Quantitative Biology 2016, 4(2): 115–128
DOI 10.1007/s40484-016-0067-0

REVIEW

Whole genome sequencing and its applications in medical genetics

Jiaxin Wu, Mengmeng Wu, Ting Chen and Rui Jiang*

MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
* Correspondence: ruijiang@tsinghua.edu.cn

Received October 30, 2015; Revised February 4, 2016; Accepted February 4, 2016

Fundamental improvement was made for genome sequencing since the next-generation sequencing (NGS) came out in the 2000s. The newer technologies make use of the power of massively-parallel short-read DNA sequencing, genome alignment and assembly methods to digitally and rapidly search the genomes on a revolutionary scale, which enable large-scale whole genome sequencing (WGS) accessible and practical for researchers. Nowadays, whole genome sequencing is more and more prevalent in detecting the genetics of diseases, studying causative relations with cancers, making genome-level comparative analysis, reconstruction of human population history, and giving clinical implications and instructions. In this review, we first give a typical pipeline of whole genome sequencing, including the lab template preparation, sequencing, genome assembling and quality control, variants calling and annotations. We compare the difference between whole genome and whole exome sequencing (WES), and explore a wide range of applications of whole genome sequencing for both mendelian diseases and complex diseases in medical genetics. We highlight the impact of whole genome sequencing in cancer studies, regulatory variant analysis, predictive medicine and precision medicine, as well as discuss the challenges of the whole genome sequencing.

Keywords: whole genome sequencing; whole exome sequencing; next-generation sequencing; non-coding; regulatory variant

INTRODUCTION

Sequencing technology has developed swiftly and thoroughly since location-specific primer extension DNA sequencing strategy was first introduced by Ray Wu and then largely improved by Frederick Sanger in the 1970s [1]. Limited to the technology and methodology, sequencing was applied to small genomes at the first time, such as the genome of the bacteriophage and viruses [2]. In the late 1980s, the automated DNA sequencing method, usually considered as the first generation sequencing, had been successfully applied for almost two decades and achieved a serious of essential accomplishments [3,4]. In 1995, Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first paper which used the whole-genome shotgun sequencing to sequence the complete genome of a free-living organism, the bacterium *Haemophilus influenza*. By the year of 2001, shotgun sequencing methods had been widely adopted to produce the draft sequence of the large genomes, especially the monumental world-wide achievement of the initial rough draft of human genome [5,6].

Despite of the steady improvement in the first generation sequencing, it remains some fatal problems, such as cost, speed, scalability and resolution [7]. Fundamental improvement was made for genome sequencing since the next-generation sequencing (NGS) came out in the 2000s [8,9]. The newer technologies make use of the power of massively-parallel short-read DNA sequencing, genome alignment and assembly methods to digitally and rapidly search the genomes on a revolutionary scale, which enable large-scale whole genome sequencing accessible and practical for researchers [10,11]. Several NGS platforms for whole genome sequencing have emerged with high speed, comparable

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

low cost and high coverage, which makes the whole genome sequencing a more and more popular way for research. Nowadays, more than 90% of the reported complete human genome sequences are produced by the platforms of two famous companies, Illumina and Complete Genomics (CG) [12]. Large-scale comparative and evolutionary studies are then allowed by the sequencing of the whole genomes of many related organisms [13,14]. Whole genome sequencing also provides solutions to complex genomic and genetic research problems by offering the most comprehensive collection of rare variants and structural variations for sequenced individuals [15]. Recently, whole genome sequencing has been successfully applied to reconstruction of human population history [1], uncovering the roles of rare variants in common diseases [16,17], and provide clinical interpretation and implications [18–21]. Even more, it is reported that whole genome sequencing is more powerful than whole-exome sequencing for detecting exome variants [22].

This review first gives a typical pipeline of whole genome sequencing, including the lab template preparation, sequencing, genome assembling and quality control, variants calling and annotations. Then we compare the difference between whole genome and whole exome sequencing. We explore a wide range of applications of whole genome sequencing for both mendelian diseases and complex diseases in medical genetics. At last, we highlight the impact of whole genome sequencing in cancer studies, regulatory variant analysis, predictive medicine and precision medicine.

A TYPICAL PIPELINE OF WHOLE GENOME SEQUENCING

We introduce a typical pipeline of whole genome sequencing (shown in Figure 1), which is largely built based on literature [7,23]. After lab preparation, a proper sequencing platform is chosen for sequencing the samples. The next steps are genome assembling and quality control, followed by variants calling and annotations. Detected variants can be further analyzed to infer the biological relevance, prioritized or filtered according to the causative relation to a concerned phenotype. Further verification tests can be applied according to results of analysis.

High quality NGS lab preparation is an essential procedure for accurate whole genome sequencing. As this step is often outsourced to sequencing companies, for more details about the lab preparation, please see the guidebook of Preparing Samples for Sequencing Genomic DNA [24] from Illumina or instructions provided in the literature [25].

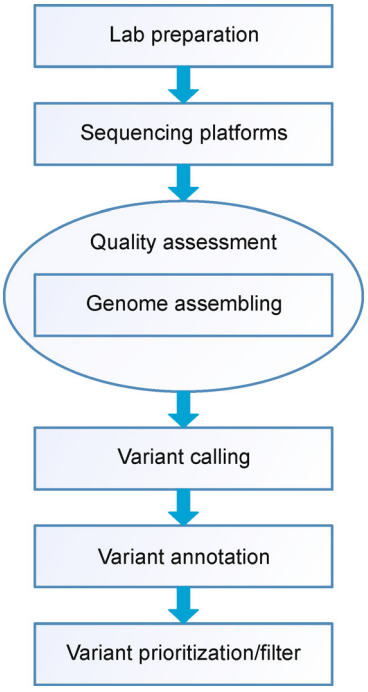

Figure 1. A typical pipeline of whole genome sequencing.

Sequencing platforms

Nowadays, many sequencing companies provide the service of whole genome sequencing, so choosing an affordable and accurate sequencing platform is also an essential step to offer reliable and wholesome sequencing outputs for further biological or bioinformatics analysis [12,26]. Table 1 shows the properties of some current sequencing platforms, which are summarized by the AllSeq Knowledge Bank [27]. Besides these platforms, Complete Genomics, the leader of whole human genome sequencing, provides high quality sequencing outputs (SNP calling rate $>90\%$ with a reference consensus accuracy of $>99.999\%$).

Alignment or genome assembling

Different next generation sequencing platform generates massive short reads of different quantity and different read length for one genome. Due to the complexity of some genomes, the most comprehensive and accurate genome assemblers are based on the pair-end reads sequenced from both ends of the DNA fragment [23]. Based on whether there exists a reference genome, two assembling approaches are dominated to integrate the short reads into longer continuous sequences after the quality assessment, and then build the draft genome

Table 1. Comparison of whole genome sequencing platforms.

