Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing

Riyue Bao[1,*], Lei Huang[1,*], Jorge Andrade[1,*], Wei Tan[2,*], Warren A. Kibbe[3], Hongmei Jiang[4,§] and Gang Feng[3,§]

[1]Center for Research Informatics, The University of Chicago, Chicago, IL, USA. [2]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. [3]Biomedical Informatics Center (NUBIC), Clinical and Translational Sciences Institute (NUCATS), Northwestern University, Chicago, IL, USA. [4]Department of Statistics, Northwestern University, Evanston, IL, USA. *These authors contributed equally to this work. §Co-corresponding authors.

ABSTRACT: The advent of next-generation sequencing technologies has greatly promoted advances in the study of human diseases at the genomic, transcriptomic, and epigenetic levels. Exome sequencing, where the coding region of the genome is captured and sequenced at a deep level, has proven to be a cost-effective method to detect disease-causing variants and discover gene targets. In this review, we outline the general framework of whole exome sequence data analysis. We focus on established bioinformatics tools and applications that support five analytical steps: raw data quality assessment, preprocessing, alignment, post-processing, and variant analysis (detection, annotation, and prioritization). We evaluate the performance of open-source alignment programs and variant calling tools using simulated and benchmark datasets, and highlight the challenges posed by the lack of concordance among variant detection tools. Based on these results, we recommend adopting multiple tools and resources to reduce false positives and increase the sensitivity of variant calling. In addition, we briefly discuss the current status and solutions for big data management, analysis, and summarization in the field of bioinformatics.

KEYWORDS: big data, InDel, next generation sequencing, sequence alignment, SNV, whole exome sequencing, variant analysis

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Bao et al. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Cancer Informatics* 2014:13(S2) 67–82 doi: 10.4137/CIN.S13779.

RECEIVED: April 22, 2014. **RESUBMITTED:** July 6, 2014. **ACCEPTED FOR PUBLICATION:** July 7, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: RB, LH and JA were supported by Biological Sciences Division, the Institute for Translational Medicine/CTSA (NIH UL1 RR024999) at the University of Chicago. HJ's works was partially supported by NSF DMS-1043080, NSF DMS-1222592, and CBC Catalyst Award # C-024. The contributions of GF are partially supported by the NCI Cancer Center Support Grant (NCI CA060553) and the National Center for Advancing Translational Sciences Grant (8UL1TR000150). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: g-feng@northwestern.edu and hongmei@northwestern.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

High-throughput "-omics" technologies, such as microarray, sequencing, and quantitative real-time polymerase chain reaction (PCR), have been widely applied in basic biological and biomedical research for more than a decade. Advances in these techniques have enabled a broad spectrum of applications in genomic, transcriptomic, proteomic, metagenomic, and metabolomic studies.[1–8]

Two important challenges accompany these technologies: (1) How do we best manage the enormous amount of "-omics" data? (2) What are the most appropriate choices among the available computational methods and analysis tools[9]? In this review, we first present an overview of next-generation sequencing (NGS) technologies, highlight some of the issues associated with NGS data analysis, and then survey the current bioinformatics strategies and computational tools for whole exome variant detection. Based on our evaluation of the major analysis tools, we present our recommendations for the selection of variant analysis tools for specific research tasks. We will

also discuss challenges in large-scale NGS data analysis and management.

First-generation sequencing such as Sanger and shotgun sequencing techniques were successfully employed for the creation of the human genome project (HGP). The HGP required 3.4 billion USD, 13 years, and collaboration of hundreds of international labs to complete the first human genome assembly. Since the introduction of NGS technologies in 2007, the price for sequencing a whole human genome has dropped rapidly. In early 2014, Illumina Inc. introduced the HiSeq X platform into the market, claiming a 1,000 USD personal genome price (experimental cost). With this new machine, 16 human genomes can be sequenced in 3 days at a depth of 30×.[10] Several new initiatives have been established to take advantage of this dramatic price drop to sequence thousands of human genomes. For instance, the Genome England project will sequence 100,000 genomes by 2017 and attach it to the medical record data as part of the UK medical system. In early March, Human Longevity Inc., founded by J. Craig Venter and two other partners, announced that it will build the world largest human genome sequencing center with the capacity of sequencing up to 40,000 human genomes per year.[11] The drop in per-base sequencing price is expected to drive the generation of immense amount of NGS data, creating a big data challenge in bioinformatics.

NGS sequencing experiments produce millions to billions of short sequence reads at a high speed. Current NGS platforms including Illumina, Ion Torrent/Life Technologies, Pacific Bioscience, Nanopore, and GenapSys can generate reads of 100–10,000 bases long,[12,13] allowing better coverage of the genome at lower cost. However, these platforms also generate huge amounts of raw data. For example, the raw data produced by Illumina HiSeq2500 platform add up to 1TB per run, covering 150–180 human whole exome sequencing (WES) samples at a depth of 50× or higher (Illumina Inc.). For tumor samples, a sequencing depth of at least 125× is recommended accounting for intra-tumor heterogeneity.[14,15]

Common NGS applications include DNA-seq, RNA-seq, ChIP-seq, and methyl-seq. DNA-seq can be applied to the whole genome sequencing (WGS), WES, or a specific targeted region of the genome. In general, the goal of DNA sequencing is to discover genomic variations in the form of single nucleotide variants (SNVs), small DNA insertions or deletions (indels), copy number variations (CNVs), or other structural variants (SVs), with the ultimate goal of associating those variations to human disease. RNA-seq that measures gene expression changes can be used to discover new transcripts including noncoding RNAs and detect transcript splicing or gene fusion events. ChIP-seq detects genome-wide transcription factor binding sites and chromatin-associated modifications. Methyl-seq is used to profile various types of DNA methylation such as 5-methylcytosine and 5-hydroxymethylcytosine at single nucleotide resolution. In this review, we focus primarily on WES techniques and data analysis. This evaluation of established bioinformatics tools covers the variant analysis workflow from the quality control (QC) of raw reads to prioritization of biologically meaningful or clinically relevant variants. We also explore current solutions for the management of big data and its applications in bioinformatics.

A typical workflow of WES analysis consists of the following steps: raw data QC, preprocessing, mapping, post-alignment processing, variant calling, annotation, and prioritization (Fig. 1).

Raw data QC and Preprocessing

FASTQ and FASTA are standard formats for representing biological sequence data. The FASTA format is a text-based representation of sequences, which begins with the sequence name followed by lines of single-letter coded nucleotides or amino acids. FASTQ format was developed to incorporate the Phred-scaled base quality scores to facilitate the assessment of sequence quality. It is widely accepted as the standard file format for NGS raw data.

Several tools have been developed to assess the quality of raw NGS data. Some of the commonly used ones include FastQC,[16] FastQ Screen,[17] FASTX-Toolkit,[18] NGS QC Toolkit,[19] PRINSEQ,[20] QC-Chain,[21] and recently published QC3.[22] FastQC is a java application that generates many useful data diagnosis and plots such as Phred score distribution along the reads, GC content distribution, read length distribution, and sequence duplication level. It also detects over-represented sequences that may be an indication of primer or adaptor contamination. With a comprehensive raw reads QC report generated by FastQC, researchers are able to determine whether any preprocessing steps such as base trimming, read filtering, or adaptor clipping are necessary prior to alignment.

Standard preprocessing procedure includes 3′ end adapter removal and trimming of low quality bases at the ends of the reads. Depending on the study design and use of the data, redundant reads and undesired sequences such as contamination from primers, adaptors, or other species may be removed at this point. Several tools are available to perform those tasks, such as Cutadapt[23] and Trimmomatic.[24] PRINSEQ[20] and QC3,[22] on the other hand, provide both QC and preprocessing functions as a suite. In addition to generic preprocessing functions as listed above, each program is equipped with their own custom features. For example, Cutadapt allows detection of adaptor sequences anywhere in the read and performs clipping afterwards. Trimmomatic is a java application that provides useful functions for handling paired-end reads. QC3 offers multi-perspective evaluation of data quality from the raw reads, mapped reads, and detected variants, with the unique feature to detect batch effects and cross-contamination.

Sequence Alignment

After raw data QC and preprocessing, the next step is to map the reads to the reference genome and with high efficiency and accuracy. Alignment mapping is a classical "string match"

