Emerging Topics in Life Sciences (2019) **3** 399–409
https://doi.org/10.1042/ETLS20190007

PORTLAND PRESS

Review Article

Variant calling and quality control of large-scale human genome sequencing data

Brandon Jew[1] and **Jae Hoon Sul**[1,2]

[1]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, U.S.A; [2]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, CA 90095, U.S.A

Correspondence: Jae Hoon Sul (jaehoonsul@mednet.ucla.edu)

Next-generation sequencing has allowed genetic studies to collect genome sequencing data from a large number of individuals. However, raw sequencing data are not usually interpretable due to fragmentation of the genome and technical biases; therefore, analysis of these data requires many computational approaches. First, for each sequenced individual, sequencing data are aligned and further processed to account for technical biases. Then, variant calling is performed to obtain information on the positions of genetic variants and their corresponding genotypes. Quality control (QC) is applied to identify individuals and genetic variants with sequencing errors. These procedures are necessary to generate accurate variant calls from sequencing data, and many computational approaches have been developed for these tasks. This review will focus on current widely used approaches for variant calling and QC.

Introduction

Genome sequencing enables discovery of nearly a complete genome sequence of an individual. While the first draft for human genome cost $2.7 billion in 2003 [1,2], the cost of genome sequencing has decreased at a rate faster than that of Moore's Law [3,4], and it has become considerably inexpensive as it currently costs less than $1000 to sequence a human genome [5]. Given the rapid decrease in cost and its ability to detect nearly all genetic variants in an individual genome, genome sequencing has become very popular in several fields of genetics such as clinical genetics [6,7], cancer genetics [8,9], population genetics [10,11], and genetic studies for complex diseases and traits [12,13]. Currently, there are several ongoing large-scale whole-genome sequencing (WGS) studies aimed at identifying genetic variants influencing a diverse range of human diseases and traits such as the Trans-Omics for Precision Medicine (TOPMed) [14], Genome Sequencing Program (GSP) [15], and whole-genome sequencing in psychiatric disorders (WGSPD) [16]. Each of these large-scale WGS studies involves at least 10 000 individuals while some have collected well over 100 000 individuals.

A primary goal of these sequencing-based studies is the identification of genetic elements that vary among individuals. These genetic elements, such as single nucleotide variants (SNVs) and small insertions or deletions (indels), associated with a disease or trait, may provide clues for understanding the genetic basis of a disease or trait and for identifying possible therapeutic targets. One key obstacle in these analyses is the technical noise associated with sequencing technologies. Specifically, technical biases from the sequencing process can introduce issues such as incorrectly reported sequences. Several platforms exist for sequencing the human genome, and each has unique technical biases, such as differing sequencing error rates [17]. These biases must be accounted for to accurately identify genetic elements from sequencing data and distinguish true genetic variation from technical noise.

This review will discuss computational approaches that have been developed to analyze genome sequencing data or next-generation sequencing (NGS) data. Specifically, it will focus on the two major topics in statistical genetics: (1) variant calling and (2) quality control (QC) of NGS data. Variant calling is a procedure to obtain genetic variants from NGS data; a variant is a position or a locus in

Received: 24 April 2019
Revised: 28 June 2019
Accepted: 16 July 2019

Version of Record published: 29 July 2019

Emerging Topics in Life Sciences (2019) **3** 399–409
https://doi.org/10.1042/ETLS20190007

the genome that differs from a reference genome. Variant calling is one of the most important steps in an analysis of NGS data because one major goal of genetic studies is to identify genetic variants that influence a phenotype of interest, and hence, it is important to discover as many true variants as possible in an analysis. QC is the next step after the variant calling procedure, and its goal is to filter out individuals and genetic variants with poor sequencing quality and to improve the quality of variant calls. Correct variant calls are critical in downstream analyses such as an association test to reduce false-positive findings. This review will discuss common practices used in variant calling and QC in genetic studies and some of the challenges in those procedures.

Variant calling

Variant calling process

This review will focus on the germline variant caller on SNVs and indels. Germline variants are present in gametes and can be passed onto offspring, while somatic variants are found in non-germline cells and cannot be inherited. Variant callers for structural variants (SVs) and somatic variants will be discussed briefly. As there are many different variant callers and pipelines, we will discuss one based on the genome analysis toolkit (GATK) pipeline [18–20] developed by the Broad Institute as it is one of the most widely used variant callers and pipelines; it is the primary variant caller for several large-scale WGS studies such as WGSPD, the Alzheimer's disease sequencing project [21], and the Genome Aggregation Database (gnomAD) [22]. This pipeline consists of several computational steps (Figure 1), which will be discussed below.

Map reads to a reference genome

When a sequencing machine processes a genome, it generates many *reads*, which are a short fragment of the genome, typically in length of 100 bp for short read sequencers [17]. Depending on the type of sequencing (e.g. whole-exome sequencing vs. WGS) and coverage (e.g. low coverage vs. high coverage), millions or billions of reads can be generated per genome. Information on reads (their sequences and quality scores) is stored in FASTQ files. We then need to find a location of the genome where each read originated from since this information is not retained during the sequencing procedure. For this purpose, we map reads to a reference genome with a known complete sequence. This procedure aims to identify a location in a reference genome where each read matches in sequence while tolerating some mismatches. There are many read mappers or aligners [23–25] such as BWA-MEM [26], a method that is often used in the GATK pipeline. BWA-MEM uses a Burrows–Wheeler transform-based algorithm [27] to map reads. Read mapping is one of the most time-consuming steps in the variant calling process. For high-coverage WGS, this step may use more than half of total computational time required for the entire process [28]. This read mapping information is stored in the binary alignment map (BAM) files.

