De novo assembly of a northern Chinese genome

Zhenglin Du[1], Na Yuan[1], Jingyao Zeng[1], Shuo Shi[1], Hongzhu Qu[2], Lili Dong[1], Jingfa Xiao[1,2]*

[1]BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

[2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

*Corresponding author: xiaojingfa@big.ac.cn

Advances in genome sequencing and assembly technology provide an opportunity to investigate the human genetic diversity across different population groups. In recent years, two *de novo* assembly genomes of southern Chinese individuals, HX1 and YH2.0, were consecutively released. Given the huge genetic diversity between the population of North and South China, here we report a *de novo* assembly of a Northern Han individual (NH1.0), using the single-molecule real-time sequencing platform (PacBio), 10x Genomics linked reads, Illumina pair-end reads and Bionano Saphyr optical mapping system. The genome of NH1.0 was assembled with a contig N50 size of 3.6Mbp and a scaffold N50 size of 46.63 Mbp (scaffold N50 size is 20.52 Mbp for YH2.0 and 21.98 Mbp for HX1), which covered 15 chromosome arms with coverage more than 85% of euchromatic regions (Fig 1). The total assembly length of the NH1.0 genome was 2.89 Gbp, including 5,581 scaffolds and 8,491 gaps, which completely closed 99 gaps in the GRCh38 genome with length adds up to 609,822 bp. It is indicated that the hybrid approach of combining PacBio and 10X Genomics technologies can highly improve the integrity of genome assembly. 2,218,371 SNPs and 18,613 structural variations (insertion and deletion) were detected in the NH1.0 genome compared with the GRCh38 genome, and 55.9% SNPs and 10.1% SVs shared with other two southern Chinese genomes (YH2.0 and HX1). This work presents the most contiguous human genome assembly for the Chinese population, with extensive investigation of population-specific genetic variation for precision medicine.

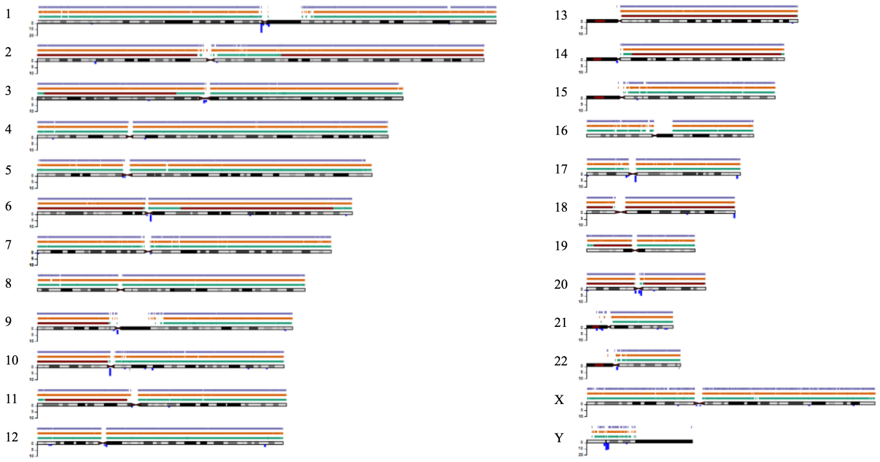

Figure 1. The genome coverage comparison of NH1.0 with GRCh38, YH2.0 and HX1. Cytobands indicate the GRCh38 reference genome. Genome coverage of assembled scaffolds are shown by three lines above the cytobands (blue for HX1, red for YH2.0 and green for NH1.0).