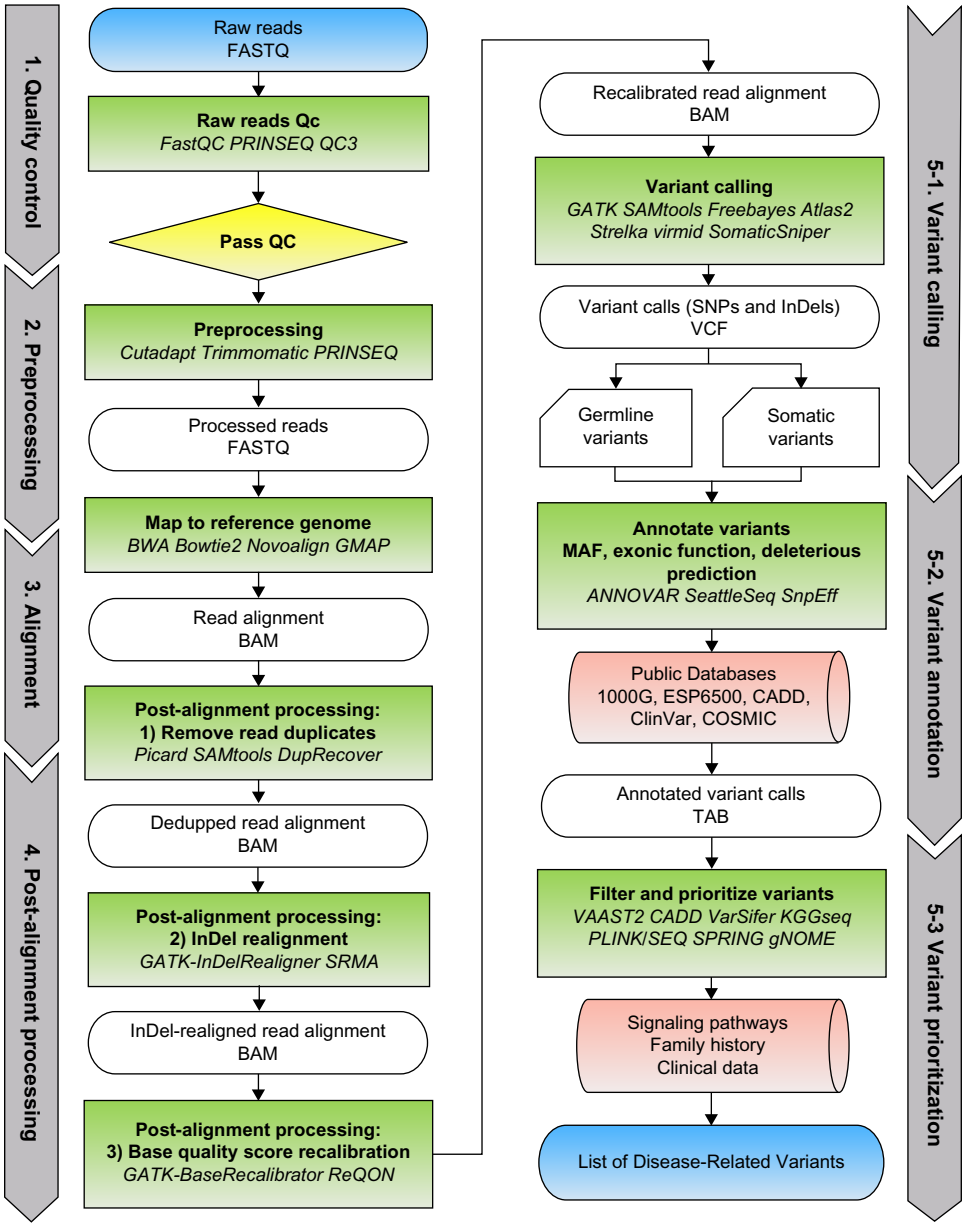

Figure 1. A general framework of WES data analysis. Five major steps are shown: raw reads QC, preprocessing, alignment, post-processing, and variant analysis (variant calling, annotation, and prioritization).

Notes: FASTQ, BAM, variant call format (VCF), and TAB (tab-delimited) refer to the standard file format of raw data, alignment, variant calls, and annotated variants, respectively. A selection of tools supporting each analysis step is shown in italic.

task in computer science. For example, most web browsers and text editors provide a "Find" function to search for the perfect matching string with a given query. However, finding the optimal alignment for a sequence read requires an alignment algorithm that is tolerant to imperfect matches, where genomic variations may occur. Moreover, the algorithm needs to be able to align millions of reads at a reasonable speed. As a first step to address this challenge, the reference genome is usually indexed in a hash table for efficient querying.

Many different tools have been developed for short reads mapping. They use Burrows–Wheeler Transformation (BWT) compression techniques, Smith–Waterman (SW) Dynamic programing algorithm or the combination of both in order to find the optimal alignment match within an acceptable computational time.

Bowtie2[25] and BWA[26] are two well-known short reads alignment tools that implement BWT algorithm. SW is a score-based dynamic programming algorithm that provides at least one optimal local alignment even though the solution might not be unique. This algorithm is tolerant to mismatches and gaps at the expense of increased computational time. MOSAIK,[27] SHRiMP2,[28] and Novoalign (http://www.novocraft.com) are implementations of SW algorithms with increased alignment accuracy. Multithreading and/or MPI implementations are employed in those mapping tools allowing significant reduction in the runtime.

Evaluation of short-read aligners. Using simulated datasets of 5 million 100-bp reads, we evaluated four commonly used alignment tools, Bowtie2, BWA, Novoalign V3, and genomic mapping and alignment program (GMAP).[25,29] Those reads were randomly generated from the human genome assembly hg19 with various types of genomic variations introduced. These include 1–5 bp SNVs, insertions only, deletions only, and mixed insertions and deletions (indels), as well as a mixture of SNVs and indels. If a read is aligned to the expected target region, we consider this as a true alignment. Sensitivity was calculated as the percentage of true alignments out of the total 5 million reads, and precision as the ratio of the number of true alignments to the total number of aligned reads.

Figure 2A and 2B shows the sensitivity and precision of the four tools with 1–5 bp variants introduced in each read. Bowtie2, Novoalign, and GMAP show high accuracy for all simulated genomic variations and are tolerant to various numbers of errors in a read. Bowtie2 and Novoalign have a similar level of sensitivity and precision between 94% and 97%. Compared with Bowtie2 and Novoalign, a 2–5% increased specificity and 4–10% reduced sensitivity were observed in GMAP alignment. The sensitivity of Novoalign drops to between 77% and 85% when a read contains 5 bp deletions. On the other hand, BWA performs well when there is at most one error in a read, and its sensitivity and specificity are reduced by 10% to 66% when attempting to align reads with more than one error.

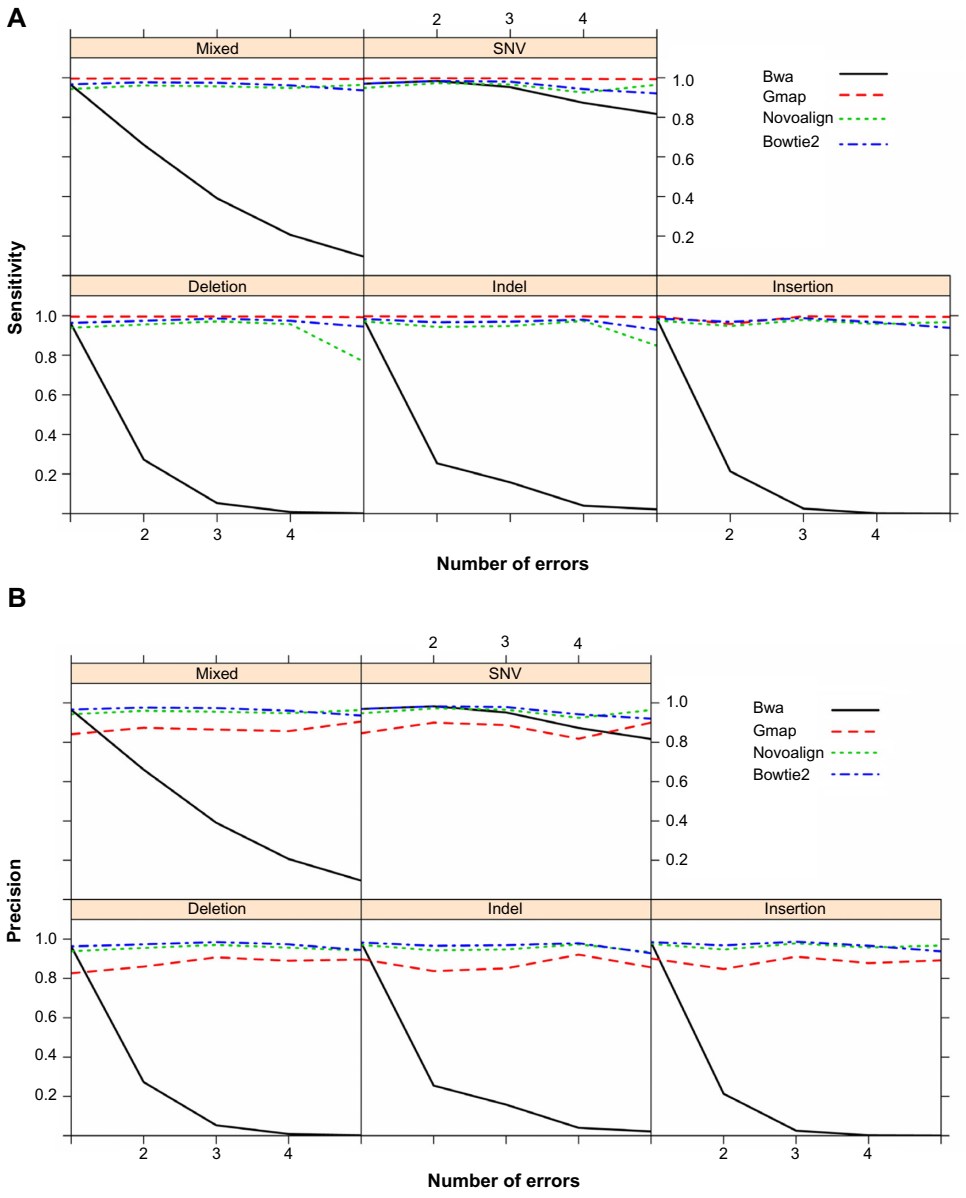

Figure 2. Comparison of alignment tools in terms of sensitivity (**A**) and precision (**B**) with 1–5 bp genomic variations per simulated read. Five sets of alignment are shown with introduction of errors categorized by types of errors (deletions only, insertions only, insertions and deletions (indels), SNVs, and a mixture of indels and SNVs (mixed)).

Notes: Sensitivity is represented by the percentage of true alignments out of all simulated reads (5 million in total), and precision is the ratio of the number of true alignments to the number of aligned reads.