Mark duplicates and recalibrate base quality scores

These two procedures are additional steps to improve the quality of variant calling in subsequent steps. The mark duplicates procedure marks reads that are duplicates, which reduces PCR duplication artifacts, and it is implemented in the Picard software. Identification of duplicate reads is important for downstream analyses that assume the measurements of a particular genomic position in a sample are independent. In addition, PCR duplicates can overrepresent certain sequences and can lead to false positives in variant calling if these sequences harbor an error. The Base Quality Recalibration Score step recalibrates a quality score associated with each base of a read, which is implemented in GATK. This processing step is meant to refine the quality score estimates produced by a sequencing machine which may be inaccurate. The GATK method utilizes known SNVs and covariates from the observed sequencing data, such as the original quality score, positions within reads, and nucleotide context, to more accurately model quality scores. Quality scores are important for variant calling, since they provide a measure of the confidence we have in a variant detected in a sequencing read; therefore, accurate quality scores are essential for these analyses. We apply the previous read alignment and these two procedures to sequence data of each individual and generate analysis-ready reads that can be used for a variant caller. These methods utilize read group information that typically indicates which sets of reads were generated from a single sequencing run. The purpose of this information is to allow for the detection of technical artifacts arising from multiple sequencing runs. We note that an indel realignment procedure was part of the GATK pipeline, which performs local realignment to correct potential mapping errors around

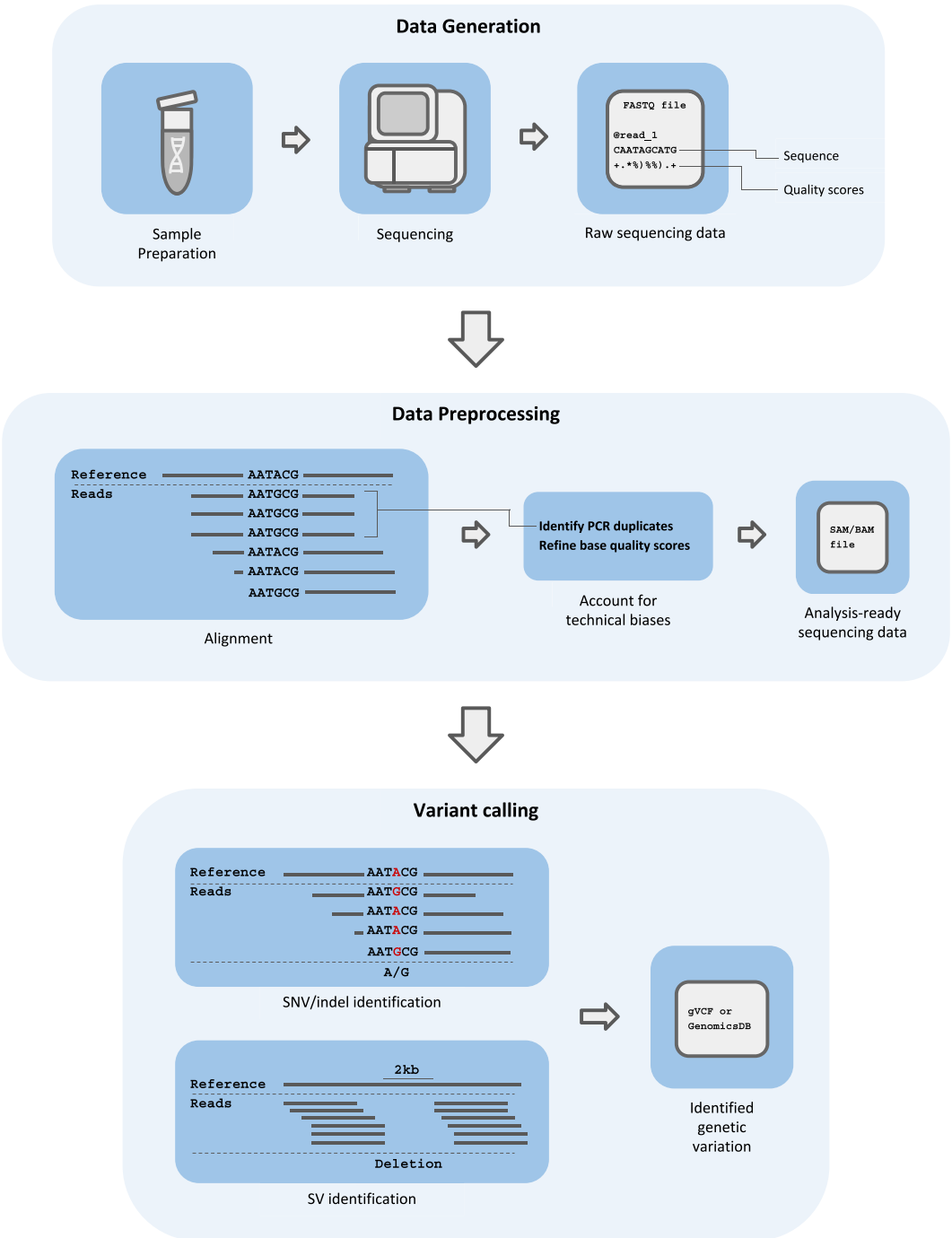

Figure 1. Generation of raw NGS data from a biological sample.

