Metadata of the chapter that will be visualized online

Chapter Title	Quality Control of Common and Rare Variants	
Copyright Year	2018	
Copyright Holder	Springer Science+Business Media, LLC, part of Springer Nature	
Corresponding Author	Family Name	**Panoutsopoulou**
	Particle	
	Given Name	**Kalliope**
	Suffix	
	Organization	Wellcome Trust Sanger Institute
	Address	Hinxton, Cambridgeshire, UK
	Email	kp6@sanger.ac.uk
Author	Family Name	**Walter**
	Particle	
	Given Name	**Klaudia**
	Suffix	
	Organization	Wellcome Trust Sanger Institute
	Address	Hinxton, Cambridgeshire, UK

Abstract

Thorough data quality control (QC) is a key step to the success of high-throughput genotyping approaches. Following extensive research several criteria and thresholds have been established for data QC at the sample and variant level. Sample QC is aimed at the identification and removal (when appropriate) of individuals with (1) low call rate, (2) discrepant sex or other identity-related information, (3) excess genome-wide heterozygosity and homozygosity, (4) relations to other samples, (5) ethnicity differences, (6) batch effects, and (7) contamination. Variant QC is aimed at identification and removal or refinement of variants with (1) low call rate, (2) call rate differences by phenotypic status, (3) gross deviation from Hardy-Weinberg Equilibrium (HWE), (4) bad genotype intensity plots, (5) batch effects, (6) differences in allele frequencies with published data sets, (7) very low minor allele counts, (8) low imputation quality score, (9) low variant quality score log-odds, and (10) few or low quality reads.

Keywords (separated by '-')

Genome-wide association study - Whole genome sequencing - Sample quality control - Variant quality control

Author's Proof

Chapter 3

Quality Control of Common and Rare Variants

Kalliope Panoutsopoulou and Klaudia Walter

AU1

Abstract

Thorough data quality control (QC) is a key step to the success of high-throughput genotyping approaches. Following extensive research several criteria and thresholds have been established for data QC at the sample and variant level. Sample QC is aimed at the identification and removal (when appropriate) of individuals with (1) low call rate, (2) discrepant sex or other identity-related information, (3) excess genome-wide heterozygosity and homozygosity, (4) relations to other samples, (5) ethnicity differences, (6) batch effects, and (7) contamination. Variant QC is aimed at identification and removal or refinement of variants with (1) low call rate, (2) call rate differences by phenotypic status, (3) gross deviation from Hardy-Weinberg Equilibrium (HWE), (4) bad genotype intensity plots, (5) batch effects, (6) differences in allele frequencies with published data sets, (7) very low minor allele counts, (8) low imputation quality score, (9) low variant quality score log-odds, and (10) few or low quality reads.

Key words Genome-wide association study, Whole genome sequencing, Sample quality control, Variant quality control

Uncorrected Proof

1 Introduction

High-throughput approaches such as genome-wide association scans (GWAS) and whole genome sequencing (WGS) technologies are used to interrogate the genotypes of tens of thousands of individuals at hundreds of thousands or millions of sites across the genome for association with diseases or other complex traits. Rigorous quality control (QC) at the sample and variant level is crucial to the success of the study because it can dramatically reduce the number of false positive or false negative findings down the line. Extensive research over the past 10 years in the field of GWAS has established several commonly accepted criteria and thresholds for sample and variant QC after the genotype calling process [1, 2]. Most of these quality control steps are applicable to sequencing data but additional filters have, and will constantly be developed as these technologies evolve. Here, we describe the most commonly applied sample and variant QC steps in datasets from GWAS and low-depth WGS studies. We recommend that most of the QC

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793,
https://doi.org/10.1007/978-1-4939-7868-7_3, © Springer Science+Business Media, LLC, part of Springer Nature 2018

Author's Proof

steps, and in particular the example thresholds that are presented here based on previous research are tested for suitability and adapted to each study.