As for the running time of each tool, Bowtie2 and BWA are dramatically faster than the other two. Bowtie2 took approximately 8–25 minutes to finish aligning 5 million reads in each alignment, which is two to four times faster than BWA. A 5- to 10-fold increase in runtime was observed for GMAP compared to Bowtie2 and BWA. Novoalign took almost 3 hours to finish, which is probably due to its implementation of SW alignment algorithm as it usually takes longer time than BWT-based aligners. All tools were tested using one single processor on Northwestern University's Quest high-performance computing cluster.

Post-alignment Processing

After mapping reads to the reference genome, a three-step post-alignment processing procedure is recommended to minimize the artifacts that may affect the quality of downstream variant calling. It consists of read duplicate removal, indel realignment, and base quality score recalibration (BQSR).

In the alignment, reads aligned with exact mapping coordinates are considered "read duplicates," which represent either true DNA materials or PCR artifacts. The two cases, however, cannot be distinguished solely based on sequence or alignment information. Before sequencing, a library of DNA fragments from genomic regions of interest is prepared. Those fragments are amplified via certain amount of PCR cycles to provide a sufficient amount of DNA materials for sequencing, while limiting the duplication level of templates introduced by rounds of amplifications. For WES analysis, it is recommended to remove duplicates before variant calling, with the purpose of eliminating PCR-introduced bias due to uneven amplification of DNA fragments. Programs such as Picard MarkDuplicates (http://picard.sourceforge.net) and SAMtools[26] determine read duplicates based on their identical 5′ mapping coordinates and orientation on the genome. 3′ Coordinates are usually not considered due to two reasons. First, the quality of bases generated by sequencers tends to drop down toward the 3′ end of a read; thus its alignment is less reliable compared to the 5′ bases. Second, if reads are trimmed at 3′ low-quality bases before alignment, they will have different read lengths resulting in different 3′ mapping coordinates. It should also be noted that read sequence information is not taken into account during duplicate removal. PCR duplicates do not necessarily have the identical sequence due to errors introduced in PCR amplification or sequencing processes. However, this may introduce bias in the calculation of variant frequency with de-duped read alignment. The bias becomes more severe in ultra-deep sequenced tumor samples (500× or higher), where removing large amounts of duplicates may affect allele frequency-based tumor subclone discovery and CNV detection. To address this issue, Zhou et al.[27] proposed a quantitative solution, DupRecover, which systematically estimates the degree of PCR-introduced bias and corrects for allele fractions.

After duplicate removal, the second step is to identify genomic regions that contain indels and improve the alignment quality in the target region. Compared to reads that contain only SNVs, mapping reads composed of indels requires gapped alignment which is more prone to noise. When aligning reads to the genome (discussed in the previous section), most short-read aligners walk through the reads one by one and the optimal alignment is determined for each read independently. As a result, the introduction of gaps in each alignment may be different among overlapping reads. The quality of the resulting gapped alignment can be improved by considering all aligned reads in the same region after mapping. Two algorithms have been developed to achieve this task: (1) local realignment of gapped reads to the reference genome or alternative candidate haplotypes; (2) local de novo assembly of the reads aligned around the target region followed by construction of a consensus sequence for indel discovery. Programs that implement one algorithm or a mixture of the two include SRMA[28] and IndelRealigner from the Genome Analysis Toolkit (GATK).[30] Some aligners and variant callers incorporate indel alignment improvement as part of the mapping or variant calling procedure. For example, Novoalign internally performs local SW alignments in addition to the hash-table based "seed" string searching to determine the optimal alignment for each read. The newly published GATK HaplotypeCaller program conducts a local de novo assembly of aligned reads prior to indel calling, which demonstrates to greatly improve the quality of indel calls.[31] Dindel implements a Bayesian approach to detect indels by calculating the posterior possibility of a haplotype after realigning reads to a number of candidate haplotypes constructed from the target region.[32] It is designed to accurately call indels correcting for mapping errors, base-calling errors, and sequencing errors in homopolymer-rich regions.

In the sequencing reads, each base is assigned with a Phred-scaled quality score generated by the sequencer, which represents the confidence of a base call. Base quality is a critical factor for accurate variant detection in the downstream analysis. However, the machine-generated scores are often inaccurate and systematically biased.[33] Therefore, BQSR is recommended to improve the accuracy of confidence scores before variant calling. It takes into account all reads per lane and analyzes covariation among the raw quality score, machine cycle, and dinucleotide content of adjacent bases. A corrected Phred-scaled quality score is reported for each base in the read alignment, assuming that all observed differences between the aligned reads and the reference genome are sequencing errors. One of the most commonly used BQSR programs is BaseRecalibrator from the GATK suite, which takes alignment files and recalibrates base scores across multiple sequencing runs. After base score recalibration, the variant calling accuracy was shown to be significantly improved and the bias was greatly reduced (GATK Best Practices recommendations[31,34]).

Other well-established programs include Recab from the NGSUtils suite,[35] which provides similar functions as GATK BaseRecalibrator,[30] and the Bioconductor package ReQON,[36] which uses logistic regression for recalibration of the base quality scores. In addition, ReQON outputs a set of diagnostic data and plots before and after recalibration to illustrate the improved accuracy. Some aligners, such as Novoalign, implement BQSR as one of the internal post-alignment processing options and the output contains aligned reads with improved base score accuracy. It is important to exclude known variants before score recalibration, as those represent true genomic variations and should not be considered as sequencing errors. Most programs take a list of known variants in addition to the alignment files for recalibration. For genomes without known variants available, it is recommended to run the variant calling without BQSR first to generate a list of variants, filter for high-quality ones, and then re-run BQSR with the list of high-confidence variants as the known genomic variations. Furthermore, in case of targeted sequencing where only a small region of the genome is sequenced, BQSR is not recommended, as it will not be able to accurately estimate the errors with limited coverage of the genome.

Variant Analysis

Variant analysis consists of genotyping, variant calling, annotation, and prioritization. Different types of genomic variants including SNVs, indels, CNVs, and large SVs can be detected from the sample by comparing the aligned reads to the reference genome. In cancer studies, it is important to distinguish somatic from germline variants as the two classes of variants often play distinct roles in tumor development. Germline variants are inherited mutations present in the germ cells, which are related to patient family history. Somatic variants are mutations that are present only in somatic cells and can be tissue-specific.

Variant calling. With a reasonable number of samples available, multiple sample variant calling is usually recommended. By taking into account all reads from one genomic region across multiple samples, it reduces the possibility of calling randomly presented sequencing errors and increases the possibility of calling alleles of low frequency or low coverage in a single sample. As a result, the accuracy and sensitivity of multi-sample variant calling are in general much higher than single sample variant calling.[37–39] In some circumstances, however, multi-sample calling becomes less practical. If the sample size is large, the requirement of computational resources and time for multi-sample variant calling increases dramatically. Furthermore, if one project is executed at multiple stages where a subset of the samples are sequenced each time, multi-sample calling would require re-running the variant calling step when new samples are added. Therefore, it may be more feasible to conduct single sample variant calling under those circumstances. Alternatively, large amounts of samples may be pooled into smaller groups and called per group. It should be noted that the results may be different from calling all the samples together.

Programs available for germline variant calling include GATK [38],[30,31] SAMtools,[26] FreeBayes,[40] and Atlas2.[41] GATK is a NGS analysis suite that employs a MapReduce framework to accelerate the processing of large amounts of sequence alignment files in parallel. It implements a simple Bayesian model to estimate the likelihood of genotype in the sample based on the observed sequence reads that cover the specified locus.[30,31,34] GATK consists of two main variant calling programs, UnifiedGenotyper and HaplotypeCaller. UnifiedGenotyper calls SNVs and indels separately with the assumption that each variant locus is independent. HaplotypeCaller simultaneously detects SNVs, indels, and some SVs with increased accuracy by performing a local de novo assembly of the aligned reads (discussed in the previous section).

SAMtools contains a set of utilities for the manipulation of aligned sequence reads in the SAM/BAM format.[26] One of the available utilities, mpileup, scans every position along the covered genome, computes possible genotypes from the aligned reads, and calculates the likelihood that each of these genotypes is truly present in the sample. Another tool, bcftools, then uses the genotype likelihoods to call the SNVs and indels. The main difference of the variant calling models between SAMtools and GATK is the estimation of the genotype likelihood of SNVs and indels. The differences also lie in the filtering step, where SAMtools uses predefined filters while GATK learns the filters from the data.

FreeBayes is a haplotype-based short polymorphism caller that can simultaneously detect SNVs, indels, multi-base mismatches, polyallelic sites, polyploidy, as well as CNVs in a single sample, pooled samples, or mixed populations.[40] It is built on a Bayesian statistical framework.

Compared to the three tools discussed above, Atlas2 is quite different in the implementation of variant calling algorithms.[41] For data generated by SOLiD™ platform, it uses logistic regression models trained on validated WES data to detected SNVs and indels. For Illumina data, it uses logistic regression models for calling indels and a mixture of logistic regression and a Bayesian model for SNV detection. Atlas2 consists of two separate programs, Atlas2-SNV and Atlas2-InDel.

Somatic variant detection. A major application of NGS variant analysis in cancer research is to distinguish somatic mutations in tumor cells from germline polymorphisms present in normal tissue. However, sequencing errors, insufficient variant coverage, sample contamination, and misclassification of germline variations often pose significant challenges in the detection of somatic variants. A number of tools have been developed to identify somatic mutations with paired tumor–normal samples, with their algorithms classified in two categories.

