat are between them. Intermediate blocks consist of fragments of sequences that are between consequent global blocks.

In addition NPG-explorer provides: (1) Multiple alignments of input chromosomes represented by a sequence of block identifiers. These alignments allow to detect chromosomal rearrangements. (2) File with consensus sequences of all blocks and file with description of all mutations with respect to consensuses. Thus, all input genome sequences can be completely reconstructed from these two files. (3) Phylogenetic trees of core blocks and of whole genomes. Core blocks are those that contain exactly one fragment of each genome. These trees are computed on the base of diagnostic positions in block alignments. (4) All gene annotations, mapped on blocks. This data are useful for detection and correction mis-annotations, gene corruption etc.

Using NPG-explorer we constructed nucleotide pangenomes of five sets of genomes: 17 complete genomes of Brucella genus (56 Mb totally), 39 partially completed genomes of Brucella genus (129 Mb totally),12 genomes of Yersinia pestis (55 Mb), 8 genomes of Rickettsia rickettsii (10 Mb), 5 genomes of Burkholderia cenocepacia (38.5 Mb).

In Brucella pangenome there are 653 stable blocks covering 91.5% of sum of lengths of all genomes. Identity within joined alignment of s-blocks is 99.2% showing high sequence similarity of all genomes. Program detected 33 global blocks. Phylogenetic tree of genomes computed by NPG-explorer by using diagnostic positions is in agreement with published data for 10 Brucella genomes [1]. The program found large translocation from first to second chromosome in Brucella suis ATCC 23445 and large inversion in chromosome 2 of Brucella abortus, also described earlier [2].

NPG-visualization tool presents interactively a list of blocks, the alignment with mapped genes, alignments of block identifiers. NPG-explorer is written in C++ and is licensed under the GNU GPL. Simple script language for program modules invocation is introduced.

The work was supported by RFBR grants 14-04-01693, 13-07-00969.

[1] Wattam et al. (2009) Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle, *J.Bacteriology*, **191**:3569-79

[2] Tsoktouridis et al. (2003) Molecular characterization of Brucella abortus chromosome II recombination, *J.Bacteriology*, **185**:6130-6