Published in final edited form as:

Curr Protoc Hum Genet. ; 90: 2.14.1–2.14.16. doi:10.1002/cphg.15.

Quality Control for the Illumina HumanExome BeadChip

Robert P. Igo Jr¹, Jessica N. Cooke Bailey¹, Jane Romm², Jonathan L. Haines^{1,3}, and Janey L. Wiggs⁴

¹Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio

²Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD

³Institute of Computational Biology, Case Western Reserve University, Cleveland, Ohio

⁴Department of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA

Abstract

The Illumina HumanExome BeadChip and other exome-based genotyping arrays offer inexpensive genotyping of some 240,000 mostly nonsynonymous coding variants across the human genome. The HumanExome chip, with its highly non-uniform distribution of markers and emphasis on rare coding variants, presents some unique challenges for quality control (QC) and data cleaning. Here, we describe QC procedures for HumanExome data, with examples of challenges specific to exome arrays from our experience cleaning a data set of approximately 7,500 samples from the NEIGHBORHOOD Consortium. We focus on standard procedures for QC of genomewide array data, including genotype calling, sex verification, sample identity verification, relationship checking and population structure, that are complicated by the HumanExome panel's enrichment in rare, exonic variation.

Introduction

Genome-wide association studies (GWAS) have successfully identified thousands of common genetic variants of small to moderate effect associated with common, heritable diseases (Hindorff et al., 2016), but have explained only a minority of heritable variation in most of these diseases (Manolio et al., 2009). It has been hypothesized that uncommon and rare genetic variation may account for much of this “missing heritability.” Whole-exome sequencing (WES) has recently become a popular approach for identifying rare, potentially causal genetic variants. The cost of whole-exome sequencing, however, is still daunting for samples of the size required for detecting single rare variants of moderate effect (Page et al., 2015). Other considerations, including data storage requirements and a labor-intensive data processing pipeline, also make WES less attractive.

Exome-based genotyping arrays combine targeted interrogation of exonic variation with the low cost and convenience of a GWAS chip. The examples used here are based on the Illumina Infinium HumanExome-12 Beadchip (HumanExome chip; Illumina, Inc., San Diego, CA) but the concepts and approaches are appropriate for any genotyping array with substantial numbers of rare variants. The Beadchip contains more than 240,000 exonic

variants, including nearly 220,000 nonsynonymous coding variants, collected from over 12,000 individuals with whole-exome or whole-genome sequence data, and contributed by 16 study centers (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). Researchers may add up to 200,000 custom variants. The chip, moreover, includes a panel of 4,388 ancestry-informative markers (3,388 AA vs. European, 1,000 Native American vs. European) and a scaffold of 5,286 markers with high minor allele frequency (MAF) in several continental populations, suitable for linkage analysis or relationship estimation.

Quality control (QC) is an often underappreciated process crucial to the success of any genetic association study. Standard procedures for QC of GWAS data have been well described (e.g., (Laurie et al., 2010; Turner et al., 2011; see Unit 1.19). Two previous publications have specifically addressed QC for exome arrays. Grove et al. published guidelines for genotyping and detailed QC screening criteria for a large sample from multiple cohorts (Grove et al., 2013). Recently, Guo et al presented detailed protocols for running Illumina GenomeStudio, Illumina's sample processing and genotype calling software, with references to accessory R and Python scripts to perform specific data-processing tasks (Guo et al., 2014). Both of these reports have some limitations: the Grove et al. study describes best practices from a single, atypically large project, and the Guo et al. protocols are highly specialized for a specific QC pipeline and certain file formats.

This unit draws primarily on our experience working with a genotyping project on approximately 7,500 samples in the NEIGHBORHOOD Consortium (Wiggs et al., 2013). Here, we emphasize procedures and potential problems not previously described, though we outline the complete process. The general concepts will also apply to the similar Affymetrix Exome Genotyping Array (Affymetrix Inc., Santa Clara, CA).

These procedures make extensive use of the program PLINK (Purcell et al., 2007), a fast and easy-to-use tool for simple quality control and analysis of large single-nucleotide-polymorphism (SNP)-based data sets. PLINK has excellent online documentation (<http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml>) and is straightforward to use; however, for speed and processing capacity, it should be run within a Linux environment. In addition, considerable text-file manipulation and processing is required, for which a scripting language such as Perl or Python is best suited. For statistical summaries and analyses that cannot be run in PLINK, we recommend the statistical programming language R, which is freely available (<https://cran.r-project.org/>).

Genotyping and Initial QC

Genotype calling and initial QC screens are conducted in Illumina GenomeStudio. GenomeStudio's GenCall algorithm is optimized for common variants (Perreault et al., 2014), and consequently, special care must be taken to verify accuracy of calling rare variants. For this reason, we verified rare variant genotypes by means of a second genotyping program, zCall (Goldstein et al., 2012). Basic GenomeStudio procedures relevant to the HumanExome chip are described in detail in Guo et al. (Guo et al., 2014). Additionally, Grove et al. has published an extensive list of QC screening criteria (Grove et

al., 2013). Here, we will briefly review the procedure and describe the particular screens we used for the NEIGHBORHOOD Exome Chip project.

Genotype calling in Illumina GenomeStudio

Genotyping from raw fluorescence data (Illumina .idat files) comprises three major steps: (a) initial calling, (b) removal of poor-quality samples, and (c) reclustering. When data are initially loaded to create a GenomeStudio project, GenomeStudio either clusters genotypes empirically, using only the data, or calls genotypes from predefined clusters in a cluster (.egt) file (see the GenomeStudio documentation for details). We used Illumina's standard clustering file to call genotypes within GenomeStudio. The CHARGE Consortium has also made available a custom genotype clustering file for the exome chip, CHARGE_ExomeChip_v1.0_Cluster_File.zip, with clusters based on a very large sample ($n = 62,266$) (Grove et al., 2013); these clusters may be preferable to Illumina's for data sets similar to the CHARGE sample. Raw fluorescence data are contained in two .idat files (for the "red" and "green" fluorescence channels) per sample, typically organized in directories representing each HumanExome chip. A Sample Sheet, provided to GenomeStudio by the user to organize the project, links the sample identifiers (IDs) provided by the investigators with the .idat file names and provides other sample information, such as recorded sex. Merging from several genotyping projects into a single GenomeStudio project involves extra steps, and requires that the arrays be exactly the same version across all data sets. After loading the first dataset, choose Load Additional Samples from the File drop-down menu and load the data as with the first dataset and repeat until all samples are loaded. Recluster SNPs and re-calculate SNP statistics before continuing the QC.

