proceedings

MDPI

Proceedings

Chromosome-Level Genome Assemblies Expanded Capabilities of Conservation Biology †

Azamat Totikov [1], Andrey Tomarovsky [1], Lorena Derezanin [2], Olga Dudchenko [3], Erez Lieberman-Aiden [3], Klaus Koepfli [4] and Sergei Kliver [5,*]

[1] Saint Petersburg State University, 7/9 Universitetskaya Emb., 199034 St Petersburg, Russia; a.totickov1@gmail.com (A.T.); andrey.tomarovsky@gmail.com (A.T.)
[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 (O.D.); erez@erez.com (E.L.-A.)
[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, 630090 Novosibirsk, Russia
* Correspondence: mahajrod@gmail.com
† Presented at the 1st International Electronic Conference on Genes: Theoretical and Applied Genomics, 2–30 November 2020; Available online: https://iecge.sciforum.net/.

Received: date; Accepted: date; Published: date

Abstract: Conservation biology usually is far from cutting-edge technologies due to very limited budgets of studies, but chromosome-level assemblies being the hype theme several years ago are slowly infiltrating even in this area. We compared genetic diversity in 7 threatened species with both new chromosome-level and old fragmented assemblies. New contiguous assemblies allowed better estimation of genetic diversity, localization and especially visualization of low heterozygosity regions in the genomes.

Keywords: conservation biology; genomics; chromosome level assemblies

1. Introduction

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. Conservation biology deals with a huge number of the nonmodel species but corresponding genomic studies usually have significantly smaller budgets than in medical or agricultural areas resulting in continuous trade off between quality of generated data and its price. Recently proposed by DNA Zoo team $1 k approach for generation of chromosome-level assemblies from short-insert Illumina PE library and in situ HiC library [1] might be a temporary solution of this problem for several next years. We compared genetic diversity in 7 threatened mammalian species (cheetah, sea otter and others) for both old highly fragmented and recently generated chromosome-level assemblies (including generated by $1k approach). New contiguous assemblies allowed better estimation of genetic diversity, localization and visualization of low heterozygosity regions in the genomes.

2. Methods

2.1. Quality Control and Filtration of the Data

Draft and chromosome level assemblies of 7 species were downloaded from the NCBI Genome and DNA Zoo databases (Table 1) [1–6]. Raw short read libraries with the following ids were obtained from NCBI SRA: SRR2712398, SRR2712418, SRR2737521, SRR2737520, SRR2737519, SRR12437584, SRR5768052, SRR11431910, SRR11286173, SRR8588180, SRR12437584 [1,2,4–8]. Raw data quality control was performed using FastQC [9] and KrATER [10]. Adapter trimming and filtration by quality was performed in two stages with initial kmer-based trimming of large adapter fragments Cookiecutter [11] followed by additional small fragment trimming and quality filtration by Trimmomatic v0.36 [12].

2.2. Alignment and Variant Calling

Alignment of the filtered reads to the corresponding reference genome assemblies was performed using BWA [13]. Read duplicates were marked with Samtools package v1.9 [14]. Variant calling was performed using Bcftools v1.10[15] with following parameters: "-d 250 -q 30 -Q 30 --adjust-MQ 50 -a AD,INFO/AD,ADF,INFO/ADF,ADR,INFO/ADR,DP,SP,SCR,INFO/SCR" for bcftools mpileup and "-m -v -f GQ,GP" for bcftools call. Low quality variants ('QUAL < 20.0 || FORMAT/SP > 60.0 || FORMAT/DP < 5.0 || FORMAT/GQ < 20.0') were removed using bcftools filter.

2.3. Heterozygosity Visualization

Filtered genetic variants were split into SNP and indel categories. All following analysis was based only on SNPs. Indels were not used due to usually low quality calls from short reads. Counts of heterozygous SNPs were calculated in non-overlapping windows of 100 kbp and 1 Mbp and scaled to SNP/kbp. Heatmaps and boxplots were drawn using custom scripts based on Matplotlib 2 library [16].

