CHROMOSOME-LEVEL GENOME ASSEMBLIES EXPANDED CAPABILITIES OF CONSERVATION BIOLOGY

Authors: Azamat Totikov [1], Andrey Tomarovsky [1], Lorena Derezanin [2], Olga Dudchenko [3], Erez Lieberman-Aiden [3], Klaus Koepfli [4], Sergei Kliver [5].
[1] Saint Petersburg State University, 7/9 Universitetskaya Emb., St Petersburg, 199034, Russia, a.totickov1@gmail.com, andrey.tomarovsky@gmail.com
[2] Leibniz Institute for Zoo and Wildlife Research (IZW), 17 Alfred Kowalke Straße, 10315 Berlin, Germany, derezanin@izw-berlin.de
[3] The Center for Genome Architecture, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA, Olga.Dudchenko@bcm.edu, erez@erez.com
[4] Smithsonian-Mason School of Conservation, 1500 Remount Road, Front Royal, VA 22630, USA, klauspeter.koepfli527@gmail.com
[5] Institute of Molecular and Cellular Biology SB RAS, 8/2 Acad. Lavrentiev Ave. 8/2, Novosibirsk 630090, Russia, mahajrod@gmail.com

Background:

Conservation biology aims to keep and restore biodiversity on genetic, species and ecosystem levels, prevent species extinction and protect their habitats. One of the important aspects of conservation is genetic diversity assessed within endangered populations or species. Reduction in sequencing costs facilitated estimation of the genetic diversity in multiple individuals on the whole genome level even with a very limited funding. However, the whole genome approach requires generation of reference genome assembly of suitable quality first. Current trend is to use chromosome-level assemblies offering a set of useful advantages. We compared genetic diversity in 7 threatened mammalian species for both old highly fragmented and recently generated chromosome-level assemblies. New contiguous assemblies allowed better estimation of genetic diversity, localization and visualization of low heterozygosity regions in the genomes.

Latin name	Red List category	Common name	2n	Assembly source or ID	Assembly type*	Length, Gbp	Ns, Mbp	N50, Mbp
Enhydra lutris	Endangered	Sea otter	38	DNAzoo	Chr	2.45	28.94	145.94
				GCA_002288905.2	Draft	2.46	29.68	38.75
Acinonyx jubatus	Vulnerable	Cheetah	38	DNAzoo	Chr	2.37	42.86	144.64
				GCA_001443585.1	Draft	2.37	42.06	3.12
Neofelis nebulosa	Vulnerable	Clouded leopard	38	DNAzoo	Chr	2.42	7.94	147.11
				DNAzoo draft	Draft	2.41	5.89	1.38
Pteronura brasiliensis	Endangered	Giant otter	38	DNAzoo	Chr	2.46	11.89	133.38
				DNAzoo draft	Draft	2.45	1.40	0.17
Ailurus fulgens	Endangered	Red panda	36	DNAzoo	Chr	2.34	34.41	143.80
				GCA_002007465.1	Draft	2.34	34.04	2.98
Aonyx cinereus	Vulnerable	Asian small-clawed otter	38	DNAzoo	Chr	2.44	15.50	130.94
				DNAzoo draft	Draft	2.42	1.35	0.10
Bison bison	Near threatened	American bison	60	DNAzoo	Chr	2.83	199.31	101.69
				GCF_000754665.1	Draft	2.83	195.77	7.19

Table 1. Mammalian species and corresponding genome assemblies used in this study.
* Assembly types: **Draft** - initial fragmented assembly, **Chr** - chromosome-level assembly based on Draft.

Results:

The simplest way to assess heterozygosity is to do it genome-wide but such an approach provides only a single value limiting data on the genetic diversity. More informative way includes calculation of mean or median heterozygosity in staking or overlapping windows of fixed size. The size of the window is a matter of choice depending on the integrity of the assembly and planned analysis and visualization but commonly used sizes fall in the 50 - 5000 kbp range. A significant part of the genome must be presented in windows to make heterozygosity estimates reliable. Among the studied species most fragmented assemblies were drafts of P. brasiliensis and A. cinereus with N50 of 0.17 and 0.1 Mbp, respectively (Table 1) which significantly affected assessment of heterozygosity distribution (Figure 1). From the lower boundary window size is limited by a reasonable number of heterozygous SNPs present in the most of windows and the number of windows that could be drawn without the mess on the plots, figures or heatmaps. In the case of mammalian genomes with typical size of 2.5 - 3.0 Gbp number of 100 kbp windows exceeds 20000 thousands for assembly of high integrity. Number of 1 Mbp windows is at least 10-fold less and in case of chromosome-level assemblies could be easily visualized on chromosomal scaffolds. Such plots are impossible for draft assemblies due to the high number of scaffolds.