The first type of algorithms treats both samples as the same type, performs multi-sample variant calling on all

samples, applies genotype-based subtraction methods with integration of sample pair information, and retrieves variants that are only present in tumor samples (somatic) or in both samples (germline). This approach is prone to false positives if a germline variant is not called due to low frequency in the normal sample and false negatives if the tumor samples possess low mutation levels that cannot be distinguished from sequencing error. However, this type of caller has been well established and has demonstrated high accuracy and sensitivity in calling variants, such as GATK,[30] SAMtools mpileup,[42] and Isaac variant caller.[43]

The other type of algorithm treats tumor and normal as paired samples from the onset and detects variants simultaneously on both samples using joint diploid genotype likelihoods or shared allele frequency between the samples. A number of somatic mutation tools have emerged in the past 2 years, including deepSNV,[44] Strelka,[45] MutationSeq,[46] MutTect,[47] QuadGT (http://www.iro.umontreal.ca/~csuros/quadgt), Seurat,[48] Shimmer[49] and SolSNP (http://sourceforge.net/projects/solsnp), jointSNVMix,[50] SomaticSniper,[51] VarScan2,[51] and Virmid.[52] Each of them is equipped with unique features and applications. For example, deepSNV[44] was specially designed for detection of subclonal variants in ultra-deep sequenced tumor samples. Most of the above tools require matching tumor–normal samples as mandatory input. We discuss five of them in detail.

JointSNVMix implements a probabilistic model for somatic mutation discovery. It utilizes the joint genotype information across the paired tumor/normal samples and treats them as being conditionally independent. This leads to an increase in specificity while the sensitivity stays nearly unchanged.[50]

Strelka is built on a novel Bayesian model where the normal sample is represented as a mixture of diploid germline variation with noises and the tumor sample is represented as a mixture of the normal sample with somatic variations through allele frequencies.[45] A score derived from the joint probability of a somatic variant and a specific genotype in the normal sample is used to make the variant call. This algorithm computes allele frequency variation in samples at any level without requiring an estimation of tumor purity.

SomaticSniper is a tool that compares the diploid genotype likelihood in the tumor and normal pair for the somatic variant calling.[51] The likelihood is computed using the germline genotyping algorithm adopted from the MAQ program with the consideration of the dependency between the tumor and normal genotypes from the sample patient.[42] A set of somatic detection filters is applied to the calls from the MAQ genotyping. The current tool does not take into account tumor purity or copy number state.

Varscan2 applies a heuristic and statistical algorithm to identify variants as germline or somatic and detects loss of heterozygosity events from the variant calls.[53] The genotype for each sample is first determined from mpileup calling of the variants, and then one-sided Fisher's exact test is applied to call somatic or germline variants. Tumor purity can be added to the program explicitly. This provides the robustness when the sample is contaminated or alternate ploidy exists in the normal sample.

Virmid uses the level of impurity in the sample to improve the somatic variation detection.[52] The algorithm estimates the sample contamination level and the disease genotypes using the maximum likelihood estimation. A Bayesian model utilizes the joint genotype probability of tumor and normal samples to call the variant. The unique design of the algorithm allows the program to recalibrate the genotype probabilities with varied contamination levels. This procedure reduces the running time and greatly improves the accuracy. The method also takes into account other types of noise such as sequencing and mapping errors, mapping bias, and CNV stage.

Evaluation of variant callers. Previous studies reported that the accuracy and robustness vary among variant calling methods.[54,55] Liu et al.[38] evaluated the performance of four variant callers, SAMtools, GATK, glftools, and Atlas2. They recommended GATK for general-purpose variant analysis. Xu et al.[56] compared the performance of five somatic SNV calling methods (GATK UnifiedGenotyper followed by simple subtraction, MuTect, Strelka, SomaticSniper, and VarScan2) for matched tumor–normal sequencing data. They used the NIST-GIAB gold standard dataset to demonstrate that the sensitivities of these methods vary in regard to the allelic fraction of the mutation in the tumor sample. Roberts et al.[57] conducted a comparison between VarScan, SomaticSniper, JointSNVMix2, and Strelka. Their results revealed substantial discordance among the called variants. To facilitate the method evaluation process, Talwalkar et al.[58] proposed a benchmarking methodology for the evaluation of the human genome variant callers. Their SMASH toolkit consists of three components: (1) short reads from NGS experiments; (2) reference genome; and (3) the validation data in standard VCF format are used to measure the algorithm performance. A set of evaluation metrics is defined to measure the accuracy of the variant calls and the computational performance of the algorithm.

We evaluated the performance of four variant callers (GATK UnifiedGenotyper, SAMtools mpileup, Atlas2, and FreeBayes) with read alignment generated by three aligners (BWA, Bowtie2, and Novoalign V3) using the NIST-GIAB benchmark genotype calls.[59] It contains 2,915,731 high-confidence SNVs, indels, and homozygous reference genotype calls for NA12878 (version 2.18) after integrating 14 datasets from five sequencing platforms. For the tool evaluation, we downloaded 100× NA12878 WES data generated by University of Washington (SRX079575). A total of 170,987,444 50 bp paired-end reads were preprocessed and mapped to the hg19 reference genome using the three aligners. On average, over 85% of the reads were mapped with mapping quality higher than 30.[42] The read alignment was post-processed for duplicate

removal using Picard and indel realignment and base score recalibration using GATK.

Twelve sets of variants were generated from the three aligners and four callers, with low-quality calls removed. Variants not located in the targeted regions and/or with read depth lower than 6× were excluded from further consideration. The variants were further filtered to exclude those located within genome regions where no confident calls could be made.[59] A total of 21,661 calls were kept and compared to the benchmark set, which consists of 23,294 on-target calls covering approximately 31.9 Mb of the human exome. 95.3% of the variants were detected by at least two callers in reads mapped by at least two aligners (referred to as filter "2Aligner×2Caller"), yielding a final variant list with high sensitivity (99.04% for SNVs and 87.28% for small indels), specificity (>99.99%), and precision rate (>99%) with overall performance better than any single algorithm (Fig. 3). When comparing variants called from the alignment generated by each of the three aligners, close to 80% of the united variant set is concordantly detected by all four callers in each alignment. Furthermore, 83–93% of the united variant set was concordantly detected in all three alignments. When comparing SNVs with indels in between callers, indels showed a lower fraction of consistently called variants, which may be due to difficulties in gapped alignment and noises from variants present in the neighborhood regions (Fig. 4A and 4B).

In summary, these results suggest that by integrating multiple aligners and callers, variants of high confidence and high sensitivity are obtained. Among all 12 variant sets, variants called by FreeBayes from Novoalign-mapped reads are associated with the highest sensitivity (SNVs: 95.97% and indels: 83.39%) and precision rate (SNVs: 99.70% and indels: 99.57%). Therefore, if only one aligner and one caller to be chosen, a combination of Novoalign and FreeBayes is recommended for data analysis.

Variant annotation. Even with the increasing amount of sequencing data, identification of disease-causing mutations from a background of random errors and polymorphisms remains challenging. After variants are detected, annotation attributes such as genomic feature, gene symbol, exonic function, and amino acid change can be attached to the variant list. Most studies focus on the non-synonymous SNVs and indels in the protein-coding regions, which account for 85% of the discovered disease-causing mutations in Mendelian disorders[60,61] and many disease-associated mutations in complex diseases. Synonymous SNVs are important for estimating the background mutation rate in the genome. For example, the background mutation rate in the melanoma patients was discovered to be approximately 10–20 mutations per MB of human exome, indicating a low mutation rate in the melanoma genome.[62,63]

In addition to the basic annotation discussed above, many programs have been developed to integrate public databases for additional information of the variants, including minor allele frequency (MAF) in normal populations, experimental evidence from clinical assays, deleterious prediction of variant function, and collection of variant and genes in known cancer or genetic disease studies. Detailed genomic contents, including tissue-specific expression, transcription factor binding sites, histone modifications, DNase I hypersensitive sites, enhancers, and eQTLs can be retrieved from the ENCODE project[64] and public databases such as RegulomeDB[65] and HaploReg[66] and added to variant annotation. While most of those tools only provide a small number of annotations, a few programs were developed to combine the annotations from numerous sources. One of the most commonly used variant annotation programs is ANNOVAR, which provides three annotation modes, gene-based, region-based, and filter-based, with a collection of over 4,000 public databases for annotation.[67] It integrates dbSNP, 1000Genomes, ESP6500, Complete Genomics personal genomes and NCI-60 human tumor cell line panel exome sequencing data for accessing MAF information, seven programs from the LJB23 database[68] plus Combined Annotation Dependent Depletion (CADD) database[69] for deleterious prediction, Phylop[70] and Genomic Evolutionary Rate Profiling (GERP)[71] score for indicating conservation of the mutated site across 29 mammalian species, as well as experimental evidence in pathogenesis of the variant from disease variant databases such as COSMIC[72] and ClinVar.[73] CADD uses a unified score for the potential deleteriousness of all 8.6 billion possible human mutations by comparing variants that survived natural selection with simulated mutations.[69] The database combines 63 annotations through a machine-learning framework. Those annotations consist of prediction scores from GERP, PolyPhen, and other programs. The simple metric provides a straightforward approach to interpret high-penetrance mutations in Mendelian disease and low-penetrance variants found in genome-wide association studies. SeattleSeq, a web-based variant annotation system, offers direct upload of variant lists through the website and annotates with public databases such as dbSNP and deleterious prediction programs such as PolyPhen.[74] It integrates databases of KEGG Pathways, CpG islands, transcription binding sites, and protein–protein interactions and provides useful information regarding the gene regulatory network and regulation of the mutated genes. Other stand-alone variant programs include VariantAnnotator from the GATK[30] and SnpEff.[75] They are equipped with various read filters or unique features such as gene set enrichment analysis for downstream analysis.