(Top) Sequencing is performed on fragments of the genome, and the data are typically stored in the FASTQ format, which includes the sequence information and machine-generated quality scores of each reported base. (Middle) Processing of NGS data includes an alignment step to recover the genomic positions of each sequenced fragment or read. Subsequent processing addresses technical biases that arise from the sequencing process, such as PCR duplicates and sequencing errors. (Bottom) Variant calling is performed on processed reads to identify sites of genetic variation. Here, we depict a simple example of detecting an SNV and a large deletion.

Emerging Topics in Life Sciences (2019) **3** 399–409
https://doi.org/10.1042/ETLS20190007

indels. This process is no longer necessary after the introduction of HaplotypeCaller that performs a haplotype assembly, which will be discussed next.

Haplotypecaller

Given a BAM file that contains mapping information of sequenced reads of a target genome that we want to sequence, the goal of a variant caller is to identify variants or positions of the genome that vary among samples. To achieve this goal, HaplotypeCaller performs a local *de novo* assembly by building a De Bruijn-like graph and identifies both SNVs and indels. Once it identifies those variants, the next step is to identify a genotype of each variant, which is to infer an allele for each chromosome. HaplotypeCaller calculates the likelihood of each possible genotype based on the read information and identifies the most likely genotype. The output file of HaplotypeCaller can be either a variant call format (VCF) file or a genomic VCF (gVCF) file. A VCF file contains information on all genetic variants detected and corresponding genotypes for individuals as well as several quality scores and depth information for genotypes. While one may include BAM files of multiple individuals and call variants together with HaplotypeCaller, which will generate a multi-sample VCF file, it is recommended that HaplotypeCaller is applied to a BAM file of each individual to generate a gVCF file for large sample sizes. A gVCF file contains not only genetic variants but also non-variant sites. Multiple gVCF files can be joint-genotyped using a GenotypeGVCFs command in GATK, which is more efficient than joint genotyping multiple BAM files using HaplotypeCaller. To further improve the efficiency, one may split many gVCF files into multiple batches where each batch of gVCF files can be combined into a multi-sample gVCF file using a CombineGVCFs command in GATK. One can then apply the GenotypeGVCFs to several multi-sample gVCF files from those batches and perform joint genotyping on a large number of individuals simultaneously. More recently, the GenomicsDB format was developed for efficient storage of variants and variant retrieval in GATK version 4. However, VCF files remain widely used.

Other variant callers and their performance

We focused on methods such as BWA-MEM and GATK due to their current dominance in variant calling pipelines for human NGS data. However, a number of other read aligners and variant callers for SNVs and indels have been developed, and several studies have compared their performance using either simulated or real genome sequencing data [29–33]. In those studies, FreeBayes [34] and Samtools [35] are variant callers that are most frequently evaluated in addition to GATK. FreeBayes is a haplotype-based variant caller for SNVs and indels which uses a Bayesian statistical framework, and Samtools also uses a Bayesian method to detect SNVs and indels. Results of those studies that compared the performance of different variant callers are somewhat discordant as some studies found that Samtools and Freebayes perform better than GATK, while other studies found the opposite results. This could be due to different genome sequencing data tested, different variant calling pipelines, or different software versions for variant callers in those studies. A more recent variant calling method, DeepVariant, utilizes deep learning to call variants from images of aligned short-reads [36]. A recent study found that the accuracy of SNV calling is similar across DeepVariant and GATK [37]; however, they observed improved precision in indel calling. As stated before, GATK remains the most widely used tool for variant calling from human genome sequencing data.

Challenges in variant calling

The key challenge in variant calling is distinguishing true genetic variation from technical artifacts. These artifacts can arise from the sequencing process or variant calling algorithms. The preprocessing steps discussed here (duplicate identification and base quality score recalibration) are some ways to address technological biases. False positives arising from variant calling algorithms are also a concern, especially for variants that are difficult to identify from short-read sequencing, such as SV. It is notoriously difficult to identify SVs accurately from short-read sequencing as each read may not span an entire SV. Numerous methods have been developed for SV calling that use different sources of information such as BreakDancer [38], cn.MOPs [39], CNVNator [40], DELLY [41], GenomeSTRiP [42], Hydra [43], LUMPY [44], and BreakSeq [45], but no single SV algorithm can identify all types of SVs with high accuracy [46]. Hence, recently a few methods have been developed to combine the results of multiple SV callers using an overlap approach or a machine learning algorithm and improve both precision and recall of SV detection [47–49]. In addition, while this review focuses on germline variant callers, there are also many variant callers for somatic variants such as those that differ between tumor and normal samples from the same individuals [50].