Species we analyzed include both well known for extremely low heterozygosity sea otter (Figure 2a) and cheetah (Figure 2b) and species with higher genetic diversity but considered to be threatened too: american bison, asian small-clawed otter and red panda (Figure 2 g,f,e). Despite significant differences in mean heterozygosity (Figure 1) all genomes showed regions with very low diversity (blue and dark blue regions on Figure 2). The most striking difference in heterozygosity between different regions of the genome was found in giant otter. Having ~2.5 times higher mean heterozygosity it demonstrated huge highly homozygous stretches (dark blue on Figure 2d) on more than half chromosomes.

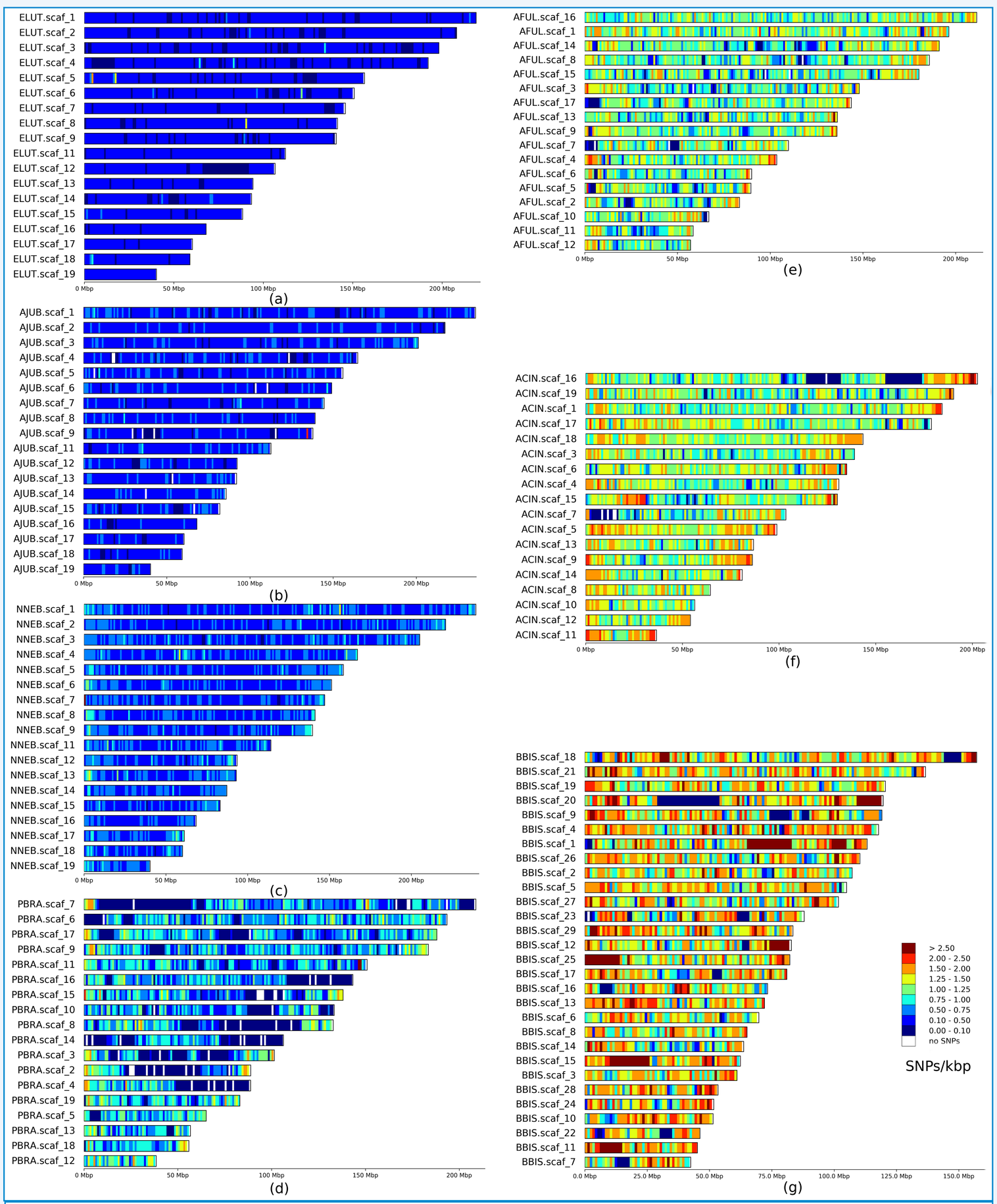

