Published in final edited form as:
Nat Protoc. 2014 May ; 9(5): 1192–1212. doi:10.1038/nprot.2014.071.

Quality control and conduct of genome-wide association meta-analyses

Thomas W Winkler[1], **Felix R Day**[2], **Damien C Croteau-Chonka**[3,4], **Andrew R Wood**[5], **Adam E Locke**[6], **Reedik Mägi**[7], **Teresa Ferreira**[8], **Tove Fall**[9,10], **Mariaelisa Graff**[11], **Anne E Justice**[11], **Jian'an Luan**[2], **Stefan Gustafsson**[9], **Joshua C Randall**[12], **Sailaja Vedantam**[13,14,15], **Tsegaselassie Workalemahu**[16], **Tuomas O Kilpeläinen**[17], **André Scherag**[18,19], **Tonu Esko**[7,13,14,15], **Zoltán Kutalik**[20,21,22], **the GIANT consortium**, **Iris M Heid**[1,*], and **Ruth JF Loos**[23,24,25,*]

[1]Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany [2]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK [3]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA [4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA [5]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK [6]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA [7]Estonian Genome Center, University of Tartu, Tartu, Estonia [8]Wellcome Trust Centre For Human Genetics, University of Oxford, Oxford, UK [9]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden [10]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden [11]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina, USA [12]Wellcome Trust Sanger Institute, Cambridge, UK [13]Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, USA [14]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA [15]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA [16]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA [17]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark [18]Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital of Essen, University of Duisburg-Essen, Essen, Germany [19]Clinical Epidemiology, Integrated Research and Treatment Center, Center for Sepsis

Corresponding authors: Ruth J F Loos, ruth.loos@mssm.edu, Icahn School of Medicine at Mount Sinai, New York, USA, Iris M Heid, iris.heid@klinik.uni-regensburg.de, University of Regensburg, Germany.

[*]These authors supervised this work jointly

AUTHOR CONTRIBUTIONS

Writing group: TWW, FRD, DCCC, ARW, AEL, RM, TFe, TOK, AS, TE, ZK, IMH, RJFL
Pipeline and procedure development:
TWW, FRD, DCCC, ARW, AEL, RM, TFe, TOK, AS, TE, ZK
Analyst contributing to the QC of recent GIANT papers:
TWW, FRD, DCCC, ARW, AEL, RM, TFe, TFa, MG, AEJ, JL, SG, JCR, SV, TW, TOK, AS, TE, ZK

COMPETING FINANCIAL INTERESTS

The authors have no competing financial interests to declare.

Control and Care (CSCC), Jena University Hospital, Jena, Germany [20]Department of Medical Genetics, University of Lausanne, Lausanne, Switzerland [21]Institute of Social and Preventive Medicine (IUMSP), Centre Hospitalier Universitaire Vaudois (CHUV), Lausanne 1010, Switzerland [22]Swiss Institute of Bioinformatics, Lausanne, Switzerland [23]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA [24]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA [25]The Genetics of Obesity and Related Metabolic Traits Program, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Abstract

Rigorous organization and quality control (QC) are necessary to facilitate successful genome-wide association meta-analyses (GWAMAs) of statistics aggregated across multiple genome-wide association studies. This protocol provides guidelines for [1] organizational aspects of GWAMAs, and for [2] QC at the study file level, the meta-level across studies, and the meta-analysis output level. Real–world examples highlight issues experienced and solutions developed by the GIANT Consortium that has conducted meta-analyses including data from 125 studies comprising more than 330,000 individuals. We provide a general protocol for conducting GWAMAs and carrying out QC to minimize errors and to guarantee maximum use of the data. We also include details for use of a powerful and flexible software package called *EasyQC*. For consortia of comparable size to the GIANT consortium, the present protocol takes a minimum of about 10 months to complete.

INTRODUCTION

Background

The genome-wide association (GWA) study approach has been extremely successful in pinpointing association of common genetic variants with diseases or disease-related quantitative phenotypes[1, 2]. However, given the small sizes of the expected effect under a polygenic model, individual GWA studies are generally too small to provide the necessary power to detect single nucleotide polymorphism (SNP) associations, while accounting for the multiple number of independent tests. Therefore, the genetics community has widely adopted the approach of combining summary statistics from multiple GWAS into a single meta-analysis to increase the statistical power of the analysis by augmenting the effective sample size[3, 4]. These GWAMAs collate data from GWA studies conducted around the world and thus require an enormous organizational effort to ensure effective communication, standardization of analytical procedures, and coordination at both the study-specific level and the meta-analysis level, followed by rigorous quality control (QC) during the meta-analysis process. Although a QC protocol for individual GWA studies has been described before[5], a comprehensive protocol describing state-of-the-art procedures to conduct and perform QC of large-scale GWAMAs is currently lacking.

The typical GWAMA approach is to design a standardized analysis plan centrally and share it with the individual study partners who will perform the GWAs according to the designated analysis plan. More specifically, the study analysts conduct study-specific GWA

QC for each SNP, and impute the genome-wide SNP array data. Next, they compute association statistics for each SNP, including effect size estimates with standard errors (or odds ratios with corresponding confidence intervals for case-control samples), allele frequencies, sample size, and P-values, and provide these summary statistics to the meta-analyses centers. Typically, data on the individual participants, alongside phenotype and genome-wide SNP genotype information, are not shared to guarantee anonymity of study participants and to conform to strict data-sharing policies. The unavailability of individual participant data at the meta-analyses centers creates unique analytical challenges for QC, requiring specific statistical and graphical tools to track errors in the study-specific analysis from the available aggregated data.

Study-specific data issues that need to be detected at the meta-analysis stage include file naming errors (e.g. female-specific files labeled as male-specific), erroneous SNP genotype data (e.g. flipped alleles, duplicate SNPs, bad imputation quality), and association issues stemming from incorrect analysis models (e.g. improper model adjustments, population stratification, and unaccounted relatedness of individuals). Although some errors impede the meta-analysis (e.g. file formatting errors), others (e.g. incorrect trait transformations and flipped alleles) limit the full contribution of a study to the meta-analysis and thus lower the power of the meta-analysis or inflate the number of false positives (type I errors, e.g. unaccounted population stratification). Issues that inflate the number of type I errors should be avoided with higher priority than issues that increase the number of false negatives (type II errors), which negatively affect the statistical power of the meta-analysis. False positives could set researchers onto the wrong track, leading them to spend time and money on misguided follow-up studies, whereas missed genetic signals can be expected to emerge in a following, larger GWAMA.

A typical GWAMA involves two stages: (i) a discovery stage, in which meta-analyzed GWA data are used to select promising variants, and (ii) a follow-up stage, in which analyses are performed on data derived either from *de novo* genotyping or from existing genome-wide data (*in silico*). This protocol focuses on the discovery stage. Although *in silico* follow-up data can generally be treated similarly to discovery GWA data for QC purposes, *de novo* genotyped data needs to be checked with a particular focus on SNP strand issues, call-rate, Hardy-Weinberg equilibrium (HWE)[5] or other technical steps related to the particular genotyping technology applied.

In recent years, GWAMAs have become more and more complex. Firstly, GWAMAs can extend from simple analysis models to more complex models including stratified[6] and interaction[7, 8] analyses. Secondly, beyond imputed genome-wide SNP arrays, new custom-designed arrays such as Metabochip[9], Immunochip[10], and Exomechip[11] are increasingly integrated into meta-analyses. Because of differing SNP densities, strand annotations, builds of the genome, and the presence of low-frequency variants, data from such arrays require additional processing and QC steps (also outlined in this protocol using the example of the Metabochip). Finally, GWAMAs involve an ever-increasing number of studies. Up to a hundred studies were involved in recent GWAMAs[12–17], often involving 1,000 to 2,000 study-specific files. Increasing the scale and complexity of GWAMAs increases the likelihood of errors by study analysts and meta-analysts, underscoring the need for more extensive and automated GWAMA QC procedures.

We present a pipeline model that provides GWAMA analysts with organizational instruments, standard analysis practices, and statistical and graphical tools to carry out QC and to conduct GWAMAs. The protocol is accompanied by an R package, *EasyQC*, a user-friendly software that implements this GWAMA QC pipeline and is flexible to accommodate additional and alternative steps.

Development of the protocol

Our protocol was developed by analysts from the GIANT Consortium, which is one of the largest global collaborations to study complex traits and diseases, currently including up to 125 studies into the meta-analysis. Established in 2006, GIANT has accumulated a lot of experience with GWAMAs. Four rounds of analyses have already been conducted, with each round incorporating new studies and chip technologies. [13, 15, 18–20]. Our work illustrates the increasing complexity of GWAMAs: we deal with multiple phenotypes (e.g. height, body mass index (BMI), waist-hip ratio (WHR), waist and hip circumference (WC and HIP), the latter three also with adjustment for BMI (WHR_{adjBMI}, WC_{adjBMI}, HIP_{adjBMI}), and body fat percentage), multiple SNP platforms (genome-wide SNP and Metabochip arrays), multiple analysis models (without and with adjustment for BMI, interaction with smoking status and with physical activity, sex- and age- stratified analyses, and various dichotomizations of the BMI distribution[6, 21]), including imputed and unimputed SNP data, and an ever-increasing number of studies per meta-analysis (16 initially and up to 125 in the current analyses). Our on-going analyses include more than 1,500 GWA input files, necessitating an efficient QC pipeline. The size and experience of the GIANT Consortium provides an ideal basis for the development of a GWAMA protocol. The protocol and tools can readily be applied by other consortia using aggregated statistics for meta-analysis, studying other quantitative traits and using other statistical models or other genotyping platforms. We have incorporated all QC steps that proved to be helpful during our GIANT work and have been known to be efficient in other consortia's work. We have also developed special tools to conduct meta-level QC and to handle the particularly large number of files.

Limitations

Firstly, this protocol has been developed for human genomic data. Although some aspects can be applied to non-human data, a detailed protocol for other species is beyond the scope of the present protocol.

Secondly, even a perfect protocol for the meta-analysis of aggregated statistics cannot fully compensate for not having access to individual participant data, which would guarantee standardized QC and analyses across studies. Advantages and disadvantages of meta-analyses using individual participant data are summarized in the "Comparison with other approaches" section, below. However, ethically motivated restrictions to sharing genome-wide genotype and phenotype data currently limit the realization of individual participant

GWAMAs, which is the reason why the aggregated statistics GWAMA – as described here – is the currently most widely applied approach.

Applications of the protocol

Generally, this protocol assumes that the study analysts have quality-controlled their study data regarding phenotype and genotype as well as accounted for ethnicity, race and familial relatedness. For these steps, there are standardized protocols available[5]. It also assumes that they either have imputed their genome-wide SNP array data – ideally with a pre-specified common reference panel – to ensure a common SNP panel across all studies, or that they have data from an unimputed custom genotype array available.

This protocol specifically focuses on the discovery stage of a GWAMA, but can be readily applied to the follow-up stage as well. Imputed *in silico* follow-up data can be treated in a similar way as the here described imputed genome-wide SNP array data, non-imputed *in silico* or *de novo* genotyped data can be treated like the Metabochip data regarding the cleaning of call rate, HWE, and strand issues.

Although this protocol has been developed for quantitative phenotypes and HapMap imputed or typed common autosomal genetic variants, it can be extended to 1000 Genomes imputed variants, dichotomous phenotypes, rare variants, gene-environment interaction (GxE) analyses and to sex chromosomal variants. A summary of directly applicable protocol steps or steps requiring adaptation is given in Table 1. Since 1000 Genomes imputed data extends to a larger SNP panel and includes structural variants (SV) and insertions or deletions (indels), the allele coding and harmonization of marker names require special consideration: (i) Additional allele codes (other than “A”,”C”,”G” or ”T”) are needed for indels and SVs (e.g., “I” and “D” for insertions and deletions). (ii) To account for the fact that some SVs and indels map to the same genomic position as SNPs, the identifier format “chr<chromosome>:<position>” would introduce duplicates. Therefore, the identifier format needs to be amended (e.g. to “chr<chromosome>:<position>:[snp|indel]”, which adds the type to the format).

