Genome Biology

RESEARCH HIGHLIGHT Open Access

One reference genome is not enough

Xiaofei Yang[1,2,3], Wan-Ping Lee[2,3,4], Kai Ye[2,5] and Charles Lee[3,4,6*]

Abstract

A recent study on human structural variation indicates insufficiencies and errors in the human reference genome, GRCh38, and argues for the construction of a human pan-genome.

Introduction

The human reference genome is a critical foundation for human genetics and biomedical research. The current human reference genome, GRCh38, blends genomic segments from a few individuals, although clones of a single individual predominate [1]. This invites criticisms of the ability of such a reference genome to present the common variants from multiple human populations accurately. In addition, the current human reference genome harbors many genomic segments that actually contain rare variants, and these impact downstream sequence analyses including read alignments and the identification of variants, especially the identification of structural variants (SVs) (that is, insertions, deletions and rearrangements) that encompass more than 50 bp of DNA. Incorporating SVs that are shared among major human populations into the current reference genome can correct for biases and improves both read alignments and the detection of variants in other individuals. Recently, a study based on deep (i.e., > 50×) long-read PacBio whole genome sequencing (WGS) data for 15 individuals from five populations led to the discovery and sequencing of a large fraction of common structural variation. These data can be used to genotype variants from other short-read sequencing datasets and ultimately to reduce biases inherent in the GRCh38 version of the human reference genome [2].

SV discovery based on long-read sequencing data

Audano et al. [2] sequenced 11 genomes (from three African, three Asian, two European and three American samples) using single-molecule, real-time (SMRT) PacBio RSII and Sequel long-read sequencing technology. They further analyzed long-read sequencing data, including data from four additional sources: CHM1 [3], CHM13 [3], AK1 [4] and HX1 [5]. Reads were aligned against the GRCh38 version of the human reference sequence using the BLASR software and SVs were detected using the SMRT-SV algorithm [6]. In total, 99,604 nonredundant SVs were identified from these 15 sequenced genomes. The analysis focused on around 95% of the human genome but excluded the pericentromeric and other regions of the genome that are enriched for repetitive DNAs (Fig. 1a). Among the 99,604 discovered SVs, the existence of 2238 'shared type' SVs (shared across all samples) and 13,053 'majority type' SVs (present in more than half of the genomes studied, but not in all samples) suggested that the current reference genome either carries a minor allele or contains an error at each of these positions. These shared and majority SVs were enriched with repetitive sequences and reflect insertions (61. 6 %), deletions (38.1%) and inversions (0.33%). Excluding analyses of the highly repetitive regions of the human genome (which probably contain many SVs), a logarithmic function conservatively suggested that adding SV data from an additional human genome would probably increase the total SV callset by 2.1%, adding 35 genomes would increase the total SV callset by 39% and, finally, adding 327 genomes would identify twice as many SVs than were identified from these 15 genomes.

Among the SVs discovered, 40.8% are novel when compared to previously described SVs from several published large-scale projects (Figure S1E in [2]). To assess the allele frequency of the discovered SVs, Audano et al. [2] went on to genotype these SVs across a total of 440 additional genomes, which were all sequenced using short-read technologies, including those of 174 individuals from the 1000 Genomes Project and 266 individuals from the Simons Genome Diversity Project [7]. The results showed that 92. 6% of the released SVs actually appeared in more than half of the samples, further confirming these biases in the GRCh38 version of the human reference genome.

* Correspondence: Charles.Lee@jax.org
[3]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
[4]Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China
Full list of author information is available at the end of the article