Figure 2. Heatmaps of heterozygous SNP densities for analyzed species based on chromosome level assemblies (sex chromosomes were excluded). Heterozygous SNPs were counted in 1 Mbp windows and scaled to SNP/kbp. A - sea otter , B - cheetah, C - clouded leopard, D - giant otter, E - red panda, F - asian small-clawed otter, G - american bison.

Conclusion:

Chromosome level genome assemblies provide better estimates of genetic diversity and new possibilities for visualization of results. It could be generated in various ways with usage of different technologies but because of limited budget short read drafts followed by HiC-scaffolding will be of first choice for conservation studies in the nearest future.

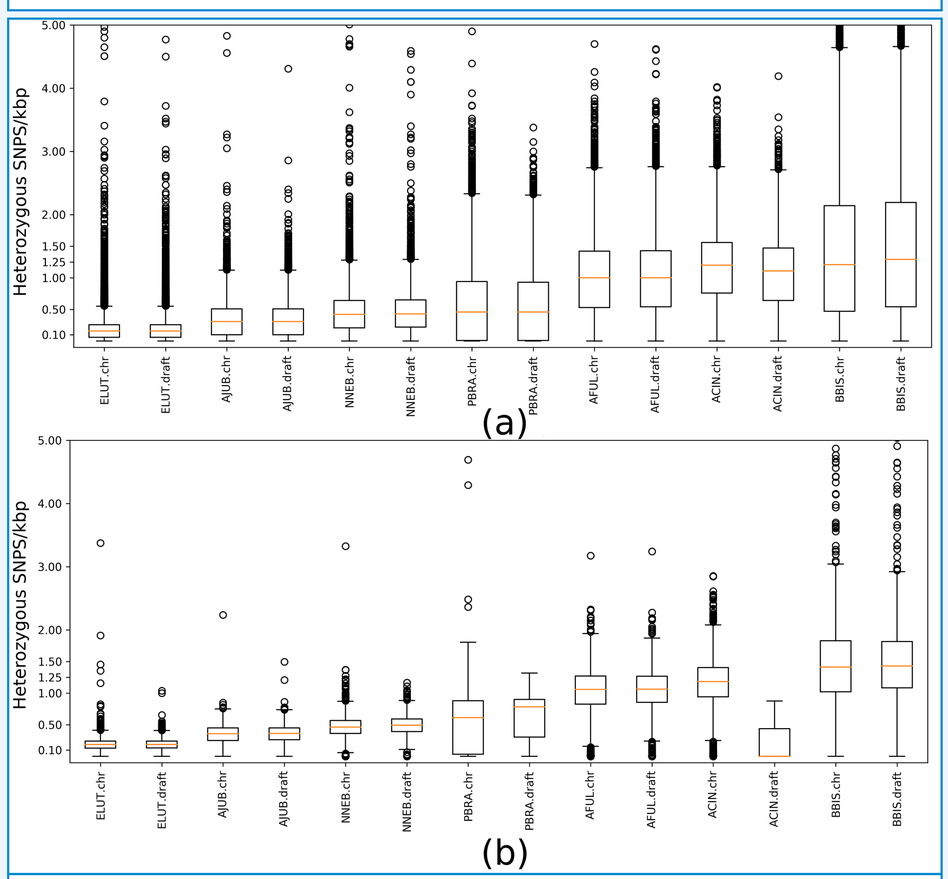

Figure 1. Comparison of distribution of mean heterozygosity in windows of 100 kb (A) and 1Mbp (B) for draft and chromosome level assemblies.

Acknowledgments: The reported study was funded by RFBR, project number 20-04-00808.