For dichotomous traits, the effective sample size needs to be computed by N_{eff}=2/(1/N_{cases} + 1/$N_{controls}$), an expression that balances the number of cases with the number of controls. Custom-array data require checks of genotype quality per case status. The analysis is usually performed using logistic instead of linear regression providing beta estimates and standard errors that enable the implementation of the same meta-analysis methods. The MAC cut-off requires more consideration: It depends on the logistic-regression-based test used and on the ratio between number of cases and controls[22].

For rare and low-frequency variants more refined considerations regarding the minimal sample size or the minimally acceptable MAC cut-off per file are required. The comparability of the study frequencies with reference data such as HapMap or 1000 Genomes is of limited use as Exomechip or custom-made chips focusing on rare variants and low-frequency tend to include novel or population-specific variants. Often, the single-variant analyses are complemented by gene-based burden tests requiring special consideration. For single-variant analyses, most of the protocol steps described herein are directly applicable.

Results for analyses models that include an interaction term can also be quality-controlled by this protocol. The main SNP effect estimates can be treated like the SNP effects without interaction. The interaction effect estimates need to be cleaned and meta-analyzed in addition. This objective can be achieved in the same fashion as the main effect estimates or implementing alternate methods[23]. As the analysis of the interaction between SNP and the environment is more and more included into GWAMA efforts, this approach will be of increased importance.

Analyses with sex-chromosomal variants require some special considerations, in particular in men. We assume that study partners have quality-controlled their data regarding rare gonosomal aberrations (X0, XXX, XYY). The potential errors in coding variants in men include differences in the coding of X-chromosomal variants (either 0|1 or 0|2 for men) or erroneous coding of pseudo-autosomal variants (should be 0|1|2). Separating the QC by X-, Y-, and pseudo-autosomal variants in men can be grasped by deflated or inflated beta-estimates (and thus standard errors) in the SE-N (i.e., inverse of the median standard error versus the square root of the sample size) plot. Generally, sex-chromosomal variants should be cleaned and analyzed in men and women separately.

Comparison with other approaches

Over the past six years, more than 100 large phenotype-driven consortia of genetic association studies have emerged[1]. Most of these consortia follow a similar framework for QC and data ‘sanity checks’ as outlined here[24].

Some consortia, such as the Uric Acid (UA) Consortium, follow slightly modified procedures, whereby study-specific QC metrics, generated by GWAStoolbox[25], were collected next to summary level association statistics[26]. This approach enables the easy detection of basic data problems even before the results are shared, but at the same time it poses an extra burden on the analysts, and its implementation does not help the necessity of meta-level checks. The Chronic Kidney Disease Genetics (CKDGen) Consortium omits filtering data based on poor imputation quality[27], whereas most consortia, including GIANT, delete badly imputed variants from the meta-analysis (see below).

Whereas most GWAMAs meta-analyze study-specific statistics, where study analysts have provided GWA results to the meta-analysis center, the Psychiatric Genomic Consortium (PGC) conducts a meta-analysis of individual participant data, as both the individual-level genotype and phenotype data of all participating studies are deposited centrally[28]. This approach has the following advantages: 1) central quality control: genotype and phenotype data can be modeled and quality-controlled centrally, eliminating the need for subsequent troubleshooting; 2) standardized study-specific analyses: fewer analysts are involved and the utilization of the same imputation and association analysis software is guaranteed; and 3) flexibility: more complex and comprehensive statistical analyses can be conducted without burdening a large number of study analysts. However, our GWAMA approach has also advantages compared to the meta-analysis of individual participant data: 1) gathering experts: the more analysts are involved, the more the network can profit from the accumulated expertise; 2) local know-how: local study analysts know their study better than a central team of meta-analysts; and 3) compliance: ethically motivated restrictions may limit the sharing of genome-wide genotype and phenotype data due to the risk of participant identification inhibit the study contribution[29–31]. In summary, the framework presented in this protocol reflects the currently most widely applied GWAMA conduct and QC approach.

Experimental Design

Organizational aspects of the conduct of a typical GWAMA (Steps 1–6)

The typical GWAMA starts with the setting up of logistics aimed at achieving a smooth communication between participating partners, analysts, and principal investigators, limiting the burden for study analysts, so as to ensure a timely delivery of results to the meta-analysis team.

Once study partners have been identified, general rules for the collaboration can be issued in a ‘memorandum of understanding’ to set out the guidelines of confidentiality, data access, publication of results, and authorship. Subsequently, collaborators and analysts are invited to join task groups and regular teleconference calls.

An analysis plan is designed centrally by the meta-analysts to describe the standardized analyses to be performed ‘locally’ and to detail phenotype transformation (e.g. to deal with non-normal phenotype distributions and to enable comparability across studies), genotype handling, imputation requirements, and association analysis methods (statistical model, adjustment, stratification). Where possible and reasonable, software scripts are provided to every participating study group to minimize the potential of errors and to alleviate the analysis burden for the study analyst. The analysis plan also defines the required aggregated association statistics (e.g. SNP identifier, effect allele, allele frequency, beta estimate, standard error, sample size, call rate or imputation quality, and P-value) and details the format in which they need to be submitted (see Box 1). In the design of the analysis plan, whether or not to provide detailed and lengthy guidelines, possibly including even software codes, needs to be weighed against providing a short and comprehensive — but potentially more error-prone — description. The less standard the requested analyses, the more details need to be provided. A general analysis plan format cannot be provided, but the GIANT analysis plan can serve as an example that has worked and has been improved through several rounds of meta-analyses (Supplementary Manual). The analysis plan is discussed with the study collaborators and then sent out to each study analyst, including a deadline and server access details for data upload.

When data from all studies have been uploaded to a password-secured file server, a data freeze ensures the integrity of the data for all meta-analysts, regardless of download time (Supplementary Figure 1).

The complete turnaround time for consortia comparable in size to GIANT (>100 studies in meta-analysis) is, at minimum, around 10 months: 2 months to set up the logistics and to develop the analysis plan, 2 months to collect the data after the analysis plan has been sent out, and 6 months to perform QC and meta-analysis.

QC—The workflow involves three QC steps: file-level QC (Steps 7–18), meta-level QC (Steps 19–26), and meta-analysis QC (Steps 29–32). The file-level QC tackles formatting issues that can be checked independently on each study file. In the meta-level QC, the study-specific statistics are compared across studies or with reference panels to detect errors in the analyses that cannot be identified by examining the study files individually. The meta-analysis QC works on the level of already aggregated meta-analysis results and helps to remove or flag suspicious SNP results. The workflow and the three QC steps are presented in Figure 1.

File-level QC (Steps 7–18): This stage involves 'cleaning' (deleting poor quality data) and 'checking' (providing summaries to judge data quality) data. Thresholds for what data to remove are typically defined *a priori* (e.g. by this protocol). Although data checking should ascertain that there are no issues left, it often reveals further issues, which require re-cleaning and re-checking. A few QC iterations may be needed before all files are fully cleaned and ready for meta-analyses. Which SNPs or study files are to be removed depends on how much the improvement in data quality weighs against loss of data. On the one hand, the stricter the QC, the more SNPs or study files are removed and thus the lower the coverage or sample size (and thus power). On the other hand, the more relaxed the QC requirements, the larger the coverage and sample size at the expense of data quality, which also decreases power.

Clearly, monomorphic SNPs or SNPs with missing (e.g. missing P-value, beta estimate, or alleles) or nonsensical information (e.g. alleles other than A, C, G, or T, P-values or allele frequencies >1 or <0, or standard errors ≤0, infinite beta estimates or standard errors) are of no help to the meta-analysis and need to be removed. Systematically missing values or errors can point towards analysis problems; thus, such data calls into question the correctness of the data and should be discussed with the study analyst. A large number of monomorphic SNPs can also point towards study-specific array problems.

If a study includes a low number of individual participants, its summary statistics can be unstable (e.g. zero or infinite standard errors, zero P-Values or extremely large beta estimates), which might drive the meta-analysis towards detecting false positives. This risk pertains in particular to low-frequency variants. The detection of false positives due to the low statistical power of the meta-analysis can be avoided by requiring a minimum sample size per study and a minimum number of minor alleles contributing to a SNP for each participating study. For example, in meta-analyses performed by the GIANT consortium, SNPs were removed from the study file if the number of individuals informative for the SNP was lower than 30 or the minor allele count was (MAC, computed as 2*MAF*N, with MAF being the minor allele frequency) equal or less than 6.

Imputed genotype data is often filtered based on imputation quality. For example, in the GIANT consortium, poorly imputed SNPs were removed according to a threshold that depended on the imputation method and on the imputation quality metric (Table 2).

Arguably, however, SNPs with poor imputation quality can be retained in the meta-analysis[27]: on one hand, a badly imputed SNP can be considered a random, non-differential error in the genotype (i.e. not systematically prioritizing one genotype and independent of the phenotype) and thus it will not tend to create a false signal and, on the other hand, a study with the SNP badly imputed will neither contribute to a true signal nor mask it. Filtering poorly imputed SNPs has the advantage that no nonsensical results are unduly decreasing the statistical significance of truly informative data.

Sex-chromosomal and autosomal SNPs require different genotype models and therefore are often studied separately from each other. To focus on autosomal SNPs and consistent genotype models across studies in its analyses the GIANT consortium has removed any sex-chromosomal SNPs.

SNP identifiers often differ between arrays and/or imputation reference panels and, therefore, often differ between studies. Their harmonization across studies is pivotal to the meta-analysis. For example, a SNP that is assigned to two different SNP identifiers (e.g. rs123 in half of the studies and rs17614680 in the other half) will appear as two different SNPs in the meta-analysis output, with the total sample size split across the two SNPs; a true signal might, therefore, be missed due to loss of statistical power. For HapMap imputed studies, a unique SNP identifier can be generated by combining the SNP’s genetic positions to generate the format “chr<chromosome>:<position>”. However for some arrays (e.g. Metabochip) not all SNPs map to a standard reference panel. In such cases, the DNA probe sequences need to be mapped to the reference genome build of interest to arrive at a common chromosome and position, which can then be used to generate the SNP identifier. This procedure will also remove SNPs that do not map uniquely to the genome. Maps with unique SNP identifiers and genomic positions (for several different genome builds) for several commercial arrays are freely available for download (see: http://www.well.ox.ac.uk/~wrayner/strand/).

Meta-level QC (Steps 19–26): This stage consists in the cross-study comparison of statistics to identify study-specific problems. This QC stage compensates for not having the individual participant data of each study available to the meta-analyst. We recommend the following plots to be included in the GWAMA QC protocol.

The SE-N plot (Steps 19–20): Several types of analytical problems can be identified by depicting, for each study file, the inverse of the median standard error of the beta estimates across all SNPs against the square root of the sample size. The inverse proportionality between the median standard error and the square root of the sample size derives from the fact that the sampling variance of a linear regression–derived beta-estimate of a specific SNP *j* depends on the variance of the phenotype, *Var(Y)*, the variance of the SNP genotype, *Var(X$_j$)*, and the sample size $N_j\!:\!SE_j^2=var(\beta_j)=\frac{var(Y)}{N_j\cdot var(X_j)}$. If the regression model is adjusted, then *Var(Y)* reflects the variance of the residuals. Thus, the average of the standard errors across all SNPs will reflect the sample size. Assuming that the sample size for a given SNP is close enough to the maximum sample size for all SNPs, $N_j = N$,, the median of the standard errors across all *m* SNPs (*j=1…m*) can be written as $median(SE_j)=(\sqrt{var(Y)}/\sqrt{N})\cdot median(1/\sqrt{var(X_j)})$ and therefore

$$c\cdot\sqrt{var(Y)}\cdot\frac{1}{median(SE_j)}=\sqrt{N}\qquad(1)$$

with $c=median(\frac{1}{\sqrt{var(X_j)}})$. The constant c can be computed per study file incorporating the genotype frequencies (for genotyped variants) or the genotype dosages and imputation quality (for imputed variants) and will depend on individuals' ethnicity, genotyping platform, imputation reference panel, and imputation quality. Ignoring the uncertainty from the imputation, c can be approximated by

$$c\sim median(\frac{1}{\sqrt{2MAF_j(1-MAF_j)}}).\qquad(2)$$

However, the computation of *c* per study is not ideal for comparing the studies with each other. Differences in the MAF distribution between any individual study and the reference would not be detected. For several standard platforms, imputation panels and ethnicities, these approximate *c* values to be used in the SE-N plot are given in Table 3. For other platforms, panels or ethnicities, *c* is to be computed from a reference study or the imputation reference panel.