Another challenge in variant calling is its computational time; it may take a couple of weeks to perform the GATK variant calling on high-coverage WGS data of one individual using one CPU core [28]. Although some procedures in the GATK variant calling pipeline such as BWA-MEM support multiple threads to improve the computational efficiency, not all procedures support multiple threads and they become a bottleneck in the variant calling process. To overcome this challenge, a few computational approaches have been developed to speed up the GATK pipeline [28,51]. Those approaches divide a genome into smaller regions such as chromosomes or regions with fixed length (e.g. 10 Mbp) and utilize a high-performance cluster or a cloud-computing resource such as Amazon Web Service or Google Cloud to simultaneously call variants in those regions and improve the efficiency of the overall variant calling process. In addition to these approaches, fast variant callers such as Playtypus [52], Strelka2 [53], and Fuwa [54] have recently been proposed.

Quality control

There are two types of QC; one is individual-level QC that identifies problematic individuals and the other is variant-level QC that identifies genetic variants with poor sequencing quality. These two procedures can be performed independently, or one procedure can be performed after the other procedure is performed (e.g. one may perform variant-level QC after removing problematic individuals if there are many such individuals).

Individual-level QC

Genotype missing rate

This metric indicates the proportion of genotypes that an individual is missing. One may measure genotype missing rate after setting genotypes with low genotype quality (GQ) to missing; for a VCF generated with the GATK pipeline, genotypes with GQ $\leq$ 20 (or higher) may be set to missing. If an individual has high genotype missing rate (>5% or >10%), it may indicate low coverage or poor sequencing quality, and hence, this individual needs to be removed from the analysis.

Genotype concordance to microarray data

Studies may have already collected microarray genotype data for sequenced individuals, and in this case, one may compare genotypes between microarray and sequencing over SNVs present in both the platforms. It is expected that genotype concordance rate between the two genotyping platforms would be high (>99%). If the rate is ~50%, it may indicate sample swap (mixed up samples), and if the rate is below 90%, it may indicate contamination, which is discussed below. SNVs that are present in both microarray and sequencing are used to calculate genotype concordance rate, and it is important to pay attention to strand issues. For example, an SNV may have A and G alleles in microarray while it has T and C alleles in sequencing due to different strands genotyped or sequenced between microarray and sequencing. Also, as microarray data are often generated much earlier than sequencing, reference human genome builds may be different such as microarray data in hg18 and sequencing data in hg19. LiftOver can be performed to change the positions of SNVs in microarray data to match the reference genome builds of sequencing data.

Contamination

During sample preparation and manipulation for sequencing, DNA from multiple individuals may be present in the same library, which represents contamination of DNA. It is important to detect this contamination and remove individuals who are heavily contaminated. One approach to detect contamination is using the genotype concordance rate as described before because contaminated individuals may have lower genotype concordance rate to microarray data. Another approach is to use a software called VerifyBamID [55] that calculates a contamination level of each sequenced individual using either sequencing data only or both sequencing and microarray data.

Sequencing statistics

These measures include a variety of statistics describing each sequenced individual. For example, they are the number of SNVs/indels (known/novel), transition/transversion ratio (Ti/TV) (known/novel), the number of singletons, and the number of multi-allelic variants that each individual carries. The known number of SNVs or known Ti/Tv ratio is calculated from SNVs present in dbSNP while the novel number of SNVs or novel Ti/Tv ratio is calculated from those not present in dbSNP. In a homogeneous population, any individual whose statistics show outlier patterns may have sequencing problems and may need to be removed. The Ti/Tv ratio is expected to

Emerging Topics in Life Sciences (2019) **3** 399–409
https://doi.org/10.1042/ETLS20190007

be around two for WGS data; the known Ti/Tv ratio may be slightly above two while the novel Ti/Tv ratio may be lower than two (e.g. 1.4–1.8). Very low novel Ti/Tv ratio may indicate problems with sequencing.

Identical-by-descent (IBD) analysis

This analysis identifies a pair of individuals who are duplicates (e.g. sequenced twice or twins) or who are related (e.g. parent/child relationship and siblings). One often uses PLINK [56] '--genome' option to calculate IBD estimates quantified with a $\hat{\pi}$ value between every pair of individuals or one may use other software to calculate relationship among pairs of individuals such as KING [57]. When calculating $\hat{\pi}$ value in PLINK, it is important to use a set of independent SNVs, since SNVs in high linkage disequilibrium (LD) will lead to common haplotypes that will be identified as IBD though they were not due to a recent common ancestor; one may perform LD-pruning with '--indep' or '--indep-pairwise' option. If $\hat{\pi}$ value is close to 1, it suggests duplicate samples, and hence, one of the duplicates will need to be removed. For $\hat{\pi}$ value close to 0.5, it suggests the first degree of the relationship while $\hat{\pi}$ value close to 0.25 suggests the second degree of relationship. For case-control studies where only unrelated individuals are expected, one individual from each pair of related individuals needs to be removed. The individual in these pairs who has the disease or trait of interest is often prioritized. For family-based studies, a study may use $\hat{\pi}$ values to check whether a known pedigree structure is consistent with the pedigree structure inferred from $\hat{\pi}$ values. If there is inconsistency, it is important to correct the issue by checking whether there is an error in the known pedigree structure or whether wrong individuals are sequenced.

