Moss phylogeny reconstructed from 24 full mitogenome sequences using new "pangenome" based approach

Denis V. Goryunov[1], Boris E. Nagaev[1,2], Michail S. Ignatov[4], Andrei V. Alexeevski[1,2,3] & Alexey V. Troitsky[1]

[1]Belozersky Institute of Physico-Chemical Biology, and [2]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University Leninskie gory 1, Moscow 119902, Russia,

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

denis.goryunov@mail.ru

[4]Tsitsin Main Botanical Garden RAS, Moscow, Russia

misha_ignatov@list.ru

Mosses (Bryophyta), is the ancient group of terrestrial plants branched off the stem of Embryophyta phylogenetic tree just after hepatics (Marchantiophyta). Bryophytes have many unique traits distinguished them from vascular plants. The peculiarity of its organization and genetic base of evolutionary novelties resulted in transformation from non-vascular plants to vascular Embryophytes are not fully recognized. Comparative genomics is a powerful tool to solve these thorny problems. That is why bryophyte genome study is quite crucial aspect of modern genomics to clarify dark spots of early stages of higher plants evolution.

To date full genome of only one moss species *Physcomitrella patens* is sequenced and assembled into 378 scaffolds [1]. Additionally, mitochondrion genomes of 23 mosses were sequenced and deposited into NCBI Refseq database: *Anomodon attenuates* (ANOAT), *Anomodon rugelii* (ANORU), *Atrichum angustatum* (ATRAN), *Bartramia pomiformis* (BARPO), *Bucklandiella orthotrichacea* (BUCOR), *Buxbaumia aphylla* (BUXAP), *Climacium americanum* (CLIAM), *Codriophorus aciculare* (CODAC), *Codriophorus laevigatus* (CODLA), *Codriophorus varius* (CODVA), *Funaria hygrometrica* (FUNHY), *Hypnum imponens* (HYPIM), *Orthotrichum rogeri* (ORTRO), *Orthotrichum speciosum* (ORTSP), *Orthotrichum stellatum* (ORTST), *Physcomitrella patens* (PHYPA), *Ptychomnion cygnisetum* (PTYCY), *Racomitrium elongatum* (RACEL), *Racomitrium emersum* (RACEM), *Racomitrium ericoides* (RACER), *Sphagnum palustre* (SPHPA), *Tetraphis pellucida*

(TETPE) and *Ulota hutchinsiae* (ULOHU).

Recently the mitochondrial genome of pleurocarpous moss *Brachythecium rivulare* (BRARI) was sequenced and analyzed (Goryunov et al., in press). It consists of 104,460 base pairs, that is in a range of such values for other studied mosses except *Sphagnum palustre* KC784957 with its largest 141,276 bp chondriome.

The high-throughput sequencing (HTS) technologies create new opportunities and drastically change methodology in life sciences. A huge amount of data constantly generated by HTS, however its analysis and biological interpretation are real bottleneck for the further development of genomics. Comparison of even relatively short genomic sequences (like mosses' mitochondrial genomes) is not trivial and typically time consuming procedure. To overcome the problem and compensate this draw back software NPG-explorer (NPGe) was created. The software designed for construction and analysis of so called 'nucleotide pangenome' (separate abstract on NPGe is submitted to MCCMB'15). In our study this approach was used for reconstructing moss phylogeny from 24 full mitogenome sequences.

NPG-explorer's representation of rearrangements is a nucleotide pangenome, a set of blocks, each block is an alignment of orthologous fragments of genomes. Alignment is guaranteed to satisfy several conditions, in particular percentage of identical positions. NPG blocks are classified into five types. Stable blocks include exactly one fragment from each genome, thus, they are free of internal rearrangements. Joined alignment of stable blocks seems to be appropriate for reconstruction of genomes phylogeny. NPGe constructs phylogenetic tree from this joined alignment using NJ method. At post-processing step NPGe creates several files with analytical information about pangenome.