After initial calling and manual review (described below), omit markers using the following criteria: call rate (proportion of successful genotypes) less than 85%, where genotypes with GenCall scores less than 0.15 are marked as missing; cluster separation less than 0.2; and heterozygosity greater than 0.8. Remove samples with call rate less than 98% over the markers passing initial QC (e.g. poorly typed samples), and perform final clustering on the remaining samples. Examine low-quality samples by means of genomewide B allele frequency (BAF) and log(R) ratio (LRR) intensity plots (see Sample Quality, below) for large chromosomal deletions that may affect overall call rate. Such samples may be usable for analysis.

Manual review and calling rare SNPs in zCall

As a means of improving genotyping for rare variants (we chose $MAF < 0.02$, but this is somewhat arbitrary), SNP genotypes are further clustered using zCall (Goldstein et al., 2012), a more sensitive approach for calling rare variant heterozygotes (Perreault et al., 2014), using a procedure developed at the Center for Inherited Disease Research (CIDR) (Romm et al., 2013). Here, final genotypes were not based on zCall output directly, but rather zCall genotypes are used as a means of identifying rare variants with true heterozygous genotypes missed by GenCall. zCall is a suite of Python and R scripts (<https://github.com/jigold/zCall/>), and therefore is more programming-intensive than GenomeStudio.

To conduct the zCall review, first generate a GenomeStudio Final Report containing genotype calls and intensity values in (X, Y) coordinates as input for zCall. Run Version 3 of zCall, using only the GenomeStudio report as input, as described in the zCall README file, with one change: to reduce false-positive heterozygote calls, raise the z-score threshold for heterozygote calls from the recommended value of $z = 7$ to $z = 21$ in step 4 of the procedure. This change increases the required distance from homozygote clusters for a genotype to be called as heterozygous. zCall generates genotype output in PLINK transposed text format (with .tped and .tfam files). Hence, both input and output files may be very large, especially the input Final Report (about 80 MB per sample).

Compare minor allele counts from GenCall and zCall by means of an allele frequency report on each version of the genotypes using PLINK's `--freq` function, with the `--counts` option to report allele counts. (See next section for exporting a GenomeStudio project in PLINK format.) Visually inspect GenCall clusters in GenomeStudio for any rare variant (GenCall MAF < 0.02) with at least two additional heterozygote calls from zCall relative to GenCall. If a distinct heterozygote cluster is evident and the cluster locations are consistent with high-quality marker data (see Guo et al. (2014) and Turner et al. (2011); see Unit 1.19 for examples), manually adjust cluster locations by repositioning or resizing the cluster ovals in the intensity plot (as described in the GenomeStudio Genotyping Module User Guide; also see Guo et al. (2014) Protocol Step 22). Otherwise, discard the marker.

Also manually review all pseudoautosomal, Y-linked and mitochondrial markers in GenomeStudio for poor clustering. Where misclustering is evident, adjust cluster locations manually where possible and discard markers without well-defined clusters.

Exporting GenomeStudio data to PLINK format and preparing for analysis

Exporting genotypes in PLINK format—After GenomeStudio QC and manual review, export samples and markers with call rates of at least 95% (for chromosome Y markers, 95% in males only) from GenomeStudio in a Final Report. The most convenient data format for further QC, as well as for statistical analysis, is PLINK (Purcell et al., 2007) binary format: genotypes for diallelic markers are stored efficiently (1 bit per genotype) in a .bed file, with additional map (.bim) and sample (.fam) information files. An extension for GenomeStudio, the PLINK Input Report Plug-in, is available for download from the Illumina Web site (http://support.illumina.com/array/array_software/genomestudio/downloads.html). A detailed protocol for extracting genotype data from GenomeStudio in PLINK plain-text format, yielding PLINK .ped and .map files, is available (Guo et al., 2014). For ease of converting genotypes to the dbSNP + strand, export genotypes on the Illumina TOP strand.

GenomeStudio's plug-in module exports genotypes to PLINK plain-text format, which can result in extremely large pedigree files. As an optional step to reduce the file size, filter out in GenomeStudio markers with a minor allele frequency of 0 in the sample, since these SNPs contain no variation useful for detecting genetic association or linkage. There are likely to be many of these monomorphic markers, unless the sample is extremely large. Do not discard monomorphic markers, however, if the exported data is to be combined with other sets of genotypes in which these markers may be polymorphic. To reduce the file size further, omit failed SNPs from the report by making sure “Include zeroed SNPs in the

report” is not chosen. Finally, omit markers mapping to chromosome 0. If necessary, correct the chromosome number of pseudoautosomal (chromosome “XY”) markers erroneously assigned to the X or Y chromosomes (see Sex-chromosome marker annotation, below). In PLINK standard format, chromosomes X, Y and XY are numbered 23, 24 and 25, respectively. Extract Y chromosome genotypes separately, with a call rate threshold somewhat below the proportion of male samples in the data set. Conduct further call rate screening for the Y chromosome using male samples only, after resolving sex-mismatch errors (described under Sample Quality, below). Convert the initial PLINK text file exported by GenomeStudio, which can easily exceed 10 GB in size, to a binary file set (.bed, .bim and .fam files) to save disk space, using PLINK's `--make-bed` function.

Converting genotypes to the + strand. This step is optional and may be taken after other QC is complete—DNA strand orientation is an important consideration when genotype data sets from different arrays are combined and when untyped markers are imputed (Unit 1.25; Porcu et al., 2013). The Illumina TOP DNA strand is assigned by a well-defined algorithm based on the SNP alleles and, if necessary, the flanking DNA sequence (http://www.illumina.com/documents/products/technotes/technote_topbot.pdf). Imputation programs, such as IMPUTE2 (Howie et al., 2009) and MaCH (Li et al., 2010), generally require genotypes on the dbSNP reference + strand, which is not predictable from the TOP strand (Unit 1.25; Porcu et al., 2013). The PLINK Format Report can report genotypes relative to the FORWARD strand, which agrees more often with the dbSNP reference (+) strand; the TOP strand, however, is more easily converted to the + strand, since the relationship between TOP/BOTTOM and +/- strands is coded explicitly in the Illumina strand report file (e.g., HumanExome-12-v1-2-B.csv, available through the Illumina Web site). The TOP/BOTTOM strand is specified under “IlmnStrand”, and the +/- strand under “RefStrand”. The Wellcome Trust Centre for Human Genetics also maintains a conversion key for several versions of the HumanExome panel (<http://www.well.ox.ac.uk/~wrayner/strand/>). Use the Linux shell script provided on the Web page to convert genotype data to the + strand (see (Guo et al., 2014) for details). Alternatively, create a list of SNP IDs whose genotypes need to be “flipped” to the + strand by means of a script in a language such as Perl, Python or R. Identify the set of SNPs in the Illumina strand report file coded as BOT and + and as TOP and -; these are the SNPs whose TOP strand is the - strand. Export the names of these SNPs, which appear under the column “IlmnID”, to a plain-text file with one name per line, and flip the genotypes for these SNPs using the `--flip` function in PLINK.