2 Sample Quality Control

The aim of performing sample quality control is to remove low-quality samples often caused by poor DNA quality and/or insufficient quantity and/or contamination; and to identify individuals with discordant information based on other sources, accidental swaps, samples that show batch effects, duplicated and related samples and ethnic outliers. It is recommended that QC at the sample level is best carried out before variant QC because it can adversely influence variant QC metrics. In addition, sample QC metrics can also be influenced by bad quality variants so variants with high missing genotype rates should not be taken into consideration when calculating these metrics. This can be achieved by pre-filtering the dataset for bad quality variants before proceeding to sample QC. With the exception of the sex determination QC all other sample QC steps are carried out using autosomal SNPs only.

Uncorrected Proof

2.1 Sample Call Rate

The proportion of missing genotypes per sample is a good indicator of DNA quality. Samples with high proportion of missing genotypes (i.e., low call rate) will typically fail other sample QC metrics and if they are not removed from the data they could lead to spurious associations. Previous GWAS studies have excluded subjects with missing genotype rate greater than 2%–5%. However, because this threshold depends on several study-specific factors an empirical threshold should be determined by examining the distribution of the missing genotype proportion per individual across all study samples.

2.2 Sex Discrepancies and Other Identity Checks

Self-reported sex is usually available from subject enrolment but the sex of an individual can also be inferred from X chromosome genetic data. Discrepancies between these two sources of information may indicate sample swaps or sample contamination or incorrect data entry for self-reported sex. These can be investigated further by feeding back conflicting sex information to the collection centers. Having the correct sex information is also important in studies where sex is included as a covariate in the analysis or to stratify males and females for calculating effect sizes in separate in studies of sexual dimorphic traits.

Before a genotyping or sequencing experiment takes place, some labs run smaller-scale marker assays in Sequenom MassARRAY iPLEX and Fluidigm platforms. Sex determination markers contained in these platforms can be used to estimate genetic sex and this can serve as a basic concordance test between genetically

Author's Proof

estimated and self-reported sex information. However, typically in GWAS or WGS experiments sex is inferred by calculating mean homozygosity across all variants on the X chromosome. Women have two copies of the X chromosome whereas males have only one copy so they cannot be heterozygous for typed variants on this chromosome. The most commonly used quality control software (PLINK) [3, 4] will call a sample male if the X chromosome homozygosity rate is more than 0.8; a female call is made if this estimate is less than 0.2. Samples that fall between these two thresholds are ambiguous and often this correlates with poor call rate and/or contamination. In rare instances this can be attributed to chromosomal abnormalities.

Further checks for sample identity can be performed by checking concordance of genotypes for the same individuals at a set of variants genotyped in more than one platform. For example, genotype concordance of a panel of variants from Sequenom/Fluidigm platforms can be checked against genotypes derived from GWAS or WGS for the same individuals at these markers. And genotypes derived from a sequencing experiment can be compared against genotypes derived from a GWAS experiment if these exist for the same or a subset of common individuals. When enough overlapping markers are available the degree of relatedness between samples can be estimated by calculating genome-wide IBD as described in the relatedness QC section.

Uncorrected Proof

2.3 Heterozygosity

Excess genome-wide heterozygosity is also a very good indicator of poor DNA quality and/or sample contamination. In the case of rare SNPs, excess heterozygosity can also be caused by differences in ethnicity of the samples assayed. On the other hand, excess genome-wide homozygosity may indicate some degree of inbreeding.

The mean genome-wide heterozygosity of a sample is the fraction or the proportion of non-missing genotypes that are heterozygous in relation to all the genotypes. This metric is platform- and sample-specific; it varies according to the marker content, the proportion of rare to common variants that have been assayed and the population examined. The threshold is therefore best determined by examining the distribution of mean genome-wide heterozygosity of all samples separately for common and rare SNPs. A reasonable approach is to remove samples that are plus or minus 3 standard deviations from the mean as shown in Fig. 1.

