New version of sleeping chironomid's genome assembly using Illumina and PacBio data

Olga Kozlova[1], S. Shigenobu[2]

[1]*Kazan Federal University, 18 Kremlyovskaya street, Kazan,* olga-sphinx@yandex.ru

[2]*National Institute for Basic Biology, Nishigonaka 38, Myodaiji, Okazaki 444-8585 Aichi, Japan,* shige@nibb.ac.jp

Oleg Gusev

Kazan Federal University, 18 Kremlyovskaya street, Kazan, gaijin.ru@gmail.com

Polypedilum vanderplanki (desiccation-tolerant midge or sleeping chironomid) presents itself as an interesting model for study different aspects of response to stress in insects. Since all biological insights obtained from the results of using bioinformatics approaches may seriously depend on the quality of draft genome assembly, there is an urgent need for relevant set of scaffolds, presenting the genome of *P. vanderplanki* with minimum rate of contamination in it. Here we present completely new version of desiccation-tolerant midge's genome assembly based on DNA sequencing of Pv11 cell line derived from embryonic tissues of this insect (Nakahara, 2010). In order to receive long, continuous scaffold sequences we combined Illumina paired-end and mate-paired libraries with the results of SMRT sequencing according to Pacific Biosciences technology. The entire set of benchmark data for *de-novo* assembly consisted of four paired-end libraries (three of them received from HiSeq2000 sequencer that gave us 137 million read pairs of 100+100 bp long and one from MiSeq sequencer with 17 million read pairs of 260+260 bp long), six mate-paired libraries (100+100 bp long, HiSeq2000) with different insert sizes (5-6 kb, 6-7 kb and 8-10 kb, two libraries in each group; approximately 85 million reads pairs of raw data in total) and about a million long reads from PacBio sequencer. The genome size of *P. vanderplanki* is about 100-120 Mbp, as it is known from previous studies and was estimated from the given data independently. On the basis of such estimation we could expect about ~500-fold coverage of the midge's genome; that is roughly the same as the coverage rate of the current version of *P.vanderplanki* assembly (Gusev, 2014).

Since there are three different approaches of using PacBio data for *de-novo* genome assembly, we used them all and thus received three sets of sequences with their own particular qualities (each had its own advantages and disadvantages). First approach, utilizing the strategy of PacBio-only assembly, gave us 401 contigs with amazingly great N50 of more than 2.2 mb and the total genome size of 133 mb (HGAP v4 assembler). The second one (hybrid *de-novo* assembly with combination of PacBio and short reads) resulted to 989 contigs with N50 of 182 kb and the total genome size of 101 mb (DBG2OLC pipeline). The last one, according to which

PacBio data is used only to scaffold contigs and close the possible gaps between them, gave us more than 83 thousand of scaffolds with N50 of 185 kb and the total genome size of 120 mb (in the given case we used Platanus assembler to receive Illumina-based contigs and then PBJelly pipeline to integrate long reads into the assembly). In order to estimate the completeness of the assemblies we used BUSCO pipeline with the set of single-copy orthologs from *Diptera* family. Interestingly, the best result in this case belonged to the last variant of assembly (Platanus+PBJelly), having 96.9% of complete protein-coding sequences from the dataset and only 0.8% of duplicated sequences. Contrariwise, PacBio-only version had only 82.7% of completeness and 4% of duplication that may correspond to slight increase of total assembly size in this case. Nevertheless, additional scaffolding and gap-closing with use of mate-paired data and SSPACE/GAPFILLER tools were performed for each assembly variant. This procedure slightly increased the values of integrity metrics while the completeness rates remained untouched.

Since each of three preliminary assemblies had its own pros and cons, we decided to combine them together instead of choosing the best one. For this purpose we used Metassembler tool, which performs iterative pairwise combination of two assemblies, and in our case the version with the leading rate of completeness, but the smallest N50 rate, was chosen as a basis. As a result of this procedure we got 400 scaffolds with N50 about ~ 1 mb and the genome size of 120 mb. Compared to the current version of *P.vanderplanki* genome assembly (80 thousand of scaffolds, including 9 thousand larger than 500 bp, and N50 of 264320) (Gusev, 2014), the new assembly looks much more relevant, so it’s reasonable to expect its fitness for subsequent bioinformatics studies, including capturing conformation of genome using HiC technique and many others.

1. Y. Nakahara et al. (2010) Cells from an anhydrobiotic chironomid survive almost complete desiccation, *Cryobiology*, **60**:138-146.
2. O. Gusev et al. (2014) Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge, *Nature Communications*, **5**:4784.