Illumina does not routinely annotate the +/- strand orientation of custom HumanExome custom in its SNP manifest files. Note carefully the relationship of + and TOP/FORWARD strand alleles of custom markers during chip design, especially in the case of ambiguous (A/T and C/G) SNPs.

Converting Illumina marker IDs to rsIDs. This step is optional, but recommended after QC is complete—Marker names appearing in HumanExome description and support files are not generally the dbSNP rsIDs commonly used in the literature. An Illumina ID to rsID conversion file, humanexome-12-v1-2-b-loci-name-to-rsid-conversion-file.zip, is available for version 1.2 of the HumanExome panel at Illumina's

Web site (https://support.illumina.com/array/array_kits/infinium_humanexome_beadchip_kit/downloads.html). Update marker names in a PLINK binary file set by changing the SNP IDs in the second column of the map information (.bim) file. A small percentage of variants may not have rsIDs. Leave these marker IDs as Illumina SNP IDs, or convert to chromosome:position format (e.g., chr18:55543071). Another, and possibly more complete, source of rsIDs for the HumanExome panel is the CHARGE Consortium's Web site (<http://www.chargeconsortium.com/main/exomechip>). The file SNPInfo_HumanExome-12v1_rev6.tsv.txt (826 MiB, uncompressed; revised Nov. 7, 2014) contains rsIDs for the HumanExome markers, in addition to a wealth of marker annotation data, for markers included in versions 1.0, 1.1 and 1.2 of the panel.

Sample Quality

Sample quality QC is concerned primarily with identifying (a) poorly typed DNA samples; (b) sample identity errors, including samples erroneously duplicated and switched; and (c) chromosome anomalies that may interfere with or confound genetic analysis. Many, but not all, poor-quality DNA samples will be identified and removed during genotype calling; thus, the sample QC procedures primarily concern sample identity. All standard sample QC objectives are readily achieved with HumanExome data, with a few complicating factors associated with the structure of the HumanExome panel. Some screens do not immediately identify samples for exclusion, but rather flag samples for additional screening and resolution. Table 1 lists sample-based screening criteria, further explained in the following sections.

Sample call rate

The primary screen for sample quality is call rate. In addition to the initial 95% call rate screen for export from GenomeStudio, we applied a final call-rate cutoff of 98.5%. The desired final call rate may vary depending on original DNA quality: DNA samples stored for many years generally yield fewer high-quality genotypes.

DNA source

Whole-genome amplification, a common method to increase abundance of DNA samples in limited quantity before genotyping, may cause anomalous results from both sample and marker QC, particularly in marker panels enriched in rare variants. Consequently, whole-genome-amplified (WGA) samples require somewhat different QC screens from non-amplified DNA, and require particular care in screens involving total fluorescence intensity or DNA copy number. WGA samples have lower call rate, on average, but screen them for call rate as for other samples. Subtle differences in fluorescence intensity may also occur between DNA extracted from blood samples and DNA from lymphoblastoid cell lines (Wellcome Trust Case Control Consortium, 2010), but these will not likely cause detectable differences in genotype frequencies in samples of moderate size, particularly among rare variants.

Sex-chromosome marker annotation

Before screening sex chromosomes for sample QC, the X- and Y-linked markers must be properly annotated. In some versions of the Illumina HumanExome annotations, as stored on the Illumina SNP Manifest (.bpm) file, numerous SNPs in the pseudoautosomal regions (PAR1, PAR2) and in the X-translocated region (XTR; PAR3) (Veerappa et al., 2013) are assigned to Chr = X or Chr = Y instead of Chr = XY, the usual chromosome label for the PARs. For best results, assign markers in these regions to Chr = XY (PLINK chromosome 25) before QC by defining a new field in the GenomeStudio SNP Table containing the chromosome number, changed to Chr = XY for all pseudoautosomal and XTR markers. Definitions for the PARs and XTR in human genome build GRCh37 appear in Table 2.

Sex check and sex-chromosome anomalies

Comparing the apparent genetic sex of each sample with the sex recorded in the clinical data can reveal DNA samples mismatched with the participants' clinical records, as well as clerical errors in sample reporting. We recommend both of two complementary procedures for screening sex-chromosome data: a plot of mean fluorescence intensity of the X vs. the Y chromosome, as an estimate of chromosomal copy number (Laurie et al., 2010); and, as a secondary screen, a plot of X chromosome heterozygosity vs. successful Y chromosome genotypes, to identify problems not observed in the fluorescence intensity plot, particularly in the case of WGA samples.

Mean fluorescence intensities—Measures of fluorescence intensity used for QC include the B allele frequency (BAF), a value between 0 and 1 representing the relative strength of the A and B allele intensities; and the log(R) ratio (LRR), a normalized measure of overall signal strength. Extract intensity data from chromosome X and Y markers passing QC by means of a GenomeStudio Final Report. To obtain intensity values over the X chromosome, filter X-lined markers by the standard QC criteria (omitting pseudoautosomal markers), plus Minor Freq > 0 (polymorphic markers only), select Final Report under the Analysis/Reports/Report Wizard menu, and choose “Remove zeroed SNPs”. Select fields SNP Name, Sample ID, Allele1 Top, Allele2 Top, B Allele Freq and Log R Ratio as the output fields and write the Final Report to a text file. Female genotypes on the Y chromosome will generally be missing. Consequently, for the Y chromosome, do not set a standard high threshold for call rate, but rather one somewhat less than the proportion of males in the study sample (e.g., 0.4 for an equal mix of males and females). Adding an *upper* limit to call rate (e.g., 0.6 for a sex-balanced sample) will ensure that no pseudoautosomal samples will be included in the Y-chromosome Final Report. From these output data, calculate the mean LRR over the entire X and Y chromosomes for each sample.

A plot of mean X and Y LRR values shows two distinct major clusters of points, representing the genetically male (with high LRR for Y, low for X, upper left) and female (lower right) samples (Fig. 1A). A tail extending toward lower intensity for the Y in males likely reflects age-related loss of the Y chromosome (Forsberg et al., 2008). A small proportion of female DNA samples may have significant loss of heterozygosity of the X chromosome (XX/XO), especially if the DNA is extracted from lymphoblastoid cell lines (Laurie et al., 2010), but a large lower tail of the X intensity in female samples is not

apparent here. A limitation of the HumanExome chip in QC is poor precision in the mean LRR values for the X chromosome compared to GWAS panels (compare Fig. 1 with Turner et al. (2011), Fig. 1.19.2 and Laurie et al. (2010), Fig. 1A).

