BMB Reports - Manuscript Submission

Manuscript Draft

Manuscript Number: BMB-16-195

Title: Birth of an Asian Cool Reference Genome: AK1

Article Type: Perspective (Invited Only)

Keywords: AK1; de novo assembly; reference genome; long reads; NGM

Corresponding Author: Changhoon Kim

Authors: Changhoon Kim[1,*]

Institution: [1]Bioinformatics Institute, Macrogen Inc. Seoul 153-023, Korea,

Perspective

Birth of an Asian Cool Reference Genome: AK1

Changhoon Kim[*]

Bioinformatics Institute, Macrogen Inc. Seoul 153-023, Korea.

The human reference genome, maintained by Genome Reference Consortium, is conceivably the most complete genome assembly ever since its first construction. It has been improved continually by incorporating corrections made to the previous assemblies thanks to various technological advances. Based on this reference genome, many population sequencing projects are ongoing to better represent human diversity, increasing hope for medical usage of genomic information, with recent maturation of the high-throughput sequencing technologies. However, the one reference genome does not fit all the populations across the globe, because of large diversity in genomic structures and technical limitations inherent to short read sequencing methods. The recent success in *de novo* construction of 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 filling some of the persistently remained gaps in the current human reference genome. Now it is expected that the AK1 genome, with more *de novo* assembled genomes coming up, will provide a chance to exploring what it is really like to use ancestry specific reference genomes instead of hg19/hg38 for population genomics toward precision medicine.

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

Keywords: AK1, *de novo* assembly, reference genome, long reads, NGM, Linked Reads, haplotype phasing.

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

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 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 from a group of other people. But there have been accumulating evidences showing that polymorphism among individuals is significantly higher than previously thought in early days of genomics. Therefore, the necessity for more human reference genomes built independently has been high and there have been many trials for *de novo* sequencing and assembly of human genomes with state-of-the-art technologies of the time. To overcome heterozygosity-related problems in the assembly, some research groups focus on hydatidiform moles, special cases of essentially haploid genomes, as in CHM1 and CHM13 genome projects.

Considering technical difficulties in building a new reference for a complex genome from the scratch, due to repeats or segmental duplications entangled with diploid structure of human genome having various level of uneven heterozygosity, it is undoubted that any *de novo* assembled genomes do not match the current human reference genome in terms of accuracy and contiguity. Thus it has been 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 contig N50 of 17.95 Mb and scaffold N50 of 44.8 Mb. This is the best contiguity ever achieved for a diploid genome. When compared with the reference genome, many of the gaps in hg38 were covered with the contigs, several chromosomal arms were almost completely covered with a single scaffold, telomeres of some chromosomal arms were covered with contigs, and thousands of structural variations that were undetected before where many of them are common in the Asian population.

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 to obtain highly contiguous phased blocks with N50 of 11.5 Mb. The complex regions including HLA and CYPD were

able to be completely phased. These phased blocks also represent the best quality *de novo* haplotype phased assembly yet achieved.

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

Second key technology was next-generation mapping from BioNano Genomics that provided much longer range information, up to several mega-bases, 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 for building physical map.

Third key technology 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 preparing Illumina HiSeq sequencing library, so that the origin of each short read can be traced back. This way, long template DNA sequence can be reconstructed using the barcodes belonging to the short reads. Again recent upgrade to the Chromium of the system also seems promising, since the genome can be covered more evenly.

The first successful generation of phased reference genome for other ethnic group will trigger to produce more reference-grade phased human genomes with the imminent application of high-throughput and single molecule sequencing technologies in precision medicine.

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

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

On the other hand, this AK1 assembly could be used as a reference material as NA12878 maintained by GIAB (Genome in a bottle) consortium led by NIST, since 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 genome analysis.

ACKNOWLEDGEMENTS

I would like to thank all the members of Macrogen Bioinformatics and Genome institutes for their constructive discussions.