3. Results and Discussion

3.1. Evaluation of Genome Assemblies

This study involved genome analysis for 7 threatened species of three different IUCN Red List categories (NT-Near threatened, VU-Vulnerable, EN-Endangered): sea otter (*Enhydra lutris*), cheetah (*Acinonyx jubatus*), clouded leopard (*Neofelis nebulosa*), giant otter (*Pteronura brasiliensis*), red panda (*Ailurus fulgens*), asian small-clawed otter (*Aonyx cinereus*), american bison (*Bison bison*) (Table 1). Each species was represented by two genome assemblies: the initial fragmented draft and chromosome-level one generated from draft using HiC-scaffolding. Included draft assemblies were generated using different sequencing and assembly approaches resulted in different quality and integrity. N50 of drafts ranged from 0.10 Mbp for A. cinereus to 38.75 Mbp for E. lutris. Total gap lengths also varied a lot from 1.4 Mbp to 195.77 in drafts. After scaffolding total gap length have not raised significantly in absolute values (maximum 14.15 Mbp were added in case of *A. cinereus*) and for E. lutris it even was decreased, probably, due to extensive correction of misassemblies preceding scaffolding stage. Used chromosome-level assemblies include as many huge scaffolds as haploid chromosomes number (1n) of the corresponding species along with a high number of small scaffolds, but there is a very high difference (1–2 decimal orders) in length between these categories. Most likely, top scaffolds represent the whole chromosomes but assignment of them to the particular chromosomes was not done yet. As included data was generated from both male and female individuals we excluded sex chromosomes from the further analysis.

Table 1. Mammalian species and corresponding genome assemblies used in this study.

Latin Name	RedList Category [1]	Common Name	Assembly Source or ID	Assembly Type [2]	Length, Gbp	Ns, Mbp	N50, Mbp
Enhydra lutris	EN	Sea otter	DNAzoo	Chr	2.45	28.94	145.94
			GCA_002288905.2	Draft	2.46	29.68	38.75
Acinonyx jubatus	VU	Cheetah	DNAzoo	Chr	2.37	42.86	144.64
			GCA_001443585.1	Draft	2.37	42.06	3.12
Neofelis nebulosa	VU	Clouded leopard	DNAzoo	Chr	2.42	7.94	147.11
			DNAzoo draft	Draft	2.41	5.89	1.38
Pteronura brasiliensis	EN	Giant otter	DNAzoo	Chr	2.46	11.89	133.38
			DNAzoo draft	Draft	2.45	1.40	0.17
Ailurus fulgens	EN	Red panda	DNAzoo	Chr	2.34	34.41	143.80
			GCA_002007465.1	Draft	2.34	34.04	2.98
Aonyx cinereus	VU	Asian small-clawed otter	DNAzoo	Chr	2.44	15.50	130.94
			DNAzoo draft	Draft	2.42	1.35	0.10
Bison bison	NT	American bison	DNAzoo	Chr	2.83	199.31	101.69
			GCF_000754665.1	Draft	2.83	195.77	7.19

[1] **EN-**Endangered, **VU-**Vulnerable, **NT-**Near threatened; [2]Assembly types: **Draft-**initial fragmented assembly, **Chr**-chromosome-level assembly based on Draft.

3.2. Heterozygosity Estimations and Visualization

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 the number of 1 Mbp and even 100 kbp windows (Table 2) and 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 20 thousand 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 (Figure 2). 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 2g,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.

Table 2. Counts of SNPs and windows for draft and chromosome-level assemblies of the analyzed genomes.

Species	Number of Variants		Number of 100 kbp Windows		Number of 1 Mbp Windows	
	Draft	Chr	Draft	Chr	Draft	Chr
Enhydra lutris	648,954	648,017	24,146	24,165	2337	2396
Acinonyx jubatus	1,147,794	1,147,409	22,861	23,609	1757	2350
Neofelis nebulosa	1,449,490	1,449,365	22,004	23,931	1194	2387
Pteronura brasiliensis	2,362,725	2,362,126	**13,589**	22,819	**32**	2262
Ailurus fulgens	2,779,501	2,779,133	22,083	23,139	1573	2298
Aonyx cinereus	3,233,877	3,233,911	**9777**	22,183	**3**	2204
Bison bison	6,515,175	6,515,068	24,286	26,213	2181	2604