Apparent sex-mismatch errors are evident as points of the opposite color within the two major clusters. Flag these samples and verify their sex in the clinical records (the “reported” sex); if no clerical error is found, the sample ID is likely to have been assigned to the wrong DNA sample, and if the DNA sample cannot be reassigned to the correct clinical data, do not use the genotype data for analysis.

Outlying points in Fig. 1A indicate potential sex-aneuploid samples, and are labeled with the most likely sex-chromosome configuration. Female samples with outlying high and low mean LRR on X are classified as XXX and as XO or XX/XO, respectively. One individual in Fig. 1A has typical intensity values for female X and male Y chromosomes, and is designated XXY. The five male samples with greatest Y intensity are tentatively labeled XYY, although these are not extreme outliers, and one individual with apparent diploid X and high Y intensity (possibly XXYY) is labeled “High Intensity.” We examined these samples more closely by means of plots of BAF and LRR vs. map position, as described below.

Whole-genome-amplified samples have more variable intensity profiles (compare Fig. 1B with 1A), and for this reason we do not recommend screening WGA samples by means of average intensity plots. The distribution of average X intensities in WGA samples encompasses the non-WGA female samples tentatively identified as sex-aneuploid (Fig. 1B). In contrast, the Y intensities are less dramatically affected in WGA samples.

An intensity-based sex check is available in the Genotyping Module of Illumina GenomeStudio. To perform the check, select all samples in the Samples Table, right-click on the selected samples, and choose “Estimate Gender for Selected Samples....” When asked, “Would you like to populated the Gender column...” choose “No” to prevent GenomeStudio from overwriting the sex with the apparent genetic sex. Like PLINK's sex check (see next paragraph), this test is not very detailed, and should not be used without confirmation via a mean intensity plot.

X chromosome heterozygosity vs. Y chromosome call rate—The `--sex-check` function in PLINK is a genotype-based approach to estimating copy number for the X chromosome (see Turner et al. (2011) Unit 1.19 for details). PLINK estimates the inbreeding coefficient F_X over all X-linked markers, and classifies the sample as male if $F_X < 0.2$, and female if $F_X > 0.8$. A value of F_X between 0.2 and 0.8, or an apparent sex that disagrees with the putative sex, is flagged as a “problem”. For best results, include only common SNPs (MAF ≥ 0.05), since rare variants will contribute little information.

Take care in conducting the PLINK sex check and interpreting the results. Because the HumanExome chip is scarce in common variants, the estimated F_X of female samples is imprecise (Fig. 2, horizontal axis), and therefore a significant proportion of truly female samples will be called as indeterminate sex. (This problem is alleviated if GWAS content is

added to the HumanExome panel, as in the HumanCoreExome chip, which includes some 250,000 genomewide common variants.) Second, PLINK does not consider the Y chromosome in calling sex, and in consequence, individuals with sex-chromosome anomalies such as XO, Turner syndrome, or XXY, Klinefelter syndrome, may be assigned as XY and XX, respectively. Third, XX/XO female individuals will appear as of indeterminate sex.

Nonetheless, X chromosome heterozygosity, when paired with chromosome Y call rate, can provide useful information that a plot of fluorescence intensity can miss, particularly for WGA samples, whose mean intensity values are not a reliable indicator of true DNA copy number. Fig. 2 displays F_X plotted against call rates over the Y chromosome. Indeed, these genotype-based sex chromosome profiles are considerably more robust to WGA status than the intensity profiles (compare Fig. 2A and 2B). One WGA sample, with a typically male profile by mean intensity, appears as XXY in a plot of Y call rate vs. F_X (Fig. 2B, “Not IDed as aneuploid”). We discuss this sample further below. In contrast, samples with high X chromosome intensity suggesting an XXX karyotype are not well resolved by F_X . Thus, the intensity- and genotype-based approaches to examining sex chromosomes have unique advantages when applied to HumanExome data.

PLINK, by default, ignores samples assigned as female when analyzing Y-linked markers, whereas call rates for all samples are required for the genotype-based sex chromosome check. To bypass PLINK's automatic screen, change the sex of female samples (sex = 2) to male (sex = 1) or missing (sex = 0) in the fifth column of the PLINK .fam file before applying PLINK's `--missing` function to calculate Y-chromosome call rates.

Plotting BAF and LRR by map position—The most informative test for sex chromosome anomalies is a plot of BAF and LRR by map position across the X and Y chromosomes (Fig. 3). The Illumina Genome Viewer in GenomeStudio generates such plots for individual chromosomes, as shown in Fig. 3A (see the Illumina Genome Viewer chapter of the GenomeStudio Framework User Guide), but these plots are difficult to create and customize within GenomeStudio. Instead, we plotted LRR and BAF from the intensity data from QC-filtered marker sets previously extracted from the GenomeStudio project for computing mean fluorescence intensities. Using R's base graphics package, we created more easily interpretable plots (Fig. 3B), with intensity ranges standardized across chromosomes and samples. An XX karyotype is evident from the presence of a band of heterozygous genotypes across the X chromosome at BAF ~ 0.5 in addition to homozygote bands at BAF ~ 0 and 1. The absence of a heterozygote band in Fig. 3B, and a mean LRR of less than zero, indicates the presence of only one X chromosome. A mean Y LRR close to zero and abundant Y genotypes at BAF near 0 and 1 reveal the presence of a Y chromosome; in the absence of a Y chromosome the BAF value would vary randomly between 0 and 1 and mean LRR would most likely fall below -2. Filtering out improperly annotated and poor-quality markers dramatically improves ease of interpreting the BAF plot (compare Fig. 3A and B, BAF panels; Fig. 3B reflects post-QC genotype data).

BAF and LRR plots of two samples with confirmed sex-chromosome aneuploidy are shown in Fig. 4. In Fig. 4A, two intermediate bands of BAF at roughly 1/3 and 2/3 are visible,

suggesting a karyotype of XXX. This sample is marked with an asterisk in Fig. 1A; note that two non-WGA female samples confirmed as XX (data not shown) had greater mean X intensity. In Fig. 4B, two X chromosomes are apparent, with loss of heterozygosity (LOH) in the central portion without loss of copy number (i.e., the mean LRR does not change in the LOH region), as well as a Y chromosome, consistent with the designation of this sample as XXY. One WGA sample had a genotype-based profile consistent with XXY, but appeared as XY in the mean intensity plot (Fig. 2, “Not IDed as aneuploid”). Fig. 4C confirms that, although the mean X LRR is typical of a single X chromosome, a band of apparent heterozygous genotypes is apparent. However, the BAF of these genotypes is more variable than expected, and closer inspection reveals that this band is off center, with an average BAF of about 0.6. An intermediate, off-center BAF band is also visible on the Y chromosome, although the mean intensity is typical of a single copy of the Y. Such a configuration suggests a mixture of two male DNA samples in approximately 3:2 proportions. This sample, which had an acceptable call rate, was not identified by VerifyIDIntensity as contaminated, likely because it was WGA; it also had an acceptable call rate.