Platform	Total output	Time	Read length	# of single reads	Run price
454 (Roche) GS FLX+	700 Mb	23 h	< 1 kb	1 M	~$6k
454 (Roche) GS Jr.	35 Mb	10 h	~700 bp	0.1 M	~$1k
Illumina Hiseq X Ten	1.8 Tb	3 d	2 × 150 bp	6 B	~$12k
Illumina Hiseq 2500 HT v4	1 Tb	6 d	2 × 125 bp	4 B	~$29k
Illumina Hiseq 2500 Rapid	180 Gb	40 h	2 × 150 bp	600 M	~$8k
Illumina NextSeq 500	129 Gb	29 h	2 × 150 bp	400 M	$4k
Illumina MiSeq	15 Gb	~65 h	2 × 300 bp	25 M	~$1.4k
Life Technologies SOLiD 5500xl	95 Gb	6 d	2 × 60 bp	800 M	~$10k
Life Technologies SOLiD 5500	48 Gb	6 d	2 × 60 bp	400 M	~$5k
Life Technologies Ion Torrent PI	~10 Gb	2–4 h	< 200 bp	< 82 M	> $1k

[23,28]. The first idea is reference based assembly, which is to align the reads to a reference genome and produce a similar sequence with affordable difference. As this method cannot generate novel sequences, which are different or absent from the reference, thus, sometimes it is combined with other methods to improve the accuracy of the assembling [29]. A more complex and popular approach is *de novo* genome assembly [30], which can discover new sequences or generate the draft genome whose related reference genome does not exist. The *de novo* genome assembly should be treated with the sequencing errors, repeat structures, and the computational complexity and speed of processing large amount of data. It is more challenging for *de novo* assembly to deal with shorter sequence reads [31]. Some popular alignment and genome assembling tools for reference based assembly or *de novo* assembly are listed in Table 2.

Quality assessment

Quality control is an essential step before and after reads alignment and genome assembling. Raw reads generated by the sequencing platforms may cause errors which are common and inevitable during sequencing, such as reads in bad qualities, base calling errors, small insertions or deletions [7]. Thus, quality assessment should be introduced to measure the quality of raw reads and remove, trim or correct the poor reads in order to avoid receiving wrong assembled sequences for further biological analysis. The quality assessment before the reads alignment and genome assembling usually includes plotting the quality score trend provided by the sequencing platforms; checking the primer contaminations, N content per base and GC bias; as well as trimming and filtering reads. As shown in Table 3, many tools have been

Table 2. Alignment and genome assembling tools.

Name	Method	Platform	Link
Bowtie2	Alignment	Illumina, 454	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	Alignment	Illumina, ABI SOLiD	http://bio-bwa.sourceforge.net/
SOAP3-DP	Alignment	Illumina	http://sourceforge.net/projects/soap3dp/
MAQ	Reference	Illumina, SOLiD	http://sourceforge.net/projects/maq/
RMAP	Reference	Illumina	http://rulai.cshl.edu/rmap/
SeqMan NGen	Ref/*De novo*	Illumina, SOLiD, 454, Ion Torrent, Sanger	http://www.dnastar.com/t-nextgen-seqman-ngen.aspx
ABySS	*De novo*	Illumina,SOLid	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATH S-LG	*De novo*	Illumina	http://www.broadinstitute.org/software/allpaths-lg/blog/
Edena	*De novo*	Illumina	http://www.genomic.ch/edena.php
Euler-sr	*De novo*	Sanger, 454, Illumina	http://cseweb.ucsd.edu/~ppevzner/software.html#EULER-short
Forge	*De novo*	Sanger, 454, Illumina	http://archive.is/KUoP0
Newbler	*De novo*	454	http://swes.cals.arizona.edu/maier_lab/kartchner/documentation/index.php/home/docs/newbler
SOAPdenovo	*De novo*	Illumina	http://soap.genomics.org.cn/soapdenovo.html
SPAdes	*De novo*	Illumina, PacBio	http://bioinf.spbau.ru/en/spades
SSAKE	*De novo*	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/ssake
Velvet	*De novo*	Illumina, 454	http://www.ebi.ac.uk/~zerbino/velvet/

Table 3. Quality assessment tools before the assembling.

Name	Platform	Link
FastQC	Illumina, SOLiD, 454, PacBio	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FASTX-Toolkit	Illumina	http://hannonlab.cshl.edu/fastx_toolkit/
HTQC	Illumina	http://sourceforge.net/projects/htqc/
NGSQC	Illumina, SOLiD	http://brainarray.mbni.med.umich.edu/brainarray/ngsqc/
NGS QC Toolkit	Illumina, 454	http://www.nipgr.res.in/ngsqctoolkit.html
PRINSEQ	Illumina, 454	http://prinseq.sourceforge.net/
SolexaQA	Illumina, 454	http://solexaqa.sourceforge.net/
TileQC	Illumina	http://denverlab.science.oregonstate.edu/tileqc/

designed to solve the error problems caused by different sequencing platforms.

Due to the complex genomes with large repeats, reads error and PCR duplicate generated by sequencing platforms, no genome assemblers can perfectly reconstruct the sequenced genome. Quality assessments are also suggested to control the assembled draft genomes and correct the errors which may lead to mistaking biological interpretations [28]. A variety of quality metrics are built to reflect different aspects of the assembling results, such as assembly size, contig numbers, N50 or N90 statistic (a statistic of a set of contigs or scaffold lengths), number of mismatches or mis-assemblies [32]. Some popular tools for evaluating the assemblers are collected in Table 4.

Variants calling

One prominent application of whole genome sequencing is to identify variants from the sequenced genome for further studying the genetic associations with diseases, detecting mutations in cancer, or characterizing heterogeneous cell populations [33]. The simple procedure includes at least two elements, an aligner and a variant caller. The aligner aligns the sequencing reads to a reference genome, and the variant caller assigns a genotype and identifies the positions of variants. According to the different types of variants, there are three types of variants calling tools, single nucleotide variation (SNV) calling tools (including the indels), copy number variation (CNV) calling tools, and structural variation (SV) calling tools. The detection of SNVs and indels is essential to discover the genetics of diseases and further help clinical diagnosis or treatments for patients [34]. As an important and special form of structural variation, more and more evidences indicate that CNVs play an important role in human diversity and disease susceptibility, especially in complex diseases [35]. Human genome has unexpectedly large amount of structural variations. Even if it is not clear the exact functions of most of the structural variations, they are not to be overlooked in study of human diseases and population genetics. Table 5 shows some popular variants calling tools.