2.4 Relatedness

Having related individuals in the data may be desirable due to the study design (for example family-based studies or isolated populations) but can also be introduced accidentally (cryptically related and/or duplicated samples). Estimating relatedness with genetic data is an important step in the QC process; the goal is to validate known (recorded) relationships, to identify pedigree errors, to

Author's Proof

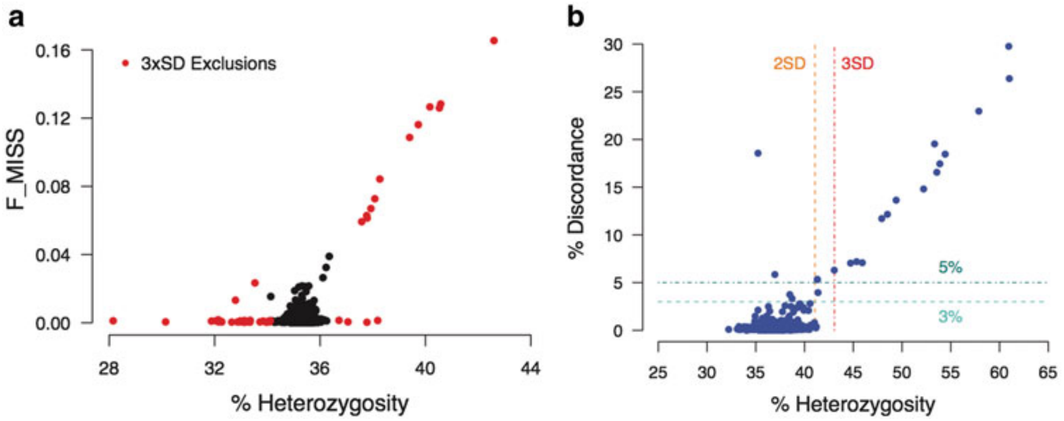

Fig. 1 (**a**) Heterozygosity versus call rate. Individuals with mean heterozygosity more or less than three standard deviations (SD) from the mean are labeled in red. (**b**) Discordance is strongly correlated with heterozygosity, where discordance (in %) is calculated from a comparison between sequenced and genotyped variants (modified from the UK10K cohorts study). The lines at 2SD and 3SD mark the two times and three times standard deviation from the sample mean of the heterozygous rate and the lines at 3% and 5% show cutoffs for the discordance rate. A threshold at 3SD would capture more or less samples with a discordance rate of >5%

Uncorrected Proof

decide on the analysis strategy that correctly accounts for related/duplicated samples, or to remove the related/duplicated pairs (usually one individual from a related pair) from downstream analysis. For family-based studies differences between recorded and estimated relationships could indicate sample swaps or adoption, misattributed paternity, etc. For case-control and population-based cohorts cryptically related or accidentally duplicated individuals can significantly inflate the significance of the association study results. These individuals are either removed from the analysis or kept in, but the data will then require analysis with approaches that appropriately account for relatedness, for example linear mixed models (LMMs).

In a homogeneous sample, the degree of relatedness between samples can be estimated by calculating genome-wide IBD (identity-by-descent) given IBS (identity-by-state) information. IBS is a term used to describe two identical alleles or two identical segments or sequences of DNA. An IBS segment is identical by descent in two or more individuals if they have inherited it from a common ancestor without recombination. Duplicated samples and monozygotic twins are expected to share 2 alleles IBD at every locus so the proportion of IBD equals 1, for parent-offspring pairs IBD is 0.5 and this value halves for second-degree (0.25), third-degree relatives (0.125), and so on. IBS/IBD calculations are affected by linkage disequilibrium (LD) so it is recommended to remove highly correlated markers by a method called LD-pruning as well as complex regions such as the MHC (Major Histocompatibility Complex) region before the IBD calculations take place. In practise,

Author's Proof

because of fluctuations that can be introduced by the LD structure and by genotyping/sequencing errors the threshold of the proportion of IBD > 0.9 is used to identify individuals that are duplicated and the threshold of the proportion of IBD > 0.2 is used to identify individuals that are second-degree or closer relatives. In outbred populations, samples that may show an unexpectedly large number of relationships with other samples at even lower IBD thresholds may indicate subtle contamination.

