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/

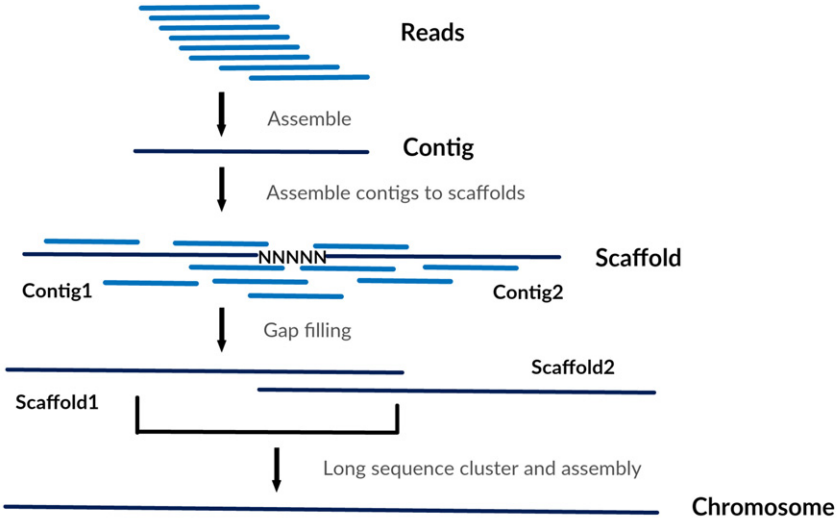

Fig. 1. The general steps of *de novo* assembly. Overlapped reads are first joined together to form contigs, then contigs were assembled to scaffolds and scaffold are assembled together to form a chromosome.

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

Table 1
Genome statistics comparison between GRCh37 and GRCh38.

	Total length	Total length	N%[a]	N%	GC%[b]	GC%	Exome%	Exome%
Chr	GRCh37	GRCh38	GRCh37	GRCh38	GRCh37	GRCh38	GRCh37	GRCh38
chr1	249,250,621	248,956,422	9.62%	7.42%	41.74%	41.72%	3.07%	3.82%
chr2	243,199,373	242,193,529	2.05%	0.68%	40.24%	40.23%	2.26%	2.87%
chr3	198,022,430	198,295,559	1.63%	0.10%	39.69%	39.67%	2.26%	2.85%
chr4	191,154,276	190,214,555	1.83%	0.24%	38.25%	38.24%	1.68%	2.05%
chr5	180,915,260	181,538,259	1.78%	0.15%	39.52%	39.51%	2.01%	2.54%
chr6	171,115,067	170,805,979	2.17%	0.43%	39.61%	39.61%	2.29%	2.70%
chr7	159,138,663	159,345,973	2.38%	0.24%	40.75%	40.70%	2.23%	2.78%
chr8	146,364,022	145,138,636	2.37%	0.26%	40.18%	40.16%	1.85%	2.33%
chr9	141,213,431	138,394,717	14.92%	12.00%	41.32%	41.28%	2.12%	2.57%
chr10	135,534,747	133,797,422	3.11%	0.40%	41.58%	41.54%	2.25%	2.69%
chr11	135,006,516	135,086,622	2.87%	0.41%	41.57%	41.54%	3.11%	4.02%
chr12	133,851,895	133,275,309	2.52%	0.10%	40.81%	40.77%	3.08%	4.12%
chr13	115,169,878	114,364,328	17.00%	14.32%	38.53%	38.55%	1.14%	1.54%
chr14	107,349,540	107,043,718	17.76%	15.39%	40.89%	40.83%	2.25%	3.14%
chr15	102,531,392	101,991,189	20.32%	17.01%	42.20%	42.03%	2.50%	3.58%
chr16	90,354,753	90,338,345	12.69%	9.44%	44.79%	44.58%	3.32%	4.64%
chr17	81,195,210	83,257,441	4.19%	0.41%	45.54%	45.31%	5.06%	6.51%
chr18	78,077,248	80,373,285	4.38%	0.35%	39.78%	39.78%	1.73%	2.24%
chr19	59,128,983	58,617,616	5.61%	0.30%	48.36%	47.94%	7.34%	9.65%
chr20	63,025,520	64,444,167	5.59%	0.78%	44.13%	43.80%	2.83%	3.39%
chr21	48,129,895	46,709,983	27.06%	14.18%	40.83%	40.94%	1.68%	2.23%
chr22	51,304,566	50,818,468	31.99%	22.94%	47.99%	47.00%	3.18%	3.97%
chrX	155,270,560	156,040,895	2.69%	0.74%	39.50%	39.53%	1.78%	2.01%
chrY	59,373,566	57,227,415	56.79%	53.84%	39.97%	40.03%	0.22%	0.25%
Total	3,095,677,412	3,088,269,832	7.57%	4.88%	40.90%	40.87%	2.43%	3.09%

[a] In reference genome, a gap or unknown region is filled with letter "N". N% denotes the percentage of Ns in whole genome.
[b] GC% was computed as number G + C divided by (total length of the genome subtracting number of Ns).