Variant filtration and prioritization. Exome sequencing of human samples with at least 100× coverage was estimated to detect approximately 20,000–30,000 SNV and indel calls on average.[54] The number of candidate variants is reduced using a three-step filtration and prioritization strategy to generate a short candidate mutation list for experimental validation.

The first step is to remove less reliable variant calls. This includes variants with low coverage, low quality, strand biased,

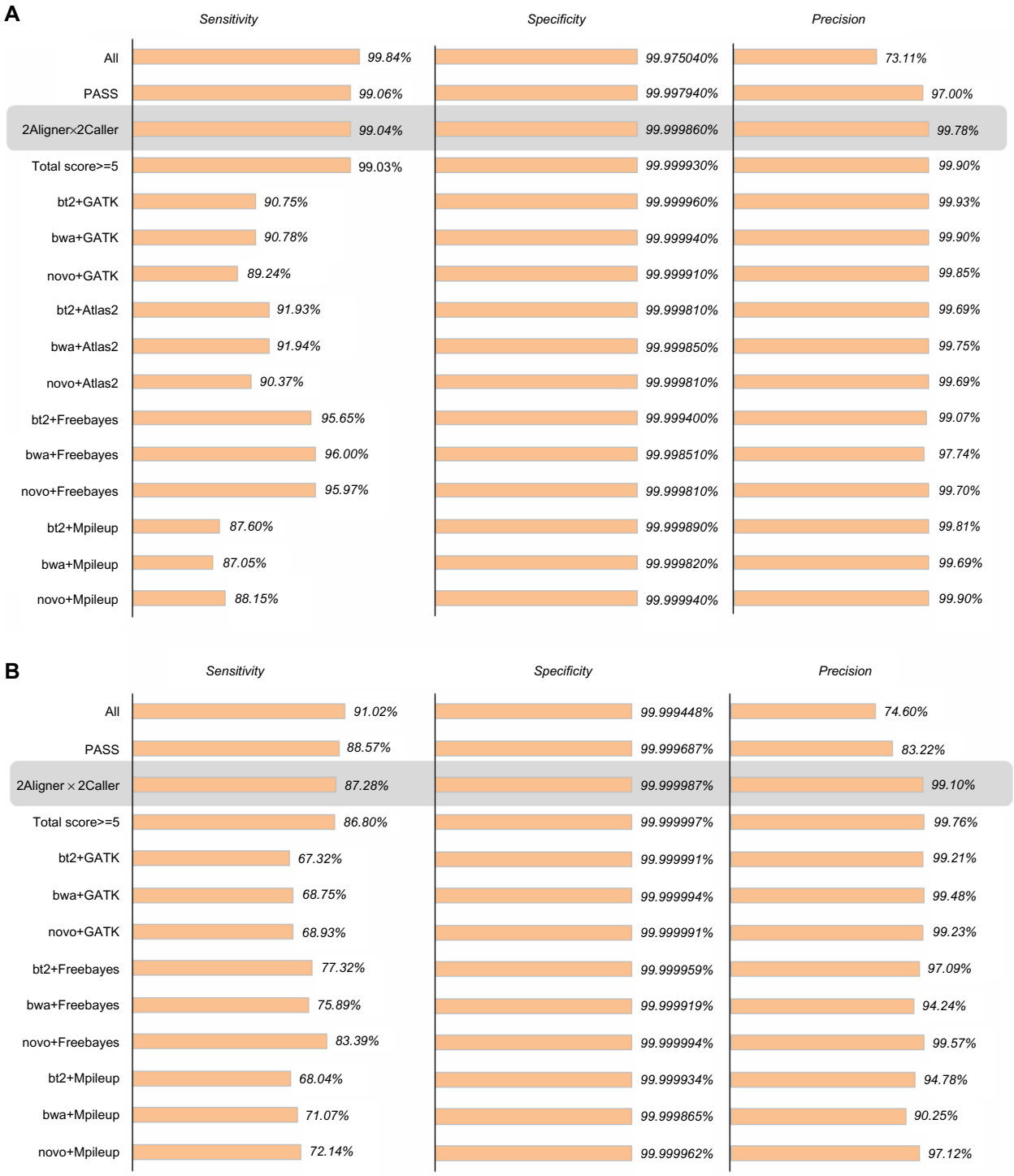

Figure 3. Evaluation results of variant callers with alignment generated by three aligners for SNVs (**A**) and indels (**B**). Aligners used for mapping the reads to the genome include Bowtie2 (bt2), BWA, and Novoalign (novo).

Notes: SNV callers include GATK UnifiedGenotyper, FreeBayes, SAMtools mpileup, and Atlas2. The first three were also used for calling indels. Gray background highlights the filter recommended for downstream variant analysis ("2Aligner × 2Caller"), which is shown to have better sensitivity than any single algorithm (99.94% for SNVs and 87.28% for indels) and high precision rate (99.78% for SNVs and 99.10% for indels). "Total score ≥5" represents variants detected in at least 5 out of 12 sets.

located in SNV clusters, and/or supported by low-confidence read alignment.[76] Depending on the quality settings, this step may reduce the variant list by 1.3× to 1.5× (unpublished data). Moreover, in the evaluation of family trios, variants detected in a child that could not have been inherited from both parents (Mendelian errors) can be identified, most of which are likely to result from sequencing artifacts.[77] Improvement in accuracy of rare or novel variant calls was observed when multiple

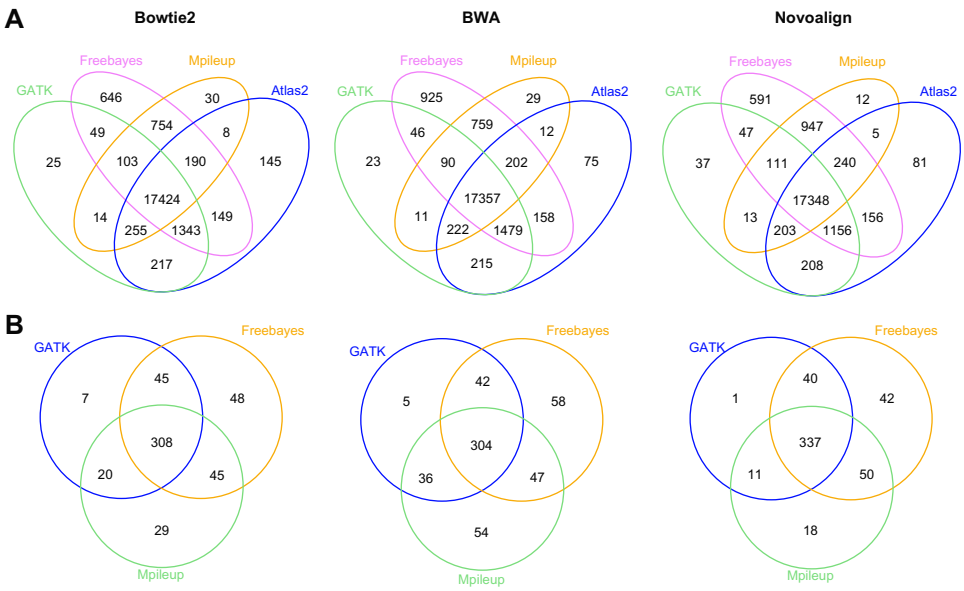

Figure 4. Counts of variants detected by four callers with alignment generated by three aligners as shown in Venn diagram of (**A**) SNVs and (**B**) indels.

family members are called simultaneously with knowledge of pedigree structure. Of note, the accuracy of error detection and variant identification increases with the number of relatives and generations sequenced per family.[78,79] Filtering out these errors should be done with extra caution as this class of variants also includes de novo mutations in the child.

The second step is to restrict variants to those of relatively low population frequency, assuming that common variants are less likely to cause disease than rare ones. The MAF threshold should be carefully chosen based on disease model of the studies. For germline variants, usually over 70% of the variants are removed at MAF <1% (unpublished data). The filter does not have as dramatic effect as on somatic variants because most of those are novel and patient specific. Detection of compound heterozygous mutations is more complicated, in which case, each of the heterozygous variants does not cause observable phenotypic changes (recessive alleles), yet together they knock out both copies of the same gene and lead to disease outcome. In addition to rare variants, common variants may also show compound heterozygous effects with other variants in the same gene. For example, Heresbach et al.[80] reported that compound heterozygotes of three common variants in *NOD2/CARD15* gene increases the risk of Crohn's disease, which is greater than any single heterozygote. Furthermore, combinatorial effect of a common variant together with a novel mutation in the *MTHFR* gene was demonstrated to cause more severe symptoms in MTHFR deficiency disorders such as hyperhomocysteinaemia.[81] Publicly accessible tools that can detect compound heterozygous mutations in familial diseases include the GeneTalk Suite, which provides a web-based interface for customized variant filtration.[82] Stand-alone programs designed for identification of recessively acting variants (homozygotes or compound heterozygotes) include the SCOREASSOC program, which detects the association between variants based on derivation from the Hardy–Weinberg equilibrium in outbred populations.[83]