2.5 Ethnicity

Population stratification can be a major confounding factor in genetic association studies. If undetected, it can lead to inflation of the test statistic and false positive associations due to the differences in allele frequency between the different populations. To guard against it, studies in outbred populations try to match individuals for broad ethnic background upon recruitment and then rely on statistical approaches to remove ethnic outliers or to correct for subtle population stratification. We present below two of the most commonly used approaches to identify and remove ethnic outliers and admixed individuals.

Ethnic outliers can be identified by principal component analysis (PCA) [5] or multidimensional scaling approaches (MDS) [3] which cluster individuals depending on their genetic similarity. Genetic data from sampled individuals can be analyzed alone or merged with genetic data from samples of known ethnicity from source populations or publically available datasets. Publically available datasets comprising samples with known ethnicities are getting larger and more diverse; the widely used 1000 Genomes Project data contains genotypes of 2504 individuals from 26 populations [6]. Clustering of samples can be visualized onto a two-dimensional projection on axes of genetic variation termed principal components. Ethnic outliers are typically removed from the dataset but more subtle population stratification may not be picked up during this step; however, it can be corrected or accounted for downstream of the QC process. For example, including principal components as covariates in the association analysis, genomic control, linear mixed models, and LD score regression are approaches that can correct for subtle population stratification.

For whole genome sequence data the number of singletons per sample can also be used to identify samples with different ancestry. In general, there is a positive correlation between the number of singletons called and the read coverage (or depth) of the sequenced fragments, where read coverage or depth means how many sequenced fragments overlap each nucleotide on average after alignment to a reference genome. However, samples from different ancestries will appear as outliers when plotting the number of singletons versus average read depth for each sample (Fig. 2).

Uncorrected Proof

Author's Proof

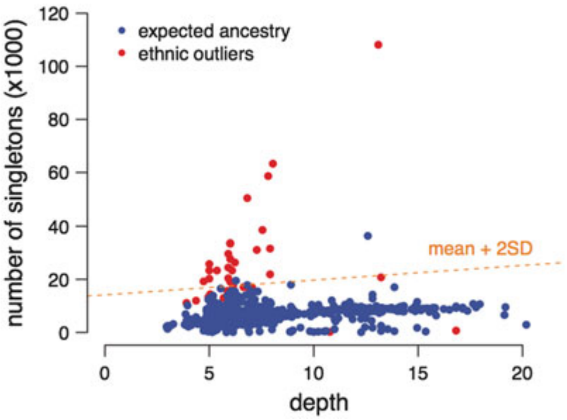

Fig. 2 Depth versus number of singletons. Samples with different ancestries, depending on the population, might be distinguished by a higher number of singletons, i.e., variants that are not shared with other samples in the cohort (modified from the UK10K cohorts study)

Uncorrected Proof

2.6 Batch Effects

Batch effects between samples in a single experiment can introduce bias in the analysis and lead to noise and false positive associations. Batch effects can be introduced by several sources, for example due to different sources of DNA (saliva vs. blood), different collections, DNA extraction, genotyping or sequencing centers, as well as different chips and sequencing platforms available. Batch effects are highly undesirable and best avoided by careful planning at the start of the study. Often, studies combine samples post-hoc and genotyping/sequencing processes are carried out in batches over a long period of time making the introduction of these effects unavoidable. QC fails partitioned per plate can identify batch effects for samples on different plates. Fortunately, gross batch effects are picked up by PCA or MDS; the principal components that are capturing this can be used as covariates in the analysis to eliminate some of this variation. Samples that have been genotyped/sequenced in duplicate could be useful to detect suspected batch effects particularly if these are more subtle. It is also possible to identify a subset of genotypes that cause this bias and remove these markers from further analysis as described in the variant QC section.