The study-specific data points of the SE-N plot will tend to describe a straight line. However, studies will deviate from the overall trend, if:

i. the study's phenotypic variance differs from other studies, which might be explained by a different study design or special study population;

ii. the study's MAFs differ from other studies, which might be explained by a diverging genotyping platform, reference panel for the imputation, or a different ethnicity;

iii. the study's SNP imputation qualities differ from those of other studies, which might reflect errors in the imputation or a different reference panel;

iv. the study's effective sample size differs from the stated sample size, which might be due to unaccounted relatedness between study participants or mis-coded sample size;

v. the study analyst has used a different statistical test; or

vi. the study analyst has mis-specified the phenotype transformation or the regression model, which results in a different phenotype variance or residual variance (see Figure 2, Anticipated Results, Supplementary Figure 2).

The P-Z plot (Steps 21–22): Analytical problems related to the study-specific computation of beta estimates, standard errors or P-values can also be revealed by a study-specific scatter plot that, for each SNP, compares the reported P-values with the P-values computed from the Z-statistics based on reported beta-estimate and standard error ($Z\ statistics = \beta_j/SE\ (\beta)_j$) (Figure 3, Anticipated Results, Supplementary Figure 3).

The EAF-plot (Steps 23–24): Plotting reported effect allele frequencies (EAF) against a reference set, such as from the HapMap[32] or 1000 Genomes[33] projects, or from one specific study, can help to visualize patterns that pinpoint strand issues, allele miscoding, or the inclusion of individuals whose self-reported ancestry did not match their genetic ancestry (Figure 4, Anticipated Results). A strand mismatch or allele miscoding may severely reduce statistical power. If, for example, a study (or several studies) reports alleles on the '−' instead of '+' strand, which cannot be corrected for 'palindromic' A/T or C/G SNPs, a true signal will be diminished, abolished, or even reversed. Although comparison of allele frequencies across studies will not detect strand issues or allele miscoding for SNPs with MAF close to 0.5, this comparison will be informative for most SNPs.

The lambda-N plot (Steps 25–26): Population stratification can either inflate or deflate association P-values and can be grasped by the genomic control (GC) inflation factor (λ_{GC})[34]. As λ_{GC} increases with sample size in the case of polygenic phenotypes[35], plotting λ_{GC} versus sample size per study file identifies inflated λ_{GC} and thus potential problems with population stratification (Figure 5, Anticipated Results). In the GIANT Consortium, analysts of studies with λ_{GC} >1.1 are contacted and asked to revisit their analyses (e.g. adjusting for principal components) and results.

Meta-Analysis and QC of meta-analysis output (Steps 27–32)—The meta-analysis combines the study-specific association results to obtain an overall estimate of the association and its P-value. The inverse-variance weighted meta-analysis using the fixed-effects model is most commonly used for GWAMAs (e.g. implemented in METAL[36]). The Q statistic and I^2 measure test and estimate between-study heterogeneity[22,37]. For SNPs with pronounced heterogeneity (I^2 > 75%), the effect estimation benefits from a random effects meta-analysis[38]. An alternative approach for deriving overall P-values is the sample size-weighted Z-score meta-analysis[39]. This approach is used when beta-estimates or standard errors are not available, or when the meta-analyzed traits are on a different scale (e.g. blood level data measured in different labs or differences in trait transformation) at the cost of losing power.

Meta-analyses are conducted by two meta-analysts independently, each uploading the results and log files on to the server (Supplementary Figure 1). Results are compared using (i) the log files that specify the study files included and the meta-analysis parameters set in the software program, (ii) descriptive statistics (min, median, max) of sample size and number of SNPs included in meta-analysis results, and (iii) correlation and scatter plot of P-values. Differences between the two analyses are resolved until agreement is reached.

To evaluate whether the statistics of the meta-analyzed effect are inflated due to population stratification accumulated across studies or due to unaccounted relatedness, the λ_{GC} is computed for the meta-analysis result (complementing the file-specific λ_{GC} values, see above). A high value (λ_{GC} >1.1) might be due to (a) an excess of association signals in large

GWAMAs for highly polygenic traits[14], (b) residual population stratification per study file accumulated across studies, (c) relatedness between individuals across strata, when the study-specific analyses have been performed separately by strata), or (d) related subjects across studies, which can more likely occur in very large GWAMAs. In the case of (c), a meta-analysis across strata per study can be conducted and a study-specific λ_{GC} >1.1 might provide insight into inflation requiring contact of the study analyst. Generally, we recommend applying the lambda GC correction on the file-level and on the meta-analysis level (*double GC correction*), but very large GWAMAs (> 200,000 individuals) on highly polygenic traits, such as height, may opt to omit the second GC correction (*single GC correction*)[35].

Finally, when all issues are resolved, one of the analysts shares the final results with the analysis task group (Supplementary Figure 1). The final results file will be used for all subsequent steps, including SNP selection for top hit identification and/or follow-up.

Special considerations for custom array data instead of genome-wide SNP array data—The GIANT consortium has worked on data genotyped using the Metabochip, a custom genotyping array that contains ~195,000 replication and fine-mapping SNPs chosen from GWAMAs of metabolic, cardiovascular and anthropometric traits[9]. Although many of the QC steps for HapMap imputed SNP data can be directly applied to the Metabochip and other customized genotype arrays, some steps need to be adjusted, which are summarized in the following section and given in the protocol as alternative route to using HapMap imputed SNP data: (i) To control genotype quality instead of imputation quality, a filter on call rate and deviation from Hardy-Weinberg equilibrium (HWE) is required; (ii) some genotyped SNPs may not be available in the HapMap reference data, which requires other references to identify strand and allele frequency errors; and (iii) to perform GC correction for chips designed to cover multiple traits, the calculation of the λ_{GC} needs to be limited to a subset of SNPs that are chosen from a trait that is uncorrelated with the trait of interest. This λ_{GC} is then to be applied to all SNPs on the array. For example, GIANT limits the λ_{GC} computation for Metabochip data to the 4,427 QT-interval SNPs as the QT-interval is uncorrelated with the GIANT traits, as recommended by the Metabochip designers[9].

Software—Using the standard, open-source and freely-available software R[40] and the graphical R package ‘Cairo’, we created a pipeline for completing this protocol into a downloadable GWAMA-QC R package called *EasyQC*. We provide code application directly in the procedure steps. The general basic usage is described in Box 2. Minimum system requirements are described in the Materials.

We provide a number of template scripts that enable to conduct multiple procedure steps at once: (i) *EasyQC* scripts ‘*1_filelevel_qc.gwa.ecf*’ and ‘*1_filelevel_qc.metabochip.ecf*’ to perform file-level QC (Steps 7–18); (ii) *EasyQC* script ‘*2_metalevel_qc.ecf*’ to perform meta-level QC (Steps 19–26); (iii) METAL-script ‘*3_metaanalysis.metal.txt*’ to perform the meta-analysis (Steps 27–28); (iv) *EasyQC* script ‘*4_metaanalysis_qc.compare.ecf*’ to compare two meta-analysis results for meta-analysis QC (Steps 29–30); (v) R script ‘*4_metaanalysis_qc.compare_logfiles.r*’ to compare two meta-analysis log-files with regards to included and excluded files for meta-analysis QC (Step 30); and (vi) *EasyQC* script '*4_metaanalysis_qc.studymeta.ecf*' to perform study-specific meta-analyses for meta-analysis QC (Steps 31–32).

Parts of the *EasyQC* template scripts and single *EasyQC* functions can also be included into other existing QC pipelines. This task can be accomplished by removing functions from the scripting interface of the template scripts (see Box 2).

Future studies, like GWAMAs using 1000 Genomes imputed data, will exhibit an increased number of variants and will include additional genetic structures such as indels or SVs. The *EasyQC* software is specifically designed to handle large datasets and can thus be used for larger SNP panels. With regard to memory requirement, *EasyQC* requires a minimum of 30 GB random access memory (RAM) for 1000 Genomes imputed data (~40M SNPs) for the file-level QC, which is the protocol part requiring the largest memory. Alternatively, the file-level QC steps can be parallelized by splitting the data into smaller parts, e.g. by chromosome or into overlapping segments of 5Mb, as recommended for 1000 Genomes imputation. To handle indels and SVs, adjustments to the scripts, like allowing for "I" (insertion) and "D" (deletion) alleles, are needed and can be made directly to the provided *EasyQC* scripts. To this end, the EasyQC package is under active development and future updates will include scripts tailored to 1000 Genomes data.

MATERIALS

EQUIPMENT

Data—

- Allele frequency reference panels: For HapMap imputed GWAs data: HapMap CEU frequencies as given in '*AlleleFreq_HapMap_CEU.v2.txt.gz*'. For typed Metabochip data: 1000 Genomes EUR frequencies as given in '*AlleleFreq_1000G_EUR_Metabochip.v1.txt.gz*'. Both files are available from the relevant website of the Department of Genetic Epidemiology, University of Regensburg http://www.genepi-regensburg.de/easyqc/.

- SNP identifier reference panel for marker harmonization: The file '*SNPID_to_ChrPosID.b36_v2.txt.gz*' (available from the website http://www.genepi-regensburg.de/easyqc/) maps ~9.1 million known different SNP-IDs (column "SNPID", which contains different versions of rs-IDs from b35, b36 or b37, as well as array-specific marker names like "SNP_1_12345") to ~4.8 million unique ChrPosIDs (column "ChrPosID"). It can be used to harmonize SNP identifier names between HapMap imputed or Metabochip data (see Step 15). It does not include sex-chromosomal SNPs. Please see the Supplementary Methods for a description of the file creation.

- QT interval SNPs for GC correction of typed Metabochip data and only for traits that are not correlated with the QT interval: '*QTSNPs_AEL_TW.txt*' (available from the website http://www.genepi-regensburg.de/easyqc/).

- Multiple summary-level association result files.

Software—

- Statistical software R (http://cran.r-project.org/)
- R Package *EasyQC* (http://www.genepi-regensburg.de/easyqc/);
- Meta-analysis software METAL (http://www.sph.umich.edu/csg/abecasis/metal/);
- Template R-, *EasyQC*- and METAL-scripts that can be used to conduct multiple procedure steps are available from http://www.genepi-regensburg.de/easyqc/

Hardware—

- Computer workstation or server with Unix or Linux operating system
- Minimum memory requirements: For performing the file-level QC (which is the most memory-intensive step due to evaluating unfiltered data) with HapMap imputed data (~2.8M SNPs) at least 4GB of random access memory (RAM) should be available

PROCEDURE

Setting up logistics of meta-analysis (Timing ~2 months)

1. Identify GWAS partners and lay out rules of cooperation (“Memorandum of understanding”, MOU). Form task groups and set up phone meetings.
2. Develop a GWAS analysis plan (Supplementary Manual), including instructions on phenotype transformation, analysis models, covariate adjustment, stratification, use of reference panels for imputation, and formatting of data submissions.
3. Set up an sftp site that will be used to collect and securely store the data and organize and label directories and sub-directories in a logical self-explanatory manner (Supplementary Figure 1).

Collecting aggregated statistics per study (Timing ~2 months)

4. Send out the analysis plan and allow for 2 months for the collaborators to provide the data.
5. In the meantime, prepare file cleaning instructions and a meta-analysis plan.
6. When all files are available (or at least files from >80% of studies), freeze the data, i.e. protect the data from further changes (Supplementary Figure 1),and start conducting the file-level QC.

File-level QC (Timing ~2 months)

Critical—The following file-level QC tasks (Steps 7–18) can be grouped by study and be assigned to a set of analysts. Check whether format and variable names included in the study file match the requested format and columns. The following example uses the terminology of the GIANT format described in Box 1 and assumes that data is provided by study collaborators in tabular (TAB) delimited text files with missing values being indicated by '.'.

7. To check variable names and format in *EasyQC*, define the requested columns and format using the DEFINE and the EASYIN functions in the ecf-header (for more information on how to use *EasyQC*, see Box 2) using option A for imputed data or option B for genotyped Metabochip data.