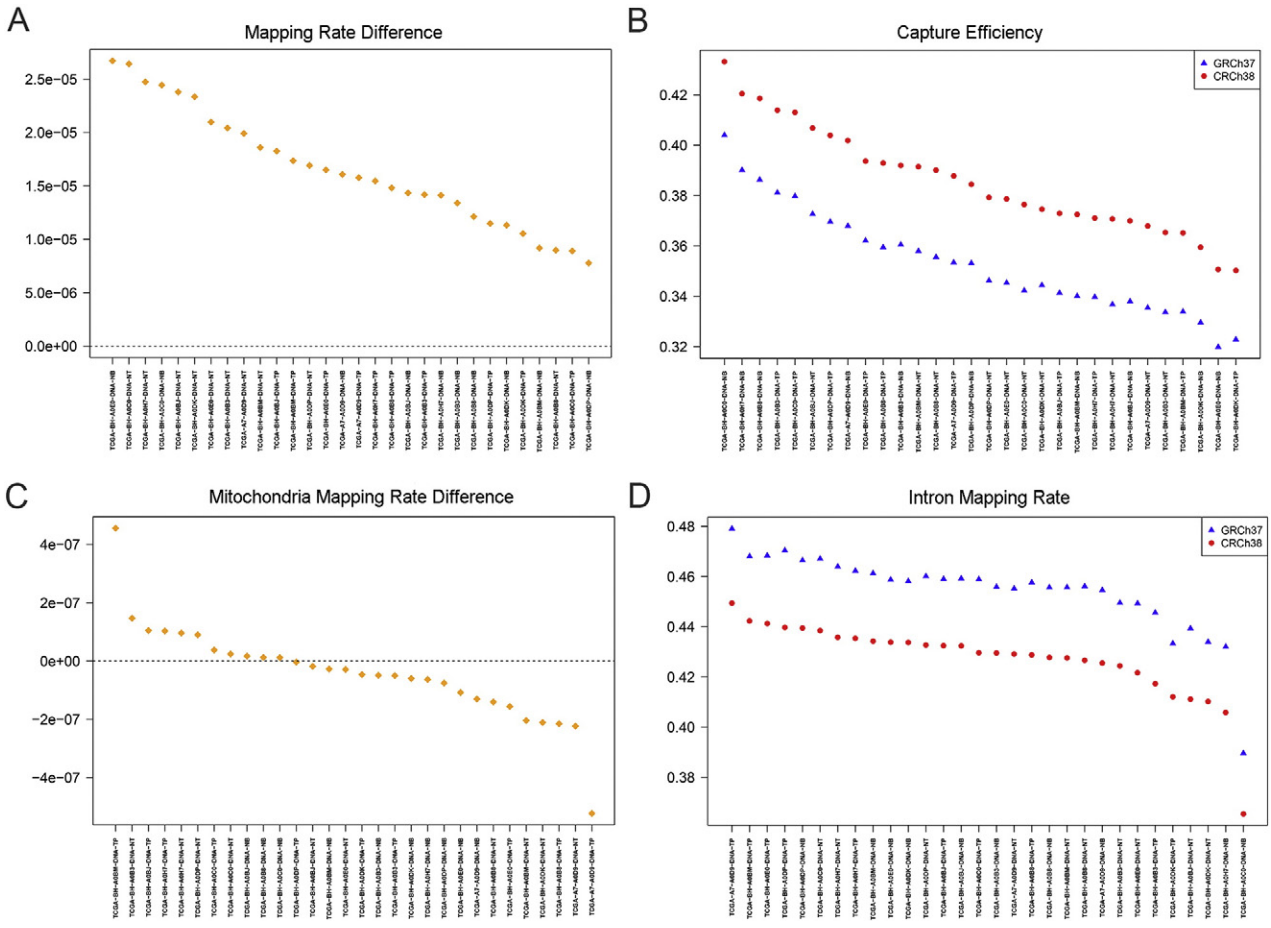

Fig. 2. Mapping rate comparisons between GRCh38 and GRCH37. A. overall mapping rate difference, computed as GRCh38 − GRCh37. All samples tested observed better mapping rate. B. Capture efficiency is defined as total number of rates mapped to exome. All samples tested observed higher capture efficiency. C. Mapping rate difference to mitochondria. More samples showed lower mapping rate to mitochondria in GRCh38. D. Mapping rate to intron. All samples tested showed lower mapping rate to intron which could be result of increased exome size in GRCh38 and better exome definitions.

Table 2
SNV categories detected.

Categories	Grch37	Grch38	Difference
Downstream	29,647	29,841	194
Exonic	46,524	46,363	−161
Exonic; splicing	11	11	0
Intergenic	2,627,378	2,585,510	−41,868
Intronic	1,651,974	1,642,466	−9508
Ncrna_exonic	14,881	15,107	226
Ncrna_exonic; splicing	5	5	0
Ncrna_intronic	215,486	228,021	12,535
Ncrna_splicing	77	76	−1
Splicing	146	138	−8
Upstream	26,203	25,884	−319
Upstream; downstream	971	1213	242
Utr3	36,065	36,031	−34
Utr5	7076	7176	100
Utr5; Utr3	17	17	0
Total	4,656,461	4,617,859	−38,602

Table 3
Categories of exonic and splicing SNVs detected.

Categories	GRCh37	GRCh38	Difference
Nonsynonymous	22,372	22,538	166
Stopgain	223	231	8
Stoploss	27	27	0
Synonymous	23,178	23,270	92
Unknown	735	308	−427
total	46,535	46,374	−161

increased size of exome from GRCh37 to GRCh38. The mapping rate to introns dropped by an average of 2.70% in GRCh38. The mapping rate to the mitochondrial genome slightly decreased by an average of 0.0005%, with 19 of the 30 samples having a slight decrease in the number of reads mapped to mitochondrial DNA. Since the mtDNA reference stayed the same between GRCh37 and 38, this decrease is likely due to modifications of nuclear copies of the mitochondrial genomes (nuMTS) [27,28].