Principal component analysis

Principal component analysis (PCA) is often applied to identify the ethnicity of sequenced individuals (Figure 2). One of the most widely used tools for PCA is EIGENSTRAT [58]. Using the reference data set such as 1000 Genomes data set [59], principal components (PCs) estimated from PCA cluster sequenced individuals according to their ethnicities. One may draw PCA plots using these PCs (e.g. PC1 vs. PC2), and these plots will allow a study to identify outliers in terms of ethnicity. Those outliers may need to be removed from further analysis. When performing PCA, it is important to perform LD-pruning and remove related individuals since both local LD structure and direct relatedness can be a stronger source of variation in the genetic data than the population-level ancestry differences that are of interest. For family-based studies, founders who are unrelated may be included in PCA.

Sex check

One can check whether sex inferred from sequencing data using X chromosome is consistent with known sex from sample annotation using '--check-sex' option in PLINK. If there is inconsistency, one may need to check possible sample swap.

Variant-level QC

There are two main types of variant-level QC; one is a filtering approach based on several filters and the other is a classification approach using a machine learning model. One may use a combination of both by including only variants that pass both types of QC.

Filtering approach

This QC calculates several statistics about each variant and removes it if it fails one of the filters. This approach has been widely used in GWAS based on microarray data [60,61]. The main advantage of this approach is its simplicity, and the main disadvantage is that the threshold for each filter is usually arbitrarily determined. The following filters are often used.

- Genotype missing rate: Similar to genotype missing rate for each individual, one may also calculate the missing rate for each variant and remove it if it is too high (e.g. >5% or >10%).
- Hardy–Weinberg Equilibrium (HWE) *P*-value: This *P*-value measures the deviation from HWE that compares the frequency of observed genotypes from sequencing data and expected the frequency of genotypes from HWE for each SNV. SNVs with low HWE *P*-values (e.g. $<1 \times 10^4$) will need to be removed since they significantly deviate from HWE. It is important to note that HWE *P*-values need to be estimated in a homogeneous population that consist of unrelated and healthy controls. If there are individuals from different populations or related individuals, minor allele frequency estimation for HWE *P*-values may not be accurate.

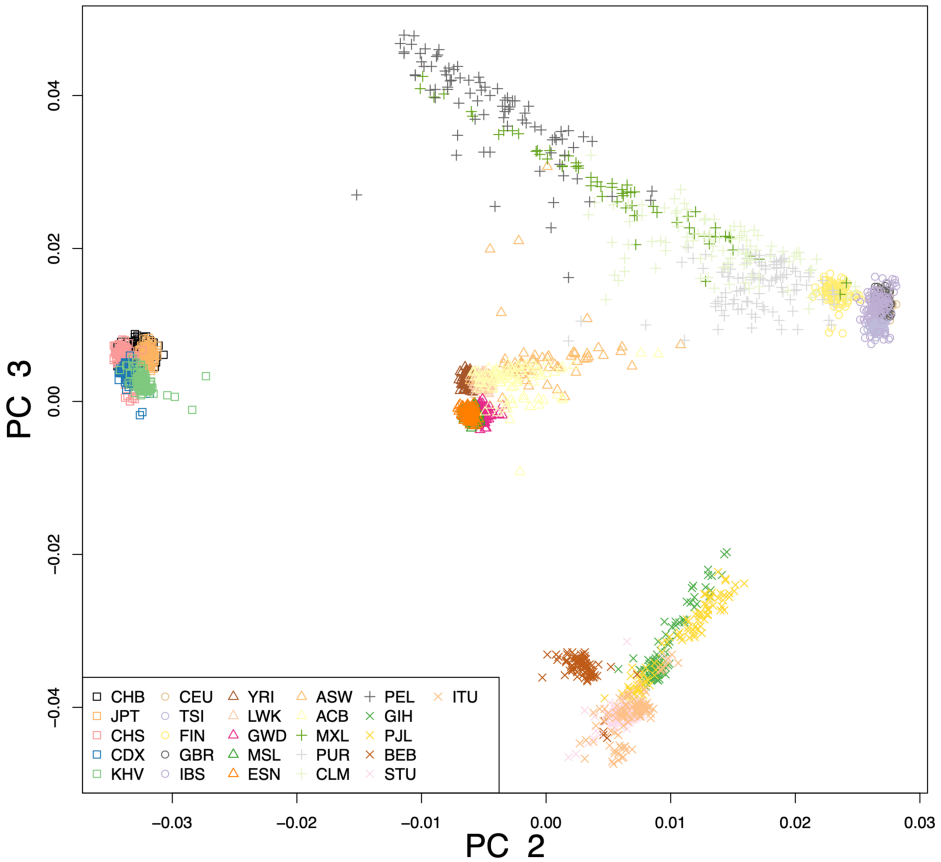

Figure 2. Principal component analysis (PCA) of 1000 Genomes data set.

PCA is performed after LD-pruning using EIGENSTRAT software. *X*-axis is PC2 and *Y*-axis is PC3. There are 26 populations from five super population codes (East Asian (EAS), European (EUR), African (AFR), Ad Mixed American (AMR), South Asian (SAS)). Each of the five super population codes is indicated with a different symbol (square for EAS, circle for EUR, triangle for AFR, + for AMR, and X for SAS).