A. Defining columns and format for imputed data

i. Type the following commands:

```
DEFINE –-acolIn
MarkerName;Strand;Chr;Pos;N;Effect_allele;Other_
allele;EAF;Information_type;Information;BETA;
SE;P
--acolInClasses
character;character;character;integer;integer;ch
aracter;character;numeric;numeric;numeric;num
eric;numeric;numeric
--strMissing .
--strSeparator TAB
EASYIN --fileIn /path2input/study.gwa.file1.txt
EASYIN --fileIn /path2input/study.gwa.file2.txt
…
```

B. Defining columns and format for genotyped Metabochip data

i. Type the following commands:

```
DEFINE -–acolIn
MarkerName;Strand;Chr;Pos;N;Effect_allele;Other_
allele;EAF;P_HWE;Callrate;BETA;SE;P
--acolInClasses
character;character;character;integer;integer;ch
aracter;character;numeric;numeric;numeric;num
eric;numeric;numeric
--strMissing .
--strSeparator TAB
EASYIN --fileIn /path2input/
study.metabochip.file1.txt
EASYIN --fileIn /path2input/
study.metabochip.file2.txt
…
```

8. (OPTIONAL) If column names were labeled wrongly, e.g. the analyst used ‘Pvalue’ instead of ‘P’, change the column names centrally, as this is more time-efficient. If any of the requested columns cannot be clearly allocated or are even missing, consult the study analyst for clarification or — if needed — ask for re-upload. *EasyQC* will only start to iterate over the defined input files if their headings and format match the requested columns and the requested format. Minor changes to the requested format, e.g. renaming column names or using a different delimiter, can be handled by *EasyQC* directly through small adjustments in the ecf-header.

?TROUBLESHOOTING

9. Filter monomorphic SNPs. Exclude and count SNPs with allele frequency = 0 or =1. In *EasyQC*, this can be done using the ‘CLEAN’ function:

```
CLEAN --rcdClean (EAF==0)|(EAF==1) --strCleanName numDrop_Monomorph
```

10. Filter SNPs with missing values. Exclude and count all SNPs with missing alleles, P-value, beta estimate, standard error, allele frequency or sample size. In *EasyQC*, this can be done using the ‘CLEAN’ function:

```
CLEAN --rcdClean is.na(Effect_allele) --strCleanName
numDrop_Missing_EA
CLEAN --rcdClean is.na(Other_allele) --strCleanName
numDrop_Missing_OA
CLEAN --rcdClean is.na(P) --strCleanName numDrop_Missing_P
CLEAN --rcdClean is.na(BETA) --strCleanName numDrop_Missing_BETA
CLEAN --rcdClean is.na(SE) --strCleanName numDrop_Missing_SE
CLEAN --rcdClean is.na(EAF) --strCleanName numDrop_Missing_EAF
CLEAN --rcdClean is.na(N) --strCleanName numDrop_Missing_N
```

11. Filter SNPs with non-sense values. Exclude and count all SNPs with alleles other than ‘A’,’C’,’G’ or ‘T’; P-values <0 or >1; negative or infinite standard errors (<=0 or =Infinity); infinite beta estimates or allele frequencies <0 or >1. In *EasyQC*, this can be done using the ‘CLEAN’ function:

```
CLEAN --rcdClean !(Effect_allele%in%c('A','C','G','T')) --
strCleanName numDrop_invalid_EA
CLEAN --rcdClean !(Other_allele%in%c('A','C','G','T')) --
strCleanName numDrop_invalid_OA
CLEAN --rcdClean P<0|P>1 --strCleanName numDrop_invalid_P
CLEAN --rcdClean SE<=0|SE==Inf --strCleanName numDrop_invalid_SE
CLEAN --rcdClean abs(BETA)==Inf --strCleanName numDrop_invalid_BETA
CLEAN --rcdClean (EAF<0)|(EAF>1) --strCleanName numDrop_invalid_EAF
```

12. Filter SNPs on allele frequency and sample size. Exclude and count SNPs with a sample size <30. Add a column called *MAC* defined as 2 times sample size times minor allele frequency and exclude and count all SNPs with *MAC*<=6. In *EasyQC*, these steps can be performed using the following *EasyQC* code:

```
CLEAN --rcdClean N<30 --strCleanName numDrop_Nlt30
ADDCOL --rcdAddCol 2*pmin(EAF,1-EAF)*N --colOut MAC
CLEAN --rcdClean MAC<=6 --strCleanName numDrop_MAClet6
```

13. Filter SNPs on genotype quality. Use option A for imputed data or option B for genotyped Metabochip data:

A. Filtering SNPs in imputed data

i. Filter SNPs due to non-sense or missingness: Exclude and count SNPs with missing *Information_type*, genotyped SNPs (indicated by *Information_type* =0) with an imputation quality less than 1 (*Information* <1), imputed SNPs (*Information_type* !=0) with missing imputation quality. In *EasyQC*, this can be done using the ‘CLEAN’ function:

```
CLEAN --rcdClean is.na(Information_type)
--strCleanName numDrop_MissingInformationType
CLEAN --rcdClean
Information_type==0&Information<1
--strCleanName numDrop_Genotyped_LowInformation
CLEAN --rcdClean (Information_type!=
0)&(is.na(Information))
--strCleanName
numDrop_Imputed_MissingInformation
```

ii. Filter SNPs on imputation quality: Exclude and count SNPs with low imputation quality using a threshold that depends on the imputation and association software used (Table 2). In *EasyQC*, this can be done using the ‘CLEAN’ function:

```
CLEAN
--rcdClean
(Information_type!=0&Information<0.3)|
(Information_type==2&Information<0.4)|
(Informati on_type==3&Information<0.8)
--strCleanName numDrop_LowInformation
```

B. Filtering SNPs in genotyped Metabochip data

i. Filter SNPs due to non-sense or missingness: Exclude and count SNPs with: (1) missing per-SNP callrate; (2) missing HWE P-values; (3) Callrate or Phwe <0 or >1. In *EasyQC*, this can be done using the 'CLEAN' function:

```
CLEAN --rcdClean is.na(Callrate) --strCleanName
numDrop_MissingCallrate
CLEAN --rcdClean is.na(P_HWE) --strCleanName
numDrop_MissingPhwe
CLEAN --rcdClean Callrate<0|Callrate>1 --
strCleanName numDrop_InvalidCallrate
CLEAN --rcdClean P_HWE<0|P_HWE>1 --strCleanName
numDrop_InvalidPhwe
```

ii. Filter SNPs on low call rate and SNPs violating the HWE: Exclude and count SNPs with Callrate<0.95 and SNPs with P_HWE<10^{-6}. In *EasyQC*, this can be done using the 'CLEAN' function:

```
CLEAN --rcdClean Callrate<0.95 --strCleanName
numDrop_LowCallrate
CLEAN --rcdClean P_HWE <1e-6 --strCleanName
numDrop_LowHwe
```

14. Filter and count SNPs on sex chromosomes. Keep the sex-chromosomal SNPs in a separate file for optional subsequent analyses. In *EasyQC*, this can be done using the 'CLEAN' function:

```
CLEAN --rcdClean !Chr%in%c(1:22,NA) --strCleanName
numDropSNP_ChrXY --blnWriteCleaned 1
```

If the chromosomal information is missing in the input file, all SNPs on the sex chromosomes will be excluded by the next step.

15. Harmonize SNP identifiers. To maximize the overlap in the number of SNPs between the study files and to ensure a proper meta-analysis, create a unique SNP-ID called *ChrPosID*, which uses the unique format "chr<chr>:<position>" (e.g. 'chr10:104207431', which only uses genetic positions on build 36). We propose two alternative approaches for this SNP-ID harmonization. Use option A, for studies that lack information on genetic positions (columns *Chr* and *Pos*). Option A was implemented in GIANT meta-analyses as the genetic positions were not available in many of the studies (in particular in those that contributed to earlier rounds of analyses). For future studies, we recommend using option B for which *Chr* and *Pos* is requested from each collaborator to allow compiling the *ChrPosID* from the provided information. Option B is the preferable, more generic approach that easily handles novel genotyping arrays (e.g. Exomechip), imputation reference panels (e.g. 1000 Genomes) or genome builds that are not depicted by the provided reference panel '*SNPID_to_ChrPosID.b36_v2.txt.gz*' (Supplementary Methods).

A. Creating ChrPosID if genetic positions are not available in the study file

i. Create a SNP identifier reference panel. Create a reference file that can be used to remap different versions of SNP names to unique *ChrPosIDs* (see Supplementary Methods for detailed descriptions on how-to create such a reference file). In case of analyzing HapMap imputed or Metabochip data on genome build 36, use the provided SNP identifier reference panel '*SNPID_to_ChrPosID.b36_v2.txt.gz*' (Supplementary Methods).

ii. Add the unique *ChrPosID* to the study file by merging the study file column *MarkerName* with the reference file column *SNPID*. In *EasyQC*, this can be done using the 'RENAMEMARKER' function:

```
RENAMEMARKER --colInMarker MarkerName
--fileRename /path2reffiles/
SNPID_to_ChrPosID.b36_v2.txt.gz
--colRenameOldMarker SNPID
--colRenameNewMarker ChrPosID
```

iii. Check the format of existing ChrPosIDs. To avoid formatting errors with existing ChrPosIDs in study files, remove all spaces from the SNP names (i.e. transform 'chr10 : 104207431' to 'chr10:104207431') and add the character string 'chr' at the beginning of the SNP name in case it was forgotten (i.e. transform '10:104207431' to 'chr10:104207431'). In *EasyQC* correcting the format of mislabeled ChrPosID SNPs can be performed using the following commands:

```
EDITCOL --rcdEditCol gsub(" ","", ChrPosID) --
colEdit ChrPosID
EDITCOL --rcdEditCol
ifelse(regexpr(":",ChrPosID)==2 |
regexpr(":",ChrPosID)==3,
paste("chr", ChrPosID,sep=""), ChrPosID)
--colEdit ChrPosID
```

B. Creating ChrPosID if genetic positions are available in the study file

i. Generate *ChrPosID* directly from the provided *Chr* and *Pos* columns by horizontally concatenating the string “chr”, column *Chr*, character “:” and column *Pos*. This approach requires genetic positions to be given in the study file. In *EasyQC*, this can be done using the ‘ADDCOL’ function:

```
ADDCOL --rcdAddCol
paste(“chr”,Chr,”:”,Pos,sep=””) --colOut
ChrPosID
```

16. Filter duplicate SNPs. To use the best candidate, exclude the duplicate with the smaller sample size. In *EasyQC*, this can be done using the ‘CLEANDUPLICATES’ function:

```
CLEANDUPLICATES --colInMarker ChrPosID --strMode samplesize --colN
N
```

17. Save cleaned files: Add the prefix “CLEANED.” to the filename, save the cleaned file and use ‘.’ as missing character. In *EasyQC*, this can be done using the ‘WRITE’ function:

```
WRITE --strPrefix CLEANED. --strMissing . -–strMode gz
```

18. To perform a file-level QC check, prepare a summary for each study file: Count and check the number of SNPs in the cleaned file and the number of exclusions for each procedure step. An example list of report variables is given in Supplementary Table 1. The number of SNPs in the cleaned file should be >2.2 million for GWAS data (if imputed to a HapMap II reference panel) and >100,000 for Metabochip data. Major departures from these expected values, generally large numbers of exclusions or any exclusions due to missing or nonsense values (Steps 10, 11, 13Ai, 13Bi) may indicate systematic issues with the file; consult the study analyst to clarify. When using *EasyQC*, open the generated summary report in Excel. The report is automatically written to the output path and carries the file extension ‘.rep’. It contains one row per input file and the QC variables - to be checked - in columns (Supplementary Table 1).