2.7 Sequence-Specific Checks for Sample Contamination

Additional checks for sample contamination can be performed for WGS data. For example, if array-based genotypes are available, it is possible to estimate the degree of the sample contamination and even to detect the source of the contamination by calculating likelihoods based on two-sample mixture models with the publicly available software VerifyBamID (http://genome.sph.umich.edu/wiki/VerifyBamID) [7]. VerifyBamID requires two input files, a file in VCF format that contains external genotypes or allele frequency information, and a file in BAM format that contains the

Author's Proof

sequenced reads. There are two options available, free-mix and chip-mix. The first option, free-mix, can be used for estimating contamination if only sequence data are available, and the second option, chip-mix, can be used for estimating contamination or sample swaps when also array-based genotype data are available. If CHIPMIX $\gg 0.02$ and/or FREEMIX $\gg 0.02$, it means that 2% or more of non-reference bases are observed in reference sites. In those cases, it is recommended to inspect the data more carefully for the possibility of contamination.

An alternative way to check for sample contamination is to compare the genotypes from the sequence data with the genotypes from existing GWAS data. If the overall discordance or the non-reference discordance (NRD) appears to be high between the two data sources, then this also points to sample contamination (Fig. 1). The NRD is calculated only from the non-reference (or alternative) genotypes, which usually represent the minor alleles, but not exclusively. In a variant call set based on sequenced reads the reference allele (REF) and the alternative allele (ALT) are clearly allocated, since reference genomes are being used for aligning the sequenced reads from next generation sequencing platforms. Mostly ALT will be the minor allele, but in some cases it will be the major allele. Often a few samples will be contaminated and they will appear as outliers. However, if the outliers appear as a smear or as a long tail of the main distribution, it might reveal a widespread low level sample contamination which should be examined more closely.

3 Variant Quality Control

Variant QC usually follows after the individuals that fail sample QC have been removed from the dataset. As with sample QC, variant QC is performed to ensure that only high-quality variants are included in downstream analysis. The main steps are described below.

3.1 Genotype Call Rate

As with sample call rate, variants with high degree of missingness across study samples constitute low-quality variants that can introduce false positive associations and hinder the identification of truly associated variants. To determine an appropriate threshold, the distribution of missing data proportion for each variant should be examined. Typically, GWAS studies exclude variants with missing call rate above 2%–5%. For low-frequency or rare variants a more stringent threshold is recommended and this is typically set at 1%.

Uncorrected Proof

Author's Proof

3.2 Call Rate Differences by Phenotypic Status

Spurious associations can be introduced when call rate differs significantly by case/control status [8]. This can be examined with a chi square test of non-random missingness in cases versus controls. Removal of variants with $p < 10^{-4}$ has been reported in the literature.

3.3 Deviation from Hardy-Weinberg Equilibrium (HWE)

In a relatively homogeneous population, gross departures from HWE can be indicative of genotyping error. This is evaluated by calculating Hardy-Weinberg test statistics for each variant using an exact test. However, departures from HWE may also be due to selection and therefore, in a case-control study this QC step is usually performed in controls. Various HWE *p*-value exact thresholds have been employed in GWAS ranging from less stringent to more stringent ($p < 5 \times 10^{-12}$ to $p < 0.0001$) and studies have chosen to either remove the variants that fail this filter or flag them for further scrutiny.

Uncorrected Proof

3.4 Genotype Cluster Plots

Genotype calling algorithms vary in their ability to call common and rare variants correctly. Therefore, for each associated variant one needs to scrutinize its genotype cluster plots. These are scatter plots of normalized probe intensities for each individual. For a bi-allelic common variant a good quality cluster plot is expected to show three clearly distinct clusters: one for the individuals who are homozygotes for the major allele, one for the heterozygotes and one for the homozygotes of the minor allele (Fig. 3). Upon