Overall fewer SNVs were identified using GRCh38. There were 4,656,461 SNVs identified using GRCh37 and only 4,617,859 SNVs identified using GRCh38. This suggests that due to an overall improved human reference genome, GRCh38 produced fewer false positive SNVs as indicated by the overall smaller number of SNVs detected. The detailed list and categories of SNVs detected can be viewed in Table 2. When focusing on the exonic nonsynonymous (nonsynonymous, stopgain, stoploss) SNVs, which are often considered the most important SNVs, GRCh38 detected more with 22,796 SNVs compared to GRCh37's 22,622 SNVs (Table 3). The increased number of nonsynonymous SNPs is caused by the increase of exome regions in GRCh38. By using the LiftOver tool to compare genomic positions, the vast majority (93.0%) of the SNVs detected were shared between the GRCh37 and GRCh38 alignment. 241,073 unique SNPs for GRCh37 and 173,969 unique SNPs for GRCh38 were also identified (Fig. 3A). For INDELs, 504,515 were identified for GRCh37, and 499,119 were identified for GRCh38, with 88.0% of the INDELs shared between the two reference genome alignments (Fig. 3B). We also examined the quality scores, and depth coverage for SNVs and INDELs and find there were no substantial difference between the GRCh37 and GRCh38 alignments (Fig. 4).

Next, we examined CNV results. There were 3702 CNVs identified through GRCh37, and 3732 CNVs identified using GRCh38, with 88.4% shared between the two reference genomes (Fig. 3C). Both reference genomes found overwhelmingly more duplications than deletions. Using GRCh37, we found 135 deletions and 3567 duplications. Using GRCh38, we found 131 deletions and 3601 duplications. These results are in agreement with our previous finding that high throughput sequencing is prone to identify more duplication than deletion CNV [29].

We compared the structure variants detected. Using GRCh37, we identified 371,558 structural variants and with GRCh38 we identified a much smaller number of 271,825 structural variants. 83.0% of the structural variants were shared between the two reference genomes (Fig. 3D). Structural variant detection is well known to be difficult and prone to a high false positive rate [30]. We hypothesized that the structural improvements in GRCh38have alleviated the false positive rate compared to GRCh37. The result from structural variant analysis showed much fewer variants identified in GRCh38 (26.8% less)

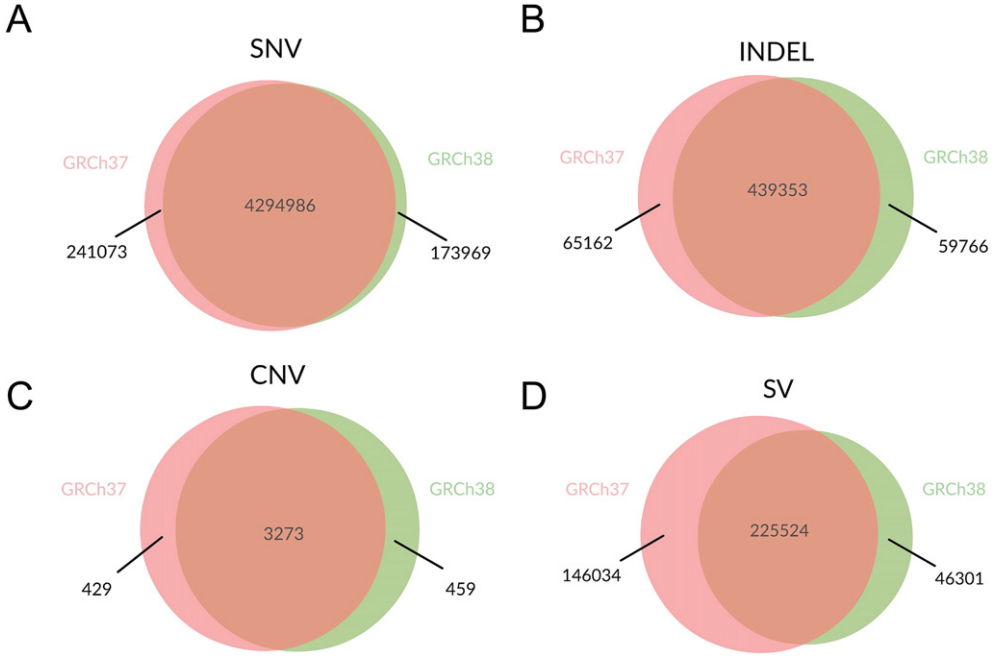

Fig. 3. A. Venn diagram of SNVs detected between using GRCh37 and GRCh38. B. Venn diagram of INDELs between GRCh37 and GRCh38. C. Venn diagram of CNVs between GRCh37 and GRCh38. D. Venn diagram of structural variants between GRCh37 and GRCh38.