Meta-level QC (Timing ~2 months)

19. **Identify analytical issues by the SE-N plot**. To check for issues with trait-transformation, the coded sample-size or file-naming, calculate the median standard error and maximum sample size of every input and produce a plot of *c/median(SE)* versus *Sqrt(max(N))* (one point for each file, Figure 2). The proportionality constant *c* depends on the genotyping platform or the imputation reference panel (Table 3). Find values for *c* for standard platforms and panels in Table 3, i.e. use 1.93 for typed Metabochip data or 1.75 for HapMap II imputed GWAS data. For platforms or panels other than those given in Table 3, the value of *c* needs to be computed *de novo* by Equation (2) for one study with the respective platform or for the imputation reference panel; this *c* can then be applied to the other studies. In *EasyQC*, calculate the statistics and create the plot using the 'CALCULATE' and 'RPLOT' functions:

```
CALCULATE --rcdCalc max(N,na.rm=T) --strCalcName Nmax
CALCULATE --rcdCalc median(SE,na.rm=T) --strCalcName SEmedian
RPLOT --rcdRPlotX sqrt(Nmax)
--rcdRPlotY [c]/SEmedian
--rcdRPlotY [c]/SEmedian
--arcdAdd2Plot abline(a=0,b=1,col="orange")
--strAxes zeroequal
--strPlotName SEN-PLOT
# Please replace [c] at --rcdRPlotY with the respective value from
Table 3.
```

20. Check whether the points follow the identity line. In case any points clearly deviate from the diagonal, consult study analyst to clarify trait transformation, sample-size coding and file-naming (Figure 2, Anticipated Results). Studies with unaccounted relatives show deviation from the identity line as the effective sample size is different from the actual sample size, but whether unaccounted relatedness is the reason for an observed deviation should be confirmed after consultation with the analyst.

21. **Identify analytical issues by the P-Z scatter plot**. To check for problems with beta estimates, standard errors and P-values, create plots comparing P-values (on the $-\log_{10}$ scale) calculated from a Z statistic ($Z=\beta/SE(\beta)$) with the P-values directly provided by study partners. (Figure 3, Supplementary Figure 3). In *EasyQC*, this can be done using the 'PZPLOT' function:

```
PZPLOT --colBeta BETA --colSe SE --colPval P
```

22. Check whether the points follow the identity line. In case any points clearly deviate from the diagonal, consult study analyst (Figure 3, Anticipated Results).

23. **Identify problems with allele frequencies or strand**. To check for strand and allele frequency issues, plot the allele frequency of each SNP and for each file against a reference allele frequency (one plot for each file) (Figure 4, Supplementary Figure 4). For HapMap imputed GWAS data, plot allele frequencies against publically available HapMap allele frequencies, which are reported in the reference file *'AlleleFreq_HapMap_CEU.v2.txt.gz'*. For genotyped Metabochip data, plot allele frequencies against publically available 1000 Genomes allele frequencies, which are reported in the reference file *'AlleleFreq_1000G_EUR_Metabochip.v1.txt.gz'*. In *EasyQC*, the AFCHECK function can be used to create these plots (please replace [reffile] in the following code with the respective reference file name):

```
AFCHECK --colInMarker ChrPosID
--colInStrand Strand
--colInA1 Effect_allele
--colInA2 Other_allele
--colInFreq EAF
--fileRef /path2reffiles/[reffile]
--colRefMarker ChrPosID
--colRefA1 A1
--colRefA2 A2
--colRefFreq Freq1
--blnMetalUseStrand 1
# Replace the path to the reference and the reference-file name at
--fileRef
```

24. The frequencies should be distributed along the identity line. Check whether there are patterns (see Figure 4, Anticipated Results) that indicate problems with strand or allele frequencies. In case you observe such patterns, contact study analyst to clarify the issue. To define the problem more precisely, it can be helpful to provide the collaborator with a list of (i) outlying SNPs, i.e. SNPs with allele frequencies that deviate >20% from the reference population, and (ii) mismatching SNPs, i.e. SNPs with alleles that do not match the reference, e.g. AC in study versus AT in reference population. The AFCHECK function automatically saves the lists of outlying or mismatching SNPs to the output path (files indicated by suffix 'AFCHECK.outlier.txt' and 'AFCHECK.mismatch.txt'). In case of problems, it can also be helpful to check the summary report variables indicated by 'AFCHECK.[variablename]' (Supplementary Table 2).

25. **Identify population stratification**. Calculate λ_{GC} for each study file – without applying the GC correction at this stage – using all SNPs for imputed GWAS data, and, for custom chip data, only a subset of SNPs that are not associated with the outcome of interest. In GIANT, 4,425 QT-interval SNPs (defined in 'QTSNPS_AEL_TW.txt') were used to derive the λ_{GC} for typed Metabochip

data. To get an overview of the λ_{GC} values across all studies and to identify studies with high λ_{GC}, produce a plot of λ_{GC} values versus the maximum sample sizes (Figure 5). In *EasyQC*, the calculation of the λ_{GC} and the plotting can be done using the 'GC' and the 'RPLOT' function:

```
GC --colPval P
--blnSuppressCorrection 1
# --fileGcSnps /path2reffiles/QTSNPs_AEL_TW.txt
# --colInMarker ChrPosID
# --colGcSnpsMarker ChrPosID
# Uncomment last three parameters for Metabochip data
RPLOT --rcdRPlotX Nmax
--rcdRPlotY Lambda.P.GC
--arcdAdd2Plot abline(h=1,col='orange');abline(h=1.1,col='red')
--strAxes lim(0,NULL,0,NULL)
--strPlotName GC-PLOT
```

26. Examine the plot and check whether λ_{GC} is above 1.1 in any of the individual studies. If this is the case, go back to the relevant study analyst to clarify potential issues with population stratification, unaccounted relatedness or duplicated samples included in the analyses (Figure 5, Anticipated Results). The summary report table created by *EasyQC* might be helpful to identify studies that exhibit high λ_{GC} (variable *GC.P.Lambda*, Supplementary Table 2).

Meta-analysis (Timing ~0.5 months)

27. Prepare scripts for an inverse variance-weighted meta-analysis using a fixed effects model with *METAL*, as follows: For quality control, we recommend that two analysts perform the meta-analysis independently. The two analysts should ensure that the order in which the studies are read into *METAL* is the same, because the first study defines the allele coding directions and the following studies are compared with this study. We advise to run *METAL* using the following column definitions and options in the *METAL* script:

```
# Input columns: MARKER ChrPosID
ALLELE Effect_allele Other_allele
EFFECT BETA
STDERRLABEL SE
FREQLABEL EAF
PVALUE P
STRAND Strand
CUSTOMVARIABLE N
LABEL N AS N
# Metal Options:
```

```
SCHEME STDERR
WEIGHT N
USESTRAND ON
AVERAGEFREQ ON
MINMAXFREQ ON
VERBOSE OFF
GENOMICCONTROL ON
# GENOMICCONTROL LIST /path2reffiles/QTSNPs_AEL_TW.txt
# Use the latter for metabochip data!
PROCESS /path2cleanedfiles/CLEANED.study1.file1.txt.gz
PROCESS /path2cleanedfiles/CLEANED.study1.file2.txt.gz
# …
PROCESS /path2cleanedfiles/CLEANED.study1.fileM.txt.gz
PROCESS /path2cleanedfiles/CLEANED.study2.file1.txt.gz
# …
PROCESS /path2cleanedfiles/CLEANED.studyN.fileM.txt.gz
OUTFILE metalout .TBL
ANALYZE HETEROGENEITY
```

To correct for file-specific population stratification, ‘GENOMICCONTROL’ should be set to ‘ON’, as this will apply GC correction to each study file. For Metabochip studies, the ‘GENOMICCONTROL LIST’ parameter can be used to limit the calculation of the λ_{GC} to the subset of QT-interval SNPs. An alternative to using *METAL* for the GC correction by study-file during the meta-analysis is provided by the *EasyQC* function ‘GC’ (see the *EasyQC* manual provided on the *EasyQC* website for further details). Implementation of this function can be added to the file-level QC to correct study-specific standard errors and P-Values in the same way METAL does. To add metrics that measure between-study heterogeneity use the command ‘ANALYZE HETEROGENEITY’ at the end of the METAL script file. We provide template METAL scripts, which include the described options and commands (‘3_metaanalysis.metal’).

28. Perform the inverse variance-weighted meta-analysis and create a METAL log-file by using the following command from the command line:

```
metal 3_metaanalysis.metal > metalout _log.txt
```

Meta-analysis QC (Timing ~ 1.5 months)

29. **Compare results from two meta-analysts**. For each of the two meta-analysis results, calculate descriptive statistics of P-values and sample sizes (length, number of missing values, minimum, maximum, median, mean and standard deviation) and the meta-level λ_{GC} (again, restrict calculation of the λ_{GC} to QT-interval SNPs for Metabochip results) and check the values for discrepancies. To compare the meta-analyzed P-values directly, merge the two data sets, create a scatter-plot of P-values (on the -$\log_{10}$ scale) and calculate their Spearman correlation coefficient. In *EasyQC*, the calculation of the statistics as well as the merging of the data sets and the creation of the plot, can be done using the following ‘EasyQC’ code:

```
DEFINE --acolIn MarkerName;P.value;N
--acolInClasses character;numeric;numeric
EASYIN --fileIn /path2metalresults/metalout.analyst1.TBL --
fileInTag A1
EASYIN --fileIn /path2metalresults/metalout.analyst1.TBL --
fileInTag A2
START EASYQC
EVALSTAT --colStat P.value
EVALSTAT --colStat N
GC --colPval P.value
--blnSuppressCorrection 1
#--fileGcSnps /path2reffiles/QTSNPs_AEL_TW.txt
#--colInMarker MarkerName
#--colGcSnpsMarker ChrPosID
# Uncomment last three parameters for metabochip data
MERGEEASYIN --colInMarker MarkerName
CALCULATE --rcdCalc
cor(P.value.A1,P.value.A2,method="spearman",use="pairwise.complete.
obs")
--strCalcName corr_Pvals
SPLOT --rcdSPlotX -log10(P.value.A1)
--rcdSPlotY -log10(P.value.A2)
--arcdAdd2Plot abline(a=0,b=1,col='orange')
STOP EASYQC
```

The summary report table created by *EasyQC* contains the descriptive values, the λ_{GC} as well as the correlation coefficient (Supplementary Table 3).

30. Examine the calculated values and the scatter plot to check for discrepancies between the two meta-analysis results. All summary statistics should be identical and the P-values should lie on the identity line. Most discrepancies observed between meta-analysts are usually explained by different file inclusions in the meta-analysis. To get a quick overview on the files included in the meta-analysis of each analyst, run the R-script ‘*4_metaanalysis_qc.compare_logfiles.r*’. This action takes the two meta-analysis log-files as inputs and creates a table that can be used to compare file inclusions.

31. (OPTIONAL) **Identify analytical issues by calculating the study-level**λ_{GC}. if the verified and agreed-on meta-analysis result displays a large meta-level λ_{GC}

(>1.1, check the λ_{GC} calculated by step 29), conduct one meta-analysis for each study (e.g. pooling strata-specific files per study) and calculate the study-level λ_{GC}. An inflated study-level λ_{GC} might pinpoint unaccounted relatedness or overlap of samples across the strata of the study; it can also pinpoint errors as simple as mis-naming the strata files (e.g. one file is labeled as men, the other as women, but the men-file was uploaded twice). A substantial fraction of the inflated meta-level λ_{GC} might be explained by such study-specific issues. In *EasyQC*, the study-specific meta-analysis as well as the calculation of the study-level λ_{GC} can be performed using the following *EasyQC* code:

```
DEFINE --acolIn ChrPosID;Effect_allele;Other_allele;BETA;SE
--acolInClasses character;character;character;numeric;numeric
EASYIN --fileIn /path2cleanedfiles/CLEANED.study1.file1.txt --
fileInTag 1
EASYIN --fileIn /path2cleanedfiles/CLEANED.study1.file2.txt --
fileInTag 2
START EASYQC
MERGEEASYIN --colInMarker ChrPosID
METAANALYSIS --acolBETAs BETA.1;BETA.2
--acolSEs SE.1;SE.2
--acolA1s Effect_allele.1;Effect_allele.2
--acolA2s Other_allele.1;Other_allele.2
--colOutBeta betaPooled
--colOutSe sePooled
--colOutP pPooled
GC --colPval pPooled
--blnSuppressCorrection 1
#--fileGcSnps /path2reffiles/QTSNPs_AEL_TW.txt
#--colInMarker ChrPosID
#--colGcSnpsMarker ChrPosID
# Uncomment last three parameters for metabochip data
STOP EASYQC
```

The summary report table created by *EasyQC* contains the study-level λ_{GC}.

32. Check the study-level λ_{GC} and consult the relevant study analyst in case of a study-level λ_{GC} >1.1. If the study analyst then flags analytical errors, re-analysis of the study data is needed and steps 7 – 32 have to be repeated for the affected files.