Resolution of apparent sex-chromosome anomalies—Resolve sex-mismatch errors, after confirming that the sex is recorded properly in the clinical data, by examining other QC measures of sex-mismatched samples, such as relationships, presence of duplicate samples and genotype concordance with known duplicates or previous genotype projects. Samples with sex-chromosome aneuploidy (e.g., XO or XXY) or other anomalies should be flagged but not necessarily removed from the study. (The sex chromosome profile from peripheral blood or buccal epithelium is not guaranteed to match the germline karyotype.) Sex-aneuploid individuals may be omitted from studies of certain phenotypes as a precaution. For example, Klinefelter syndrome is a known cause of learning disabilities (Geschwind et al., 2000).

Sample contamination

A contaminated sample contains DNA from two or more individuals erroneously mixed together. (In contrast, a *pooled* DNA sample is an intentional mixture from multiple individuals in equal proportions.) Contaminated samples are often, but not always, removed early during QC on account of low sample call rate or problems with genotype clustering, resulting in low overall GenCall scores.

Identify contaminated samples with high call rate by two screens. First, calculate the heterozygosity over common (MAF ≥ 0.05) autosomal markers (Guo et al., 2012; Laurie et al., 2010), using the PLINK `--het` function. Examine samples with outlying high autosomal heterozygosity (> 3 SD above the mean) carefully for evidence of contamination, as by plots of BAF and LRR vs. map position over the autosomes. However, the low abundance of common variants in the HumanExome chip makes this test less sensitive than for a GWAS panel. In addition to autosomal heterozygosity, therefore, we recommend directly estimating the level of contamination in all samples by means of VerifyIDIntensity (<http://genome.sph.umich.edu/wiki/VerifyIDIntensity>), the implementation of VerifyBamID for genotype array data (Jun et al., 2012). Examine autosomal BAF and LRR plots of samples with significant estimated contamination (defined for our project as greater than 10%) to

confirm contamination displaying as unusual BAF distributions (e.g., Fig. 4C). VerifyIDIntensity may have difficulty estimating contamination of WGA samples, as a consequence of uneven amplification across the genome, although contamination can still be verified my means of BAF plots by map position.

Concordance with genotypes from previous assay

If samples to be genotyped for the HumanExome chip have been previously typed for a GWAS panel, genotype concordance will reveal sex-compatible sample identity errors with high sensitivity. Because only a few hundred reliable common variants are necessary to ascertain sample identity, most medium- and high-density GWAS chips will have plenty of overlap with the HumanExome panel to test compatibility, despite the scarcity of common variants on the exome chip. In our example, we compared the HumanExome genotypes of our NEIGHBORHOOD samples with those from the Illumina 660K chip, available for about 80% of the samples (Wiggs et al., 2012).

Apply stringent selection criteria in selecting an overlapping set of markers between two different SNP arrays, especially if produced by different manufacturers. To prevent spurious errors from DNA strand uncertainty, omit ambiguous (A/T and C/G) SNPs, and compare alleles of all markers for compatibility. Make sure that all markers from both panels are polymorphic in the sample, to verify strand orientation, and preferably common. Ensure that the markers and samples from both panels are of high quality; for example, we used a call rate threshold of 99% for both. In addition, omit markers that show poor (< 95%) concordance across all sample pairs, and SNPs with genotypes available for both members of < 90% of sample pairs, as these SNPs are likely poorly typed in one or both data sets. Despite these strict criteria, we identified 11,310 eligible autosomal SNPs overlapping between the HumanExome and Illumina 660K arrays in our cohort in both panels.

Considerable text-file processing may be necessary to identify eligible markers for comparison and to compare alleles to find the markers whose DNA strand needs to be flipped in one panel to make them compatible. Because the annotation information for SNPs is periodically updated, SNPs having the same rsID in the two chips are not guaranteed to have compatible alleles (e.g., the same rsID may have alleles A/C in one panel and A/G in the other). Markers also need to have the same map positions in the different datasets, as the same marker may have more than one rsID. If the previously genotyped panel is annotated to a different build of the human genome from the HumanExome chip (currently GRCh37), update the PLINK map (.bim) file to carry compatible map positions.

To measure genotype concordance across panels using PLINK, first extract genotypes for the overlapping SNPs from each panel, flipping the DNA strand of one panel as necessary to align all strand assignments. Ensure that the sample IDs are different (but similar) for the same individual in each panel so that the individual's genotypes from the two panels will not be merged when the two data sets are merged. Perform PLINK's identity-in-state (IBS) calculation on the merged HumanExome and previous genotypes and report an IBS matrix (`--cluster --matrix` function). Genotype concordances, expressed as a proportion of compatible genotypes, will appear in the matrix where the two IDs representing the same individual intersect. Omit samples with unusually low concordance between the

HumanExome and the previous GWAS chip as discordant samples. The threshold for exclusion may depend on the genotyping quality of the previous project and the technology used for the previous panel (an early chip such as the Affymetrix 500K may give several percent discordance even for identical samples). Examine the distribution of genotype concordance over all putatively identical sample pairs and determine the cutoff by the location of outlying points. A sample-wide threshold of 95% is likely conservative for high-quality genotype data.

Relatedness and population structure

Sample relatedness—Relatedness is important to check even in unrelated case-control samples, because it identifies (a) unknown (cryptic) relationships between individuals thought to be unrelated, (b) sample identity errors in family data, (c) improperly duplicated samples and (d) contaminated samples that may have escaped other screens. Relationship checking comprises estimation of identity-by-descent (IBD) sharing between all pairs of samples, and results are expressed as estimated proportions of the autosomal genome shared 0, 1, and 2 alleles IBD (Z_0 , Z_1 and Z_2 , respectively), and an overall estimate of genome-wide sharing that determines degree of relationship ($\hat{\pi}$; twice the kinship coefficient in the absence of inbreeding). Testing pairwise relationships using PLINK using a subset of independent markers, and interpreting the results, have been described (Laurie et al., 2010; Turner et al., 2011; see Unit 1.19). For the NEIGHBORHOOD Exome Chip project, we used the program KING (Manichaikul et al., 2010) (<http://people.virginia.edu/~wc9c/KING/index.html>) rather than PLINK, since IBD estimation by the method implemented in PLINK tends to be biased upwards in the presence of rare variants (Milligan, 2003) and population stratification (Manichaikul et al., 2010). KING accepts genotype input in PLINK binary file format. Prepare HumanExome data for KING by excluding SNPs in PLINK for $MAF < 0.001$ and call rate $< 98\%$, and samples for call rate $< 98\%$. Obtain pairwise kinship coefficients by means of KING's robust `--kinship` function, and IBD-sharing estimates under the assumption of homogeneity using the `--homo --showibd` option.