In addition, genotypes among case individuals may not follow HWE, so their genotypes should not be included in the estimation of HWE *P*-values.

- Genotype concordance rate to microarray data: If microarray data are available, one may calculate genotype concordance rate between sequencing and microarray data for each SNV and remove those with low concordance rate. This filter, however, can only test those SNVs present in both sequencing and microarray data.
- Mendelian error rate: In family-based studies where there are trios or pairs of sequenced individuals, one may calculate Mendelian error rate of each SNV that indicates whether transmission of alleles from parents to offspring follows Mendelian inheritance patterns. SNVs with high Mendelian error rate will need to be removed (Figure 3).
- Allele balance of heterozygous calls (ABHet): This statistic measures the balance between reads supporting a heterozygous genotype. Specifically, it is calculated as the number of reads with a reference (or alternative) allele from an individual divided by the total number of reads from the individual for a heterozygous genotype (e.g. A/C and G/T). Ideally, ABHet should be near 0.5, and a study may remove an SNV with many individuals who have heterozygous genotypes with ABHet values much greater or smaller than 0.5 (Figure 3).

Classification approach

This approach attempts to determine whether a specific variant has high or low sequencing quality using a machine learning model. One example is Variant Quality Score Recalibration (VQSR) from GATK that uses a

Emerging Topics in Life Sciences (2019) **3** 399–409
https://doi.org/10.1042/ETLS20190007

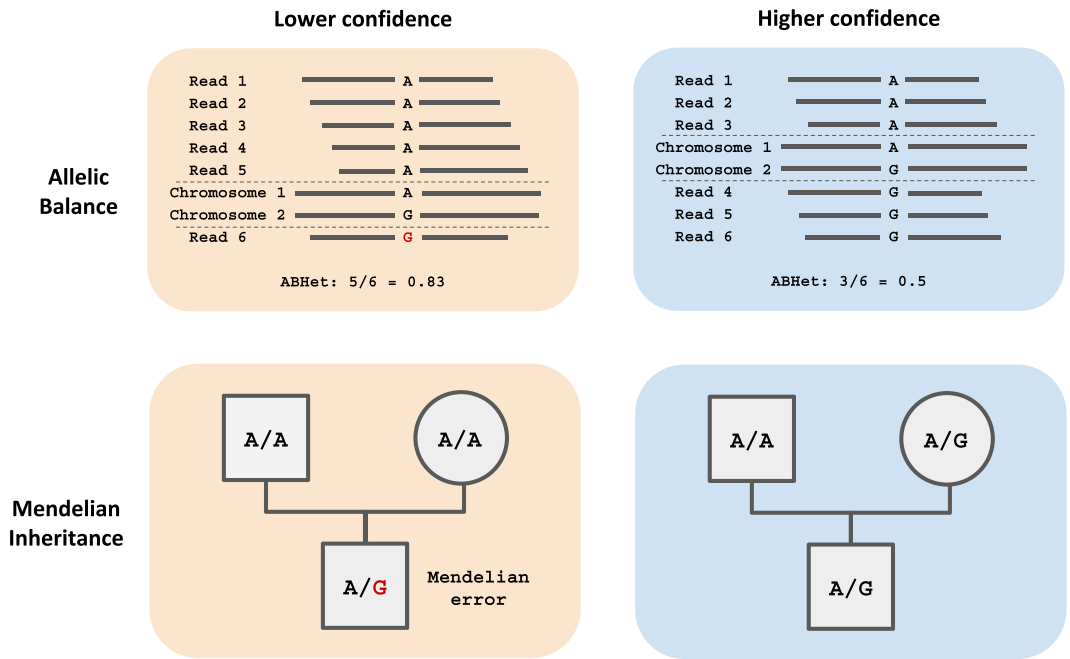

Figure 3. Quality control measures for genetic variants.

(Top left) A genomic position with a possible heterozygous A/G genotype. Its ABHet value, a measure of allelic balance, deviates from 0.5. As a result, this variant may be filtered out since it is likely due to a sequencing error. (Top right) The same genotype call but with ABHet exactly equal to 0.5, indicating that we are more confident in this identified variant. (Bottom left) An identified variant that does not follow Mendelian inheritance and is thus labeled a Mendelian error since the G allele could not be inherited from either parent. This variant could be either a sequencing error or *de novo* mutation and may be filtered out. (Bottom right) An identified variant that follows Mendelian inheritance, which provides more confidence in this call.

Gaussian mixture model [18]. To train their model, it needs a training set that contains true positives (variants that are highly validated to be true), and it uses SNVs found in known databases such as HapMap [62], Omni 2.5M SNP Chip, and 1000 Genomes [59]. After training the model and applying it to sequence data, a study may filter out variants depending on its desired sensitivity level. One potential issue with the classification approach is that the known databases may not be accurate, which may cause inaccurate classification of variants. Another issue is that those databases may not be available for certain species other than human. Lastly, the classification approach based on VQSR may require more computational resources than the filtering approach.