33. **Finalize the meta-analysis**. After passing all meta-analysis quality checks, upload the final meta-analysis results file to the ftp site and freeze the upload directory (Supplementary Figure 1). Use the agreed-on result files to extract significant SNPs, to create plots, e.g. Manhattan- or QQ-plots, and for further evaluation. If a replication of the findings using independent follow-up data is planned, all steps of the Procedure can be repeated for the follow-up meta-analysis.

TROUBLESHOOTING

Step 8: It is likely that data from some studies may have been uploaded in a format that differs from the requested one. If the format of an input file does not match the requested format, *EasyQC* stops with an error message before it starts to iterate over all input files. Issues such as completely missing columns may require contacting the study analyst. Some obvious problems, such as different column names (e.g. ‘Pvalue’ instead of ‘P’), different column separators (e.g. ‘,’ instead of TAB) or missing characters (e.g. ‘NaN’ instead of ‘.’) can instead be fixed by *EasyQC* directly (by overwriting the DEFINE parameters at the respective EASYIN statement):

```
EASYIN --fileIn /home/fileWithDifferentFormat.txt
–-acolIn
MarkerName;Strand;Chr;Pos;N;Effect_allele;Other_allele;EAF;Information_type;I
nformation;BE
TA;SE;Pvalue
--acolInClasses
character;character;character;integer;integer;character;character;numeric;num
eric;numeric; numeric;numeric;numeric
–-acolNewName
MarkerName;Strand;Chr;Pos;N;Effect_allele;Other_allele;EAF;Information_type;I
nformation;BE
TA;SE;P
--strMissing NaN
--strSeparator COMMA
```

TIMING

The timing of the whole QC and GWAMA pipeline depends on the number of studies involved and also on the experience of the analysts. The estimates reported below are based on the assumption that an existing pipeline of QC and meta-analysis is available (as given by this protocol). The original GIANT conduct and QC has taken longer due to the exploratory nature of the effort. The estimates provided are realistic as they are given by experienced meta-analysts. For a consortium of comparable size to GIANT’s, we estimate the timing of each procedure step to be as follows:

Steps 1–3, Setting up logistics of meta-analysis: ~2 months

Steps 4–6, Collecting aggregated statistics per study: ~2 months

Steps 7–18, File-level QC: ~2 months

Steps 19–26, Meta-level QC: ~2 months

Steps 27–28, Meta-analysis: ~0.5 months

Steps 29–32, Meta-analysis QC: ~1.5 months

ANTICIPATED RESULTS

Meta-level QC - Identification of analytical issues by the SE-N plot (Steps 19–20)

In the case of an inverse normal transformed phenotype, forcing the phenotype into the Standard Normal distribution, N(0,1), the data points on the SE-N plot should tend to describe a straight line on the diagonal, the identity line. Figure 2a illustrates a major deviation of a cluster of GIANT studies from the identity for HIP_{adjBMI} in the initial round of meta-level QC.

To investigate the reason for this deviation, we surveyed the way each study analyst performed the phenotype transformation. Whether the analyst adjusted the phenotype for age, age^2, study-specific covariates, and BMI by sex according to the analysis plan and then subjected to the inverse Normal transformation, again separately by sex. This survey revealed that the studies in the cluster above the identity line *first* (instead of *last*) applied the inverse Normal transformation and then adjusted the phenotype for the covariates; a few studies had done the adjustment and/or transformation in men and women combined (instead of by sex), and separated the data by sex afterwards.

Subsequent explorations revealed that the SE-N plot identified this problem for phenotypes adjusted for BMI (such as HIP_{adjBMI}), but not the BMI-unadjusted phenotypes, since the adjustment for BMI after the inverse Normal transformation had disrupted the N(0,1) distribution of the phenotype (Supplementary Figure 2a). Further explorations revealed that such type of trait transformation issue would result in a loss of power (QQ-plot, Supplementary Figure 2b) and in estimates biased towards the *null* (Supplementary Figure 2c).

Other transformation errors that we were able to identify using the SE-N plot (not shown) include (i) lack of inverse Normal transformation, (ii) the stratification by sex conducted after the adjustment and inverse Normal transform, (iii) miscoded sample size (e.g. stating full sample size rather than the sample size used for the analysis).

Meta-level QC - Identification of analytical issues by the P-Z scatter plot (Steps 21–22)

Occasionally, for a large proportion of SNPs, we observed a discrepancy between the P-value reported by an analysis software and the P-value calculated manually from the Z statistic based on the reported beta estimates and standard errors ($Z=\beta/SE(\beta)$). In the GIANT Consortium, we observed such discrepancies caused by the "--score" option in the SNPtest software. The P-Z plots can detect such issues (Figure 3a) and asking study analyst to re-analyze the data, using the requested and (in our case) correct "--expected" option, resolved these issues (Figure 3b). Panels of such plots for each file in the meta-analysis can provide a quick overview across files and studies (Supplementary Figure 3).

Meta-level QC - Identification of problems with allele-frequencies or strand (Steps 23– 24)

Heterogeneity in allelic patterns may be observed when study allele frequencies are plotted against a reference set, whether derived from HapMap, 1000 Genomes, or the meta-analysis mean allele frequency. Figure 4 shows patterns observed in data submitted to the GIANT Consortium. Deviations from the reference frequencies are expected for studies of different ancestry and for studies that have incorrectly coded effect alleles, allele frequencies or strand. Creating a panel displaying such plots for each study-file at once provides a quick overview and can identify studies with any of the above issues (Supplementary Figure 4).

Meta-level QC - Identification of population stratification (Steps 25–26)

To detect studies with population substructure, the file-specific λ_{GC} versus the square root of the sample size can be plotted (Figure 5). Study analysts with λ_{GC} >1.1 should be contacted asking them for re-analysis, e.g. including principal components in their analysis model. The *EasyQC* report may be of help to identify which studies exhibit a high λ_{GC} (Supplementary Table 2).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the BMBF (01ER1206 for IMH), the Leenaards Foundation and the Swiss National Science Foundation (31003A-143914 for ZK), the National Institutes of Health (DK078150, T32 HL007427 for DCCC; R01DK075787 for TE), the UK Medical Research Council (MRC; U106179471, U106179472 for F.R.D.), the European Research Council (SZ-245 50371-GLUCOSEGENES-FP7-IDEAS-ERC for A.R.W.), the Targeted Financing from the Estonian Ministry of Science and Education (SF0180142s08 for TE), the Development Fund of the University of Tartu (SP1GVARENG for TE), the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN, 3.2.0304.11-0312 for TE) and by FP7 (313010 for TE).

We are also thankful for the GIANT Consortium and the many participating research groups that have allowed us to develop this protocol.

REFERENCES

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:9362–9367. [PubMed: 19474294]
2. McCarthy MI, Hirschhorn JN. Genome-wide association studies: past, present and future. Human molecular genetics. 2008; 17:R100–R101. [PubMed: 18852196]
3. Hirschhorn JN, Gajdos ZK. Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annual review of medicine. 2011; 62:11–24.
4. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. American journal of human genetics. 2012; 90:7–24. [PubMed: 22243964]
5. Anderson CA, et al. Data quality control in genetic case-control association studies. Nature protocols. 2010; 5:1564–1573.
6. Randall JC, et al. Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits. PLoS genetics. 2013; 9:e1003500. [PubMed: 23754948]
7. Surakka I, et al. A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol. PLoS genetics. 2011; 7:e1002333. [PubMed: 22028671]
8. Manning AK, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nature genetics. 2012; 44:659–669. [PubMed: 22581228]

9. Voight BF, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS genetics. 2012; 8:e1002793. [PubMed: 22876189]
10. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis research & therapy. 2011; 13:101. [PubMed: 21345260]
11. Huyghe JR, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nature genetics. 2013; 45:197–201. [PubMed: 23263489]
12. Teslovich TM, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010; 466:707–713. [PubMed: 20686565]
13. Heid IM, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nature genetics. 2010; 42:949–960. [PubMed: 20935629]
14. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]
15. Speliotes EK, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics. 2010; 42:937–948. [PubMed: 20935630]
16. Scott RA, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nature genetics. 2012; 44:991–1005. [PubMed: 22885924]
17. Schunkert H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics. 2011; 43:333–338. [PubMed: 21378990]
18. Loos RJ, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nature genetics. 2008; 40:768–775. [PubMed: 18454148]
19. Willer CJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature genetics. 2009; 41:25–34. [PubMed: 19079261]
20. Lindgren CM, et al. Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. PLoS genetics. 2009; 5:e1000508. [PubMed: 19557161]
21. Berndt SI, et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nature genetics. 2013; 45:501–512. [PubMed: 23563607]
22. Cochran WG. The Combination of Estimates from Different Experiments. Biometrics. 1954; 10:101–129.
23. Manning AK, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. Genetic epidemiology. 2011; 35:11–18. [PubMed: 21181894]
24. de Bakker PI, et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human molecular genetics. 2008; 17:R122–R128. [PubMed: 18852200]
25. Fuchsberger C, Taliun D, Pramstaller PP, Pattaro C, consortium CK. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. Bioinformatics. 2012; 28:444–445. [PubMed: 22155946]
26. Kottgen A, et al. Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nature genetics. 2013; 45:145–154. [PubMed: 23263486]
27. Kottgen A, et al. New loci associated with kidney function and chronic kidney disease. Nature genetics. 2010; 42:376–384. [PubMed: 20383146]
28. Schizophrenia Psychiatric Genome-Wide Association Study, C. Genome-wide association study identifies five new schizophrenia loci. Nature genetics. 2011; 43:969–976. [PubMed: 21926974]
29. Knoppers BM, Dove ES, Litton JE, Nietfeld JJ. Questioning the limits of genomic privacy. American journal of human genetics. 2012; 91:577–578. author reply 579. [PubMed: 22958905]
30. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013; 339:321–324. [PubMed: 23329047]
31. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS genetics. 2009; 5:e1000628. [PubMed: 19798439]
32. International HapMap C, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]

33. Genomes Project C, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

34. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]

35. Yang J, et al. Genomic inflation factors under polygenic inheritance. European journal of human genetics : EJHG. 2011; 19:807–812. [PubMed: 21407268]

36. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190–2191. [PubMed: 20616382]

37. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003; 327:557–560. [PubMed: 12958120]

38. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986; 7:177–188. [PubMed: 3802833]

39. Whitlock MC. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. Journal of evolutionary biology. 2005; 18:1368–1373. [PubMed: 16135132]

40. R Core Team R: A Language and Environment for Statistical Computing.

BOX 1 – Study-specific GWAS results – Columns as requested by GIANT

Stated are the columns requested by GIANT from the study partners for each GWAS to ensure uniform study-specific files:

“MarkerName” – character string; the SNP identifier of the marker analyzed.

“Strand” – a single character “−“ or “+”; Strand on which the allelles are reported.

“Chr” – character; Chromosome.

“Pos” – integer; Base position of the SNP.

“N” – positive integer; The effective number of subjects analyzed.

“Effect_allele” – a single upper case character “A”, “C”, “G”, or “T”; The allele associated with phenotypic traits (corresponding to change in beta estimates).

“Other_allele” – a single upper-case character "A" "C" "G" or "T"; Indicating the other (non-effect) allele.

“EAF” – numeric; Effect allele frequency (range 0–1).

“BETA” – numeric; Estimate of the effect size.

“SE” – numeric; Estimated standard error on the estimate of the effect size.

“P” – numeric; Significance of the variant association, uncorrected for genomic control.

Only for genotyped data:

“P_HWE” – numeric; Exact HWE P-value for the sample analyzed.

“Callrate” – numeric; Call rate for this SNP across all subjects. Perfectly genotyped (100%) data will have a Callrate = 1.000.

Only for imputed data:

“Information_type” – integer; Code indicating the type of data in the “Information” column (i.e. the type of the imputation and analysis software used):

0 if the SNP was not tested using imputation/genotyping uncertainty, in which case the following column “Information” should be missing (e.g. for directly genotyped SNPs);

1 for a MACH imputed SNP, whereas the following column “Information” either contains “r2_Hat” from MACH2DAT/MACH2QTL OR “INFO” from PLINK (in case you have used PLINK for the association with MACH imputed SNP data)

2 if the following column “Information” contains “proper_info” from SNPTEST;

3 for a PLINK imputed SNP, i.e. the following column “Information” contains “INFO” from PLINK (in case the SNP was imputed using PLINK as well)

4 if the following “Information” column contains “rSqHat” from QUICKTEST.