Variant annotation

Variant annotation is a crucial procedure in the analysis of genome sequencing data, which provides functional information for DNA variants and give implications and evidence for biological analysis and disease studies [36]. With the dramatic increase in variant amount and complexity given by the whole genome sequencing, predicting the functional impact of variants becomes a new challenge rather than the sequencing or variant calling. There are many types of annotations ranging from the context, conservation metrics, functional genomic properties, transcript information, to the protein structural and functional predictions. Most of the variant annotation tools are available for comprehensively analyzing, prioritizing or filtering SNVs or small indels from many aspects, such as CADD [37], dbNSFP [38], GATK [39], GEMINI [40], and SPRING [41]. Although it is more complex for predicting the function of structural variants, recently some annotations tools are available to analysis structural variants, especially CNVs, including AnnTools [42], ANNOVAR [43], CNVannotator [44] and VEP [45]. For a comparable complete list of variant annotations tools and their usage hints, please see the literature [7] for details.

Table 4. Quality assessment tools for genome assemblers.

Name	Properties	Link
ALE	Reference-independent, statistical measure	http://sc932.github.io/ALE/about.html
Picard	A set of tools for processing and analyzing Illumina sequence data	http://broadinstitute.github.io/picard/index.html
QUAST	With and without a reference genome	http://bioinf.spbau.ru/quast
REAPR	Assemblers using paired end reads, without a reference genome	http://www.sanger.ac.uk/resources/software/reapr/

Table 5. Variants calling tools for detect SNVs, CNVs or SVs.

Name	Type	Link
Bambino	SNVs, indels	https://cgwb.nci.nih.gov/goldenPath/bamview/documentation/index.html
CORTEX	SNVs, indels	http://cortexassembler.sourceforge.net/index.html
GATK	SNVs, indels	https://www.broadinstitute.org/gatk/
glfTools	SNVs	http://csg.sph.umich.edu/abecasis/glfTools/
SAMtools	SNVs, indels	http://samtools.sourceforge.net/
SNVer	SNVs, indels	http://snver.sourceforge.net/
SomaticSniper	SNVs	http://gmt.genome.wustl.edu/packages/somatic-sniper/
VarScan 2	SNVs, CNVs	http://varscan.sourceforge.net/
AS-GENSENG	CNVs	http://sourceforge.net/projects/asgenseng/
cn.mops	CNVs	http://bioconductor.org/packages/2.12/bioc/html/cn.mops.html
CNV-seq	CNVs	http://tiger.dbs.nus.edu.sg/cnv-seq/
CNVrd2	CNVs	http://www.bioconductor.org/packages/devel/bioc/html/CNVrd2.html
CopySeq	CNVs	http://www.embl.de/~korbel/CopySeq/
GENSENG	CNVs	http://sourceforge.net/projects/genseng/
GROM-RD	CNVs	http://grigoriev.rutgers.edu/software/
modSaRa	CNVs	http://c2s2.yale.edu/software/modSaRa/
RDXplorer	CNVs	http://rdxplorer.sourceforge.net/
QDNAseq	CNVs	http://www.bioconductor.org/packages/release/bioc/html/QDNAseq.html
BreakDancer	SVs	http://gmt.genome.wustl.edu/packages/breakdancer/
ClipCrop	SVs	https://github.com/shinout/clipcrop
GASV	SVs	http://code.google.com/p/gasv/
Pindel	SVs	https://github.com/genome/pindel
SLOPE	SVs	http://www-genepi.med.utah.edu/suppl/SLOPE/index.html
TIGRA	SVs	http://bioinformatics.mdanderson.org/main/TIGRA
VariationHunter	SVs	http://compbio.cs.sfu.ca/software-variation-hunter

COMPARISON BETWEEN WHOLE GENOME AND WHOLE EXOME SEQUENCING

With the rapid development of sequencing technology and lower cost of each run of sequencing, whole genome sequencing is more and more prevalent in the detecting genetics of diseases, studying causative relations with cancers, making genome-level comparative analysis, and giving clinical implications and instructions [46,47]. Apparently, whole genome sequencing is superior to whole exome sequencing if there is no limitations of resources and time. Compare to the whole exome sequencing, whole genome sequencing provides examinations of SNVs, indels, CNVs and SVs in both coding (~1% part of the genome) and non-coding regions of the genome. Whole genome sequencing has more reliable and unified sequence coverage, no limitations of sequencing read length, no requirement of PCR amplification in library preparation or reference genome for assembling [46]. Whole genome sequencing possesses more advantages for sequencing a species other than human. Even more, it is reported that whole genome sequencing is more powerful than whole exome sequencing for detecting exome variants [22]. However, whole genome sequencing do suffering some problems of cost and time-consuming (see Table 6 for exact numbers), and it is more difficult to accurately interpret a huge amount and variety of detected variants [48].

WHOLE GENOME SEQUENCING FOR MENDELIAN DISEASES

Mendelian diseases refer to those disorders caused by single gene, and make up the largest proportion of human inherited diseases. According to OMIM database [49], the largest collection of Mendelian diseases, about 7,000 different diseases are characterized, of which ~3,500 disorders own unknown genetic causes. Traditional approaches to pinpoint the causal genes for Mendelian diseases are mainly based on linkage analysis [50], which measures the segregation degree between genomic regions and disease status. Those identified linked regions usually contain hundreds of candidate genes, and those

Table 6. Comparison between whole genome and whole exome sequencing.

	Whole genome sequencing	Whole exome sequencing
Time	6–8 weeks	1–3 weeks
Cost	$795–$4150/sample	$390–$1050/sample
Sequencing depth	Usually 30×	Usually > 50×
Sequenced region	Coding and non-coding regions of the genome	Exomes, promoters and enhancers

candidate genes are further validated and investigated by Sanger sequencing. Despite of its successful cases for identifying causal genes for some diseases, several drawbacks of this strategy prevent it from being widely used now. For example, linkage analysis is only effective for those familial diseases with enough sample size [51]. Whole exome sequencing has emerged as a powerful and popular approach to elucidate the genetic determinants of Mendelian diseases [52]. With acceptable cost and easy interpretation, WES has identified causal genes for many Mendelian diseases [51,53–55]. Recent evidences suggest the advantages of WGS over WES on detecting exonic variants [22] from technical perspectives. For the same task of detecting variants, including SNVs and indels in coding regions, WGS can identify more variants that are missed by WES than variants that are only captured by WES but missed by WGS. This fact makes WGS a preferable alternative to WES without consideration of cost and time-consuming. Besides coding variants, WGS provides more insights into genomic structural variants and noncoding variants. Recently, WGS has also been successfully applied to identify causal mutations in rare Mendelian diseases [56,57].