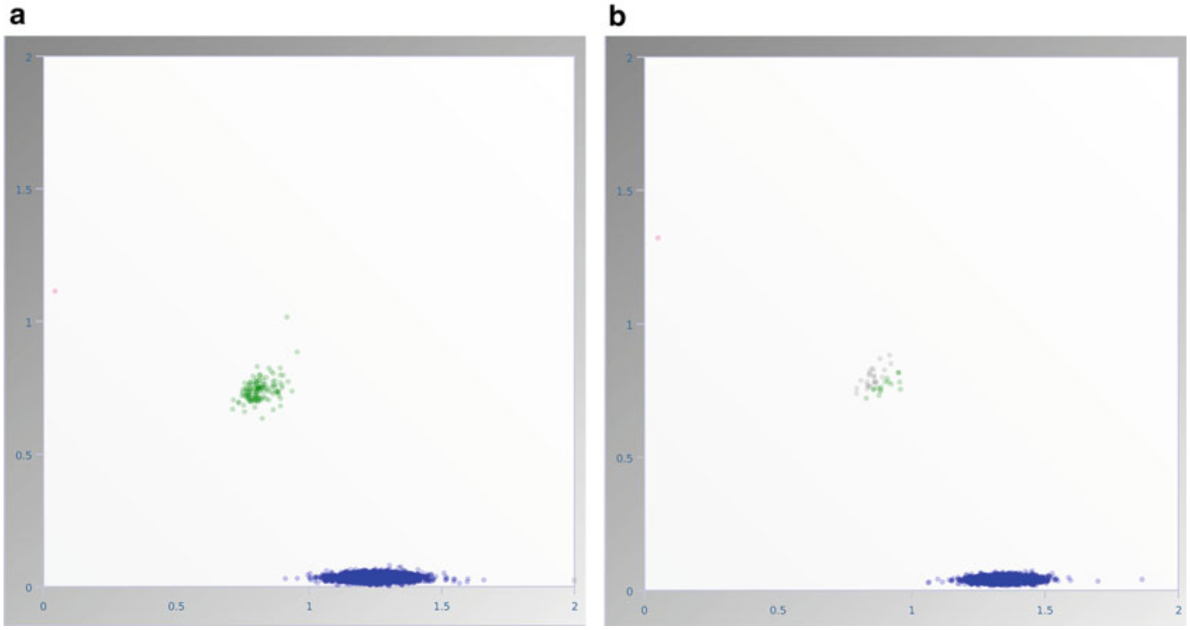

Fig. 3 Genotype intensity (cluster) plots for a rare variant. Depicted in blue are the individuals that are homozygotes for the major allele (AA), in green are the heterozygotes (AB), and in red is the homozygote for the minor allele (BB). Missing calls are depicted in gray. (**a**) Shows a good cluster plot (**b**) shows a bad cluster plot where several heterozygotes have not been called

Author's Proof

visual inspection variants with overlapping clusters and/or samples that have not been called or have been incorrectly assigned to a cluster should be removed from the analysis. Genotype calling is even more problematic for rare variants. The minor allele cluster may be composed of none or a few calls and any missing or incorrectly assigned calls for rare variants will have a bigger effect on the apparent association with a trait or disease. Therefore, it is recommended that removal of rare variants based on imperfect clustering is more stringent than for common variants.

3.5 Variants Causing Batch Effects

As discussed in the sample QC section there are instances where batch effects could be alleviated by removing the variants that cause them, obviating the need for correcting for batch effects on a genome-wide scale. As examples we present two different approaches that were used to remove batch effects in two high profile GWAS and WGS studies, the African Genome Variation Project [9] and the UK10K project [10]. In the former, principal component analysis showed clear batch effects between samples that were typed on two versions of the Illumina HumanOmni 2.5 M platform, the octo and the quad Beadchips. The principal components that captured this separation were identified and SNP loadings were calculated along these principal components in order to remove highly weighted SNPs. The authors checked the correlation of SNP weights and genotype discrepancy between a subset of samples that were typed on both platforms and found this to be highly correlated. Subtle chip effects and/or chip effects at rare variants may not be picked up by the PCA approach. For example, panel A of Fig. 3 shows the genotype calls of cases that were typed on one version of the Illumina Human CoreExome Beadchip (v1.0) and panel B shows the genotype calls of controls that were typed on the next version of the same chip (v1.1). In panel B several heterozygotes have not been called. A genotype concordance test where missing calls are not taken into account will not pick this up this either. A stringent threshold for call rate differences by phenotypic status should remove most of these variants but the best way to ensure that these have been called accurately is by examining the genotype intensity plots.