IBD-sharing profiles of the second-degree relatives half siblings, avuncular and grandparent-grandchild pairs are identical ($Z_0 = Z_1 = 1/2$, $Z_2 = 0$) when measured by summation over single markers considered independently. If age data are available, these relative pairs may be distinguished with reasonable specificity by comparing the age of relatives to guess the difference in generations. Different types of second-degree relative pairs do differ, nonetheless, in the pattern of IBD chromosomal sharing across the genome, and methods exist that can differentiate between them when provided with high-density SNP data (e.g., Zeng et al. (2016)).

Population structure and ethnic background—Examining population structure is important, particularly in multiethnic samples, because it can (a) identify sample mislabeling and (b) verify self-reported ancestry. The standard approach to determining ethnic background is to conduct a principal components analysis (PCA) on normalized genotype data, usually by means of the smartpca program in EIGENSOFT (Patterson et al., 2006). The first two principal components (PCs) roughly describe global ancestry (Guo et al., 2014), also see Unit 1.22 (Edwards and Gao, 2012).

In estimating global ancestry, include “anchor” samples from the three HapMap I continental samples (European = CEU, African = YRI, East Asian = CHB + JPT) in the PCA. Ideally, HapMap controls from all of these samples will have been included in the genotyping project. Alternatively, the CHARGE Consortium has made available HumanExome genotypes for 96 HapMap control individuals, and genotypes for the overlap between the HumanExome and 1000 Genomes dataset for the 1,092 Phase I samples may be found in PLINK binary format at the VANGARD Exome Chip Processing page (<https://github.com/slzhao/ExonChipProcessing/>).

The HumanExome chip includes approximately 2,950 ancestry-informative markers (AIMs) scattered across the autosomes and the X chromosome (for a complete list, see Guo et al. (2014) Supplementary Table 6). However, in our sample the first two PCs from EIGENSOFT, based on only the panel of AIMs, do not define ancestry very precisely (Fig. 5A): the YRI controls are poorly clustered, the African American samples are unusually scattered with varying extent of Asian admixture, and Hispanic samples display unusually high apparent Asian admixture. Supplementing the AIMS with additional common HumanExome markers did not improve precision (data not shown). The clustering of HapMap controls can be improved and better estimates of ethnic admixture generated by means of the SNPweights software package (Chen et al., 2013). SNPweights uses precomputed loadings for more than 600,000 AIMs to calculate two PCs and to estimate admixture proportions based on 310 individuals from the HapMap CEU, YRI, CHB and JPT samples used as ancestral continental populations. In the plot of the first two PCs from SNPweights (Fig. 5B), the HapMap populations form tight clusters at the three vertices of the population “triangle,” and the Hispanic and African American subsamples are distributed as expected.

SNPweights is limited to reporting two PCs. Consequently, if PCs are required to adjust for population structure that may confound tests of genetic association, perform a second PCA over a subset of independent common variants using a program such as EIGENSOFT, which can report as many PCs as desired. Although it is tempting to take advantage of the bounty of rare variants in the HumanExome chip for PCA, using rare variants for PCA does not appear to improve detection of potentially confounding population structure (Babron et al., 2012; Lee et al., 2014; Zhang et al., 2013).

Resolving sample identity errors

Attempt to resolve identity errors where possible, and thereby salvage samples with such errors, by drawing together data from sex-incompatibility errors, relationship errors (including samplewide Mendelian incompatibility errors), self-reported ethnicity, and, where available, inconsistency with previous genotyping projects. Identifying samples responsible for multiple relationship errors sometimes presents a challenge, if most of the errors are sex-compatible. The major exception is false paternity errors, which change putative full-sibling relationships to apparent half-sibling, and one parent-child relationship to apparent unrelated. Otherwise, in pedigrees with multiple relationship errors, search for a pair of individuals that, if their DNA samples were swapped, could resolve all or most of the errors. Similarly, where multiple samples are inconsistent with blind duplicates or with the same

individuals from previous genotype projects, try to resolve inconsistencies by swapping DNA samples for a pair of individuals. An unexpected duplicate pair of samples may correspond in this way to an unexpected non-duplicate pair.

The relationship check may reveal or confirm contaminated DNA samples. A contaminated sample sometimes appears to have a moderately close (second- or third-degree) relationship with most of the samples in the data set. Run a quick scan for such samples, using PLINK's relationship estimation (`--genome`) with a moderately high cutoff for output of estimates (only report relationships with $\hat{r} \geq 0.10$, i.e., the `--min 0.1` option), and screen the output for sample IDs that appear far more often than predicted from the sample pedigree structure.

Resolving pairs of inadvertently exchanged samples is feasible only if relatives or duplicates of the affected samples are present in the data set, or in some cases if the samples belong to different continental populations easily resolved by PCA. Generally, independent (singleton) samples with sex-mismatch errors, family members apparently unrelated to the rest of the pedigree, and unexpected duplicate (and non-duplicate) samples must be discarded if a DNA sample switch cannot be identified.

Marker Quality

Several screens for marker quality beyond those applied during genotype calling in GenomeStudio, can be implemented as summarized in Table 3. Most of these tests are straightforward and available in PLINK. We suggest a final call rate threshold for SNPs of 98%. The test for departure from Hardy-Weinberg proportions (dHWP; the `--hwe` screen or `--hardy` report in PLINK) has very low power for rare variants, and is not recommended for rare variants in a case-control sample where they are expected to be of high penetrance if causal. Genotypes of HapMap control samples should be tested against their known genotypes available from dbSNP, and any SNP with one or more genotype discrepancies should be omitted (publicly available datasets for HapMap and 1000 Genomes samples are available, as described above). Checks for differences in male and female allele frequencies and male heterozygosity for X-linked markers should be run as indicated in Table 3, by calculating allele frequencies in multiple subsets of the data set in a single PLINK analysis incorporating a text file containing family ID, individual ID and a subgroup indicator (such as "M" and "F"), and using PLINK's `--within <filename>` option in the `--freq` function.

In this example screening individual markers for Mendelian transmission errors was not possible. A large number of Mendelian incompatible genotypes, where parent-child data are available, can indicate poor-quality common variants, but Mendelian errors in a rare variant will be uncommon and may indicate *de novo* mutations.

The HumanExome panel (version 1.1) contains 754 pairs of duplicated markers (98 on the same DNA strand, 656 on complementary strands) and 45 triallelic SNPs, each represented by two diallelic markers (Guo et al. (2014) Suppl. Table 7). Before analysis, screen duplicate markers pairs and prune to one marker per unique map position, since multiple markers at identical map positions may cause errors in PLINK and other analysis software programs. A

reasonable approach is to retain the copy of each duplicate pair that has the higher call rate. Do not use triallelic markers in ordinary GWAS, since the true genotypes will differ from the apparent SNP genotypes. Moreover, PLINK is specialized for panels of diallelic markers, and therefore cannot make use of the additional information from the third allele.