The widely used workflow for identifying disease-causing variants from exome sequencing in Mendelian diseases involves combination of biological information about genes, predicting functional consequence of variants, variant frequency in well-known large databases (e.g., 1000 G, ESP) and evolutionary conservation (e.g., GERP [58]). The rationale behind this workflow assumes that disease-causing variants for Mendelian diseases tend to be rare variants that alter protein functions on disease-related genes. Although successful applications of this strategy in some studies, it is suspected to be powerless when the available sample size is limited. Because normal individuals without phenotypes for studied diseases could also carry some such rare functional variants, thus, additional variants in other samples or statistical evidence are needed for establishing pathogenicity [59]. This problem becomes even more difficult when whole genome sequencing is applied for studying Mendelian diseases. Due to the largely increased number of variants compared with WES, the list of candidate variants that need functional follow-up or manual investigation becomes more time-consuming even if some filters are applied. Additionally, it is harder to evaluate the functional consequence of noncoding variants than coding variants since coding regions are more well-studied than noncoding part. The large number of candidate variants and interpretative difficulty for non-coding variants pose great challenges for applying WGS in clinical testing and medical research. Although hindered by such difficulties, WGS is believed to play an important role in genetics with the development of sequencing technologies and increased understanding about human genome, especially noncoding regions. For example, with the increased number of sequenced genomes, such as 1000 G Project and others, the filters based on variant frequency will become more powerful with more complete catalogue of human genetic variants. With the efforts of large constoria, such as ENCODE and Roadmap, deeper understanding about noncoding genomes and regulatory elements will enable the development of computational methods to assess regulatory impact of noncoding variants more precisely.

Functional prediction of sequence variants provides a fast assessment of deleterious effect of variants, and is widely used in sequencing based studies as filters. Many tools have been developed for analyzing variants locating in protein-coding region, such as SIFT [60], PolyPhen2 [61], but available methods for predicting functional effects of noncoding sequence variants are relatively limited. Recently, several computational methods for whole genome variants have been developed, including CADD [62], DANN [63], FATHMM-MKL [64], Funseq [65,66], SInBaD [67], deltaSVM [68], GWAVA [69]. Most of those predictors utilize machine learning approaches to discriminate harmful variants from normal variants with various genomic annotations (e.g., ENCODE) as features. Of those predictors, deltaSVM is the only one to consider cell type specificity. Training a gkm-SVM, a method for modeling DNA sequences, on cell type specific regulatory sequences and discovering corresponding sequence vocabularies, deltaSVM has the ability to evaluate the regulatory effect of sequence variants under different cell lines. It is expected that the development of computational prediction for noncoding sequence variants will be an active area of research.

Besides SNVs, noncoding CNVs are believed to play important roles in Mendelian diseases and complex diseases [15,70–73]. CNVs refer to large alterations happened in the genome, including deletions and

duplications, and are believed to cause severe consequences since large proportions of genes or regulatory elements are affected. Researchers have developed many tools for detecting both of coding CNVs [74,75] and noncoding CNVs [76]. However, how to elucidate the effect and predict the consequence of CNVs, especially noncoding CNVs, remains elusive.

WHOLE GENOME SEQUENCING FOR COMPLEX DISEASES

Common or complex diseases refer to those diseases affected by more than one gene or one variant, which makes it unsuitable for those methods used in Mendelian diseases to apply to complex diseases. Usually, the variants that contribute to disease susceptibility of complex diseases have modest effect size, thus genome-wide association studies (GWAS) are designed to study complex diseases [77], in which large cohort is required to ensure the power. However, GWAS is suspected for its rationality due to two important issues without effective solution. The first is called "missing heritability" [78], in which associated common variants only explain limited heritability, and rare variant is considered as sources for missing heritability [79]. GWAS only genotypes common variants, but WGS overcomes this limitation by sequencing all variants, including common and rare variants. The second problem arises from the existence of linkage disequilibrium (LD, defined as the non-random association of alleles at different loci), and associated variants detected in GWAS are usually not the functional variants but just in LD with true functional variants, which lead to the prosperous development of fine mapping methods [80]. General approaches for fine mapping include dense genotyping and imputation, while WGS guarantees that the real functional variants are sequenced. We demonstrate the usage of GWAS for overcoming the two important issues in detail as follows.

Association Mapping

Association mapping has been successfully applied to discover variants associated with diseases or traits of interest, and it will continue to be a powerful approach for studying complex diseases or traits in WGS setting. Recently, researcher utilized WGS to discover two loci associated with major depressive disorder [81], providing evidence to support the effectiveness of low-coverage WGS. In this study, association signals of common variants (MAF $>$ 1%) were calculated with linear mixed model [82,83], which was proved to be an effective method for association mapping and controlling population structure. Although success of WGS for association mapping is observed, several issues should be considered and handled properly in the future. Due to consideration of cost, coverage of WGS for large-scale cohorts is low, which lead to potential quality problem in variant calling and imputation. Care must be taken to ensure the quality of variants called, and methodological development is needed to improve accuracy. In addition, WGS discovers many rare variants besides common variants, therefore, how to utilize those variants and related rare variant association method, like SKAT [84], to find biologically meaningful associations pose a challenge.

Genetic architecture analysis

Understanding the genetic architecture (e.g., heritability) of complex diseases provides important insights about them. Traditional approach to study the genetic architecture of complex diseases is usually achieved with GWAS. Considering cost and efficiency, tagging SNPs in LD blocks are genotyped with genotyping platforms. Although thousands of significant loci have been discovered through GWAS, those associated SNPs only account for small proportion of variance of traits, which is also called "missing heritability". Limited ability for interpreting heritability makes research communities suspect about GWAS, and figure out that the missing heritability has become an important problem [78] in recent years. WGS has the ability to genotype each loci, thus hold the promise to figure out "missing heritability". Recently, Taylor *et al.* [85] uses WGS to study thyroid function, and identifies more heritability than previous GWAS does. Alanna *et al.* [86] studied the genetic architecture of HDL-C, (shorts for high-density lipoprotein cholesterol) with whole genome sequence data of 962 individuals, and revealed that common variants accounted for more heritability than rare variants for this complex traits, providing some insights and evidence about the argument between common and rare variants [87]. Those studies highlight the utility of association tests for rare variants [84] and linear mixed model for estimating heritability of complex traits [88]. Since high-depth whole genome sequencing is not feasible now, low-depth WGS represents the major candidate for large-scale analysis. Special care must be taken to deal with artifacts owing to low depth, and strict quality control is essential to guarantee eliminating false positives [89].

Fine mapping

GWAS has identified thousands of disease- or trait-associated common variants, which provide insights about complex diseases or traits. Considering cost, GWAS usually only tag several SNPs within a haplotype block that could be up to several thousands of base pairs in distance. Thus, associated variants are usually in LD

with real functional variants and fine-mapping is needed to uncover the real functional sites. All variants within the associated loci are required to be genotyped in typical fine-mapping studies, in which targeted sequencing or imputation based on large population data (e.g., 1000 G, HapMap) are needed [90]. The accumulation of uncertainty across those steps could undermine the identification of causal variants underlying GWAS loci. However, WGS has the ability to sequence all variants along the whole genome, thus holds the promise to solve this problem and facilitates the progress towards discovering causal variants underlying associations. Although significant differences exist between GWAS and WGS, the methodological development for refining GWAS results can also be beneficial for refining WGS results [80,91,92]. The increasing number of genotyped variants in WGS also poses a greater challenge for fine mapping than GWAS.

