BMB
Reports

BMB Rep. 2016; 49(12): 653-654
www.bmbreports.org

Perspective

Birth of an 'Asian cool' reference genome: AK1

*Changhoon Kim**
Bioinformatics Institute, Macrogen Inc., Seoul 08511, Korea

The human reference genome, maintained by the Genome Reference Consortium, is conceivably the most complete genome assembly ever, since its first construction. It has continually been improved by incorporating corrections made to the previous assemblies, thanks to various technological advances. Many currently-ongoing population sequencing projects have been based on this reference genome, heightening hopes of the development of useful medical applications of genomic information, thanks to the recent maturation of high-throughput sequencing technologies. However, just one reference genome does not fit all the populations across the globe, because of the large diversity in genomic structures and technical limitations inherent to short read sequencing methods. The recent success in *de novo* construction of the highly contiguous Asian diploid genome AK1, by combining single molecule technologies with routine sequencing data without resorting to traditional clone-by-clone sequencing and physical mapping, reveals the nature of genomic structure variation by detecting thousands of novel structural variations and by finally filling in some of the prior gaps which had persistently remained in the current human reference genome. Now it is expected that the AK1 genome, soon to be paired with more upcoming *de novo* assembled genomes, will provide a chance to explore what it is really like to use ancestry-specific reference genomes instead of hg19/hg38 for population genomics. This is a major step towards the furthering of genetically-based precision medicine. [BMB Reports 2016; 49(12): 653-654]

*Corresponding author. E-mail: kimchan@macrogen.com

https://doi.org/10.5483/BMBRep.2016.49.12.195

Received 22 November 2016

Keywords: AK1, De novo assembly, Haplotype phasing, Linked Reads, Long reads, NGM, Reference genome

Abbreviations: AK1, Altaic Korean One; BAC, Bacterial Artificial Chromosome; NGM, Next-Generation Map; NIST, National Institute of Standards and Technologies; PacBio, Pacific Biosciences; SMRT, Single Molecule Real Time

Perspective to: Seo, J-S et al., 2016, De novo assembly and phasing of a Korean human genome. Nature; 538:243 http://dx.doi.org/10.1038/nature20098

It was once believed that just one reference genome might be good enough for all human beings − even if it was derived mainly from RP11 (~70%), a male donor of likely African-European admixed ancestry (and partly derived from a group of other people). But there has been an accumulating amount of evidence showing that polymorphism among individuals is significantly higher than had previously been thought in the early days of genomics. Therefore, the necessity for more human reference genomes to be built independently has been high. Thus, there have been many trials for *de novo* sequencing and assembly of human genomes using state-of-the-art technologies of the time. To overcome heterozygosity-related problems in the assembly, some research groups did focus on hydatidiform moles − special cases of essentially haploid genomes − as in CHM1 and CHM13 genome projects.

Considering the technical difficulties in building a new reference for a complex genome from scratch − due to technical challenges such as repeats or segmental duplications entangled with the diploid structure of a human genome having various levels of uneven heterozygosity − it had been doubted that any *de novo* assembled genomes would match the current human reference genome in terms of accuracy and contiguity. Thus, it has remained the most complete assembly − even if its sequence is far from complete, with a number of gaps still remaining.

For the first time in genomics, however, it was shown that a high-resolution reference genome for a diploid state human genome can be generated without resorting to extremely expensive clone-by-clone approaches for sequencing and physical map construction. Seo and his colleagues have *de novo*-assembled and haplotype-phased the Korean AK1 diploid genome by integrating PacBio SMRT long reads, BioNano Genomics next-generation maps, Illumina HiSeq reads, 10X Genomics GemCode linked reads, and BAC clone sequencing.

The combination of PacBio long reads and BioNano Genomics genome maps could yield the assembly with a contig N50 of 17.95 Mb and a scaffold N50 of 44.8 Mb. This is the best contiguity ever achieved for a single diploid genome.

When compared with the reference genome, many of the gaps in hg38 were covered with the AK1 contigs. Several chromosomal arms were almost completely covered with a single scaffold. The telomeres of some chromosomal arms were covered with contigs. Moreover, thousands of structural variations which had previously remained undetected (despite

ISSN: 1976-670X (electronic edition)

the finding as to their being commonly found in the Asian population) were discovered.

Furthermore, the assembly was haplotype phased using reads of (1) PacBio SMRT platform and (2) 10X Genomics Gemcode Platform whole genome sequencing and (3) Illumina HiSeq reads from BAC clones – all to obtain highly contiguous phased blocks with a N50 of 11.5 Mb. Notably complex regions, including HLA and CYP2D6, were able to be almost completely phased. These phased blocks also represent the best quality *de novo* haplotype phased assembly yet achieved.

The first key technology for this project was the PacBio long read SMRT sequencing technology, a so called 'third-generation' sequencing technology that can enable the capture of long-range information of up to tens of kilo-bases. It can do this without bias, since – unlike other second-generation high-throughput sequencing technologies – it does not involve PCR steps. It has repeatedly been proven that relatively high error rate of ~15% in raw sequencing reads can easily be overcome by taking the consensus of aligned reads, since the errors are randomly distributed. It is also important to note that effective assemblers of error-prone long reads – assemblers such as Jason Chin's FALCON – have been critical. PacBio SMRT sequencing will be more affordable and attractive for routine sequencing of large genomes due to the recent upgrade of the system from RSII to Sequel.

The second key technology for this project was next-generation mapping from BioNano Genomics that provided much longer-range information – up to several mega-bases in length – in terms of restriction patterns instead of sequence information. The genome maps assembled from NGMs can, by far, effectively replace the role of BAC clones in building the physical map.

The third key technology for this project was 10X Genomics GemCode linked reads used to generate large phased blocks. Template DNA molecules can be partitioned into small emulsions with reaction components and then bar-coded before Illumina HiSeq sequencing library preparation, so that the origin of each short read can be traced back to its source. This way, long template DNA sequence can be reconstructed using the barcodes belonging to the short reads. Again, recent upgrades to the Chromium of the system provide promising options—since with these upgrades, the genome can be covered more evenly.

The first successful generation of a phased reference genome for another ethnic group is envisioned as being a trigger for the future production of more reference-grade phased human genomes with an imminent application of high-throughput and single molecule sequencing technologies in precision genomic medicine.

Those assemblies will be crucial for practicing precision medicine globally, since direct comparison of the AK1 assembly with the reference genome reveals ethnic-specific structural variations that had previously remained undetected with conventional re-sequencing based approaches. This information could help society head towards an era of precision genomic medicine, in which healthcare will be tailored for the genetic makeup of each individual.

Although many people have had their genomes sequenced with high-throughput sequencing technologies so far, the interpretations of such genomic information heavily rely on mapping of the sequenced reads onto the reference genome, using bioinformatics tools such as BWA and Bowtie. Therefore, integrated mapping results for multiple reference genomes would be much more powerfully informative for precision genomic medicine.

On the other hand, this AK1 assembly could be used as reference material as is NA12878 maintained by GIAB (Genome in a bottle) consortium led by NIST, since the AK1 cell line is available with high-quality assembly and raw sequencing read data sets produced from different platforms, which will help development of better experimental and computational tools for genomic analysis.

ACKNOWLEDGEMENTS

In thanks for their constructive discussions, I would like to thank all those members of Macrogen Bioinformatics and Genome Institutes who contributed to this study.