Isolated chromosome sequencing for evolutionary genomics

Alexey Makunin

Institute of Molecular and Cell Biology SB RAS, Novosibirsk, Russia, alex@mcb.nsc.ru

Rapid development of high-throughput sequencing technologies allowed to establish a multitude of genome projects more or less uniformly covering the vertebrate phylogeny under big coordination projects such as Genome10K [1]. However, generation of chromosome-level genome assemblies remains a costly and labor-intensive procedure, which is omitted for many species. As a result, studies of chromosome evolution at the genomic level are limited to the high-quality genome assemblies, most of which represent model and human-associated organisms.

We are interested in the most rapidly evolving karyotype elements: B-chromosomes (or supernumeraries) and sex chromosomes (which have a high turnover rate in fish and reptiles). B-chromosomes (Bs) are not necessary for the host, but persist in a significant fraction of populations in some species, often demonstrating non-mendelian behavior and accumulation over generations. For example, Bs are present in some 70 mammalian species (mostly rodents), none of which have publicly available genomes. Presence of protein-conding gene on mammalian Bs was first demontstrated for fox just a decade ago [2]. Still, details on genetic content of B-chromosomes content remained unknown.

Generation of DNA libraries for separate chromosomes is well established and used in comparative cytogenetics studies. At first step, chromosomes can be isolated from cell cultures with flow sorting resulting in hundreds to millions copies (depending on the number of the input cells) and from any metaphase plate spreads using chromosome microdissection (one to several chromosomes at once). Then, amplification with DOP-PCR [3] or randomly primed whole-genome amplification (WGA) is used to generate libraries suitable for subsequent fluorescence *in situ* hybridization (FISH).

In the first study, we sequenced chromosome libraries derived from flow sorting with DOP-PCR amplification on Illumina MiSeq [4]. The studied samples included chromosomes of well studied organisms (cow and dog), and B chromosomes of two cervid species (Siberian roe deer and gray brocket deer). In the study of the non-repetitive sequences we aligned the reads to the reference genomes (dog for dog chromosome and cow for cow and cervid chromosomes). As DOP-PCR amplicons covered reference genome non-uniformely, we utilized distances between consecutive amplicons as a metric for detection of regions present on chromosomes. The predictions agreed with FISH data for dog and cow chromosomes, and allowed to identify 2 regions on Siberian roe deer Bs (2 Mbp), and 26 regions on grey brocket deer Bs (9 Mbp). In general, regions present on B-chromosomes of mammals generally are up to 2-4 Mbp in size. The procedures leading to the predictions of the regions present on chromosomes were unified in the pipeline DOPseq_analyzer (https://github.com/ilyakichigin/DOPseq_analyzer).

Our further efforts on Illumina MiSeq sequencing of both flow-sorted and microdissected chromosome amplified with DOP-PCR and WGA demonstrated the capabilities of the developed method. Alignment to the reference genome assembled to the chromosomal level allows to identify regions present on chromosomes starting from 10-20 kbp in size, and the increase of distance between the studied and reference genomes results in the decreasing resolution of the identified regions. Our method is more likely to produce false-negative results (i.e., deletions within regions entirely present on chromosomes), although putative false-positives occurred: gene families with significant sequence homology were included in chromosomes. We also examined sequence variation patterns and repeat composition for the chromosomes studied. Both seemed to require additional validation due to amplification biases.

Aside from B-chromosome studies, this approach can be used for characterization of sex chromosomes and assignment of scaffolds to previously unassembled chromosomes [5], of within-species (clinical) and between-species (evolutionary) chromosome rearrangements

with quite high resolution. Data on repetitive DNA content and sequence variation can be used in subsequent of more specific chromosome features.

This study was supported by RSF grant 16-14-10009.

[1] D. Haussler *et al.*, “Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species,” *J. Hered.*, vol. 100, no. 6, pp. 659–674, 2009.
[2] A. S. Graphodatsky *et al.*, “The proto-oncogene C-KIT maps to canid B-chromosomes,” *Chromosome Res.*, vol. 13, no. 2, pp. 113–122, 2005.
[3] H. Telenius, N. P. Carter, C. E. Bebb, M. Nordenskjo¨ld, B. A. J. Ponder, and A. Tunnacliffe, “Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer,” *Genomics*, vol. 13, no. 3, pp. 718–725, Jul. 1992.
[4] A. I. Makunin *et al.*, “Contrasting origin of B chromosomes in two cervids (Siberian roe deer and grey brocket deer) unravelled by chromosome-specific DNA sequencing,” *BMC Genomics*, vol. 17, p. 618, 2016.
[5] I. G. Kichigin *et al.*, “Evolutionary dynamics of Anolis sex chromosomes revealed by sequencing of flow sorting-derived microchromosome-specific DNA,” *Mol. Genet. Genomics*, vol. 291, no. 5, pp. 1955–1966, 2016.