WHOLE GENOME SEQUENCING FOR CANCER

As an important type of complex disease, cancer is a genetic disease and accounts for many death worldwide each year. The popular platform now for analyzing cancer genome is whole exome sequencing for its acceptable cost and highly interpretation. Recently, several international groups, like TCGA and ICGC, have paid much attention to characterize multiple types of cancers by using WES, such as prostate cancer [93] and gastric cancer [94]. Somatic mutations detected from more than 1 million cancer samples are accumulated and deposited in COSMIC database, which collects the largest number of somatic mutations thus far. Although WES is the primary approach for cancer research, it focuses on protein-coding regions and ignore noncoding regions, which limit more deep understanding about cancer [95]. Several studies reveal the pathogenic impact of noncoding mutations on cancer genome, especially promoter mutations in TERT gene, which is a catalytic subunit of the enzyme telomerase and comprises the most important unit of the telomerase complex [96–98].

With the increasing number of sequenced cancer genomes, systematic analysis of large-scale cancer whole genome sequences could identify noncoding regions of interest, which are frequently mutated across different cancer types. Integrating whole genome sequence data from multiple cancer samples with regulatory annotations or expression profiles emerges as an effective approach to study somatic mutations in noncoding regions [99–101]. Fredriksson [100] proposed a method to identify the associations between regulatory regions containing somatic mutations and gene expression, and highlighted TERT promoter region with highest statistically significant association with TERT gene expression. Due to the low frequency of somatic mutations, regional association test is used, which borrows the methodology from association test for rare population variants [84,102,103]. Weinhold [99] performed three distinct analysis, including hotspot analysis, regional recurrence analysis and transcription factor analysis, to identify functionally important somatic mutations in enhancer, promoter, 5′UTR and 3′UTR. Melton [101] integrated 436 whole genome sequencing data and regulatory annotations from ENCODE to identify significantly mutated regulatory regions. All the three studies [99–101] detect TERT promoter mutations as significant mutated region across multiple cancer types.

Similar to Mendelian diseases, computational methods for identifying deleterious variants are also important for analysis of cancer genomes. However, variants that disrupt protein-coding regions or regulatory elements are not necessarily driver mutations whose effects lead to tumor progression. Although several methods have been developed specifically for cancer mutations [104], their performance is far from satisfactory, suggesting further improvement is needed.

WHOLE GENOME SEQUENCING FOR REGULATORY VARIANT ANALYSIS

Despite the variants that disrupt the 1% protein-coding regions tend to have large deleterious effect, variants in the remaining 99% noncoding regions are also believed to play important roles in human diseases. Several reviews [105,106] have discussed about regulatory variants. Mulin *et al.* [106] focused on general regulatory variants analysis, including genetic mapping, prediction, prioritization, and functional validation. Frank *et al.* [105] discussed the role that regulatory variants played in human complex traits and disease, especially the molecular nature of regulatory variants and their influence on transcriptome and proteome. Ward *et al.* [107] discussed the interpretation of noncoding variants discovered in GWAS, with focus on enrichment analysis of regulatory annotations among discovered loci. Here, we highlight the recent development, especially integrative analysis, for interpretation and systems-level analysis of regulatory variants discovered by WGS.

QTL refers to the regions of genome containing sequence variants that can affect molecular quantitative traits, such as gene expression (eQTL), chromatin accessibility (dsQTL), alterative splicing (sQTL) (see Table 7 for more details). Studies on QTL can provide insights about the molecular mechanisms by which causal variants exert their effect to affect disease status. The typical eQTL studies require two types of data to test associations between variants and gene expression. One is

Table 7. Various types of QTLs.

Name	Molecular trait	Other techniques	Ref.
eQTL	Gene expression	RNA-seq or microarray	Rockman [108]
dsQTL	Open chromatin	DNaseI-seq	Degner [109]
sQTL	RNA splicing	RNA-seq	Monlong [110]
rtQTL	DNA replication timing	FACS-sorting	Amnon [111]
haQTL	Histone acetylation	ChIP-seq	Rosario [112]
metQTL	DNA methylation	Bisulfite-seq	Gibbs [113]

the genotype of a recruited individual, which is often obtained through genotyping array, and the other is gene expression, which is often measured by microarray and RNA sequencing. The genotyping array provides a cost-effective solution to obtain genotypes, while this method only assays the certain loci, which probably result in missing hits for those stronger associations between ungenotyped loci and gene expression. Whole genome sequencing overcomes this problem through discovering all sequence variants and allowing identification all possible genetic associations between sequence variants and gene expression. For example, a recent WGS based eQTL mapping [114] found that indels (short insertions and deletions) may play a more important role in cis-eQTL than SNPs. This study fully sequenced 462 individuals and discovered all types of sequence variants, so it provided more insights than traditional eQTL studies.

Integration of variants discovered by WGS and functional annotations tend to be a promising approach for dissecting regulatory variants. Although most attention is paid on analyzing disease-associated loci discovered by GWAS with integration of regulatory annotations, the similar idea or methodology can be also applied in WGS settings. Since WGS has the ability to discover every variant along the genome, it is expected to find more associations than GWAS, which poses greater challenge for integrative analysis. Recent studies find that disease or trait associated variants are enriched in DHS regions [115,116], and these regions could be used for marking regulatory elements with functional potential. Such annotations will help to elucidate molecular mechanisms underlying disease etiology and refine mapping of associated variants. Several studies also reveal the importance of TF binding in etiology of disease and disease-associated variants may contribute to the pathogenesis through disrupting the TF binding, such as PolII and NFkB [116–118].

Possible computational issues

With the increasing number of sequenced genomes, several computational issues need to be considered in order to facilitate the application of WGS to studies of diseases, gene regulation, and genomics etc. The first issue is the speed of data processing of WGS data. It usually takes long time to perform read mapping and variant calling for WGS data, and this issue becomes more severe when the number of samples is large. Recently, an ultra-fast WGS pipline called SpeedSeq [119] is developed, which greatly speed up the data-processing procedure. How to further speed up the process and guarantee the accuracy at the same time will be an important computational issue that needs to be solved. The second issue is how to quantify the impact of variants detected from WGS on disease or trait of interest. We have reviewed several methods for this task on different scenarios, like Mendelian diseases, complex diseases, cancers and regulatory variants. The common strategy underlying those methods is integration of WGS data with information obtained from other sequencing technology, like ChIP-seq [120], DNase-seq [121], RNA-seq [117] and ATAC-seq [122]. How to integrate those genomics data into WGS will be an important field of research.

WHOLE GENOME SEQUENCING FOR PREDICTIVE MEDICINE AND PRECISION MEDICINE

Benefit from the high-throughput sequencing technologies with high speed and low cost, personal whole genome sequencing or whole exome sequencing becomes more and more available for customers. The genotype of a person can be achieved from the sequencing data, and compared to known disease databases or related published literature to determine likelihood of trait expression and the risk of some diseases. Our research group developed a database of human whole-genome single nucleotide variants and their functional predictions, namely dbWGFP [123]. This database contains functional predictions and annotations of nearly 8.58 billion possible human whole-genome single nucleotide variants, with each of them described by 48 functional predictions and 44 valuable annotations. Specifically, the 48 prediction scores include 32 functional predictions calculated by 13 popular computational methods, 15 conservation features

derived from 4 conservation calculation approaches, and 1 sensitivity measurement. The 44 annotations are obtained from the ENCODE project. dbWGFP is helpful in the capture of causative variants from massive candidate variants derived from whole-genome or whole-exome sequencing data.