The third step is to prioritize variants relative to the disease. In general, SNVs can be ordered by their coding effect, in which case splice mutations (SNVs that occurred at splice donor or receptor sites) and nonsense mutations are in general more damaging than missense mutations. Indels, on the other hand, can be ordered based on whether they cause splice disruption or frameshift of the coding sequence. Furthermore, candidate disease-causing mutations are identified using discovery-based and hypothesis-based approaches. If little is known about the disease, the discovery-based approach is usually employed to filter and prioritize variants based on the mutation frequency across patient groups if the patients are genetically unrelated or on the inheritance pattern of mutations with the presence of a pedigree. Other criteria, including MAF values, deleterious function prediction, experimental evidence from published studies, and pathway information are also important factors for variant prioritization. In cancer genomics, computational approaches have been developed to distinguish driver mutations from passenger ones based on high mutation frequency in the patient cohort and highly damaging amino acid changes on the protein function or structure.[84] Moreover, variants that predispose an individual toward cancer may be identified from the germline variants and interpreted together with the pattern of somatic mutations.[85] In the hypothesis-based approach, a disease model is drawn based on the family history or known studies, and the discovery of candidate pathogenic mutations is driven by the proposed hypothesis. For Mendelian disorders, this may be recessive, dominant or compound heterozygous mutation model, or a combination of the models for the explanation of the inherited traits.[86] Moreover, pathways discovered in previous studies may guide discovery of new mutations in

the known pathways, yet it is limited by the prior knowledge of the disease. It should be noted that new pathways can be established by linking newly identified mutations to known diseases. For example, six novel mutations in *C5ORF42* were reported to be causative in the development of Joubert syndrome in French Canadian population, although the pathogenic mechanism remains unknown.[87] In most cases, more than one approach is employed to identify the most interesting candidate variants.

Several tools have been developed to systematically filter, evaluate, and prioritize thousands of variants all at once, taking into account the annotation results of the variants, patient family information, as well as phenotypes and disease subtype information. VAAST2[88] implements an aggregative variant test combining the amino acid changes, allele frequencies, and phylogenetic conservation. The program generates a ranking list of variants sorted by its importance for the disease, which is useful for analysis of complex genetic diseases and rare Mendelian disorders. Other tools that are publicly available include VarSifer,[89] KGGseq,[90] PLINK/SEQ,[91] and SPRING.[92] A newly developed GUI tool, gNOME, allows direct upload of the variant file and performs streamline variant annotation, filtration, and prioritization, with its output summarized at variant, gene, or genome level.[93] One unique feature of gNOME is that it takes group information and detects disease-related genes or variants enriched in cases but not controls. One commercial variant analysis software, Ingenuity® Variant Analysis™ (http://www.ingenuity.com/variants), utilizes public databases such as those derived from the 1000 Genomes Project and the NHLBI Exome Sequencing Project (ESP6500), as well as experimental data collected from the literature and signaling pathways to filter and prioritize variants driven by data or with a specified disease model.

The above filter-based approaches have a higher success rate in identification of inherited or novel deleterious mutations in familial diseases, whereas with unrelated subjects, its power becomes limited. For rare diseases, one solution is to study unrelated probands with the same syndrome and focus on rare mutations that occurred in multiple individuals. The approach is known to be effective if the variant is clearly deleterious and disease related. Using this approach, Hood et al.[94] discovered truncating mutations in the C-terminal of SNF2-related CREBBP activator protein as the causative mutations for Floating-Harbour syndrome in five unrelated patients. In other cases, statistical tests are often necessary to discover mutations or target genes that contribute to a disease.[86] This is particularly important for addressing the effects of multiple rare variants that cause functional damage in a combinatorial manner. Previous studies identified rare predisposing variants that are significantly associated with complex traits such as colorectal adenomas,[95] high-density lipoprotein cholesterol,[96] and schizophrenia.[97] There are two major types of variant association tests: (1) burden tests, which include collapsing methods such as CAST,[98] CMC,[99] RareCover,[100] aSum,[101] and aggregation methods such as WSS,[102] KBAC, and[103] RBT[104]) (2) non-burden test based methods such as VT,[105] C-alpha,[106] EREC,[107] and SKAT.[108] For example, Wu et al.[108] applied SKAT to Dallas Heart Study data on 93 variants in three genes to test the association between log-transformed serum triglyceride levels and rare variants. The results showed that SKAT was a very powerful test for the dichotomous trait and performed comparably with burden-test-based methods for continuous trait. It should be noted that SKAT could be applied to both common and rare variants. Furthermore, the combination of genetic linkage, association analysis, and WES can serve as a useful approach to reduce the search space for rare variants in complex diseases with increased discovery power.[109]

With all the tools available and new ones emerging monthly, variant filtration and prioritization are becoming more automated like other parts of variant analysis such as the detection and annotation. Regardless, a deep understanding of the biological questions being asked and the etiology of the disease being studied is crucial for properly choosing tools and parameters that suit a study the best. Integration with clinical data such as patient diagnosis and family history is often helpful for identifying the variants that are responsible for the symptoms.[110]

NGS Data Management

The technology evolution in molecular biology, especially in NGS, has moved biology into the big data era. For example, European Bioinformatics Institute (EBI) currently stores 20 petabytes of data, 2 petabyte of which is genomics related.[111] While the capacity of computing hardware doubles every 18 months, new biological data are doubling every 9 months. With this trend, the challenges faced by life scientists have been shifted from data acquisition to data management, processing, and knowledge extraction. While many studies have recognized the big data challenge, few systematically present approaches to tackle it. In this section, we propose a framework to address the big data challenge faced by biological scientists which consists of data, computation, workflow, and knowledge.

Data. As the volume of biological data generated by NGS instrumentation grows from hundreds of terabytes to petabytes, it has moved beyond the storage capabilities typically handled by individual scientists. Only a few organizations, such as the EBI and NIH NCBI have the capacity and the mandate to store large datasets and provide public access. The sustainability of this model is not clear as we move into the 100,000 genomes world. Big data hosted in the cloud seems to be a promising solution, one that many institutions are exploring. As an example, Amazon Simple Storage Service (Amazon S3) offers a cloud-based file system, with virtually unlimited capacity and charged by usage. Currently, NCBI stores a subset of the human genomics 1000 data (about 200 TB) in S3 (http://aws.amazon.com/1000genomes).

Commercial vendors are coming into this space too. In the USA, Illumina's NGS environment is cloud based with data hosted by Amazon AWS. Illumina also has established Clinical Services Laboratory to provide genome-based testing to quickly access genomic aberrations and assist medical diagnosis. Another example is Genewiz, a DNA services company based in New Jersey, which provides similar services. International examples also exist, including Beijing Genomic Institute (BGI, Shenzhen), who acquired Complete Genomics last year, that has built solid connections with many healthcare providers. BGI also has set up two data centers, Bio-Data Centre (CliMB) and Biol-Cloud Computing centre (BGI Cloud), offering cloud-based data service, sequence alignment, and many other features. Meanwhile, healthcare providers are also launching genome-sequencing programs, such as Massachusetts General Hospital, Geisinger Health System, Scripps Health, and Inova Health System. For instance, the Falls Church, Virgina-based Inova Health System, has performed "1,500 complete DNA (whole genome) sequences as part of the first clinical study aimed at unraveling the mysteries of pre-term birth" in its Inova Translational Medicine Institute. In the next 2 years, they plan to produce 20,000 family-based whole genome sequences. In the foreseeable future, we expect to see more public or private biology data stored in cloud.

Despite the massive capacity provided by cloud computing, storing the huge amount of genome sequencing data is still challenging. To make the storage economic, various data compression techniques have been developed, including naive bit encoding, dictionary-based, statistical, and referential approaches.[112]

Computation. A key to reduce the latency of analyzing big data is to move computation to the data. This has two meanings in a cloud environment. First, computation should occur in the cloud, rather than moving data out of the cloud in order to compute on it. Cloud providers such as Amazon usually offer computation clusters collocated with the storage clusters and intra-cloud data movement is usually fast and free of charge. Second, when data reach the computation cluster, the computation framework should address the issue of data locality. Amazon AWS has offered a comprehensive suite of tools to process large volume of genomics data (http://aws.amazon.com/genomics/). For another example, Hadoop is an open-source software framework that parallelizes the computation and makes it easier to co-localize data and parallel computation. It has been well adopted in the field of bioinformatics,[113] including a few sequence alignment tools such as CloudBurst, Crossbow, SeqMapReduce, and CloudAligner.[114]

Workflow. New findings in biological sciences usually come out of multi-step data pipelines, also known as workflows. Biologists already used a few workflow tools in the pre-cloud and pre-big-data era. Synapse (http://aws.amazon.com/swf/testimonials/swfsagebio) and Galaxy[115,116] are two cloud-based workflow tools dealing with big data. However, it is still necessary to globally optimize the data flow in an overall multi-step workflow in order to eliminate unnecessary data movement and redundant computation.

Knowledge embedded in big data, as well as the routines and workflow used to derive it, needs to be captured properly. New journals such as *Nature Protocols* are promoting the documentation of "recipe" style of step-by-step procedures leading to a discovery. In the bioinformatics domain, some scientists are promoting the idea of embedding a provenance workflow inside a publication to preserve both the data and the routine (http://tridentworkflow.codeplex.com). Since scientific publications are the carriers of new knowledge, we expect that knowledge-embedded data and workflows should be an integral part of future scientific publications.

Conclusion

The unprecedented reduction in the cost of high-throughput sequencing has made it possible to conduct ever-larger studies on human diseases. The bottleneck of NGS has shifted from producing sequence data to data management, analysis, and summarization. In this review, we examined bioinformatics software available for whole exome data analysis, including data preprocessing, alignment, post-alignment processing, variant calling, annotation, and prioritization tools. We compared the performance of alignment tools and variant calling programs using simulated and benchmark datasets. Along the way, we attempted to highlight frequent considerations and procedures necessary for the identification of causal variants of the disease.