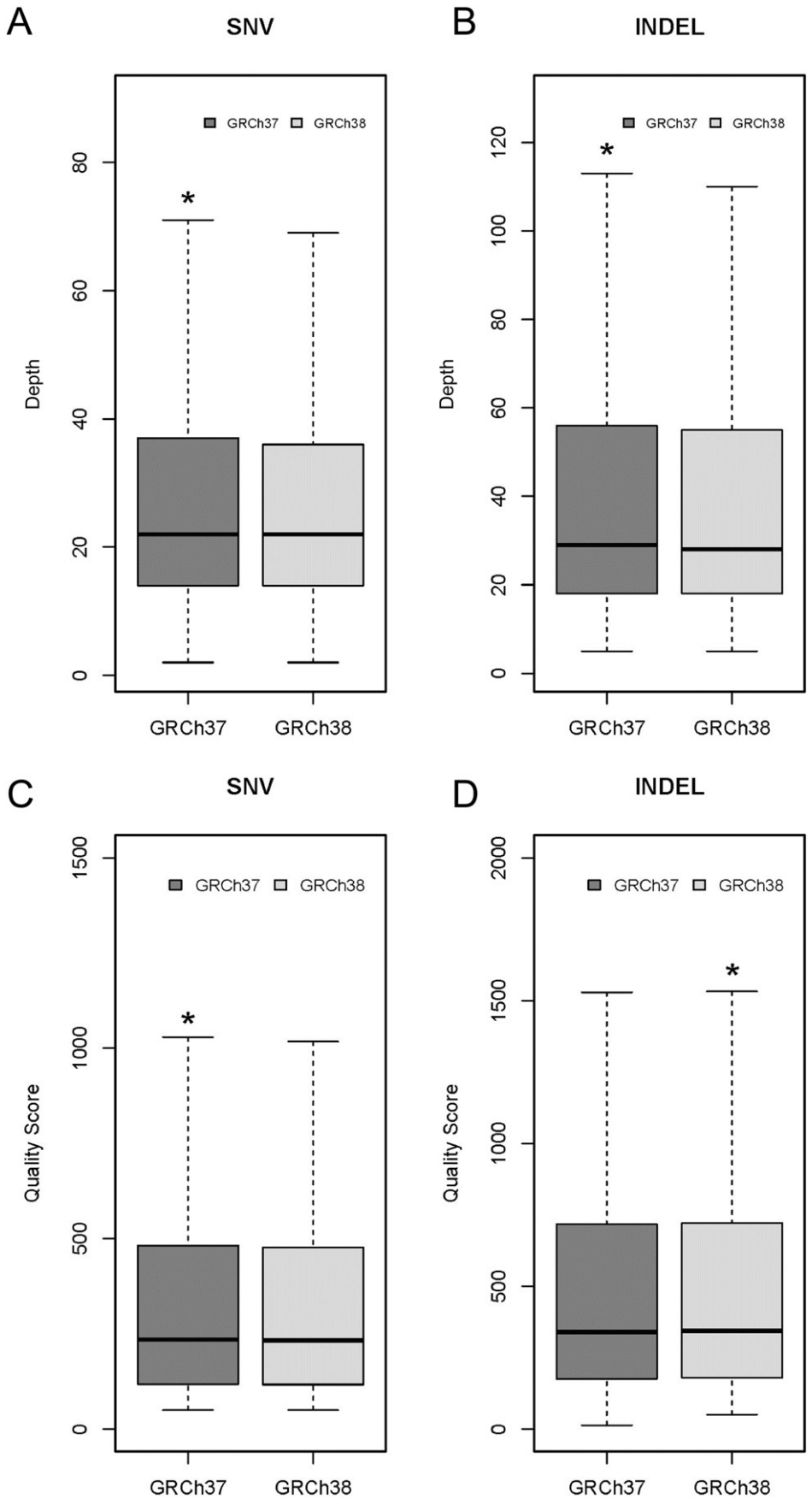

Fig. 4. A. Boxplot of depth for all SNVs detected between GRCh37 and GRCh38. B. Boxplot of depth for all INDELs detected between GRCh37 and GRCh38. C. Boxplot of quality score for all SNVs detected between GRCh37 and GRCh38. D. Boxplot off quality score for all INDELs detected between GRCh37 and GRCh38. There is no substantial difference between depth and quality score of SNVs and INDELs detected between GRCh37 and GRCh38.

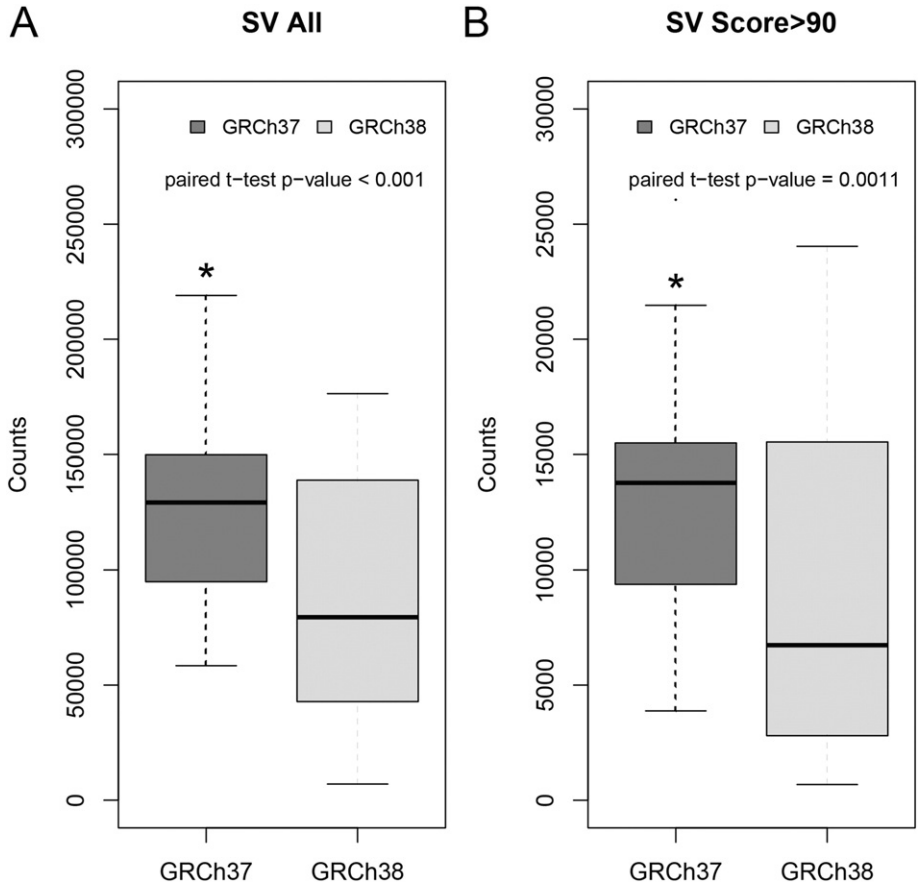

Fig. 5. A. Structural variants distribution comparison between GRCh37 and GRCh38 without quality filter. B. Structural variants distribution comparison between GRCh37 and GRCh38 with quality score >90. Quality score was reported by BreakDancer with range from 1 to 100. Higher is better.