In the UK10K project [10] where ~4000 samples from two cohorts were sequenced in two different centers batch effects were visualized in a multidimensional scaling analysis by labeling the samples by cohort and sequencing center (Fig. 4). Then logistic regression models were fitted using sequencing center as the case/control status to test for allele frequency differences between the two centers and by treating the cohort of origin as a covariate. Variants that showed a significant association with sequencing center were removed from further analysis. However, this approach can be only used for variants that are not too rare (e.g., MAF > 1%).

Uncorrected Proof

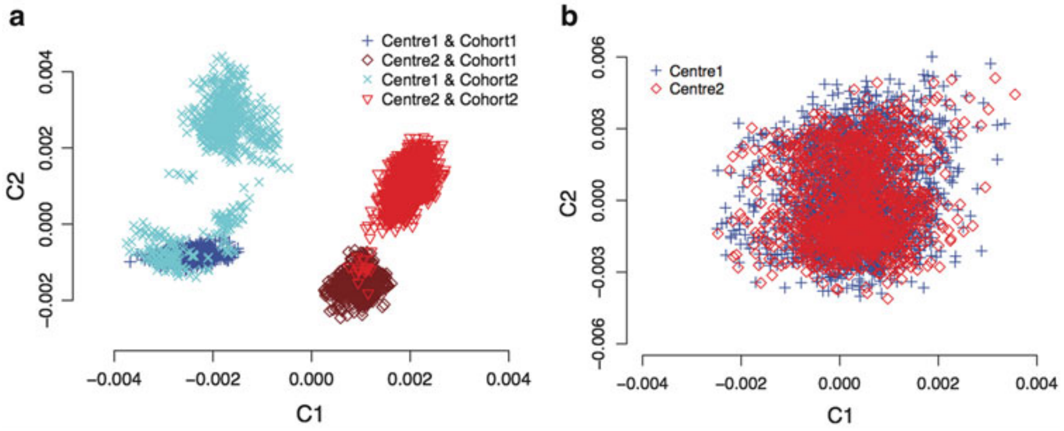

Fig. 4 Sample batch effects. (**a**) A multi-dimensional scaling analysis (MDS) highlights the sample batch effects for two cohorts sequenced at two different centers over some period of time by plotting the first component against the second component. (**b**) The first two MDS components after removing the batch effect. Both panels show data adapted from the UK10K cohorts study

Uncorrected Proof

3.6 Allele Frequency Comparisons with 1000G and UK10K

To analyze the overall quality of the variant calling, the allele frequencies of the call set can be compared to an existing published data set such as the 1000 Genomes Project [6] or the UK10K Project [10] in a case-control analysis manner. Variant sites that differ greatly in allele frequencies could be removed to boost the quality of the call set. Additionally, common variants with allele frequency greater than 5% should be mostly shared with those large-scale sequencing data sets.

3.7 MAF and Minor Allele Count (MAC) Filters

MAF filters are optional but they can eliminate a lot of noise in the data. This is particularly important for studies that have been typed on older genotyping platforms and called with earlier versions of genotype calling algorithms with poor performance at calling rare variants. Imposing a MAF filter of less than 1% across all samples is strongly recommended if the data is to be used for imputation. Minor allele count filters for cases and controls in separate are more robust to study sample size and are more effective filters for particularly unbalanced case/control designs. In an unbalanced study design the MAC but not necessarily the MAF will be different in cases and controls which can invalidate the assumptions of the association test, inflate the test statistic, and lead to spurious associations at low frequency or rare variants [11].

3.8 Imputation to Fill in Missing Genotypes and Post Imputation QC

A large proportion of the genotypes that will be removed by the variant quality control steps above will be captured by genotype imputation [12–14]. In addition, imputation using the latest reference panel by the HRC Consortium (McCarthy et al. 2016) (comprising 64,976 haplotypes at 39,235,157 SNPs constructed using whole genome sequence data from 20 studies of predominantly European ancestry) will lead to accurate genotype imputation at AU2

Author's Proof

minor allele frequencies as low as 0.1%. Imputation is a probabilistic approach and the accuracy depends on many factors including the density and content of the platform used to genotype the SNPs, as well as the ethnicity of the study population. The most widely used metric for imputation accuracy is the imputation information score which ranges from 0 to 1. Variants with imputation information score <0.3–0.4 are considered low quality and are typically removed from downstream analysis. In practise these filters are best determined by sequential filtering and examination of the inflation in a quantile-quantile (QQ) plot.