“Information” – numeric; A value (range 0–1; PLINK values can exceed 1) corresponding to the information content output from the association testing (according to the data type specified in the "InformationType" column above).

BOX 2 - Easy QC programming

Generally *EasyQC* is started by calling the *EasyQC* function at the R-prompt and with an ecf-file as parameter:

```
> library(EasyQC)
> EasyQC(“/path2ecffile/examplescript.ecf”)
```

Every data input/output (I/O) and the conducted pipeline is defined in the ecf-file. *EasyQC*’s ecf-files are modularized and each step can be conducted separately. An ecf-file consists of two parts: (i) a header or config-section at the beginning that defines data I/O using the DEFINE and EASYIN functions; followed by (ii) a scripting interface which defines the QC steps being executed.

Structure of an ecf-file:

```
Header to define I/O:
[DEFINE, EASYIN]
Scripting interface with EasyQC function steps:
[CLEAN,GETNUM,ADDCOL …]
```

Several example scripts and templates that combine multiple steps described in this protocol are available from our website http://www.genepi-regensburg.de/easyqc/.

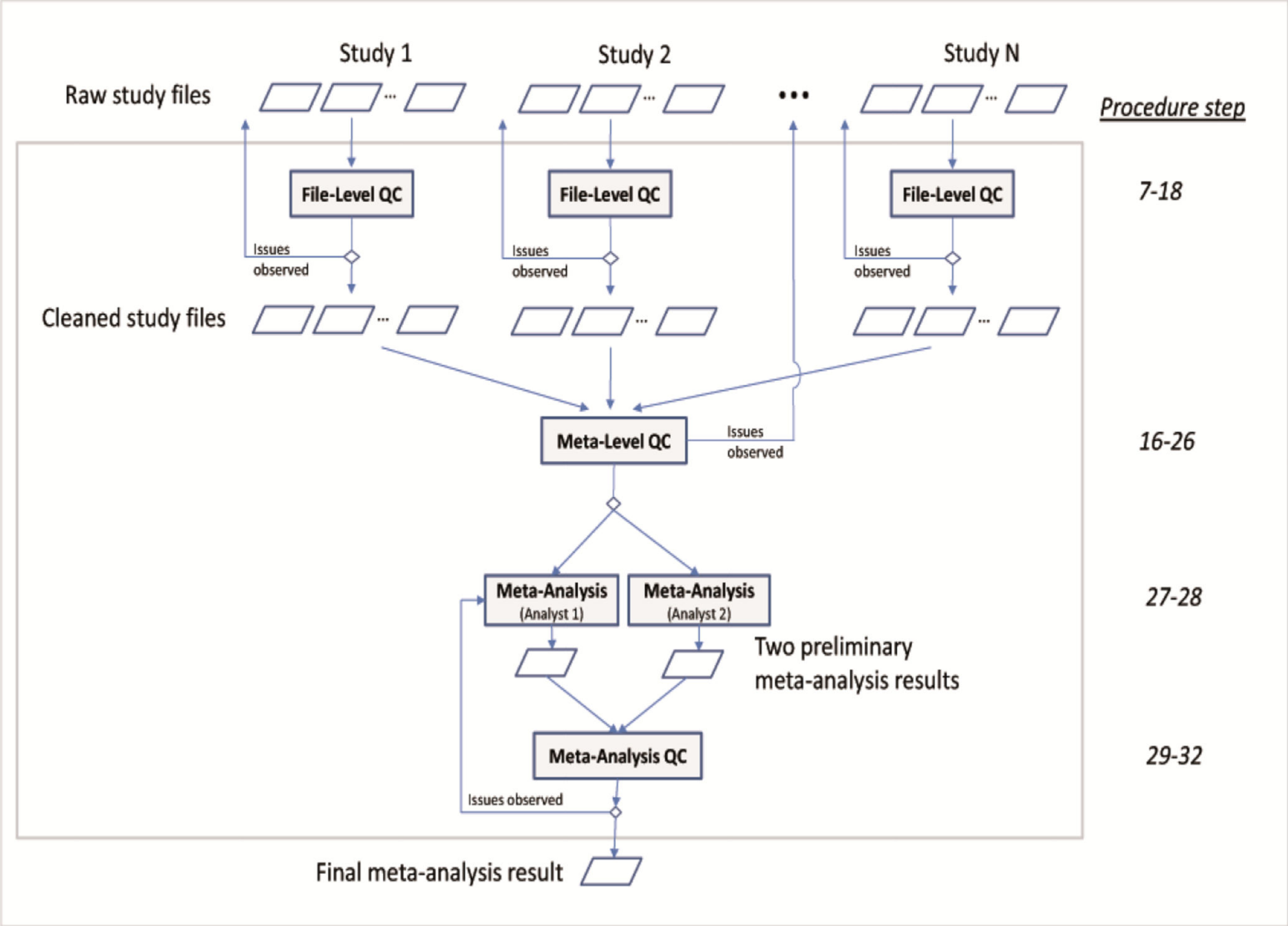

Figure 1. Workflow of the QC and the meta-analysis

A typical GWAMA includes four major stages: (i) The *File-level QC* (Steps 7–18) includes the QC of each study file to ensure validity. This stage involves file cleaning (e.g. adjustments of column headings, file format changes, SNP exclusions based on certain criteria, or adding columns) and file checks (e.g. checking overall characteristics of the file or the number of SNP exclusions), usually in an iterative fashion. Typically this task is divided by study among analysts of the meta-analysis team. Files that pass the file-level QC are labeled as “CLEANED”. Any issues observed with particular files should be clarified with the respective study analyst directly. (ii) The *Meta-level QC* (Steps 19–26) addresses the comparison of file-specific statistics across files in order to depict study-specific issues yet undetected. In case issues of specific studies cannot be resolved centrally, the relevant study analyst should be contacted for clarification.. (iii) Meta-analysis (Steps 27–28) is the stage at which the meta-analysis is actually conducted, a task typically performed by two analysts independently. (iv) Meta-analysis QC (Steps 29–32) involves the checking the meta-analysis results and includes the comparison of the two meta-analyses performed by the different analysts and the quality control of the meta-analysis result.

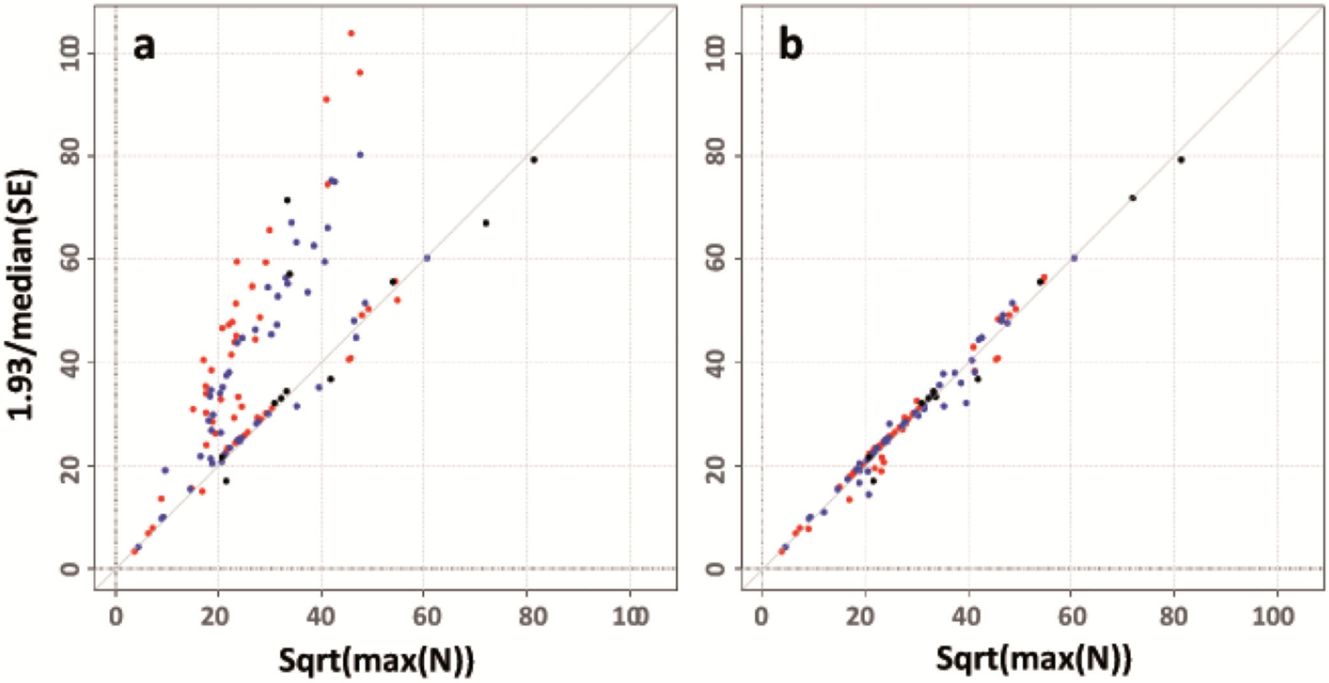

Figure 2. SE-N plots to reveal issues with trait transformations

SE-N plots to detect issues with trait transformations contrasting the study-specific standard errors with sample sizes for GIANT studies typed on Metabochip and tested for association with $\text{HIP}_{\text{adjBMI}}$ (N=81,000): (a) before QC: a number of studies (in fact the majority of studies) revealed errors by clustering above the identity line, and (b) after QC: the same plot after having gone back to the relevant study analysts and having resolved all trait transformation issues. Different colors for the points in the plot indicate men-specific (blue), women-specific (red) or sex-combined (black) association results.

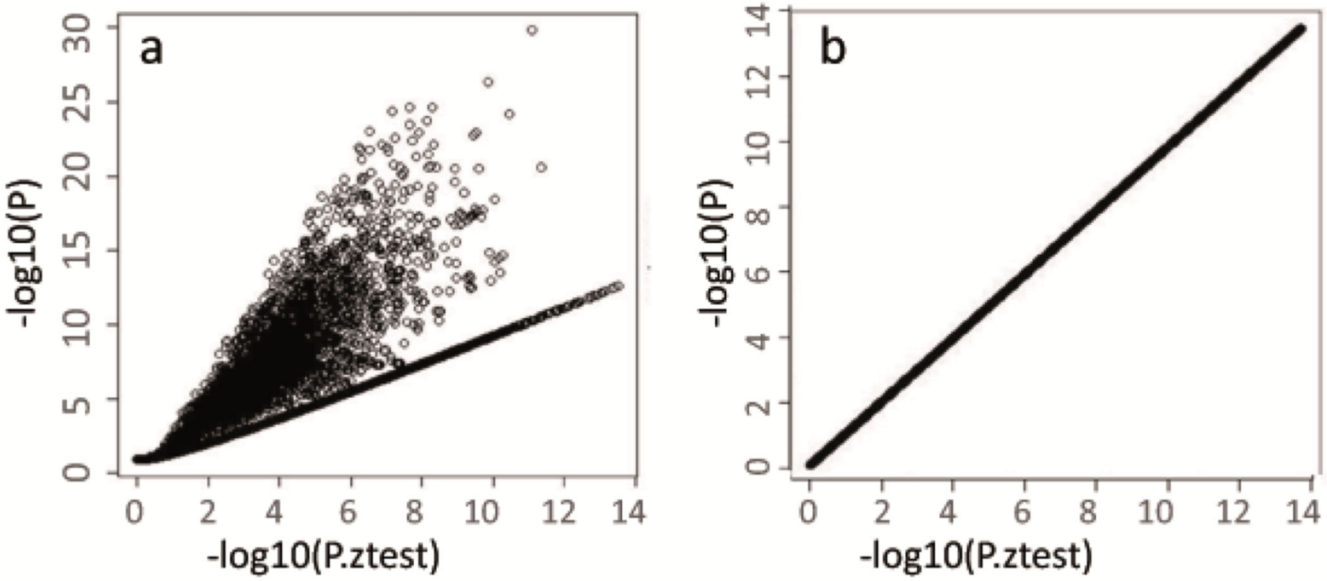

Figure 3. P-Z plot to reveal analytical issues with beta, standard error and P-values

Plots to reveal issues with beta estimates, standard errors and P-values for (a) an uncleaned study file showing severe deviations from the identity line and (b) the cleaned dataset showing perfect concordance. The plots compare P-values reported in the association result file to P-values calculated from Z statistics derived from the reported beta and standard error from an example GIANT file. The uncleaned study file contained a large number of highly significant but erroneous (reported) P values.