(Fig. 5). Although, we do not have a gold standard to compute the true positive and true negative rates, the smaller amount of variants detected is a strong indication of a reduction in false positives using the new reference genome.

Finally, we studied GRCh38's improvement on centromere. We obtained the coordinates for centromeres in GRCh38 from UCSC Genome Browser. The centromere's coordinates for GRCh37 is incomplete for GRCh37. We extracted all reads that mapped to GRCh38's centromere regions. Then mapped these centromere reads back to GRCh37. Between 41.5% to 53.3% of the reads were successfully mapped to GRCh37. However, the mapping locations were not necessarily centromere regions. The complete results are presented in Table S3.

3. Discussion

Reconstructing the complete human reference genome is a laborious and complicate task. There have now been 20 releases of the human reference genome by the year of 2013. Considerable improvements have been made with each release of the reference. One of the major novel technical advances GRCh38 used is the application of hydatidiform moles, which is a type of abnormal pregnancy that occurs when a sperm fertilizes an egg without nuclear DNA. The sperm then replicates its own DNA, resulting in two identical copies of each chromosome, which therefore does do not have any allelic variation and is ideal for generating unambiguous reads in highly homologous regions.

The latest release GRCh38 is of course still not a perfect representation of the human reference genome. One of the major categories of unresolved regions in GRCh38 are the centromeres, which are millions of bases long and highly repetitive. In GRCh37, the centromeres were represented as gaps. Modeled centromeres were used to fill the centromere gaps in GRCh38. The models starts by identifying alpha-satellite centromere sequences from reads. Then models were built to represent the approximate repeats for each of the alpha-satellite sequences [31]. The representation of centromeres in this way is only an approximation, but is still a huge step forward compared to GRCh37 which essentially ignored these genomic regions.

The human reference genome only represent one allele of the human genome at each genomic site. For diploid genomic regions, there are two alleles presented for any individual. Further complicating matters, the phenomenon of multi-allelic positions in an individual have been observed in nuclear regions due to copy number increases [32], and in mtDNA due to heteroplasmy [1,33]. At the population level, it is also possible to observe multi-allelic positions, especially among mixed race populations. The reference allele might not represent the major allele (the allele with allele frequency >50% within the population), because the reference allele was determined from a very small group of individuals. In some extreme cases, the population of subjects of interest might not even have the reference allele present at all. Earlier version of SNP detecting tools such as GATK only allowed two possible alleles at a genomic position. The later versions allow multiple alleles at a genomic positions which better representing the true human population.

Based on our comparative exome sequencing data analysis between GRCh37 and GRCh38, we can safely conclude that GRCh38 is an improvement over GRCh37 and these improvements resulted in more accurate genomic analysis results. One of the noticeable differences is the increase of exome size as the result of improved annotation of the exome, which can also positively impact RNAseq data analysis. From the alignment perspective, more reads were aligned and fewer reads were unaligned to GRCh38, which indicates better structural or sequence representation in GRCh38. The improvement of accuracy of GRCh38 can also be indirectly observed from the results of structural

variant analysis which identified substantially fewer variants using GRCh38 than GRCh37. In summary, GRCh38 is a big step forward from GRCh37 in term of genome structural accuracy and high throughput sequencing analysis based on GRCh38 will benefit significantly from this increase of genome structural accuracy.

4. Materials and methods

Aligned sequencing data (BAM) from thirty exome sequencing samples were downloaded from The Cancer Genome Atlas project's breast cancer consortium. The re-processing of these thirty samples (Table 1) were described in a previous study [34]. In short, alignment was performed using BWA [35]. We followed GATK's processing pipeline [36] including the steps of marking duplicate reads, realignment, and recalibration. Quality control of the data was performed at multiple stages [37] of the sequencing analysis using QC3 [38]. SNVs and INDEL were inferred using GATK's HaplotypeCaller, and filtered based on GATK's best practice recommendation, and annotation was done using ANNOVAR [39]. Copy number variations were detected using cn.MOPs [40]. Structure variants including large deletions, insertions, transversions, and inversions were detected using BreakDancer [41]. During comparison, the genomic coordinates between GRCh38 and GRCh37 were converted using the LiftOver tool provided by the USCS genome browser. A CNV or structural variant was considered to be the same between both reference genomes if they shared a starting position after the genome lift over.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2017.01.005.

Acknowledgement

This study was supported by P30 CA68485. We would also like to thank Stephanie Page Hoskins for editorial support.