Constructing NPG of Bryophyta mitochondria we used 0.7 as a threshold of the fraction of conserved positions within a block and 100 bp as a threshold of block length. 142 stable blocks were detected, they cover 44% of NPG length. Fraction of conserved positions within joined alignment of all stable blocks is 79%. Phylogenetic tree constructed by NPGe is in a good agreement with moss phylogeny reconstructed from 8 nuclear and organellar genes [2] as well as from forty mt protein-coding genes [3] (Fig. 1).

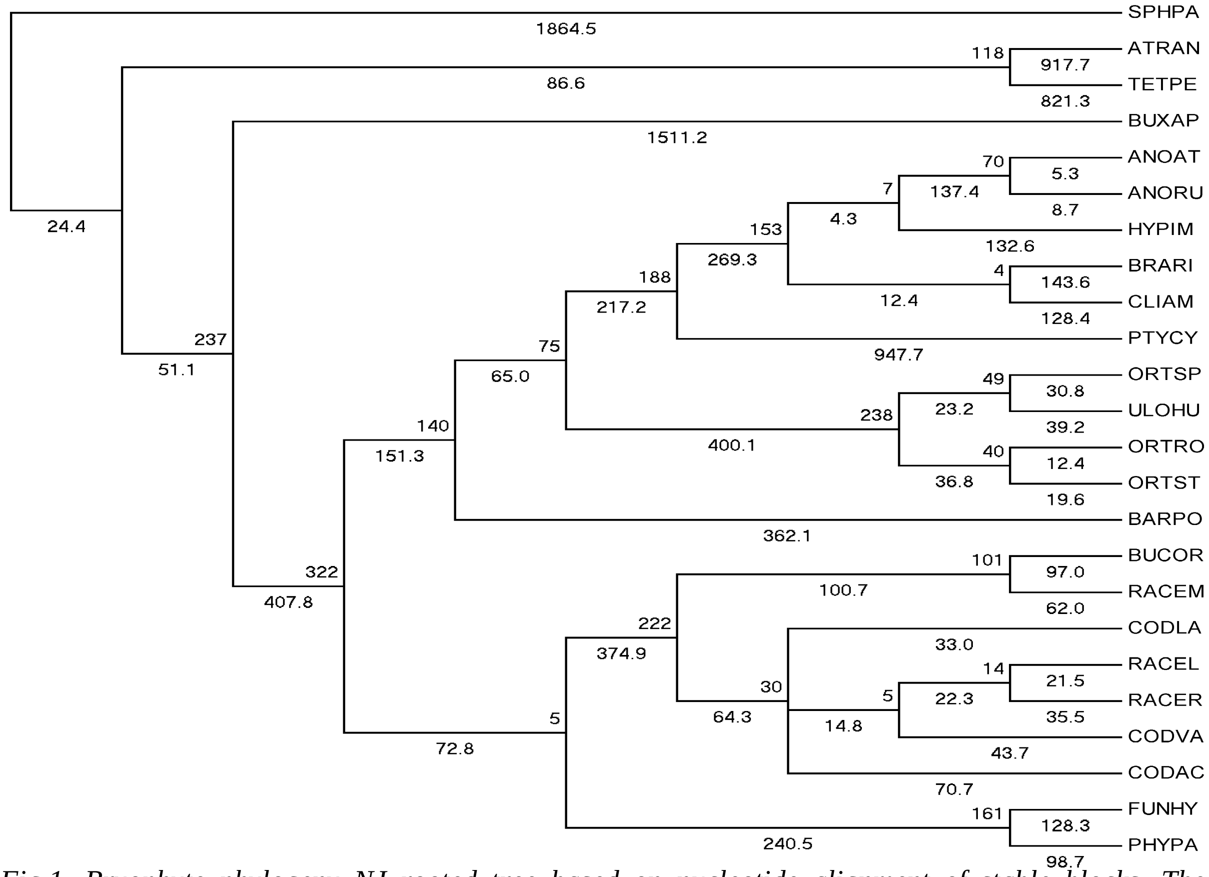
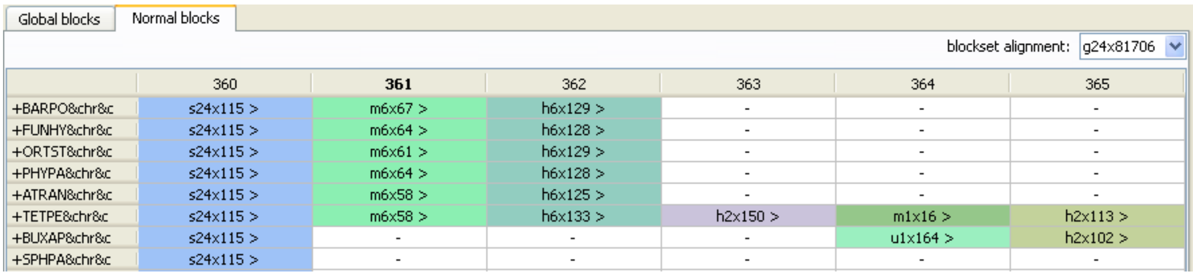