Validating variants

Even with these preprocessing and QC methods, false positives can still occur. There are several experiments for validating identified genetic variants. For example, PCR amplification of regions containing a putative SNV, indel, or SV breakpoints can be verified by Sanger sequencing. This sequencing method is slow and expensive compared with NGS; however, it has lower error rates and, furthermore, the targeted PCR amplification yields high coverage for validating a variant. As another example, fluorescence *in situ* hybridization can be used to validate SVs, such as CNVs. This method utilizes fluorescent probes that hybridize with genomic sequences and can be visualized at the chromosomal level.

Summary

- Genome sequencing has become increasingly popular in several fields of genetics due to the rapid decrease in sequencing cost and its ability to discover nearly a complete genome sequence of an individual genome.

- Variant calling from sequencing data involves many processing steps to accurately identify genetic variants. The GATK best practice pipeline is one of the most widely used variant calling approaches, which consists of read alignment, duplicate read identification, base quality score recalibration, and HaplotypeCaller.
- Quality control (QC) is performed after a variant calling to detect and remove individuals and genetic variants with sequencing errors.
- Individual-level QC removes individuals with high genotype missing rate, low genotype concordance rate to microarray data, contamination, outlier patterns in sequencing statistics, or wrong sex. It also identifies related individuals using the IBD analysis and population outliers using PCA.
- Variant-level QC can be performed using the traditional filtering approach and/or the classification approach using machine learning algorithms. The filtering approach uses several filters such as genotype missing rate, HWE *P*-values, genotype concordance rate to microarray data, Mendelian error rate, and ABHet.

Abbreviations

ABHet, allele balance of heterozygous; ADSP, Alzheimer's Disease Sequencing Project; BAM, Binary Alignment Map; FISH, fluorescence *in situ* hybridization; GATK, Genome Analysis Toolkit; GQ, genotype quality; GSP, Genome Sequencing Program; HWE, Hardy–Weinberg Equilibrium; LD, linkage disequilibrium; NGS, next-generation sequencing; PCA, principal component analysis; PCs, principal components; QC, quality control; SNVs, single nucleotide variants; SVs, structural variants; VCF, variant call format; VQSR, Variant Quality Score Recalibration; WGS, whole-genome sequencing; WGSPD, Whole-Genome Sequencing in Psychiatric Disorders.

Funding

B.J. was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650604. J.H.S. was supported by the National Institute of Environmental Health Sciences (NIEHS) grant [K01ES028064], National Institute of Neurological Disorders and Stroke (NINDS) [R01NS102371], and the National Science Foundation grant [#1705197].

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

References

1 The Human Genome Project Completion: Frequently Asked Questions. (2010) https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/
2 Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 https://doi.org/10.1038/35057062
3 DNA Sequencing Costs: Data. (2018) https://www.genome.gov/sequencingcostsdata/
4 Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L. et al. (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 https://doi.org/10.1186/s13059-016-0917-0
5 Heusel, J. and Richards, N. (2018) Now we can cheaply sequence DNA, how do we store all that data? https://www.wired.co.uk/article/precision-medicine
6 Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A. et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 https://doi.org/10.1056/NEJMoa1306555
7 Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. et al. (2011) Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 https://doi.org/10.1038/nrg3031
8 International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C. et al. (2010) International network of cancer genome projects. *Nature* **464**, 993–998 https://doi.org/10.1038/nature08987
9 Mwenifumbo, J.C. and Marra, M.A. (2013) Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14**, 321–332 https://doi.org/10.1038/nrg3445

