Genomics 109 (2017) 83–90

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis

Yan Guo [a,*], Yulin Dai [a], Hui Yu [a], Shilin Zhao [a], David C. Samuels [b], Yu Shyr [c,*]

[a] Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA
[b] Vanderbilt Genetics Institute, Dept. of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, TN, USA
[c] Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 8 September 2016
Received in revised form 16 January 2017
Accepted 24 January 2017
Available online 26 January 2017

Keywords:
Human reference genome
GRCh37
GRCh38
High throughput sequencing
SNP
Copy number variation
Structural variant

ABSTRACT

Analyses of high throughput sequencing data starts with alignment against a reference genome, which is the foundation for all re-sequencing data analyses. Each new release of the human reference genome has been augmented with improved accuracy and completeness. It is presumed that the latest release of human reference genome, GRCh38 will contribute more to high throughput sequencing data analysis by providing more accuracy. But the amount of improvement has not yet been quantified. We conducted a study to compare the genomic analysis results between the GRCh38 reference and its predecessor GRCh37. Through analyses of alignment, single nucleotide polymorphisms, small insertion/deletions, copy number and structural variants, we show that GRCh38 offers overall more accurate analysis of human sequencing data. More importantly, GRCh38 produced fewer false positive structural variants. In conclusion, GRCh38 is an improvement over GRCh37 not only from the genome assembly aspect, but also yields more reliable genomic analysis results.

1. Introduction

The complete human genome consists of 22 diploid chromosomes (1 − 22), two sex chromosomes (X and Y) and maternally inherited mitochondrial DNA (mtDNA). Variant alleles have different features depending on what part of this genome they occur in. The diploid chromosomes are the simplest case, where two alleles are presented at any genomic position, with one inherited from each parent. The mtDNA is maternally inherited, thus only 1 allele should be present at any given genomic position. However, with the rise of high throughput sequencing technology, the phenomenon of heteroplasmy has been consistently detected in humans [1–4]. Heteroplasmy can produce an essentially continuous distribution of mtDNA allele frequency in a single individual. The ploidy of sex chromosomes differs by gender. Males have both X and Y chromosomes, and females have two X chromosomes without the Y chromosome. Therefore, we should not observe any heterozygous genotypes in the X chromosome for males, and we should not observe any genotype for Y chromosome variants for females.

The human reference genome is the fundamental necessity for almost all high throughput re-sequencing based biomedical research. The assembly of a reference genome is usually referred as *de novo* assembly. To reconstruct a reference genome, DNA fragments of the targeted specie are sequenced in high quantity, resulting the sequenced reads to theoretically cover the entire genome. By aligning and merging the sequenced reads, based on their overlapping nucleotides, contiguous segments (contig) DNA sequences can be assembled. A contig is a contiguous length of genomic sequence in which the order of bases is known to a high confidence level. Multiple contigs can be assembled together to form a scaffold based on the paired read information. A scaffold is a portion of the genome sequences composed of contigs but which might contains gaps between these contigs. Various tools have been developed to perform genome assembly from short reads [5–7] and to close gaps between scaffolds [8–10]. Finally, multiple scaffolds can be joined together to form a chromosome (Fig. 1).

In practice, there are many challenges associated with reconstructing a complete and correct human reference genome. The best known challenges include repetitive DNA regions such as telomeres [11], which can considerably convolute the consensus sequence; limitations on high throughput sequencing read length where longer reads are preferable since they will result in larger overlapping segments, and thus less ambiguity in joining reads [12]; and uneven representation of the genome due to sequencing sensitivity to GC bias [13,14] which can cause gaps between scaffolds. Researchers have been actively tackling these challenges and have gradually improved the human reference genome. The very first human genome reference was assembled by The Human Genome Project in 2001 [15].

In 2009, the Genome Reference Consortium (GRC) released human reference genome version GRCh37 which is also often refereed as HG19 because it was the 19th release. GRCh37 was released around

[*] Corresponding authors.
E-mail address: yan.guo@vanderbilt.edu (Y. Guo).

http://dx.doi.org/10.1016/j.ygeno.2017.01.005
0888-7543/© 2017 Published by Elsevier Inc.

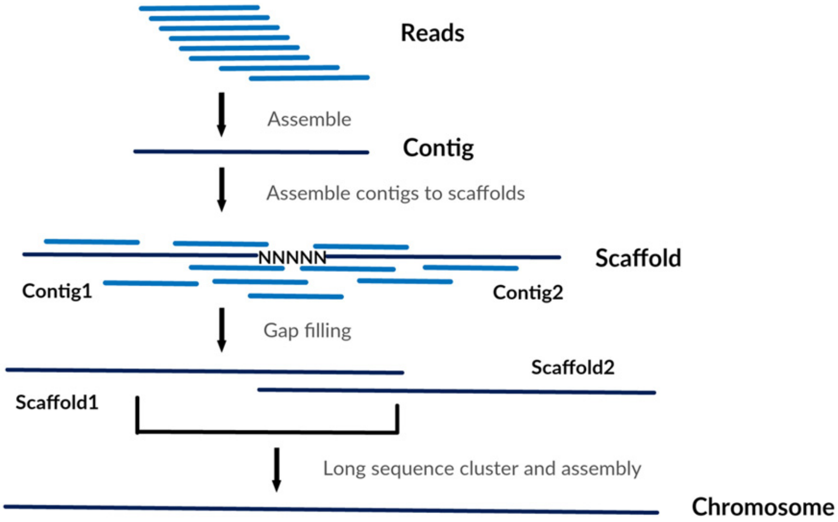

Fig. 1. The general steps of *de novo* assembly. Overlapped reads are first joined together to form contigs, then contigs were assembled to scaffolds and scaffold are assembled together to form a chromosome.

