Building the set of orthologous genes for 66 Gammaridae transcriptomes

Sergey Naumenko

Laboratory of Evolutionary Genomics, Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye Gory 1-73, Moscow 119992; *Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, Bolshoi Karetny pereulok 19, Moscow 127994, Russia*

sergey.naumenko@yahoo.com

Ksenia Lezhnina

Laboratory of Evolutionary Genomics, Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye Gory 1-73, Moscow 119992;

oxia.com@gmail.com

65 transcriptomes of Lake Baikal *Gammaridae* and 1 transcriptome of *Pandorites podoceroides* (Kaspian *Gammaridae* representative) were sequenced in the laboratory of evolutionary genomics (Logacheva, Klepikova, Penin). These samples were collected during two field expeditions to lake Baikal (Yampolsky, Etingova, Logacheva, Naumenko).

We created the high quality dataset of orthologous genes for these species, including the information of specie's polymorphism inferred from sample's heterozygocity.

The problem of orthobase creation in that case turned to be not trivial: the usage of well-known and novel methods such as agalma pipeline [1], orthomcl [2], phylogenomic dataset construction [3] resulted in a few clusters of orthologous genes with number of species approaching 60. Due to differential expression of genes it is diffucult to obtain the complete sequence of a transcriptome, so some genes are absent in the sample, and this set of absent genes is unique to the sample. Additionally some genes are sequenced partially and the completeness of transcriptomes is varying by sample. Thus clusterization algorithm experiences difficulties in building orthogroups.

To overcome this problem we developed a mixed approach. First, we clustered genes of 3 species of high quality transcriptomes representing 3 major clades of *Gammaridae* using the latest version of mcl [4] constructing the reference orthologous genes dataset. Then for each reference cluster we were mapping reads from all other species, calling SNPs and changing reference sequences towards target sequences. We repeated iterations of mapping – SNP calling – consensus building for each sequence until no new SNPs were detected during

the iteration. This approach arises out of the model of molecular evolution where the sequence walks through the space of possible variants [5]. Starting from 3 diverse points in the space of Gammaridae's genes we arrive to the target point under the pression of reads sequenced from the target. Because we are using 3 starting points representing different clades we are able to find the path for any specie from our set.

This approach, while computationally intensive, allowed us to obtain ~ 1000 clusters of orthologous genes with average species number of ~ 45 and >90% of gapless nucleotide sites. We hope that such a dataset will form the basis of research of molecular evolution, codon evolution, parallel evolution, epistasis and speciation of lake Baikal Gammaridae.

Two examples of such an analyses are presented on Figures 1,2.

On Figure 1 the phylogeny of 66 *Gammaridae* is depicted. The philogeny is in an agreement with previously published one based on mtDNA sequnces [6]. *G.lacustris* is Kaspian specie and it is an outgroup to all Baikal species. Two big clades represent probably two invasion events. *G.lacustris*, which is common to Siberia, here is an outgroup to the *Acanthogammaridae* group. This fact underlines the independence of *Micruropidae* invasion to Baikal. While the major clades here resolved in an agreement with morphology, the short branches with low bootstrap values in Eulimnogammaridae need further work. We plan to resolve this branches using only orthologs from closely related species for specific branch.

Figure 2 shows the phylogeny of 12 species of *Acanthogammaridae.* The branch lengths represent dN (number of nonsynonymous substitutions per nonsynonymos site) (a), dS (number of synonymous substitutions per synonymous site) (b) and dN/dS ratio (c). These figures reveal the uniqueness of the *Gammaridae* dataset for studies in molecular evolution: there is no other *Eukaryotic* dataset with such a short branch lengths.

This the joint research project with Georgii Bazykin, Maria Logacheva, Alexey Penin (FBB MSU, IITP RAS), Anna Etingova (Baikal Museum, the Irkutsk Scientific Center of RAS), Anna Klepikova, Michael Shelkunov (FBB MSU), Lev Yampolsky (Department of Biological Sciences, East Tennessee University), Alexey S. Kondrashov (FBB MSU, Deparment of Ecology and Evolutionary Biology, Michigan State University).