References

[1] P. Zhang, D.C. Samuels, B. Lehmann, T. Stricker, J. Pietenpol, Y. Shyr, Y. Guo, Mitochondria sequence mapping strategies and practicability of mitochondria variant detection from exome and RNA sequencing data, Brief. Bioinform. 17 (2016) 224–232.
[2] Y. Guo, J. Li, C.I. Li, Y. Shyr, D.C. Samuels, MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis, Bioinformatics 29 (2013) 1210–1211.
[3] Y. Guo, Q. Cai, D.C. Samuels, F. Ye, J. Long, C.I. Li, J.F. Winther, E.J. Tawn, M. Stovall, P. Lahteenmaki, N. Malila, S. Levy, C. Shaffer, Y. Shyr, X.O. Shu, J.D. Boice Jr., The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation, Mutat. Res. 744 (2012) 154–160.
[4] F. Ye, D.C. Samuels, T. Clark, Y. Guo, High-throughput sequencing in mitochondrial DNA research, Mitochondrion 17 (2014) 157–163.
[5] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.
[6] J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J.M. Jones, I. Birol, ABySS: a parallel assembler for short read sequence data, Genome Res. 19 (2009) 1117–1123.
[7] R.Q. Li, W. Fan, G. Tian, H.M. Zhu, L. He, J. Cai, Q.F. Huang, Q.L. Cai, B. Li, Y.Q. Bai, Z.H. Zhang, Y.P. Zhang, W. Wang, J. Li, F.W. Wei, H. Li, M. Jian, J.W. Li, Z.L. Zhang, R. Nielsen, D.W. Li, W.J. Gu, Z.T. Yang, Z.L. Xuan, O.A. Ryder, F.C.C. Leung, Y. Zhou, J.J. Cao, X. Sun, Y.G. Fu, X.D. Fang, X.S. Guo, B. Wang, R. Hou, F.J. Shen, B. Mu, P.X. Ni, R.M. Lin, W.B. Qian, G.D. Wang, C. Yu, W.H. Nie, J.H. Wang, Z.G. Wu, H.Q. Liang, J.M. Min, Q. Wu, S.F. Cheng, J. Ruan, M.W. Wang, Z.B. Shi, M. Wen, B.H. Liu, X.L. Ren, H.S. Zheng, D. Dong, K. Cook, G. Shan, H. Zhang, C. Kosiol, X.Y. Xie, Z.H. Lu, H.C. Zheng, Y.R. Li, C.C. Steiner, T.T.Y. Lam, S.Y. Lin, Q.H. Zhang, G.Q. Li, J. Tian, T.M. Gong, H.D. Liu, D.J. Zhang, L. Fang, C. Ye, J.B. Zhang, W.B. Hu, A.L. Xu, Y.Y. Ren, G.J. Zhang, M.W. Bruford, Q.B. Li, L.J. Ma, Y.R. Guo, N. An, Y.J. Hu, Y. Zheng, Y.Y. Shi, Z.Q. Li, Q. Liu, Y.L. Chen, J. Zhao, N. Qu, S.C. Zhao, F. Tian, X.L. Wang, H.Y. Wang, L.Z. Xu, X. Liu, T. Vinar, Y.J. Wang, T.W. Lam, S.M. Yiu, S.P. Liu, H.M. Zhang, D.S. Li, Y. Huang, X. Wang, G.H. Yang, Z. Jiang, J.Y. Wang, N. Qin, L. Li, J.X. Li, L. Bolund, K. Kristiansen, G.K.S. Wong, M. Olson, X.Q. Zhang, S.G. Li, H.M. Yang, J. Wang, J. Wang, The sequence and de novo assembly of the giant panda genome, Nature 463 (2010) 311–317.
[8] D. Paulino, R.L. Warren, B.P. Vandervalk, A. Raymond, S.D. Jackman, I. Birol, Sealer: a scalable gap-closing application for finishing draft genomes, Bmc Bioinformatics 16 (2015) 230.
[9] M. Pop, D.S. Kosack, S.L. Salzberg, Hierarchical scaffolding with Bambus, Genome Res. 14 (2004) 149–159.
[10] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D.W. Cheung, S.M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.W. Lam, J. Wang, SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience 1 (2012) 18.
[11] R.K. Moyzis, J.M. Buckingham, L.S. Cram, M. Dani, L.L. Deaven, M.D. Jones, J. Meyne, R.L. Ratliff, J.R. Wu, A highly conserved repetitive DNA-sequence, (Ttaggg)N, present at the telomeres of human-chromosomes, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 6622–6626.
[12] K. Berlin, S. Koren, C.S. Chin, J.P. Drake, J.M. Landolin, A.M. Phillippy, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, Nat. Biotechnol. 33 (2015) 623–630.
[13] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, J.M. Boutell, J. Bryant, R.J. Carter, R.K. Cheetham, A.J. Cox, D.J. Ellis, M.R. Flatbush, N.A. Gormley, S.J. Humphray, L.J. Irving, M.S. Karbelashvili, S.M. Kirk, H. Li, X.H. Liu, K.S. Maisinger, L.J. Murray, B. Obradovic, T. Ost, M.L. Parkinson, M.R. Pratt, I.M.J. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M.T. Ross, A. Sabot, S.V. Sankar, A. Scally, G.P. Schroth, M.E. Smith, V.P. Smith, A. Spiridou, P.E. Torrance, S.S. Tzonev, E.H. Vermaas, K. Walter, X.L. Wu, L. Zhang, M.D. Alam, C. Anastasi, I.C. Aniebo, D.M.D. Bailey, I.R. Bancarz, S. Banerjee, S.G. Barbour, P.A. Baybayan, V.A. Benoit, K.F. Benson, C. Bevis, P.J. Black, A. Boodhun, J.S. Brennan, J.A. Bridgham, R.C. Brown, A.A. Brown, D.H. Buermann, A.A. Bundu, J.C. Burrows, N.P. Carter, N. Castillo, M.C.E. Catenazzi, S. Chang, R.N. Cooley, N.R. Crake, O.O. Dada, K.D. Diakoumakos, B. Dominguez-Fernandez, D.J. Earnshaw, U.C. Egbujor, D.W. Elmore, S.S. Etchin, M.R. Ewan, M. Fedurco, L.J. Fraser, K.V.F. Fajardo, W.S. Furey, D. George, K.J. Gietzen, C.P. Goddard, G.S. Golda, P.A. Granieri, D.E. Green, D.L. Gustafson, N.F. Hansen, K. Harnish, C.D. Haudenschild, N.I. Heyer, M.M. Hims, J.T. Ho, A.M. Horgan, K. Hoschler, S. Hurwitz, D.V. Ivanov, M.Q. Johnson, T. James, T.A.H. Jones, G.D. Kang, T.H. Kerelska, A.D. Kersey, I. Khrebtukova, A.P. Kindwall, Z. Kingsbury, P.I. Kokko-Gonzales, A. Kumar, M.A. Laurent, C.T. Lawley, S.E. Lee, X. Lee, A.K. Liao, J.A. Loch, M. Lok, S.J. Luo, R.M. Mammen, J.W. Martin, P.G. McCauley, P. McNitt, P. Mehta, K.W. Moon, J.W. Mullens, T. Newington, Z.M. Ning, B.L. Ng, S.M. Novo, M.J. O'Neill, M.A. Osborne, A. Osnowski, O. Ostadan, L.L. Paraschos, L. Pickering, A.C. Pike, A.C. Pike, D.C. Pinkard, D.P. Pliskin, J. Podhasky, V.J. Quijano, C. Raczy, V.H. Rae, S.R. Rawlings, A.C. Rodriguez, P.M. Roe, J. Rogers, M.C.R. Bacigalupo, N. Romanov, A. Romieu, R.K. Roth, N.J. Rourke, S.T. Ruediger, E. Rusman, R.M. Sanches-Kuiper, M.R. Schenker, J.M. Seoane, R.J. Shaw, M.K. Shiver, S.W. Short, N.L. Sizto, J.P. Sluis, M.A. Smith, J.E.S. Sohna, E.J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C.L. Tregidgo, G. Turcatti, S. van