Predictive medicine is a field of medicine which may take advantage of genetic information generated from personal whole genome sequencing to predict the probability of disease and what medical treatments are appropriate for a particular individual [124–126]. Precision medicine is a medical model that formulates personalized healthcare, including disease prevention, medical decisions and therapies [127,128]. Example of application of predictive and precision medicine includes selecting appropriate drugs for a patient to maximize the effect of drugs and minimize the side effects, or giving a tailor therapy to a patient to accelerate the recovery [129,130].

CONCLUSION AND DISCUSSION

As cost of the whole genome sequencing decreases rapidly and approaches $1000, WGS are increasingly used for revealing the genetic basis of Mendelian or complex diseases, explicating novel disease biology, helping clinical diagnosis and treatment. Whole genome sequencing provides exceptional coverage of genomic regions, including exonic, intronic and other unexplored noncoding regions, and a large collection of rare variants and comprehensive structural variants. Associated with other type of data and annotations, WGS also successfully helps to interpret the genetics and biology underlying the cancer genome. In the future of predictive medicine and precision medicine, WGS will be an important tool to guide therapeutic prevention and treatment.

Although the introduction of WGS has successfully applied in many researches, there exist some problems to be solved in the future. Next-generation sequencing technologies can generate tremendous amounts of data, in the mean while they are suffering from the sequencing errors, such as bias of GC/AT rich genomes and context specific error. The amplification, which is a necessary step for some platforms, may also bring errors. In addition, most of the WGS studies could not provide sufficient coverage, which may lead to some mistakes by genome assembling and variant calling steps. Furthermore, different sequencing platform may provide different analysis results, especially for potential loss-of-function mutations, or rare variants which are likely to be pathogenic [47]. Even if the cost of whole genome sequencing of a sample has dropped dramatically, the sequencing of a comparable large number of samples with high coverage is still unaffordable for most of researchers.

The main challenge in WGS studies is the processing and interpreting whole genome sequencing data. Even if introducing the step of quality control, there still exist errors in the process of genome assembling, such as insufficient read coverage or mis-assembly [131]. Another more important step is to interpret the sequencing data, discover the relationship from genotype to phenotype, and link the analyzed data to clinically applications [132]. The volume of information contained in a genome sequence is so vast that it is hard to wholesomely and accurately explain all the hidden knowledge. The role of most of variants, genes and non-coding factors in the human genomes is still unclear or incompletely known [133,134]. Although a lot of bioinformatics approaches have been developed to deal with the sequencing data for different applications, most of the predicted or examined results remain to be testified. The pathogenic mechanisms for some diseases, such as cancers, are so complex, that they require the analysis of much more WGS data in a larger sample set and combining with other data, such as multi-omics data, functional data and clinicpathological data [95].

ACKNOWLEDGEMENTS

This research was partially supported by the National Basic Research Program of China (No. 2012CB316504), the National High Technology Research and Development Program of China (No. 2012AA020401), the National Natural Science Foundation of China (Nos. 61573207 and 61175002) and Beijing Collaborative Innovation Center for Cardiovascular Disorders.

COMPLIANCE WITH ETHICS GUIDELINES

All authors confirm the absence of previous similar or simultaneous publications, their inspection of the manuscript, their substantial contribution to the work, and their agreement to submission.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Veeramah, K. R. and Hammer, M. F. (2014) The impact of whole-genome sequencing on the reconstruction of human population history. Nat. Rev. Genet., 15, 149–162
2. Hendrix, R. W. (2003) Bacteriophage genomics. Curr. Opin. Microbiol., 6, 506–511
3. Metzker, M. L. (2010) Sequencing technologies—the next generation. Nat. Rev. Genet., 11, 31–46
4. Zimmermann, J., Voss, H., Schwager, C., Stegemann, J. and Ansorge, W. (1988) Automated Sanger dideoxy sequencing reaction protocol. FEBS Lett., 233, 432–436
5. Watson, J. D. (1990) The human genome project: past, present, and future. Science, 248, 44–49
6. Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science, 269, 496–512
7. Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M.,

Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. and Trajanoski, Z. (2014) A surveyof tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform., 15, 256–278

8. Liu, L., Li, Y. H., Li, S. L., Hu, N., He, Y. M., Pong, R., Lin, D. N., Lu, L. H. and Law, M. (2012) Comparison of next-generation sequencing systems. J. BioMed. Biotech., 251364
9. Voelkerding, K. V., Dames, S. A. and Durtschi, J. D. (2009) Next-generation sequencing: from basic research to diagnostics. Clin. Chem., 55, 641–658
10. Ng, P. C. and Kirkness, E. F. (2010) Whole Genome Sequencing. In Genetic Variation, pp. 215–226, Springer
11. Hurd, P. J. and Nelson, C. J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief. Funct. Genomics, 8, 174– 183
12. Lam, H. Y., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., *et al.* (2012) Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol., 30, 78–82
13. Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., *et al.* (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature, 419, 512–519
14. Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R., *et al.* (2006) Comparative genome sequencing of *Escherichia coli* allows observationof bacterial evolution on a laboratory timescale. Nat. Genet., 38, 1406–1412
15. Lupski, J. R. (2015) Structural variation mutagenesis of the human genome: impact on disease and evolution. Environ. Mol. Mutagen., 56, 419–436
16. Saunders, C. J., Miller, N. A., Soden, S. E., Dinwiddie, D. L., Noll, A., Alnadi, N. A., Andraws, N., Patterson, M. L., Krivohlavek L. A., Fellis, J., *et al.* (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci. Trans. Med. 4, 154ra135
17. Cirulli, E. T. and Goldstein, D. B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet., 11, 415–425
18. Foley, S. B., Rios, J. J., Mgbemena, V. E., Robinson, L. S., Hampel, H. L., Toland, A. E., Durham, L. and Ross, T. S. (2015) Use of whole genome sequencing for diagnosis and discovery in the cancer genetics clinic. EBioMedicine, 2, 74–81
19. Chen, K. and Meric-Bernstam, F. (2015) Whole genome sequencing in cancer clinics. EBioMedicine, 2, 15–16
20. Berg, J. S., Khoury, M. J. and Evans, J. P. (2011) Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. Genet.Med., 13, 499–504
21. Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., Merker, J. D., Goldfeder, R. L., Enns, G. M., David, S. P., *et al.* (2014) Clinical interpretation and implications of whole-genome sequencing. JAMA, 311, 1035–1045
22. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L. and Abel, L. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. USA, 112, 5473–5478
23. Ekblom, R. and Wolf, J. B. (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl., 7, 1026–1042
24. https://support.illumina.com/downloads/genomic_dna_sample_-prep_guide_1003806.html
25. van Dijk, E. L., Jaszczyszyn, Y. and Thermes, C. (2014) Library preparation methods for next-generation sequencing: tone down the bias. Exp. Cell Res., 322, 12–20
26. Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., Iida, T., Yasunaga, T., Horii, T., Arakawa, K., *et al.*(2014) Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics, 15, 699
27. http://allseq.com/knowledgebank
28. Bao, S., Jiang, R., Kwan, W. K., Wang, B. B., Ma, X. and Song, Y.-Q. (2011) Evaluation of next-generation sequencing software in mapping and assembly. J. Hum. Genet., 56, 406–414
29. Swindell, S. R. and Plasterer, T. N. (1997) SEQMAN. In Sequence Data Analysis Guidebook, pp. 75–89, New York: Springer
30. Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. and Batzoglou, S. (2007) Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS One, 2, e484
31. Paszkiewicz, K. and Studholme, D. J. (2010) *De novo* assembly of short sequence reads. Brief. Bioinform., 11, 457–472
32. Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M. and Otto, T. D. (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol., 14, R47
33. Liu, X., Han, S., Wang, Z., Gelernter, J. and Yang, B.-Z. (2013) Variant callers for next-generation sequencing data: a comparison study. PLoS One, 8, e75619
34. Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B. and Müller-Myhsok, B. (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum. Genet., 131, 1541–1554
35. Pirooznia, M., Goes, F. S. and Zandi, P. P (2015) Whole-genome CNV analysis: advances in computational approaches. Front.Genet., 6, 138
36. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet., 43, 491–498
37. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46, 310–315
38. Liu, X., Jian, X. and Boerwinkle, E. (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat., 32, 894–899
39. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencingdata. Genome Res., 20, 1297–1303
40. Paila, U., Chapman, B. A., Kirchner, R. and Quinlan, A. R. (2013) GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput. Biol., 9, e1003153
41. Wu, J., Li, Y. and Jiang, R. (2014) Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. PLoS Genet., 10, e1004237