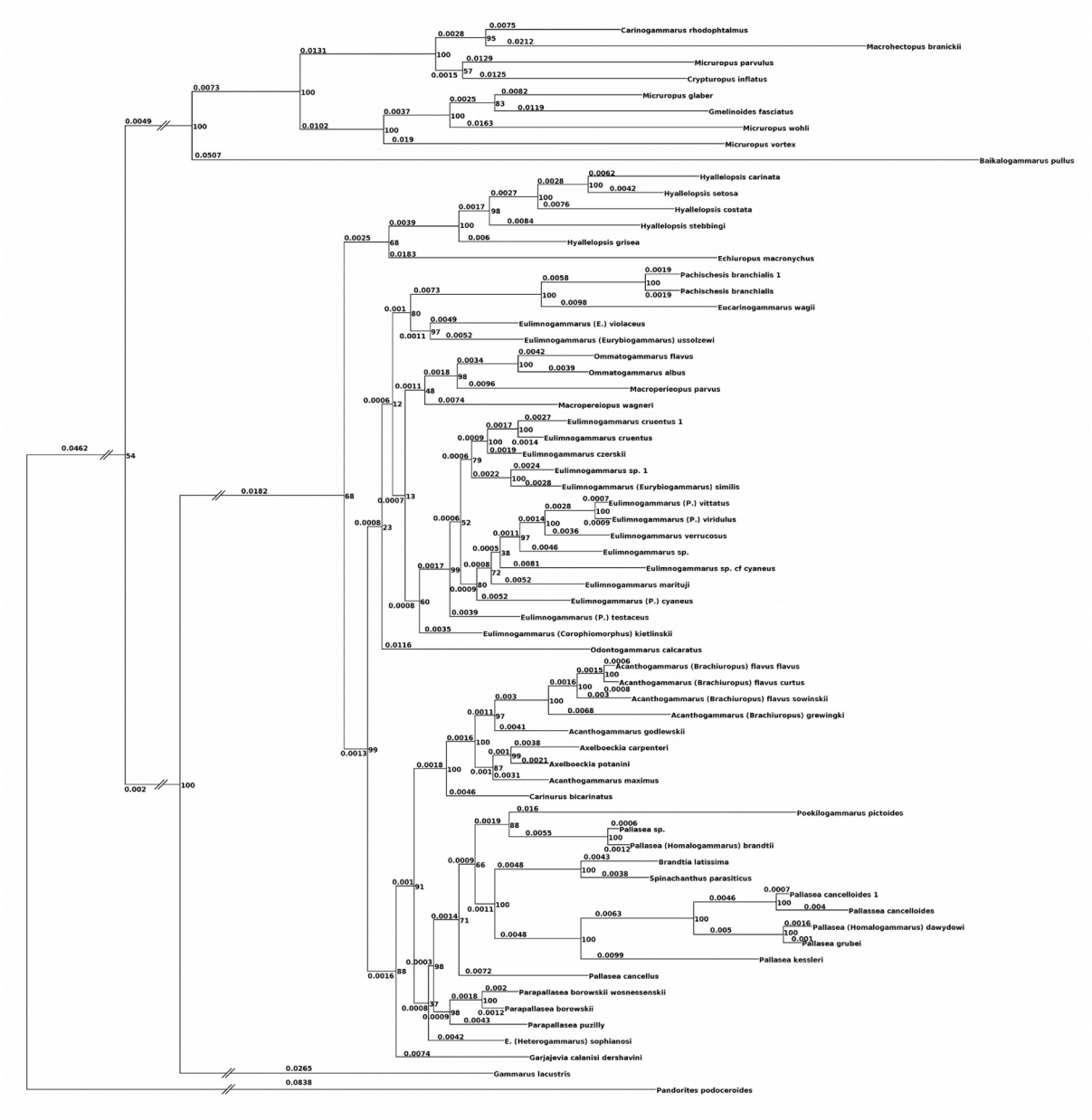
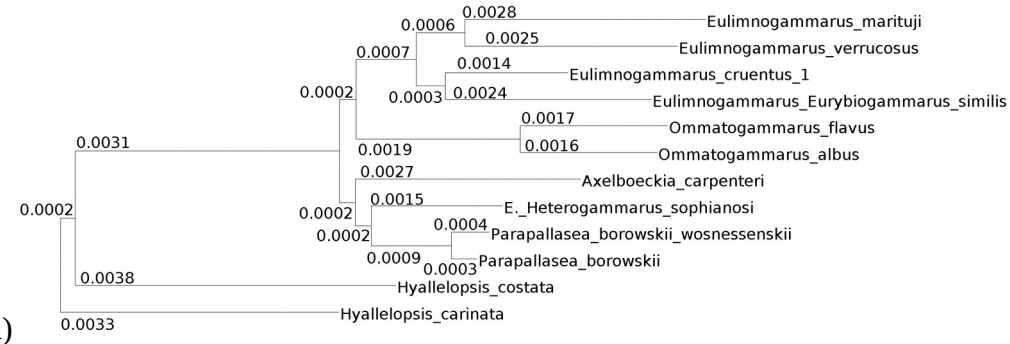
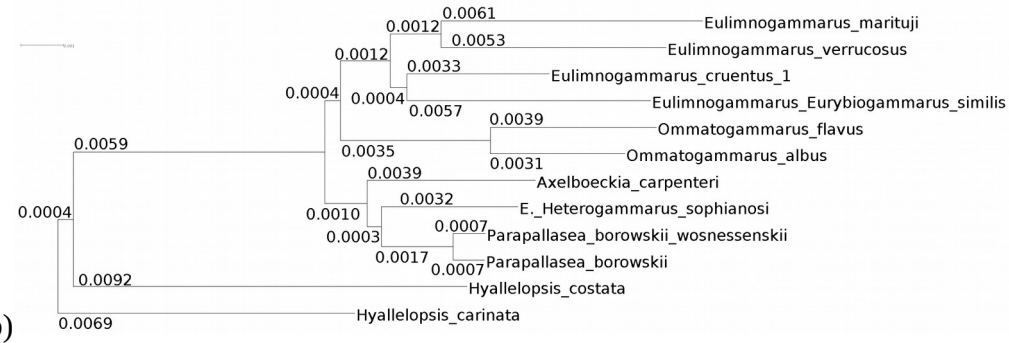