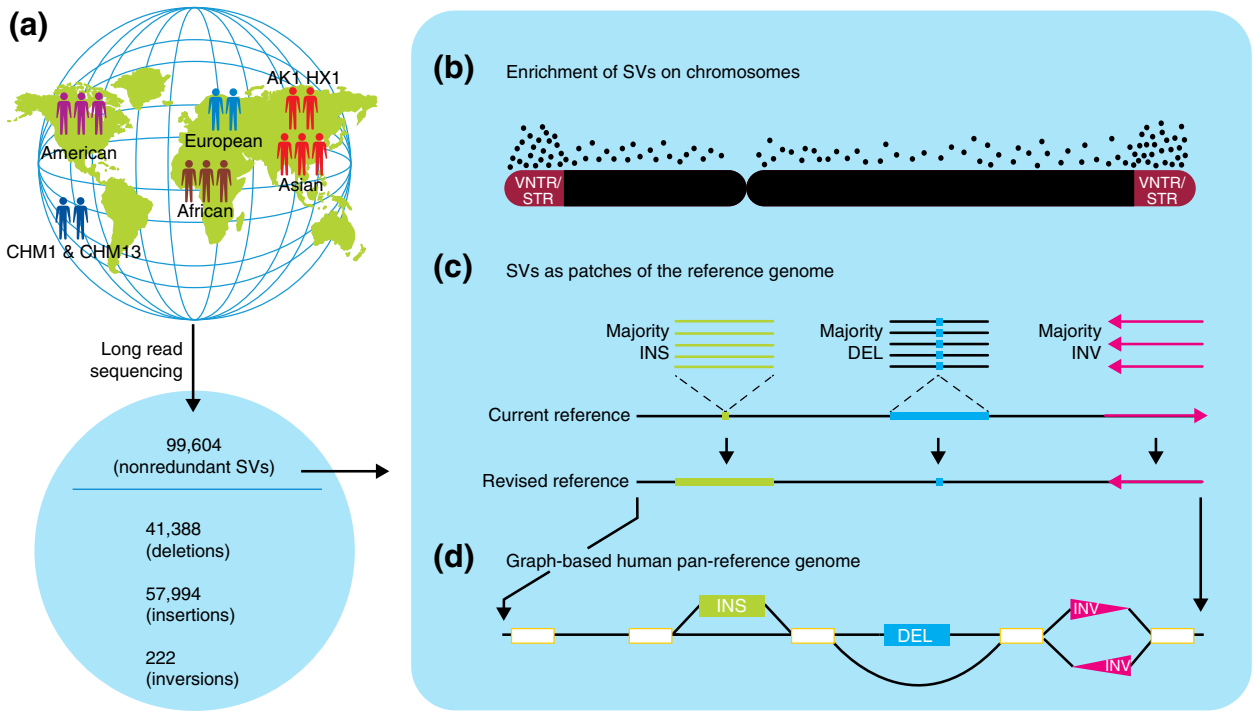

Fig. 1 The human genome structural variant (SV) resource. **a** The detection of 99,604 nonredundant SVs in 15 samples from five populations using a long-read sequencing technology. AK1 [4] and HX1 [5] are Asian individuals whose genomes were previously sequenced. **b** The subtelomeric regions of human chromosomes are particularly enriched for SVs of the variable number of tandem repeats (VNTR) and short tandem repeat (STR) types. Here, the frequency of black dots along the length of the chromosome indicates the relative density of SVs. **c** About 15% of the discovered SVs can be found in more than 50% of the samples studied, indicating that these sites actually harbor minor alleles or errors in the current reference genome. **d** Ultimately, a human pan-reference genome can be developed using genome graphs (or other methods) to represent common SVs accurately. *DEL* deletion, *INS* insertion, *INV* inversion

SVs enriched with tandem repeat sequences

Audano et al. [2] found that SVs are not randomly distributed across the genome, and in fact, there was as much as a nine-fold increase in SV density within the subtelomeric regions (the last 5 Mb) of human chromosomes. In addition, SVs in these subtelomeric regions were significantly enriched with tandem repeats, particularly for VNTRs (variable number of tandem repeats) and STRs (short tandem repeats), rather than retrotransposons (Fig. 1b). There was also a positive correlation between the abundance of STRs (R = 0.27) and VNTRs (particularly larger VNTRs; R = 0.48) with known hotspots of meiotic double strand breaks (DSBs), suggesting a potential role for DSBs in the formation of SVs in these genomic regions.