42. Makarov, V., O'Grady, T., Cai, G., Lihm, J., Buxbaum, J. D. and Yoon, S. (2012) AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics, 28, 724–725
43. Wang, K., Li, M. and Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164
44. Zhao, M. and Zhao, Z. (2013) CNVannotator: a comprehensive annotation server for copy number variation in the human genome. 8, e80170
45. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics, 26, 2069–2070
46. Bick, D. and Dimmock, D. (2011) Whole exome and whole genome sequencing. Curr. Opin. Pediatr., 23, 594–600
47. Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A. Jr and Hewlett, E. L. (2015) Whole-genome sequencing in outbreak analysis. Clin. Microbiol. Rev., 28, 541–563
48. http://blog.genohub.com/whole-genome-sequencing-wgs-vs-whole-exome-sequencing-wes/
49. Online Mendelian Inheritance in Man. Johns Hopkins University (Baltimore,MD), http://omim.org/
50. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: pastsuccesses for mendelian disease, future approaches for complex disease. Nat. Genet., 33, 228–237
51. Ku, C.-S., Naidoo, N. and Pawitan, Y. (2011) Revisiting Mendelian disorders through exome sequencing. Hum. Genet., 129, 351–370
52. Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. and Shendure, J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet., 12, 745–755
53. Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet., 42, 30–35
54. Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., Braxton, A., Beuten, J., Xia, F., Niu, Z., *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med., 369, 1502–1511
55. Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H.I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C., *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet., 42, 790–793
56. Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science, 328, 636–639
57. Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., *et al.* (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. N. Engl. J. Med., 362, 1181–1191
58. Cooper, G. M., *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res., 15, 901–913
59. Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., *et al.* (2015) Factors influencing successof clinical genome sequencing across a broad spectrum of disorders. Nat. Genet., 47, 717–726
60. Kumar, P., Henikoff, S. and Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc., 4, 1073–1081
61. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. Nat. Methods, 7, 248–249
62. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46, 310–315
63. Quang, D., Chen, Y. and Xie, X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, 31, 761–763
64. Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., Gaunt, T. R. and Campbell C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, 10.1093/bioinformatics/btv009
65. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., Khurana, E. and Gerstein, M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol., 15, 480
66. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science, 342, 1235587
67. Lehmann, K.-V. and Chen, T. (2013) Exploring functional variant discovery in non-coding regions with SInBaD. Nucleic Acids Res., 41, e7
68. Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S. and Beer, M. A. (2015) A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet., 47, 955–961
69. Ritchie, G. R., Dunham, I., Zeggini, E. and Flicek, P. (2014) Functional annotation of noncoding sequence variants. Nat. Methods, 11, 294–296
70. Zhang, F. and Lupski, J. R. (2015) Noncoding genetic variants in human disease. Hum. Mol. Genet.,
71. Lupski, J. R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet., 14, 417–422
72. Lee, C. and Morton, C. C. (2008) Structural genomic variation and personalized medicine. N. Engl. J. Med., 358, 740–741
73. Lupski, J. R. (2009) Genomic disorders ten years on. Genome Med., 1, 42
74. Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J. and Nelson, S. F. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics, 27, 2648–2654
75. Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., Handsaker, R. E., McCarroll, S. A., O'Donovan, M. C., Owen, M. J., *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet., 91, 597–607
76. Zong, C., Lu, S., Chapman, A. R. and Xie, X. S. (2012) Genome-wide detectionof single-nucleotide and copy-number variations of a single humancell. Science, 338, 1622–1626
77. Cardon, L. R. and Bell, J. I. (2001) Association study designs for