DNA source

Although the standard QC procedures presented above are sufficient to test the quality of WGA samples, allele frequencies for a small percentage of markers may differ between WGA samples and other (non-WGA) samples. We found that such differences primarily occur in rare variants, with allele frequencies sometimes being more than 10-fold different. To locate such variants, conduct a simple comparison of allele frequencies in WGA samples compared to all other samples using PLINK by means of a Fisher exact test for association (`--fisher` function) using WGA status as a binary phenotype. In the NEIGHBORHOOD Exome Chip example we discovered 358 SNPs out of a total of 180,320 polymorphic SNPs passing QC with highly significantly different allele frequencies by DNA source (Fisher exact $p < 0.001$). Of these, 229 were rare ($MAF < 0.01$) in non-WGA samples, and the median fold difference in allele frequencies of these rare SNPs was 11.9, with the MAF nearly always greater in the WGA samples. Flag markers with significant differences in allele frequency in WGA and non-WGA samples for close inspection, or omit from tests of association for rare variants if the cases and controls have different proportions of WGA samples. Alternatively, add a covariate to the genetic analysis to control for DNA source (WGA vs. non-WGA), although it is not certain that this adjustment will sufficiently account for differences in allele frequencies according to DNA source, especially if WGA samples are highly concentrated within one phenotype class or if allele frequencies are widely different in the two types of DNA.

Acknowledgments

This work was supported in part by National Eye Institute grant RO1EY022305. J.N.C.B. is supported in part by a PhRMA Informatics Fellowship. Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I.

Literature Cited

- Babron MC, de Tayrac M, Rutledge DN, Zeggini E, Génin E. Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS ONE*. 2012; 7:e46519. [PubMed: 23071581]
- Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. *Bioinformatics*. 2013; 29:1399–1406. [PubMed: 23539302]
- Edwards TL, Gao X. Methods for Detecting and Correcting for Population Stratification. *Curr Protoc Hum Genet*. 2012; 73:1.22.1–1.22.14.
- Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, de Ståhl TD, Zaghlool A, Giedraitis V, Lannfelt L, Score J, Cross NCP, Absher D, Janson ET, Lindgren CM, Morris AP, Ingelsson E, Lind L, Dumanski JP. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet*. 2008; 46:624–628.
- Geschwind DH, Boone KB, Miller BL, Swerdloff RS. Neurobehavioral phenotype of Klinefelter syndrome. *Ment Retard Dev Disabil Res Rev*. 2000; 6:107–116. [PubMed: 10899803]

- Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, O'Dushlaine C, Moran JL, Chambert K, Stevens C, Swedish Schizophrenia Consortium, ARRA Autism Sequencing Collaboration. Sklar P, Hultman CM, Purcell S, McCarroll SA, Sullivan PF, Daly MJ, Neale BM. zCall: a rare variant caller for array-based genotyping. *Bioinformatics*. 2012; 28:2543–2545. [PubMed: 22843986]
- Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, Taylor KD, Hansen M, Borecki IB, Cupples LA, Fornage M, Gudnason V, Harris TB, Kathiresan S, Kraaij R, Launer LJ, Levy D, Liu Y, Mosley T, Peloso GM, Psaty BM, Rich SS, Rivadeneira F, Siscovick D, Smith AV, Uitterlinden A, van Duijn C, Wilson JG, O'Donnell CJ, Rotter JI, Boerwinkle E. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS ONE*. 2013; 8:e68095. CHARGE Consortium quality control criteria for genotype calling on a combined set of 62,000 individuals genotyped at multiple study centers, with links to CHARGE Consortium resources for the HumanExome chip, including genotypes for HapMap samples and a custom GenomeStudio clustering (.egt) file. [PubMed: 23874508]
- Guo X, Liu Z, Wang X, Zhang H. Genetic association test for multiple traits at gene level. *Genet Epidemiol*. 2012; 37:122–129. Detailed protocols for processing and quality control of Illumina HumanExome chip data, with links to supporting scripts in R and Python and many useful details for running software packages GenomeStudio, zCall and PLINK. [PubMed: 23032486]
- Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, Samuels DC, Shyr Y, Long J. Illumina human exome genotyping array clustering and quality control. *Nat Protoc*. 2014; 9:2643–2662. [PubMed: 25321409]
- Hindorf LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA. A catalog of published genome-wide association studies. 2016. <http://www.genome.gov/gwastudies>
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genomewide association studies. *PLoS Genet*. 2009; 5:e1000529. [PubMed: 19543373]
- Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*. 2012; 91:839–848. [PubMed: 23103226]
- Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS. GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*. 2010; 34:591–602. [PubMed: 20718045]
- Lee S, Abecasis G, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014; 95:5–23. [PubMed: 24995866]
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 2010; 34:816–834. [PubMed: 21058334]
- Manichaikul A, Mychaleckyj J, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010; 26:2867–2873. [PubMed: 20926424]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461:747–753. [PubMed: 19812666]
- Milligan BG. Maximum-likelihood estimation of relatedness. *Genetics*. 2003; 163:1153–1167. [PubMed: 12663552]
- Page CM, Baranzini SE, Mevik BH, Bos SD, Harbo HF, Andreassen BK. Assessing the power of exome chips. *PLoS ONE*. 2015; 10:e0139642. [PubMed: 26437075]
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006; 2:e190. [PubMed: 17194218]
- Perreault LPL, Legault MA, Barhdadi A, Provost S, Normand V, Tardif JC, Dubé MP. Comparison of genotype clustering tools with rare variants. *BMC Bioinform*. 2014; 15:52.