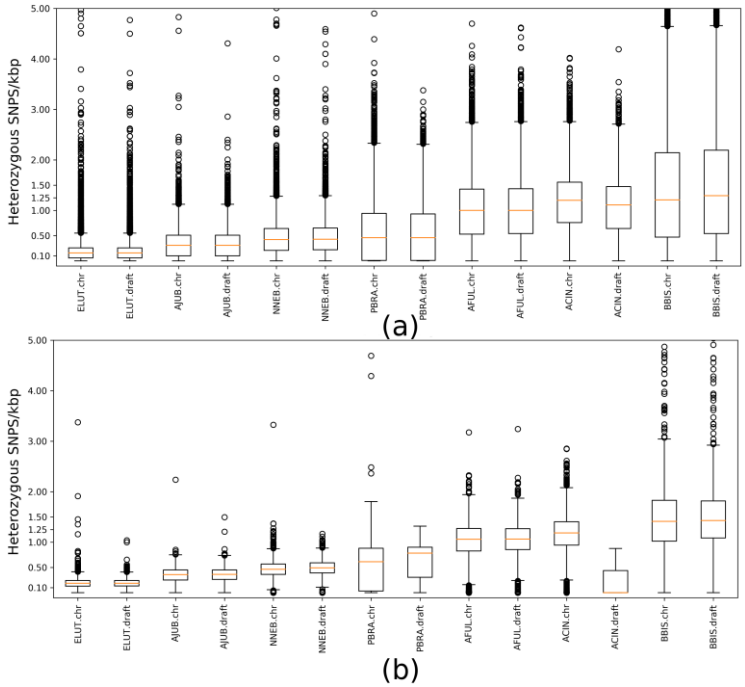

Figure 1. Comparison of distribution of mean heterozygosity in windows of 100 kb (**a**) and 1 Mbp (**b**) for draft and chromosome level assemblies.