Despite the large number of data analysis options, the complexity of the human genome and the lack of concordant results from different variant detection tools[54] highlight the urgency for developing community standardized protocols, tools, and benchmarks. As one of the efforts, The Cancer Genome Atlas (TCGA) and the International Cancer Genomics Consortium (ICGC) launched the ICGC-TCGA DREAM Somatic Mutation Calling Challenge, a competition of the best tools/pipelines for the detection of the cancer genome mutations using NGS data.[117] The organizers hope the challenge will accelerate the adoption of the best somatic variant identification techniques and help answer other key questions in cancer genome research.

The rapid growth of NGS technologies affords new opportunities to conduct large-scale patient sequencing projects, which facilitates discovery of tumor-specific mutations as potential targets for development of precision medicine.[45] As an example of the power of genomic techniques in understanding human disease and identifying new treatment options, J. Carpten's group discovered a unique pattern in the tumor mutation profile of 14 metastatic triple-negative breast cancer (TNBC) patients based on whole genome and transcriptome sequence data. They suggested that TNBC should be treated as a genetically different disease instead of a subtype of breast cancer.[118] They identified somatic mutations in the RAS/RAF/MEK and PI3K/AKT/mTOR signaling pathways, which

have led to a clinical trial combining agents targeting of MEK and mTOR genes with encouraging results. Furthermore, by performing WES on a metastatic, castration-resistant prostate cancer patient, Van Allen et al. discovered somatic genomic mutations in the PI3K pathway and a *BRCA2* germline variant that may predispose individuals to cancer.[119] An increasing number of studies employ whole genome and exome sequencing to successfully identify disease-causing mutations that provide attractive treatment targets using gene therapy, therapeutic drugs, or transplantation.[120–122]

In the next five to twenty years, whole exome/genome analysis may be adopted as a routine procedure as part of the clinical laboratory for disease treatment. Universities and institutions including University of Pennsylvania (http://www.pennmedicine.org/personalized-diagnostics), Emory University (http://genetics.emory.edu/egl), and University of Washington (http://depts.washington.edu/labweb/Divisions/MolDiag/MolDiagGen/index.htm) have already established clinical laboratories that offer genetic testing that employ streamline NGS technologies and data analysis, with the goal of efficient molecular diagnosis of human diseases. Two years ago, the NGS: Standardization of Clinical Testing (Nex-StoCT) workgroup, led by The US Centers for Disease Control and Prevention, published general principles and guidelines for quality practices of NGS data used in clinical testing.[123] The next challenge will be data integration on millions of genomic variants, clinical records, and patient information, allowing novel discovery of variants contributing to disease, rapid retrieval of this information, and user-friendly visualization and decision support for care givers.

Establishment of disease-specific variant databases and web servers to store, retrieve, and view the genome-wide contents is an ongoing task at universities and institutions, including the integration of visualization tools such as GBrowse[124] and Bioconductor packages such as Gviz[125] and Shiny.[126] International efforts to build a federated cancer genome resource where genomics and clinical data can be easily searched and visualized via a web-based interface include the ICGC Cancer Genome Portal (http://icgc.org), which was established based on the ICGC cancer genomes and Oncomine,[127] which integrates both microarray and NGS data with a large collection of pre-analyzed results. NCI also has the cancer Genomics Data Commons effort (http://ocg.cancer.gov/news/genomics-data-commons) focused on TCGA and TARGET and an affiliated effort to build cloud computing capabilities around TCGA data (http://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots). Because of protected clinical information, some of the data in these resources require permission to access. Standards for patient consent, what is freely sharable, and what is available as limited datasets are under active debate. The upcoming changes to the USA Common Rule (http://www.hhs.gov/ohrp/humansubjects/anprm2011page.html) and some of the proposed activities by the Global Alliance for Genomics and Health (http://genomicsandhealth.org/) are examples of activities that are attempting to streamline and standardized data access for clinically derived genomic data. At some point in the near future, systems will be designed that enable the full processing of NGS data, from QA through to variant calling. Such a system might even incorporate pieces or concepts from existing frameworks like the Galaxy project.[115,128,129]

As whole exome and genome sequencing are now starting to be integrated into research and clinical practice, some ethical and legal issues have arisen regarding the resulting genomic data. Particularly, what procedures are needed to protect patient's privacy, how do we interpret the causality of identified variants in human disease and return the reports to the research participants or patients, how should we report incidental findings of pathologic mutations that were not originally ordered by the clinician.

For privacy-related issues, careful control has to be exerted in order to prevent the identity of individuals (de-identified datasets). Informed consent should be acquired to address return of individual research results. However, like other health information data, technological solutions cannot completely resolve confidentiality problems. Even if data are anonymized, individuals can be re-identified if phenotype and genotype data are combined.[130] To increase the confidence of identified disease-causing variants, standards of data interpretation need to be imposed. Clinician should convey the patient with any information about variants that is clinically important but not necessarily with information of variants with unknown significance.[131] Recently, MacArthur et al.[132] proposed a list of guidelines for interpreting and reporting sequence variants in human disease. For incidental findings, the American College of Medical Genetics and Genomics (ACMG) has published a policy statement that provides recommendations by an appointed working group for reporting incidental findings in clinical sequencing.[133] The working group developed a minimum list of 56 genes (23 cancer susceptibility genes) which should be reported as incidental findings. Due to the limitations of current technology, the disorders associated with the gene list are restricted to those caused by SNVs and small indels. The list will be refined and updated at least annually as the technologies evolve. The ACMG recommendation also describes the responsibility of the ordering clinician who provides comprehensive pre- and post-test genetic counseling to the patient.

As the costs of NGS continue to fall, we expect WGS will eventually overtake WES as the mainstream tool for human genomics and disease studies. The principles outlined in this review are generally applicable to WGS as well.

Author Contributions

Conceived and designed the experiments: GF, HJ, JA, RB, LH. Analyzed the data: RB, LH, GF, HJ. Wrote the first draft of the manuscript: RB, LH, JA, WT, WAK, HJ, GF. Contributed to the writing of the manuscript: RB, LH, JA,

WT, WAK, HJ, GF. Agree with manuscript results and conclusions: RB, LH, JA, WT, WAK, HJ, GF. Jointly developed the structure and arguments for the paper: RB, LH, JA, WT, HJ, GF. Made critical revisions and approved final version: RB, LH, JA, WT, HJ, GF. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7.
2. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474:609–15.
3. Fertig EJ, Slebos R, Chung CH. Application of genomic and proteomic technologies in biomarker discovery. *Am Soc Clin Oncol Educ Book*. 2012;32:377–82.
4. Ansari NA, Bao R, Voichita C, Draghici S. Detecting phenotype-specific interactions between biological processes from microarray data and annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 2012;9:1399–409.
5. Huang L, Zhao S, Frasor JM, Dai Y. An integrated bioinformatics approach identifies elevated cyclin E2 expression and E2F activity as distinct features of tamoxifen resistant breast tumors. *PLoS One*. 2011;6:e22274.
6. Shi Y, Sha G, Sun X. Genome-wide study of NAGNAG alternative splicing in *Arabidopsis*. *Planta*. 2014;239:127–38.
7. Sheppard S, Lawson ND, Zhu LJ. Accurate identification of polyadenylation sites from 3' end deep sequencing using a naive Bayes classifier. *Bioinformatics*. 2013;29:2564–71.
8. Jiang H, An L, Lin SM, Feng G, Qiu Y. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. *PLoS One*. 2012;7:e46450.
9. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2013:bbs086.
10. Hayden E. Is the $1,000 genome for real? January 15, 2014. *Nature*. doi:10.1038/nature.2014.14530.
11. Craig Venter Wants to Build the World's Biggest Database of Genome and Physiological Information. Available at http://www.technologyreview.com/news/525416/microbes-and-metabolites-fuel-an-ambitious-aging-project/, March 11, 2014.
12. Kowalczyk SW, Wells DB, Aksimentiev A, Dekker C. Slowing down DNA translocation through a nanopore in lithium chloride. *Nano Lett*. 2012;12:1038–44.
13. Esfandyarpour H. Genapsys 100X solution: label-free fully-integrated "personal genomixer. *J Biomol Tech*. 2012;23(suppl):S9.
14. Wagle N, Allen EMV, Treacy DJ, et al. MAP kinase pathway alterations in BRAF-mutant melanoma patients with acquired resistance to combined RAF/MEK inhibition. *Cancer Discov*. 2013;4(1):1–8.
15. Londin ER, Keller MA, D'Andrea MR, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics*. 2011;12:464.
16. Andrews S. *FastQC*. Babraham Bioinformatics; 2012. Cambridge, UK.
17. Babraham Bioinformatics. *FastQ Screen*. Babraham Bioinformatics; 2013. Cambridge, UK.
18. Hannon Lab. *FASTX-Toolkit*. Hannon Lab; 2010. Cold Spring Harbor, NY.
19. Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*. 2012;7:e30619.
20. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27:863–4.
21. Zhou Q, Su X, Wang A, Xu J, Ning K. QC-chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One*. 2013;8:e60234.
22. Guo Y, Zhao S, Sheng Q, et al. Multi-perspective quality control of Illumina exome sequencing data using QC3. *Genomics*. 2014;103(5–6):323–8.
23. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;17:10–2.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20.
25. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010;26:873–81.
26. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
27. Zhou W, Chen T, Zhao H, et al. Bias from removing read duplication in ultra-deep sequencing experiments. *Bioinformatics*. 2014;30(8):1073–80.
28. Homer N, Nelson SF. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol*. 2010;11:R99.
29. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005;21:1859–75.
30. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
31. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics. Curr. Protoc. *Bioinform*. 2013;43:11.10.1–33
32. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. *Genome Res*. 2011;21:961–73.
33. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol*. 2011;12:R112.
34. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
35. Breese MR, Liu Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics*. 2013;29:494–6.
36. Cabanski CR, Cavin K, Bizon C, et al. ReQON: a bioconductor package for recalibrating quality scores from next-generation sequencing data. *BMC Bioinformatics*. 2012;13:221.
37. Cheng AY, Teo YY, Ong RT. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*. 2014;30:1707–13.
38. Liu X, Han S, Wang Z, Gelernter J, Yang B-Z. Variant callers for next-generation sequencing data: a comparison study. *PLoS One*. 2013;8:e75619.
39. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC Genomics*. 2012;13(suppl 8):S8.
40. Garrison E, Marth G: Haplotype-based variant detection from short-read sequencing. *ArXiv*12073907 *Q-Bio* 2012.
41. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012;13:8.
42. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18:1851–8.
43. Raczy C, Petrovski R, Saunders CT, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29:2041–3.
44. Gerstung M, Beisel C, Rechsteiner M, et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*. 2012;3:811.
45. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*. 2012;28:1811–7.
46. Ding J, Bashashati A, Roth A, et al. Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data. *Bioinformatics*. 2012;28:167–75.
47. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31:213–9.
48. Christoforides A, Carpten JD, Weiss GJ, Demeure MJ, Hoff DDV, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics*. 2013;14:302.
49. Hansen NF, Gartner JJ, Mei L, Samuels Y, Mullikin JC. Shimmer: detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics*. 2013;29:1498–503.
50. Roth A, Ding J, Morin R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*. 2012;28:907–13.
51. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28:311–7.
52. Kim S, Jeong K, Bhutani K, et al. Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol*. 2013;14:R90.
53. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568–76.
54. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. 2013;5:28.
55. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2011;8:61–5.
56. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*. 2014;15:244.
57. Roberts ND, Kortschak RD, Parker WT, et al. A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*. 2013;29:2223–30.