the time when Illumina's high throughput sequencing technology started to take over the market of high throughput biomedical research. A few years later, Illumina's high throughput sequencing technology completely replaced microarray hybridization based gene expression profiling [16–20] and its applications in sequencing of DNA had increased exponentially [21]. The GRCh37 reference was used extensively in sequencing data analysis for many years. Even after the 20th release of the human reference genome GRCh38 in 2013, GRCh37 was still being used to some extent. There are several reasons for the hesitation of researchers to switch to the latest reference build, including the initial lack of annotation tools and resistance to altering existing working pipelines.

According to The GRC press release, GRCh38 is the most accurately sequenced human genome in the world. It was constructed from many donors instead of a few, and the sequencing was performed using the gold standard Sanger sequencing, which can produce reads as long as 1000 nucleotides and 10 times more accurate than high throughput short read sequencing. Compared to GRCh37, GRCh38 altered 8000 nucleotides, corrected several misassembled regions, filled in gaps, added sequence for centromeres, and substantially improved the diversity of the reference by including 261 alternate loci across 178 regions.

On paper, GRCh38 has been touted as a major improvement from GRCh37. These improvements should in theory be translated to more accurate bioinformatics and genomic analysis. To evaluate how much improvement the new reference genome can provide, we designed a study to quantitatively access the difference between analysis results based on GRCh37 and GRCh38.

2. Results

To access the impact of GRCh38 compared to GRCh37 on genomics analysis, we analyzed a dataset of exome sequences ($N = 30$) using both human reference genomes. The comparisons between GRCh38 and GRCh37 were performed from multiple perspectives including basic chromosome statistics, alignment, single nucleotide variables (SNV), small insertions and deletions (INDEL), copy number variations (CNV) and structural variants.

The comparison result of basic statistics between GRCh38 and GRCh37 can be viewed in Table 1 and Table S1. GRCh37 has a total of 3,095,677,412 nucleotides, and GRCh38 has a total of 3,088,269,832 nucleotides, a decrease of 7,407,580 nucleotides when counting chromosome 1 to 22, X and Y. The mitochondrial genome was ignored in the counting because the most used mitochondria reference is the revised Cambridge mitochondrial reference sequence (rCRS) [22] which has not changed since 1999. Of the 24 chromosomes examined, 16 have decreased and 8 have increased nucleotide counts for GRCh38. The letter “N” was used in the reference genome (FASTA file) to represent a sequence gap or unannotated regions. There are total of 234,350,281 Ns in GRCh37, and 150,630,719 Ns in GRCh38, a large decrease of 83,719,562 Ns. All 24 chromosomes showed decreased number of Ns. GC content is the percentage of G and C nucleotides in the genome. It has been shown that the GC content can affect Illumina sequencing's efficiency [23] and influence subsequent analysis such as CNV detection, which is heavily dependent on depth of coverage [24]. The GC content percentage varies by regions of the human genome [25]. The overall number of GC sites increased from GRCh37's 1,170,371,008 to GRCh38's 1,200,551,672 by 30,180,664 nucleotides. When we computed the GC%, we subtracted the number of Ns from the denominator. Because GRCh38 has much less number of Ns, seventeen of the 24 chromosomes have decreased GC%.

The exome is arguably the most important component of the human genome due to its encoding of protein coding sequences. It is the intended target of exome sequencing [26]. The definition of the exome depends critically on genome annotation. We examined the size of the exome from the latest Gene Feature Format (GTF) files downloaded from Ensembl (GRCh37 v37.75, GRCh38 v38.82). The exome size increased significantly from GRCh37's 75,231,228 to GRCh38’s 95,505,476 by 20,274,248 nucleotides, a 26.9.0% increase. All chromosomes increased in exome size. Percentage wise, 2.43% of GRCh37 is exome as compared to 3.09% of CRCh38. The increase in exome size can be attributed to several reasons. First, the total number of distinct exons increased from 327,058 to 457,748 in GRCh38 and the median number of exons per gene also increased from 13 to 19 in GRCh38, while the median number of nucleotide per exon increased slightly almost from 140 to 146 in GRCh38. These combined factors explain why the increase in the exome% in GRCh38.

Alignment is the very first step for conducting sequencing data analysis. It is well known that a small percentage of the sequenced reads will not align to the human genome and it has been suggested that improving the human reference genome may also improve the alignment rate [21]. Thus, we examined the mapping rate of the 30 exome sequencing samples, and all 30 samples showed an improved mapping rate (Fig. 2, Table S2). The average of the improvement is 0.0017%. All samples also showed increased mapping rate to the exome by an average of 3.22%. The increased mapping rate to the exome can be explained by the