de Vondele, Y. Verhovsky, S.M. Virk, S. Wakelin, G.C. Walcott, J.W. Wang, G.J. Worsley, J.Y. Yan, L. Yau, M. Zuerlein, J. Rogers, J.C. Mullikin, M.E. Hurles, N.J. McCooke, J.S. West, F.L. Oaks, P.L. Lundberg, D. Klenerman, R. Durbin, A.J. Smith, Accurate whole human genome sequencing using reversible terminator chemistry, Nature, 456 (2008) 53–59.
[14] Y. Guo, S.L. Zhao, B.D. Lehmann, Q.H. Sheng, T.M. Shaver, T.P. Stricker, J.A. Pietenpol, Y. Shyr, Detection of internal exon deletion with exon Del, Bmc Bioinformatics 15 (2014).
[15] J. Szustakowski, I.H.G.S. Consor, Initial sequencing and analysis of the human genome (vol 409, pg 860, 2001), Nature 411 (2001) 720.
[16] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet 10 (2009) 57–63.
[17] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, Genome Res. 18 (2008) 1509–1517.
[18] Y.W. Asmann, E.W. Klee, E.A. Thompson, E.A. Perez, S. Middha, A.L. Oberg, T.M. Therneau, D.I. Smith, G.A. Poland, E.D. Wieben, J.P. Kocher, 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer, BMC Genomics 10 (2009) 531.
[19] N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, A.J. Robertson, A.C. Perkins, S.J. Bruce, C.C. Lee, S.S. Ranade, H.E. Peckham, J.M. Manning, K.J. McKernan, S.M. Grimmond, Stem cell transcriptome profiling via massive-scale mRNA sequencing, Nat. Methods 5 (2008) 613–619.
[20] Y. Guo, Q. Sheng, J. Li, F. Ye, D.C. Samuels, Y. Shyr, Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data, PLoS One 8 (2013), e71462.
[21] D.C. Samuels, L. Han, J. Li, S. Quanghu, T.A. Clark, Y. Shyr, Y. Guo, Finding the lost treasures in exome sequencing data, Trends in genetics: TIG 29 (2013) 593–599.
[22] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[23] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, J.M. Boutell, J. Bryant, R.J. Carter, R. Keira Cheetham, A.J. Cox, D.J. Ellis, M.R. Flatbush, N.A. Gormley, S.J. Humphray, L.J. Irving, M.S. Karbelashvili, S.M. Kirk, H. Li, X. Liu, K.S. Maisinger, L.J. Murray, B. Obradovic, T. Ost, M.L. Parkinson, M.R. Pratt, I.M. Rasolonjatovo, M.T. Reed, R. Rigatti, C. Rodighiero, M.T. Ross, A. Sabot, S.V. Sankar, A. Scally, G.P. Schroth, M.E. Smith, V.P. Smith, A. Spiridou, P.E. Torrance, S.S. Tzonev, E.H. Vermaas, K. Walter, X. Wu, L. Zhang, M.D. Alam, C. Anastasi, I.C. Aniebo, D.M. Bailey, I.R. Bancarz, S. Banerjee, S.G. Barbour, P.A. Baybayan, V.A. Benoit, K.F. Benson, C. Bevis, P.J. Black, A. Boodhun, J.S. Brennan, J.A. Bridgham, R.C. Brown, A.A. Brown, D.H. Buermann, A.A. Bundu, J.C. Burrows, N.P. Carter, N. Castillo, E.C.M. Chiara, S. Chang, R. Neil Cooley, N.R. Crake, O.O. Dada, K.D. Diakoumakos, B. Dominguez-Fernandez, D.J. Earnshaw, U.C. Egbujor, D.W. Elmore, S.S. Etchin, M.R. Ewan, M. Fedurco, L.J. Fraser, K.V. Fuentes Fajardo, W. Scott Furey, D. George, K.J. Gietzen, C.P. Goddard, G.S. Golda, P.A. Granieri, D.E. Green, D.L. Gustafson, N.F. Hansen, K. Harnish, C.D. Haudenschild, N.I. Heyer, M.M. Hims, J.T. Ho, A.M. Horgan, K. Hoschler, S. Hurwitz, D.V. Ivanov, M.Q. Johnson, T. James, T.A. Huw Jones, G.D. Kang, T.H. Kerelska, A.D. Kersey, I. Khrebtukova, A.P. Kindwall, Z. Kingsbury, P.I. Kokko-Gonzales, A. Kumar, M.A. Laurent, C.T. Lawley, S.E. Lee, X. Lee, A.K. Liao, J.A. Loch, M. Lok, S. Luo, R.M. Mammen, J.W. Martin, P.G. McCauley, P. McNitt, P. Mehta, K.W. Moon, J.W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S.M. Novo, M.J. O'Neill, M.A. Osborne, A. Osnowski, O. Ostadan, L.L. Paraschos, L. Pickering, A.C. Pike, D. Chris Pinkard, D.P. Pliskin, J. Podhasky, V.J. Quijano, C. Raczy, V.H. Rae, S.R. Rawlings, A. Chiva Rodriguez, P.M. Roe, J. Rogers, M.C. Rogert Bacigalupo, N. Romanov, A. Romieu, R.K. Roth, N.J. Rourke, S.T. Ruediger, E. Rusman, R.M. Sanches-Kuiper, M.R. Schenker, J.M. Seoane, R.J. Shaw, M.K. Shiver, S.W. Short, N.L. Sizto, J.P. Sluis, M.A. Smith, J. Ernest Sohna Sohna, E.J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C.L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S.M. Virk, S. Wakelin, G.C. Walcott, J. Wang, G.J. Worsley, J. Yan, L. Yau, M. Zuerlein, J.C. Mullikin, M.E. Hurles, N.J. McCooke, J.S. West, F.L. Oaks, P.L. Lundberg, D. Klenerman, R. Durbin, A.J. Smith, Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.
[24] Y. Guo, S. Zhao, B.D. Lehmann, Q. Sheng, T.M. Shaver, T.P. Stricker, J.A. Pietenpol, Y. Shyr, Detection of internal exon deletion with exon Del, Bmc Bioinformatics 15 (2014) 332.