58. Talwalkar A, Liptrap J, Newcomb J, et al. SMaSH: a benchmarking toolkit for human genome variant calling. *ArXiv*13108420 *Q-Bio* 2013.
59. Zook JM, Chapman B, Wang J, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32:246–51.
60. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet*. 2003;33:228–37.
61. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet*. 2014;59:5–15.
62. Shull AY, Latham-Schwark A, Ramasamy P, et al. Novel somatic mutations to PI3 K pathway genes in metastatic melanoma. *PLoS One*. 2012;7:e43369.
63. Wei X, Walia V, Lin JC, et al. Exome sequencing identifies GRIN2 A as frequently mutated in melanoma. *Nat Genet*. 2011;43:442–6.
64. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
65. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22:1790–7.
66. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40:D930–4.
67. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164–e164.
68. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat*. 2013;34:E2393–402.
69. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310–5.
70. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20:110–21.
71. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*. 2010;6:e1001025.
72. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2011;39(suppl 1):D945–50.
73. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980–85.
74. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009;461:272–6.
75. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso–2; iso–3. *Fly (Austin)*. 2012;6:80–92.
76. Li H: Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* 2014:btu356.
77. Patel ZH, Kottyan LC, Lazaro S, et al. The struggle to find reliable results in exome sequencing data: filtering out Mendelian errors. *Appl Genet Epidemiol*. 2014;5:16.
78. Li B, Chen W, Zhan X, et al. A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet*. 2012;8:e1002944.
79. Peng G, Fan Y, Palculict TB, et al. Rare variant detection using family-based sequencing analysis. *Proc Natl Acad Sci U S A*. 2013;110:3985–90.
80. Heresbach D, Gicquel-Douabin V, Birebent B, et al. NOD2/CARD15 gene polymorphisms in Crohn's disease: a genotype- phenotype analysis. *Eur J Gastroenterol Hepatol*. 2004;16:55–62.
81. Rummel T, Suormala T, Häberle J, et al. Intermediate hyperhomocysteinaemia and compound heterozygosity for the common variant c.677C > T and a MTHFR gene mutation. *J Inherit Metab Dis*. 2007;30:401.
82. Kamphans T, Sabri P, Zhu N, et al. Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees. *PLoS One*. 2013;8:e70151.
83. Curtis D. Approaches to the detection of recessive effects using next generation sequencing data from outbred populations. *Adv Appl Bioinform Chem*. 2013;6:29–35.
84. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*. 2014;6:5.
85. Frank SA. Genetic predisposition to cancer–insights from population genetics. *Nat Rev Genet*. 2004;5:764–72.
86. Ionita-Laza I, Makarov V, Yoon S, et al. Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am J Hum Genet*. 2011;89:701–12.
87. Srour M, Schwartzentruber J, Hamdan FF, et al. Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population. *Am J Hum Genet*. 2012;90:693–700.
88. Hu H, Huff CD, Moore B, Flygare S, Reese MG, Yandell M. VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol*. 2013;37:622–34.
89. Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics*. 2012;28:599–600.
90. Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucleic Acids Res*. 2012;40:e53–e53.
91. ATGU. *PLINK/SEQ*. Massachusetts General Hospital; 2012.
92. Wu J, Li Y, Jiang R. Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*. 2014;10:e1004237.
93. Lee I-H, Lee K, Hsing M, et al. Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. *Hum Mutat*. 2014;35(5):537–47.
94. Hood RL, Lines MA, Nikkel SM, et al. Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause floating-harbor syndrome. *Am J Hum Genet*. 2012;90:308–13.
95. Azzopardi D, Dallosso AR, Eliason K, et al. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res*. 2008;68:358–63.
96. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004;305:869–72.
97. Walsh T, McClellan JM, McCarthy SE, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008;320:539–43.
98. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*. 2007;615:28–56.
99. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83:311–21.
100. Bhatia G, Bansal V, Harismendy O, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol*. 2010;6:e1000954.
101. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010;70:42–54.
102. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5:e1000384.
103. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*. 2010;6:e1001156.
104. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet*. 2011;7:e1001289.
105. Price AL, Kryukov GV, de Bakker PIW, et al. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*. 2010;86:832–8.
106. Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011;7:e1001322.
107. Lin D-Y, Tang Z-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet*. 2011;89:354–67.
108. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89:82–93.
109. Rosenthal EA, Ranchalis J, Crosslin DR, et al. Joint linkage and association analysis with exome sequence data implicates SLC25 A40 in hypertriglyceridemia. *Am J Hum Genet*. 2013;93:1035–45.
110. Van El CG, Cornel MC, Borry P, et al. Whole-genome sequencing in health care. *Eur J Hum Genet*. 2013;21:580–4.
111. Marx V. Biology: the big challenges of big data. *Nature*. 2013;498:255–60.
112. Wandelt S, Bux M, Leser U. Trends in genome compression. *Curr Bioinf*. 2013;9:1–12.
113. Russell J: Hadoop's rise in life sciences. *BioIT World* 2012.
114. Wandelt S, Rheinländer A, Bux M, Thalheim L, Haldemann B, Leser U. Data management challenges in next generation sequencing. *Datenbank-Spektrum*. 2012;12:161–71.
115. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010;11:R86.
116. Afgan E, Baker D, Coraor N, et al. Harnessing cloud computing with galaxy cloud. *Nat Biotechnol*. 2011;29:972–4.
117. Boutros PC, Ewing AD, Ellrott K, et al. Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nat Genet*. 2014;46:318–9.

118. Craig DW, O'Shaughnessy JA, Kiefer JA, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. *Mol Cancer Ther*. 2013;12:104–16.
119. Van Allen EM, Foye A, Wagle N, et al. Successful whole-exome sequencing from a prostate cancer bone metastasis biopsy. *Prostate Cancer Prostatic Dis*. 2014;17:23–7.
120. Chiang P-W, Wang J, Chen Y, et al. Exome sequencing identifies NMNAT1 mutations as a cause of Leber congenital amaurosis. *Nat Genet*. 2012;44:972–4.
121. Goh V, Helbling D, Biank V, Jarzembowski J, Dimmock D. Next-generation sequencing facilitates the diagnosis in a child with twinkle mutations causing cholestatic liver failure. *J Pediatr Gastroenterol Nutr*. 2012;54:291–4.
122. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med*. 2011;13:255–62.
123. Gargis AS, Kalman L, Berry MW, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012;30:1033–6.
124. Stein LD, Mungall C, Shu S, et al. The generic genome browser: a building block for a model organism system database. *Genome Res*. 2002;12:1599–610.
125. Hahne F, Durinck S, Ivanek R, Mueller A, Lianoglou S, Tan G. *Gviz: Plotting Data and Annotation Information along Genomic Coordinates*. 2014.
126. RStudio Inc.: Shiny: Web Application Framework for R. 2014.
127. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 1 8,000 cancer gene expression profiles. *Neoplasia*. 2007;9:166–80.
128. Blankenberg D, Von Kuster G, Coraor N, et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*. 2010, Chapter 19:Unit 19.10.1–21.
129. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*. 2005;15:1451–5.
130. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339:321–4.
131. Institute of Medicine. Integrating Large-Scale Genomic Information into Clinical Practice: Workshop Summary. Washington, DC: The National Academies Press; 2012.
132. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508:469–76.
133. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*. 2013;15:565–74.