Fig.1. Bryophyte phylogeny NJ rooted tree based on nucleotide alignment of stable blocks. The number of non-identical positions in the alignment is 11016. The p-distance was applied for pair-wise distance computation. See text for species names abbreviations. The integers and floats on the tree nodes indicate numbers of diagnostic positions and branches lengths, respectively.

Stable blocks (Fig. 2) are separated either with minor blocks (short blocks with low fraction of conserved positions) or with hemi-stable blocks arising due to long deletions in certain genomes or with blocks with genomic duplications or with unique sequences, i.e. sequences in one genome having no BLAST hits with appropriate length and similarity in all genomes. Here we show examples of comparison analysis using NPGe. (i) We found five blocks with repeats, they appeared due to duplications in genomes *S. palustre*, *R. emersum*, *B. orthotrichacea*. (ii) Enlarging of *S. palustre* genome is mainly due to unique sequences, they comprise 43489 bp totally. (iii) Hemi-stable block demonstrate long deletions in genomes, that lack fragments in it. *B. aphylla* revealed multiple long deletions but they are compensated with a lot of unique sequences, 20661 bp totally. (iv) All stable blocks are collinear. Nevertheless, we found several rearrangements of hemi-stable blocks.



blockset alignment: g24x81706

	360	**361**	362	363	364	365
+BARPO&chr&c	s24x115 >	m6x67 >	h6x129 >	-	-	-
+FUNHY&chr&c	s24x115 >	m6x64 >	h6x128 >	-	-	-
+ORTST&chr&c	s24x115 >	m6x61 >	h6x129 >	-	-	-
+PHYPA&chr&c	s24x115 >	m6x64 >	h6x128 >	-	-	-
+ATRAN&chr&c	s24x115 >	m6x58 >	h6x125 >	-	-	-
+TETPE&chr&c	s24x115 >	m6x58 >	h6x133 >	h2x150 >	m1x16 >	h2x113 >
+BUXAP&chr&c	s24x115 >	-	-	-	u1x164 >	h2x102 >
+SPHPA&chr&c	s24x115 >	-	-	-	-	-

Fig.2. Part of NPG visualization screen. One row corresponds to one sequence; one cell corresponds to one fragment of a block. Fragments of the same block are colored similarly. s24x115 denote fragment of length 115 in blue block of 24 sequences - stable block. "m" is for minor blocks, "h" for hemi-stable blocks, "u" is for unique sequence. Second fragment of the block h2x150 is located out of presented part of the screen.

This study was supported by the grants № 14-04-31731, 15-04-06027, 14-04-01693 from Russian Foundation for Basic Research.

1. S.A. Rensing et al. (2008), The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants, *Science,* **319**(5859):64-69.

2.C.J. Cox et al. (2004), Phylogenetic relationships among the mosses based on heterogeneous Bayesian analysis of multiple genes from multiple genomic compartments, *Syst. Bot.,* **29**:234–250.

3. L. Yang et al. (2014), 350 My of Mitochondrial Genome Stasis in Mosses, an Early Land Plant Lineage, *Mol. Biol. Evol.*, doi:10.1093/molbev/msu199