10 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 https://doi.org/10.1038/nrg3625
11 Jobling, M.A. and Tyler-Smith, C. (2017) Human Y-chromosome variation in the genome-sequencing era. *Nat. Rev. Genet.* **18**, 485–497 https://doi.org/10.1038/nrg.2017.36
12 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 https://doi.org/10.1038/nrg2779
13 Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J. et al. (2012) Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623–630 https://doi.org/10.1038/ng.2303
14 Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R. et al. (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *bioRxiv* https://doi.org/10.1101/563866
15 The NHGRI Genome Sequencing Program (GSP). (2018) https://www.genome.gov/10001691/nhgri-genome-sequencing-program-gsp/
16 Sanders, S.J., Neale, B.M., Huang, H., Werling, D.M., An, J.Y., Dong, S. et al. (2017) Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 https://doi.org/10.1038/s41593-017-0017-9
17 Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 https://doi.org/10.1038/nrg.2016.49
18 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 https://doi.org/10.1101/gr.107524.110
19 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C. et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 https://doi.org/10.1038/ng.806
20 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A. et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.0.1–11.0.33 https://doi.org/10.1002/0471250953.bi1110s43
21 Beecham, G.W., Bis, J.C., Martin, E.R., Choi, S.H., DeStefano, A.L., van Duijn, C.M. et al. (2017) The Alzheimer's Disease Sequencing Project: study design and sample selection. *Neurol. Genet.* **3**, e194 https://doi.org/10.1212/NXG.0000000000000194
22 Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q. et al. (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* https://doi.org/10.1101/531210[]
23 Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 https://doi.org/10.1186/gb-2009-10-3-r25
24 Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 https://doi.org/10.1093/bioinformatics/btn025
25 Liu, Y., Schmidt, B. and Maskell, D.L. (2012) CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. *Bioinformatics* **28**, 1830–1837 https://doi.org/10.1093/bioinformatics/bts276
26 Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 https://doi.org/10.1093/bioinformatics/btp324
27 Burrows, M. and Wheeler, D. (1994) *Technical Report 124*, Digital Equipment Corporation, Palo Alto, CA
28 Kelly, B.J., Fitch, J.R., Hu, Y., Corsmeier, D.J., Zhong, H., Wetzel, A.N. et al. (2015) Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol.* **16**, 6 https://doi.org/10.1186/s13059-014-0577-x
29 Cornish, A. and Guda, C. (2015) A comparison of variant calling pipelines using genome in a bottle as a reference. *Biomed. Res. Int.* **2015**, 456479 https://doi.org/10.1155/2015/456479
30 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 https://doi.org/10.1371/journal.pone.0075619
31 Highnam, G., Wang, J.J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N. et al. (2015) An analytical framework for optimizing variant discovery from personal genomes. *Nat. Commun.* **6**, 6275 https://doi.org/10.1038/ncomms7275
32 Ni, G., Strom, T.M., Pausch, H., Reimer, C., Preisinger, R., Simianer, H. et al. (2015) Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. *BMC Genomics* **16**, 824 https://doi.org/10.1186/s12864-015-2059-2
33 Hwang, S., Kim, E., Lee, I. and Marcotte, E.M. (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5**, 17875 https://doi.org/10.1038/srep17875
34 Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv 12073907
35 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 https://doi.org/10.1093/bioinformatics/btp352
36 Poplin, R., Chang, P.C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A. et al. (2018) Creating a universal SNP and small indel variant caller with deep neural networks. *bioRxiv* https://doi.org/10.1101/092890[]
37 Supernat, A., Vidarsson, O.V., Steen, V.M. and Stokowy, T. (2018) Comparison of three variant callers for human whole genome sequencing. *Sci. Rep.* **8**, 17851 https://doi.org/10.1038/s41598-018-36177-7
38 Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S. et al. (2009) Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6**, 677–681 https://doi.org/10.1038/nmeth.1363
39 Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U. et al. (2012) Cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 https://doi.org/10.1093/nar/gks003
40 Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 https://doi.org/10.1101/gr.114876.110
41 Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V. and Korbel, J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 https://doi.org/10.1093/bioinformatics/bts378

42 Handsaker, R.E., Korn, J.M., Nemesh, J. and McCarroll, S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 https://doi.org/10.1038/ng.768
43 Quinlan, A.R., Clark, R.A., Sokolova, S., Leibowitz, M.L., Zhang, Y., Hurles, M.E. et al. (2010) Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* **20**, 623–635 https://doi.org/10.1101/gr.102970.109
44 Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 https://doi.org/10.1186/gb-2014-15-6-r84
45 Lam, H.Y., Mu, X.J., Stutz, A.M., Tanzer, A., Cayting, P.D., Snyder, M. et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 https://doi.org/10.1038/nbt.1600
46 Tattini, L., D'Aurizio, R. and Magi, A. (2015) Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol.* **3**, 92 https://doi.org/10.3389/fbioe.2015.00092
47 Becker, T., Lee, W.P., Leone, J., Zhu, Q., Zhang, C., Liu, S. et al. (2018) FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.* **19**, 38 https://doi.org/10.1186/s13059-018-1404-6
48 Zarate, S., Carroll, A., Krasheninina, O., Sedlazeck, F.J., Jun, G., Salerno, W. et al. (2018) Parliament2: fast structural variant calling using optimized combinations of callers. *bioRxiv* 424267 https://doi.org/10.1101/424267
49 Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W.L. et al. (2018) A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am. J. Hum. Genet.* **102**, 142–155 https://doi.org/10.1016/j.ajhg.2017.12.007
50 Xu, C. (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 https://doi.org/10.1016/j.csbj.2018.01.003
51 Lam, H.Y., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R. et al. (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat. Biotechnol.* **30**, 226–229 https://doi.org/10.1038/nbt.2134
52 Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Consortium, W.G.S. et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 https://doi.org/10.1038/ng.3036
53 Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Kallberg, M. et al. (2018) Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 https://doi.org/10.1038/s41592-018-0051-x
54 Li, Z., Wang, Y. and Wang, F. (2018) A study on fast calling variants from next-generation sequencing data using decision tree. *BMC Bioinformatics* **19**, 145 https://doi.org/10.1186/s12859-018-2147-9
55 Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R. et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet.* **91**, 839–848 https://doi.org/10.1016/j.ajhg.2012.09.004
56 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 https://doi.org/10.1086/519795
57 Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 https://doi.org/10.1093/bioinformatics/btq559
58 Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 https://doi.org/10.1038/ng1847
59 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M. et al. (2015) A global reference for human genetic variation. *Nature* **526**, 68–74 https://doi.org/10.1038/nature15393
60 Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. (2010) Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 https://doi.org/10.1038/nprot.2010.116
61 Turner, S., Armstrong, L.L., Bradford, Y., Carlson, C.S., Crawford, D.C., Crenshaw, A.T. et al. (2011) Quality control procedures for genome-wide association studies. *Curr. Protoc. Hum. Genet.* **68**, 1.19.1–1.19.18 https://doi.org/10.1002/0471142905.hg0119s68[]
62 International HapMap Consotium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L. et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 https://doi.org/10.1038/nature06258