[25] J. Wang, L. Raskin, D.C. Samuels, Y. Shyr, Y. Guo, Genome measures used for quality control are dependent on gene function and ancestry, Bioinformatics 31 (2015) 318–323.
[26] Y. Guo, J. Long, J. He, C.I. Li, Q. Cai, X.O. Shu, W. Zheng, C. Li, Exome sequencing generates high quality data in non-target regions, BMC Genomics 13 (2012) 194.
[27] E. Hazkani-Covo, R.M. Zeller, W. Martin, Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes, PLoS Genet. 6 (2010), e1000834.
[28] M. Li, R. Schroeder, A. Ko, M. Stoneking, Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs, Nucleic Acids Res. 40 (2012), e137.
[29] Y. Guo, Q. Sheng, D.C. Samuels, B. Lehmann, J.A. Bauer, J. Pietenpol, Y. Shyr, Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control, Biomed. Res. Int. 2013 (2013) 915636.
[30] H.J. Abel, E.J. Duncavage, Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches, Cancer Genet-Ny 206 (2013) 432–440.
[31] K.H. Miga, Y. Newton, M. Jain, N. Altemose, H.F. Willard, W.J. Kent, Centromere reference models for human chromosomes X and Y satellite arrays, Genome Res. 24 (2014) 697–707.
[32] R.E. Handsaker, V. Van Doren, J.R. Berman, G. Genovese, S. Kashin, L.M. Boettger, S.A. McCarroll, Large multiallelic copy number variations in humans, Nat. Genet. 47 (2015) 296–303.
[33] P. Zhang, D.C. Samuels, J. Wang, S. Zhao, Y. Shyr, Y. Guo, Mitochondria single nucleotide variation across six blood cell types, Mitochondrion 28 (2016) 16–22.
[34] Q.H. Sheng, S.L. Zhao, C.I. Li, Y. Shyr, Y. Guo, Practicability of detecting somatic point mutation from RNA high throughput sequencing data, Genomics 107 (2016) 163–169.
[35] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[36] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, M.J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) (491-+).
[37] Y. Guo, F. Ye, Q. Sheng, T. Clark, D.C. Samuels, Three-stage quality control strategies for DNA re-sequencing data, Brief. Bioinform. 15 (2014) 879–889.
[38] Y. Guo, S. Zhao, Q. Sheng, F. Ye, J. Li, B. Lehmann, J. Pietenpol, D.C. Samuels, Y. Shyr, Multi-perspective quality control of Illumina exome sequencing data using QC3, Genomics 103 (2014) 323–328.
[39] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010), e164.
[40] G. Klambauer, K. Schwarzbauer, A. Mayr, D.A. Clevert, A. Mitterecker, U. Bodenhofer, S. Hochreiter, cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate, Nucleic Acids Res. 40 (2012), e69.
[41] K. Chen, J.W. Wallis, M.D. McLellan, D.E. Larson, J.M. Kalicki, C.S. Pohl, S.D. McGrath, M.C. Wendl, Q. Zhang, D.P. Locke, X. Shi, R.S. Fulton, T.J. Ley, R.K. Wilson, L. Ding, E.R. Mardis, BreakDancer: an algorithm for high-resolution mapping of genomic structural variation, Nat. Methods 6 (2009) 677–681.