3.9 Sequence-Based Variant Quality Score QC

The procedure of the variant quality score recalibration (VQSR) aims at calculating a new quality score VQSLOD (variant quality score log-odds) that is supposed to be well calibrated and therefore allows fine-tuning of the specificity and sensitivity of the variant call set (https://software.broadinstitute.org/gatk). In other words, fine-tuning the specificity and sensitivity means maximizing the number of variants called and minimizing the false positive rate at the same time. The VQSR method uses machine learning algorithms, i.e., Gaussian mixture models, to help distinguish between true and false variants by combining annotations from several sources (e.g., read depth, mapping quality, and inbreeding coefficient) and by training them against a trustworthy set of variants. This approach results in determining a threshold for the VQSLOD score from the sensitivity/specificity of the variant call set against the training set to filter out the low-quality variants.

3.10 Imputation Refinement for Low-Depth Sequencing Data

For cost reasons most whole genome sequencing studies so far were sequenced at low read depth, i.e., less than ~10× (a read depth of 10× means that each nucleotide was covered on average by 10 sequenced reads). To improve the quality of variants in regions that were covered only by a few or low-quality reads, the idea is to borrow information from other samples. Therefore, it is customary to add a genotype imputation step, which helps in refining the genotypes by phasing them into haplotypes first and then filling in missing or low-quality genotypes by searching for similar haplotypes. This approach is based on Hidden Markov Models (HMM) that calculate a probability of each genotype for each of the missing genotypes [12–14].

Uncorrected Proof

References

1. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145):661–678. https://doi.org/10.1038/nature05911
2. Anderson CA, Pettersson FH, Clarke GM et al (2010) Data quality control in genetic case-control association studies. Nat Protoc 5 (9):1564–1573. https://doi.org/10.1038/nprot.2010.116
3. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome

Author's Proof

association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. https://doi.org/10.1086/519795

4. Chang CC, Chow CC, Tellier LC et al (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8
5. Reich D, Price AL, Patterson N (2008) Principal component analysis of genetic data. Nat Genet 40(5):491–492. https://doi.org/10.1038/ng0508-491
6. 1000 Genomes Project Consortium, Auton A, Brooks LD et al (2015) A global reference for human genetic variation. Nature 526 (7571):68–74. https://doi.org/10.1038/nature15393
7. Jun G, Flickinger M, Hetrick KN et al (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91(5):839–848. https://doi.org/10.1016/j.ajhg.2012.09.004
8. Clayton DG, Walker NM, Smyth DJ et al (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37 (11):1243–1246. https://doi.org/10.1038/ng1653
9. Gurdasani D, Carstensen T, Tekola-Ayele F et al (2015) The African genome variation project shapes medical genetics in Africa. Nature 517(7534):327–332. https://doi.org/10.1038/nature13997
10. Walter K, Min JL, Huang J et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526(7571):82–90. https://doi.org/10.1038/nature14962
11. Ma C, Blackwell T, Boehnke M et al (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37 (6):539–550. https://doi.org/10.1002/gepi.21742
12. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84(2):210–223. https://doi.org/10.1016/j.ajhg.2009.01.005
13. Fuchsberger C, Abecasis GR, Hinds DA (2015) minimac2: faster genotype imputation. Bioinformatics 31(5):782–784. https://doi.org/10.1093/bioinformatics/btu704
14. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1(6):457–470. https://doi.org/10.1534/g3.111.001198

Uncorrected Proof

Author's Proof

Author Queries

Chapter No.: 3	394545_1_En

Query Refs.	Details Required	Author's response
AU1	Please check whether the affiliation and correspondence details are presented correctly.	
AU2	Please provide complete details for Ref. "McCarthy et al. 2016".	

Uncorrected Proof