- Porcu E, Sanna S, Fuchsberger C, Fritsche LG. Genotype Imputation in Genome-Wide Association Studies. *Curr Protoc Hum Genet*. 2013; 78:1.25.1–1.25.14.
- Purcell S, Neale B, Todd-Brow K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007; 81:559–575. [PubMed: 17701901]
- Romm J, McMullen I, Jewell M, Zhang J, Zhang P, Ling H, Pugh E, Doheny KF. Using GenTrain and zCall to identify problematic SNPs in rare variant genotype calling. Presented at the American Society of Human Genetics 2013 Annual Meeting. 2013
- Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet*. 2011; 68:1.19.11–1.19.18. Comprehensive overview of common quality-control screens for genomewide association data.
- Veerappa AM, Padakannaya P, Ramachandra NB. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Functional & Integrative Genomics*. 2013; 13:285–293. [PubMed: 23708688]
- Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*. 2010; 464:713–722. [PubMed: 20360734]
- Wiggs JL, Hauser MA, Abdrabou W, Allingham RR, Budenz DL, DelBono E, Friedman DS, Kang JH, Gaasterland D, Gaasterland T, Lee RK, Lichter PR, Loomis S, Liu Y, McCarty C, Medeiros FA, Moroi SE, Olson LM, Realini A, Richards JE, Rozsa F, Schuman JS, Singh K, Stein JD, Vollrath D, Weinreb RN, Wollstein G, Yaspan BL, Yoneyama S, Zack D, Zhang K, Pericak-Vance MA, Pasquale LR, Haines JL. the NEIGHBOR Consortium primary open angle glaucoma genome-wide association study: rationale, study design and clinical variables. *J Glaucoma*. 2013; 22:517–525. [PubMed: 22828004]
- Wiggs JL, Yaspan BL, Hauser MA, Kang JH, Allingham RR, Olson LM, Abdrabou W, Fan BJ, Wang DY, Brodeur W, Budenz DL, Caprioli J, Crenshaw A, Crooks K, DelBono E, Doheny KF, Friedman DS, Gaasterland D, Gaasterland T, Laurie C, Lee RK, Lichter PR, Loomis S, Liu Y, Medeiros FA, McCarty C, Mirel D, Moroi SE, Musch DC, Realini A, Rozsa FW, Schuman JS, Scott K, Singh K, Stein JD, Trager EH, VanVeldhuisen P, Vollrath D, Wollstein G, Yoneyama S, Zhang K, Weinreb RN, Ernst J, Kellis M, Masuda T, Zack D, Richards JE, Pericak-Vance MA, Pasquale LR, Haines JL. Common variants at 9p21 and 8q22 are associated with increased susceptibility to optic nerve degeneration in glaucoma. *PLoS Genet*. 2012; 8:e1002654. [PubMed: 22570617]
- Zeng Z, Weeks DE, Chen W, Mukhopadhyay N, Feingold E. A pipeline for classifying relationships using dense SNP/SNV data and putative pedigree information. *Genet Epidemiol*. 2016; doi: 10.1002/gepi.21948
- Zhang Y, Shen X, Pan W. Adjusting for population stratification in a fine scale with principal components and sequencing data. *Genet Epidemiol*. 2013; 37:787–801. [PubMed: 24123217]

Internet Resources

- http://genome.sph.umich.edu/wiki/Exome_Chip_Design Description of the development and contents of the exome-based marker array implemented as the Illumina HumanExome chip.
- <ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/annotatedList.txt> List of proposed markers for the HumanExome chip with role of variants in the panel (nonsynonymous coding variant, splice variant, AIM, linkage grid marker, etc.).
- <http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml> Source for PLINK software download, documentation and tutorials.
- <https://cran.r-project.org> Source for the R software and R packages.
- <http://www.chargeconsortium.com/main/exomechip> Resources for Illumina HumanExome data processing developed by the CHARGE Consortium.

<https://github.com/jigold/zCall> Source for zCall scripts and protocols.

http://support.illumina.com/array/array_software/genomestudio/downloads.html Source for extensions to the GenomeStudio software for Illumina genotype calling and QC.

http://www.illumina.com/documents/products/technotes/technote_topbot.pdf Illumina, Inc. algorithm for assigning the TOP strand.

<http://www.well.ox.ac.uk/~wrayner/strand/> Correspondence between the Illumina TOP strand and the dbSNP + strand for HumanExome markers.

https://support.illumina.com/array/array_kits/infinium_humanexome_beadchip_kit/downloads.html Accessory files for the HumanExome Beadchip, including cluster (.egt) files, gene annotation data, rsID information and changes made between array versions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Plot of mean LRR of Y vs. X for recorded (putative) males (blue squares) and females (magenta circles) without (A) and including (B) WGA samples. Outlying values are circled and annotated with putative sex chromosome anomalies. An asterisk indicates a confirmed XXX individual. The XO and XXX individuals are not well resolved when WGA samples are present.

Figure 2.

Plot of Y chromosome call rate vs. estimated inbreeding for chromosome X (F_X) without (A) and including (B) WGA samples. Recorded sex, and potentially sex-aneuploid samples, are indicated as in Fig. 1. One WGA sample, scored as negative for contamination and sex chromosome aneuploidy by mean intensity values, is indicated (“Not IDed as aneuploid”).

Figure 3. Comparison of unfiltered (**A**) and filtered (**B**) BAF and LRR plots for CEPH individual NA07029 (male). X and Y chromosomes are indicated; horizontal lines in bottom LRR plots show mean chromosomal LRR.

Figure 4. BAF and LRR plots from selected individuals screened as sex-chromosome aneuploid: (A) XXX; and (B) XXY; (C) XY by mean intensity and XXY by chromosome X heterozygosity.

Figure 5. First two principal components from PCA using only the HumanExome AIMS (A) and (B) 9,647 AIMS in SNPweights. HapMap control samples are colored as indicated. In panel B, the African American (AA) and Hispanic (H) subsamples are indicated.

Table 1
Sample Quality Control Screens

Screens for exclusion	
Screen	Exclusion Criterion
Overall call rate	< 98.5%
Contamination (VerifyIDIntensity)	< 10%
Autosomal heterozygosity	> 4 SD above cohort mean
Screens to flag for manual resolution	
Screen	Flag Criterion
Mendelian inconsistency	Inconsistent with parent/child
Genotype concordance with blind duplicate	< 99%
Genotype concordance with previous genotyping project (if available)	< 95%
Sex mismatch	Disagreement between apparent genetic sex and sex on record
Relationship check	Incorrect relationship within family; first- or second-degree relative pair between families or singletons

Table 2
Physical Locations of Pseudoautosomal and X-translocation Regions

Region	GRCh37 Coordinates	
	X Chromosome	Y Chromosome
PAR1 ^a	60,001 – 2,699,520	1,001 – 2,649,250
PAR2 ^a	154,931,044 – 155,260,560	59,034,050 – 59,363,566
XTR ^b	88,395,830 – 92,583,067	2,926,393 – 6,543,100

^aThese coordinates and GRCh38.p5 coordinates are available from the Genome Reference Consortium, <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/>.

^bSee Veerappa et al. (2013).

Table 3
Marker Quality Control Screens

Screen	Exclusion Criterion
Cluster separation	$< 0.2^a$
Manual review	Failed ^a
Call rate	$< 98\%$ (except Y chrom.)
Heterozygosity	> 0.8
HapMap replicate errors	> 0
Absolute difference in male and female allele frequencies	> 0.0781 (0.07 for X chrom.)
Male heterozygosity, X chromosome	> 0.01
Exact test of Hardy-Weinberg proportions	$p < 10^{-6}$
Fisher test for difference in allele frequency between WGA and non-WGA samples	$p < 0.0001$ (flagged only)

^aScreen performed in GenomeStudio.