SVs affect gene structures and regulatory elements

How do the discovered SVs interfere with gene expression? To address this question, Audano et al. [2] annotated the shared and majority SVs using RefSeq. The analysis showed that 7550 of these SVs intersect with gene regions (including coding regions, untranslated regions (UTRs), introns, and 2-kb flanking regions), and 1033 of these SVs intersect with known regulatory elements. Some of the SVs disrupted gene structures: 841 intersected RefSeq-annotated coding regions and 667 intersected RefSeq-annotated noncoding RNA regions. For example, a 1.6-kb insertion was located in the 5′ UTR of *UBEQ2L1* and extended into its promoter. In another case, a 1.06-kbp GC-rich insertion was located at the 3′ UTR of *ADARB1* and incorporated motifs that may promote the formation of a quadruplex structure. Examples of SVs located in gene regulatory elements included a 1.2-kb and a 1.4-kb fragment inserted upstream of *KDM6B* and *FGFR1OP*, respectively. These insertions intersected with H3K4Me3 and H3K27Ac sites. Audano et al. [2] further investigated the impact of SVs on gene expression using RNA-seq data from 376 European cell lines and found that the expression of 411 genes was significantly associated with the discovered SVs.

The discovered SVs can be helpful for re-constructing a canonical human reference genome

GRCh38 currently contains 819 gaps, including minor alleles or actual errors. Audano et al. [2] proposed that the SVs discovered in their work could be included to

correct the reference genome (Fig. 1c). They found 34 shared insertions that intersect with scaffold switch-points of the GRCh38 version of the reference genome and the new data could be used to correct possible misassemblies in GRCh38. For instance, a 2159-bp shared insertion overlaps with a switch-point in the *NUTM1* gene and indicates a misassembly by stitching two contigs together. Additional sequencing clones from BAC libraries confirmed the misassembly. Adding the discovered SV contigs to the reference genome could rescue 2.62% of unmapped Illumina short reads, and 1.24% of the SV-contig-mapped reads show increased mapping quality, thus improving variant detection. This effect is most pronounced for insertions, for which 25.68% of the reads show increased mapping quality when compared to the reference genome. Furthermore, GATK was able to identify a substantial amount of variation within SV insertions (i.e., 68,656 alternative alleles across the 30 whole-genome haplotypes) where no reference sequence previously existed. Taken together, these data proved to be useful in re-constructing a more precise canonical human reference genome.

Concluding remarks

Audano et al. [2] provided a sequence-resolved SV callset from analysis of 15 human genomes. They found the reported SVs to be significantly enriched with VNTRs and STRs and correlated with DSB. In addition, they found that certain SVs impact gene regulatory elements and affect gene expression, opening a door for additional future studies correlating SVs with gene expression. They further patched errors and biases in the current human reference genome assembly using their SV callset, significantly improving the quality of future short-read alignments and variant calling. This study also promotes the concept of a pan-genome (Fig. 1d), which incorporates SVs into the reference genome and can be applied to recently published graph genome tools [8, 9]. The next steps will involve phasing human genomes to reduce false negatives [10] and discovering complex SVs and indels that map to large repetitive regions of the human genome.

Abbreviations

DSB: Double strand break; SMRT: Single-molecule, real-time; STR: Short tandem repeat; SV: Structural variant; UTR: Untranslated region; VNTR: Variable number of tandem repeats

Acknowledgments

We would like to thank Winnie Jane Cha for her work on Fig. 1.

Authors' contributions

XY, WL, KY and CL contributed to the writing of this article. All authors read and approved the final manuscript.

Funding

XY and KY are supported by the National Science Foundation of China (61702406 and 31671372), the National Science and Technology Major Project of China (grant number 2018ZX10302205), and the National Key R&D Program of China (2018YFC0910400 and 2017YFC0907500). WL and CL are partially supported by a grant from the National Institutes of Health (NIH) USA (U41HG007497) and CL is a distinguished Ewha Womans University Professor, supported in part by the Ewha Womans University research grant of 2018–9.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science and Technology, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [2]MOE Key Lab for Intelligent Networks & Networks Security, School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China. [3]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. [4]Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China. [5]Genome Institute, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China. [6]Department of Life Sciences, Ewha Womans University, Seoul 03760, South Korea.

Published online: 24 May 2019

References

1. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27:849–64.
2. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the major structural variant alleles of the human genome. Cell. 2019;176:663–75.
3. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017;27:677–85.
4. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538:243–7.
5. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and de novo assembly of a Chinese genome. Nat Commun. 2016;7:12065.
6. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517:608–11.
7. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.
8. Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet. 2019;51:354–62.
9. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9.
10. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10:1784.

Publisher's Note