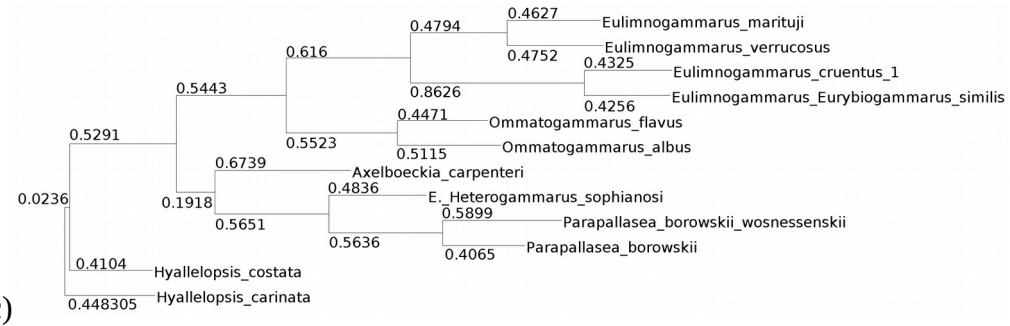
Figure 1. Phylogeny for 66 species of *Gammaridae*. Branch lengths are measured in amino acid substitutions per site.



(a)



(b)



(c)

Figure 2. dN's (a) and dS's (b) for 12 species of *Acanthogammaridae*.

1. C.W. Dunn, M. Howison, F. Zapata (2013) *BMC Bioinformatics*, **14**(1): 330.

2. L. Li, C.J. Stoekert Jr., and D. S. Roos (2003) *Genome Research*, **13**(9): 2178-2189.

3. Y. Yang, and S.A. Smith (2014) *Molecular Biology and Evolution*, **31**(11): 3081-92.

4. A.J.Enright, S.Van Dongen, C.A. Ouzounis (2002) *Nucleic Acids Research*, **30**(7):1575-84.

5. J. Maynard Smith (1970) *Nature*, **225**:563-564.

6. K.S. Macdonald III, L. Yampolsky, J.E. Duffy (2005), *Molecular Phylogenetics and Evolution*, **35**: 323-343.