complex diseases. Nat. Rev. Genet., 2, 91–99

78. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* (2009) Finding the missing heritability of complex diseases. Nature, 461, 747–753
79. Schork, N. J., Murray, S. S., Frazer, K. A. and Topol, E. J. (2009) Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev., 19, 212–219
80. Spain, S. L. and Barrett, J. C. (2015) Strategies for fine-mapping complex traits. Hum. Mol. Genet.,
81. CONVERGE consortium (2015) Sparse whole-genome sequencing identifiestwo loci for major depressive disorder. Nature, 523, 588–591.
82. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. and Heckerman, D. (2011) FaST linear mixed models for genome-wide association studies. Nat. Methods, 8, 833–835
83. Widmer, C., Lippert, C., Weissbrod, O., Fusi, N., Kadie, C., Davidson, R., Listgarten, J. and Heckerman, D. (2014) Further improvements to linear mixed models for genome-wide association studies. Sci. Rep., 4, 6874
84. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., 89, 82–93
85. Taylor, P. N., Porcu, E.,Chew, S., Campbell, P. J., Traglia, M., Brown, S. J., Mullin, B. H., Shihab, H. A., Min, J., Walter, K., *et al.* (2015) Whole-genome sequence-based analysis of thyroid function. Nat. Commun., 6, 5681
86. Morrison, A. C., Voorman, A., Johnson, A. D., Liu, X. M., Yu, J., Li, A.,Muzny, D.,Yu, F. L., Rice, K., Zhu, C. S., *et al.* (2013) Whole genome sequence-based analysis of a model complex trait, high density lipoprotein cholesterol. Nat. Genet., 45, 899–901
87. Gibson, G. (2012) Rare and common variants: twenty arguments. Nat. Rev. Genet., 13, 135–145
88. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42, 565–569
89. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. and Abecasis, G. R. (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res., 21, 940–951
90. Edwards, S. L., Beesley, J., French, J. D. and Dunning, A. M. (2013) Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet., 93, 779–797
91. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature, 518, 337–343
92. Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., *et al.* (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat. Genet., 44, 1294–1301
93. Barbieri, C. E., Baca, S. C., Lawrence, M. S., Demichelis, F., Blattner, M., Theurillat, J.-P., White, T. A., Stojanov, P., Van Allen, E., Stransky, N., *et al.* (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet., 44, 685–689
94. Wang, K., Kan, J., Yuen, S. T., Shi, S. T., Chu, K. M., Law, S., Chan, T. L., Kan, Z., Chan, A. S. Y., Tsui, W. Y., *et al.* (2011) Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat. Genet., 43, 1219–1223
95. Nakagawa, H., Wardell, C. P., Furuta, M., Taniguchi, H. and Fujimoto, A. (2015) Cancer whole-genome sequencing: present and future. Oncogene, 34, 5943–5950
96. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. and Garraway, L. A. (2013) Highly recurrent TERT promoter mutations in human melanoma. Science, 339, 957–959
97. Vinagre, J.,Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., *et al.* (2013) Frequency of *TERT* promoter mutations in human cancers. Nat. Commun., 4, 2185
98. Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., *et al.* (2014) An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science, 346, 1373–1377
99. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. and Lee, W. (2014) Genome-wide analysis of noncoding regulatory mutationsin cancer. Nat. Genet., 46, 1160–1165
100. Fredriksson, N. J., Ny, L., Nilsson, J. A. and Larsson, E. (2014) Systematic analysisof noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat. Genet., 46, 1258–1263
101. Melton, C., Reuter, J. A., Spacek, D. V. and Snyder, M. (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat. Genet., 47, 710–716
102. Li, B. and Leal, S. M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet., 83, 311–321
103. Lin, D.-Y. and Tang, Z.-Z. (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet., 89, 354–367
104. Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. Nat. Methods, 10, 723–729
105. Albert, F. W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. Nat. Rev. Genet., 16, 197–212
106. Li, M. J., Yan, B., Sham, P. C., and Wang, J.W. (2014) Exploring the function of genetic variants in the non-codinggenomic regions: approaches for identifying human regulatory variants affecting gene expression. Brief. Bioinform., 16, 393–412
107. Ward, L. D. and Kellis, M. (2012) Interpreting noncoding genetic variation in complex traitsand human disease. Nat. Biotechnol., 30, 1095–1106
108. Rockman, M. V. and Kruglyak, L. (2006) Genetics of global gene expression. Nat. Rev. Genet., 7, 862–872
109. Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., *et al.* (2012) DNase I sensitivity QTLs area major determinant of human expression variation. Nature, 482, 390–394
110. Monlong, J., Calvo, M.,Ferreira, P. G. and Guigó, R. (2014) Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat. Commun., 5, 4698
111. Koren, A., Handsaker, R. E., Kamitaki, N., Karlić, R., Ghosh, S.,

Polak, P., Eggan, K. and McCarroll, S. A. (2014) Genetic variation in human DNA replication timing. Cell, 159, 1015–1026

112. del Rosario, R. C.-H., Poschmann, J., Rouam, S. L., Png, E., Khor, C. C., Hibberd, M. L. and Prabhakar, S. (2015) Sensitive detection of chromatin-altering polymorphisms reveals autoimmune diseasemechanisms. Nat. Methods, 12, 458–464
113. Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet., 6, e1000952
114. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.*, (2013) Transcriptome and genome sequencing uncovers functional variationin humans. Nature, 501, 506–511
115. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science, 337, 1190–1195
116. Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R. and Stamatoyannopoulos, J. A. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*. Nat. Genet., 47, 1393–1401
117. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, Lukas., Rozowsky, J., Shi, M., Urban, A. E., *et al.* (2010) Variation in transcription factor binding among humans. Science 328, 232–235
118. Karczewski, K. J., Dudley, J. T., Kukurba, K. R., Chen, R., Butte, A. J., Montgomery, S. B. and Snyder, M. (2013) Systematic functional regulatory assessment of disease-associated variants. Proc. Natl. Acad. Sci. USA, 110, 9607–9612
119. Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R. and Hall, I. M. (2014) SpeedSeq: Ultra-fast personal genome analysis and interpretation. Nat. Meth., 12, 966–968
120. Park, P. J. (2009) ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680
121. Song, L. and Crawford, G. E. (2010) DNase-seq: a high-resolution technique for mapping activegene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. pdb. prot5384
122. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods, 10, 1213–1218
123. dbWGFP: http://bioinfo.au.tsinghua.edu.cn/dbwgfp
124. Biesecker, L. G. (2013) Hypothesis-generating research and predictive medicine. Genome Res., 23, 1051–1053
125. Simon, R. (2011) Genomic biomarkers in predictive medicine. An interim analysis. EMBO Mol. Med., 3, 429–435
126. Matsui, S., Simon, R., Qu, P., Shaughnessy, J. D., Barlogie, B. and Crowley, J. (2012) Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. Clin. Cancer Res., 18, 6065–6073
127. Collins, F. S. and Varmus, H. (2015) A new initiative on precision medicine. N. Engl. J. Med., 372, 793–795
128. Rubin, M. A. (2015) Health: Make precision medicine work for cancer care. Nature, 520, 290–291
129. Bellmunt, J., Orsola, A. and Sonpavde, G. (2015) Precision and predictive medicine in urothelial cancer: Are we making progress? Eur. Urol., 68, 547–549
130. Geschwind, D. H. and State, M. W. (2015) Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol., 14, 1109–1120
131. Mak, H. C. (2012) Genome interpretation and assembly—recent progress and next steps. Nat. Biotechnol., 30, 1081–1083
132. Shendure, J. and Aiden, E. L. (2012) The expanding scope of DNA sequencing. Nat. Biotechnol., 30, 1084–1094
133. Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K., Abe, T., Boroevich, K. A., Nagasaki, M., Yamaguchi, R., Shibuya, T., Kubo, M., *et al.* (2010) Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat. Genet., 42, 931–936
134. Gonzaga-Jauregui, C., Lupski, J. R. and Gibbs, R. A. (2012) Human genome sequencing in health and disease. Annu. Rev. Med., 63, 35–61