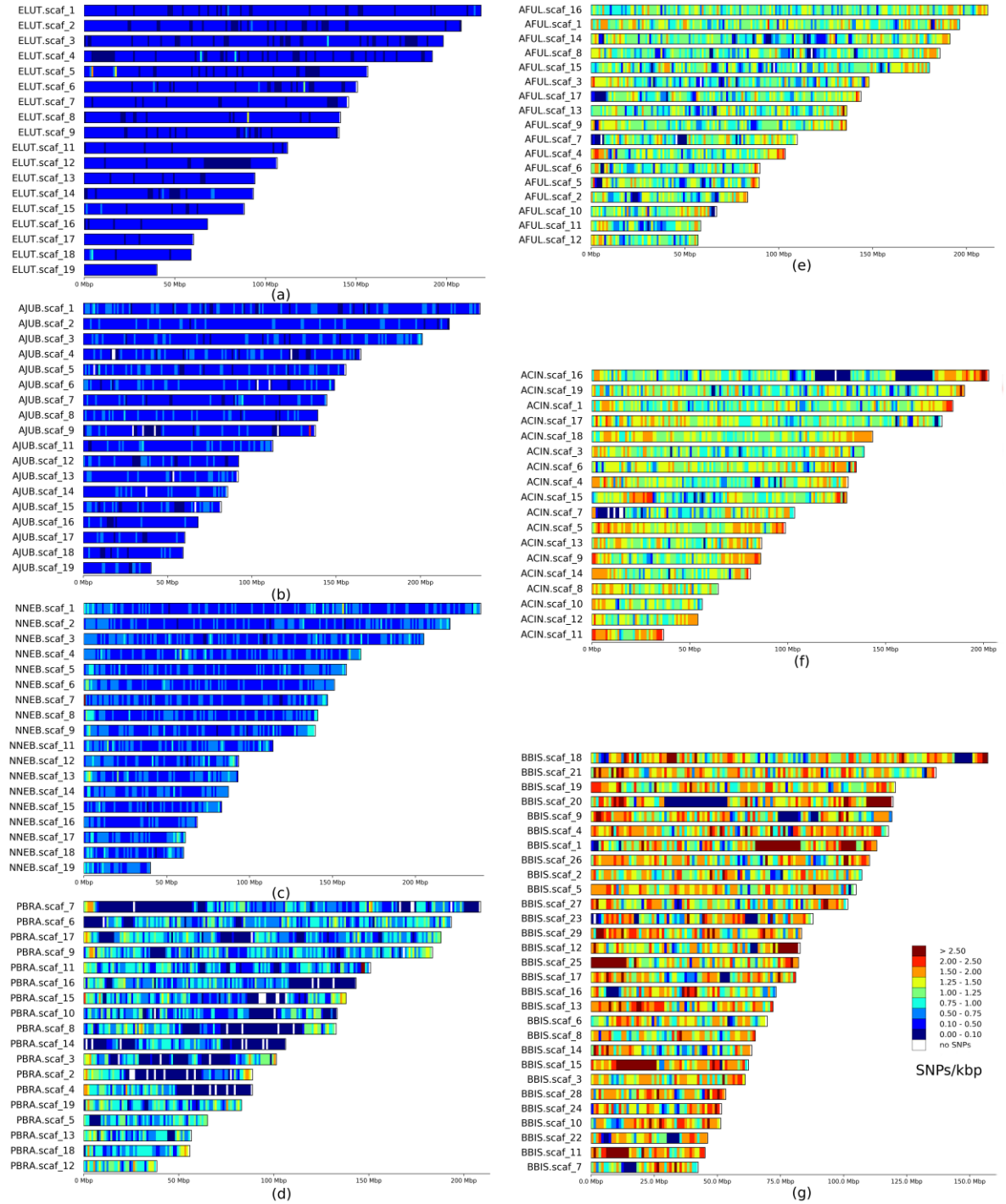

Figure 2. Heatmaps of heterozygous SNP densities for analyzed species based on chromosome level assemblies (sex chromosomes were excluded). Heterozygous SNPs were counted in 1 Mbp windows and scaled to SNP/kbp. (**a**)-sea otter, (**b**)-cheetah, (**c**)-clouded leopard, (**d**)-giant otter, (**e**)-red panda, (**f**)-asian small-clawed otter, (**g**)-american bison.

4. Conclusions

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.

Funding:

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

References

1. Dudchenko, O.; Shamim, M.S.; Batra, S.S.; Durand, N.C.; Musial, N.T.; Mostofa, R.; Pham, M.; Glenn St Hilaire, B.; Yao, W.; Stamenova, E.; et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *Genomics* **2018**, 254797, doi:10.1101/254797.
2. Dobrynin, P.; Liu, S.; Tamazian, G.; Xiong, Z.; Yurchenko, A.A.; Krasheninnikova, K.; Kliver, S.; Schmidt-Küntzel, A.; Koepfli, K.-P.; Johnson, W.; et al. Genomic legacy of the African cheetah, Acinonyx jubatus. *Genome Biol.* **2015**, *16*, 277, doi:10.1186/s13059-015-0837-4.
3. Dobson, L.K. Sequencing the Genome of the North American Bison. Thesis, 2015.
4. Hu, Y.; Wu, Q.; Ma, S.; Ma, T.; Shan, L.; Wang, X.; Nie, Y.; Ning, Z.; Yan, L.; Xiu, Y.; et al. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1081–1086, doi:10.1073/pnas.1613870114.
5. Jones, S.J.; Haulena, M.; Taylor, G.A.; Chan, S.; Bilobram, S.; Warren, R.L.; Hammond, S.A.; Mungall, K.L.; Choo, C.; Kirk, H.; et al. The Genome of the Northern Sea Otter (Enhydra lutris kenyoni). *Genes* **2017**, *8*, 379, doi:10.3390/genes8120379.
6. de Manuel, M.; Barnett, R.; Sandoval-Velasco, M.; Yamaguchi, N.; Garrett Vieira, F.; Zepeda Mendoza, M.L.; Liu, S.; Martin, M.D.; Sinding, M.-H.S.; Mak, S.S.T.; et al. The evolutionary history of extinct and living lions. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 10927–10934, doi:10.1073/pnas.1919423117.
7. Beichman, A.C.; Koepfli, K.-P.; Li, G.; Murphy, W.; Dobrynin, P.; Kliver, S.; Tinker, M.T.; Murray, M.J.; Johnson, J.; Lindblad-Toh, K.; et al. Aquatic Adaptation and Depleted Diversity: A Deep Dive into the Genomes of the Sea Otter and Giant Otter. *Mol. Biol. Evol.* **2019**, msz101, doi:10.1093/molbev/msz101.
8. Hoff, J.L.; Decker, J.E.; Schnabel, R.D.; Taylor, J.F. Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC Genom.* **2017**, *18*, 799, doi:10.1186/s12864-017-4196-2.
9. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Bioinformatics, Babraham Institute: Cambridge, UK, 2010.
10. Kliver, S. *KrATER: K-mer Analysis Tool Easy to Run*; 2017.
11. Starostina, E.; Tamazian, G.; Dobrynin, P.; O'Brien, S.; Komissarov, A. Cookiecutter: A tool for kmer-based read filtering and extraction. *Bioinformatics* **2015**, 024679, doi:10.1101/024679.
12. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120, doi:10.1093/bioinformatics/btu170.
13. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760, doi:10.1093/bioinformatics/btp324.
14. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **2009**, *25*, 2078–2079, doi:10.1093/bioinformatics/btp352.
15. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993, doi:10.1093/bioinformatics/btr509.
16. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95, doi:10.1109/MCSE.2007.55.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).