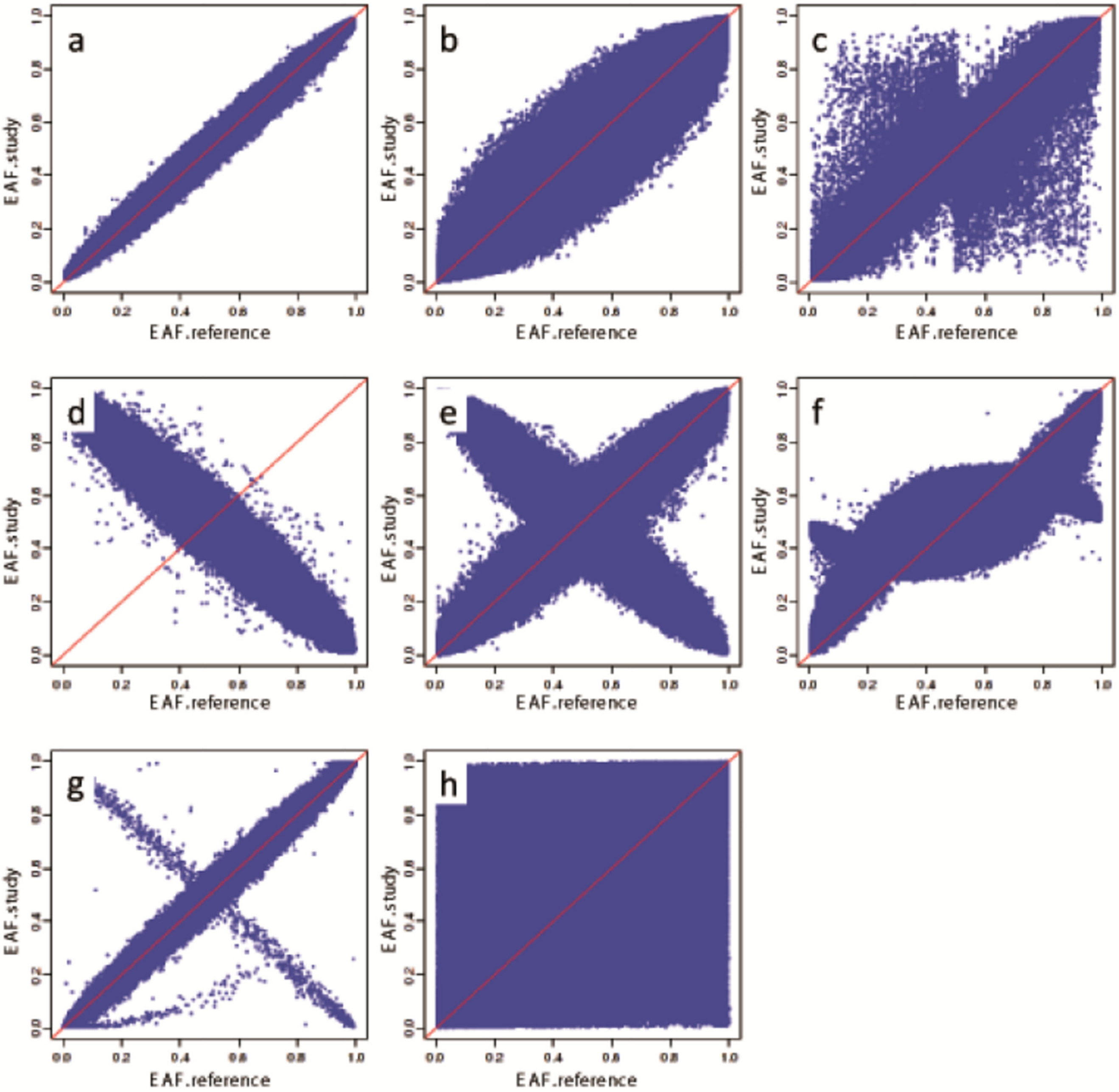

Figure 4. Different patterns of allele frequencies in the EAF plot

These different patterns have been observed during the QC checks performed by the GIANT analysts. In the graphs the observed (study-specific) allele frequencies reported on the y-axis are plotted against the expected (HapMap or 1000 Genomes) allele frequencies, reported on the x-axis. The plots (a)– (c) represent data from studies where allele frequencies and strand annotation are correct but participants exhibit different ancestries compared to the reference, which includes mostly samples of European ancestry: (a) study in which data are relatively consistent with the reference; (b) study in which participants had slightly different ancestry to the reference, resulting in a thicker band across the diagonal; (c) study involving participants of non-European ancestry resulting in substantial deviation from the reference. Plots (d)–(h) pertain to studies with errors in coding the effect allele, the effect allele frequency, and/or strand annotation: (d) a study in which the wrong allele was consistently labeled as effect allele; (e) a study in which a fraction of the effect alleles was mis-specified, e.g. from stating the MAF instead of the effect allele frequency, or from incorrectly assigning strand due to data management or wrong strand reference (sometimes specific to “palindromic” SNPs A/T or C/G); (f)–(h) all represent studies with other data management or analytical errors in calculating the allele frequencies.

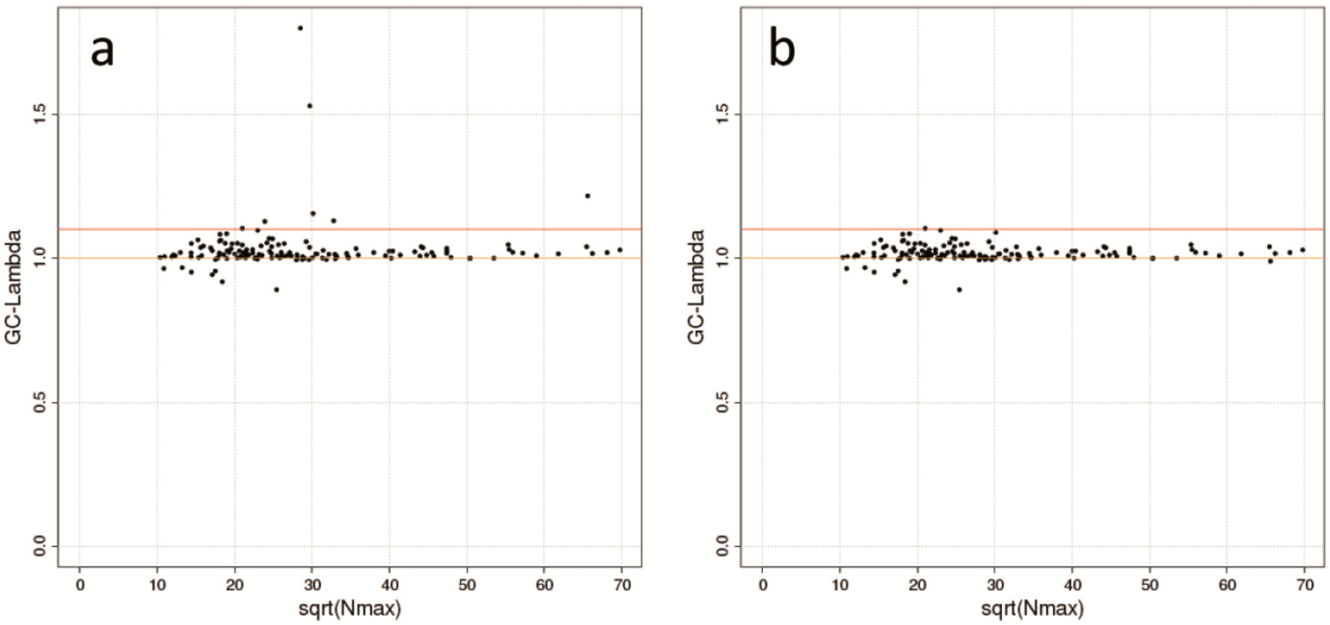

Figure 5. Lambda-N plot to reveal issues with population stratification

Plot to detect issues with population stratification contrasting the study-specific $\lambda_{GC}$with sample sizes for GIANT studies typed on Metabochip and tested for association with HIP_{adjBMI} (N=81,000): (a) before QC: a number of studies displayed high λ_{GC} values, and (b) after QC: the same plot after having gone back to study analyst and having resolved all issues. The orange line indicates the optimal λ_{GC}=1. Dots above the red line, which visualizes the threshold λ_{GC}=1.1, represent problematic studies.

Table 1

Expandability of the protocol to 1000 Genomes imputed data, dichotomous traits, rare variants, SNP x environment (E) Interactions, and x-chromosomal variants.

Procedure step	Step No.(s)	1000 G	Dichotomous trait	Rare Variant Analyses	SNP x E Interaction	Analyses of the sex chromosomes
Setting up logistics of Meta-analysis	1–3	DA	DA	DA	DA	DA
Collecting aggregated statistics per study	4–6	DA	DA	DA	DA	DA
File-Level QC	7–18	AA, allow for indels and SVs; adjust SNP name harmonization	AA, calculate β = In(OR); filter on effective N, adjust MAC cut-off	AA, adjust MAC cutoff	AA, add checks on β_{gxe}	AA, extra checks for pseudo-autosomal variants in men
Identification of analytical issues by the SE-N plot	19,20	DA	DA	AA, Adjust c	AA, add checks on β_{gxe}	AA, separately for men and women, extra considerations of coding errors in male X and Y
Identification of analytical issues by the P-Z plot	21,22	DA	DA	DA	AA, add checks on β_{gxe}	DA
Identification of problems with allele-frequencies or strand	23,24	AA, use 1000 Genomes allele frequencies as reference	AA, limit checks to control group	AA, update allele frequency reference	DA	AA, separately for men and women
Identification of population stratification	25,26	DA	DA	DA	AA, add λ_{GC} from P_{gxe}	AA, use autosomal variant to compute λ_{GC}
Meta-Analysis	27,28	DA	DA	DA for single variant analyses	AA, add meta-analysis of beta GxE	DA
Meta-analysis QC – Compare results from two analysts	29,30	DA	DA	DA	DA	DA
Meta-analysis QC – Identify analytical issues by calculating the study-level λ_{GC}	31,32	DA	DA	DA	AA, add λ_{GC} from P_{gxe}	AA, use autosomal variant to compute λ_{GC}
Finalizing meta-Analysis	33	DA	DA	DA	DA	DA

β= SNP effect on outcome; P = association P-Value; OR = Odds ratio; SE = Standard error; βgxe = SNPxE interaction effect; Pgxe = SNPxE Interaction P-Value; indels, insertions or deletions; SVs, structural variants; DA= directly applicable, AA=applicable with adaptation.

Table 2

Imputation quality metrics for different combinations of imputation and analysis software packages as observed in GIANT.

		Association Software				
		Mach2qtl	**PLINK**	**SNPtest (--expected)**	**QUICKtest**	**Other**
Imputation Software	MACH	0.3 (r2_hat)	0.3 (INFO)	0.4 (proper_info)[a]	0.3 (rSqHat)	0.3 (rSqHat)
	IMPUTE	-	-	0.4 (proper_info)[a]	0.3 (rSqHat)	-
	PLINK	-	0.8 (INFO)	-	-	-

[a] Newer versions of SNPtest output column 'info' instead of 'proper_info'

Table 3

SE-N Plot calibration factors for various genotyping platforms, imputation reference panels, and ethnicities.

Genotyping Platform	Imputation reference panel	Ethnicity	Calibration factor (*c*)
GWAS chip	HapMap	CEU	1.75
GWAS chip	HapMap	YRI	1.83
GWAS chip	1000 Genomes	ALL	8.86
Metabochip	-	EUR	1.93
Metabochip	-	AFR	2.18

The calibration factors were estimated from the publically available HapMap and 1000 Genomes reference data. Only autosomal and non-monomorphic SNPs were used in the estimation. The Metabochip *c* factors were estimated from 179,000 overlapping SNPs from 1000 Genomes reference data frequencies.