SUPPLEMENTARY MATERIAL

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

Winkler TW et al

CONTENTS

Supplementary Methods. Creation of the SNP identifier reference panel.

To harmonize marker names across studies and between Hapmap imputed and genotyped Metabochip data, in GIANT, we have applied the reference panel '*SNPID_to_ChrPosID.b36_v2.txt.gz'* to assign unique ChrPos-IDs (of the format "chr<chromosome>:<base-position>", see column *ChrPosID*) to different versions of SNP-IDs (e.g. different versions of rs-IDs or array-specific marker names like "SNP_1_12345", see column *SNPID*). The idea behind the remapping file is to map every known SNP-ID to its unique ChrPos-ID (using genome build 36, b36) and as such to maximize the overlap in the number of SNPs between study files. The remapping file maps ~9.1 million different SNP-IDs to ~4.8 million unique b36 ChrPos-IDs. The compilation of the file involved four major steps:

1) *Preparing an initial b36 remapping file.* At first, we created a remapping file for all SNPs contained in the CEU Hapmap2 r22 imputation reference panel[1]. We extracted rs-IDs and respective genomic positions (chromosome and position) from mapping files that are available from http://www.sph.umich.edu/csg/abecasis/MACH/download/HapMap-r22.html . We copied the rs-IDs into column *SNPID* and created the column *ChrPosID* by horizontally concatenating string "chr", the chromosomal position, the character ":" and the base position of the respective SNP.

2) *Adding non-standard SNP-IDs.* For non-standard SNP-IDs, particularly array-specific SNP-IDs, we recommend re-mapping the probe sequences from the array to the reference genome of interest. The probe sequences can be found in .cdf files for Affy arrays and .bpm files for Illumina arrays. These files are usually available for download from the Affymetrix (http://www.affymetrix.com/support/technical/index.affx) or from the Illumina (http://support.illumina.com/downloads.ilmn) website. Re-mapping ensures that the probe used on the array maps uniquely, and identifies the exact chr:pos location where the probe matches. Non-unique probes, mapping to multiple locations in the genome, should be excluded. For a great many arrays this re-mapping has already been done by several groups. Sites such as http://www.well.ox.ac.uk/~wrayner/strand/ have freely available files that have already accurately mapped array-specific SNP-IDs, such as Metabochip-IDs, Exomechip-IDs, or Affy-IDs, to chr:pos locations for multiple genome builds. Again, we used

b36 locations of the array-specific SNP-IDs to compile their respective ChrPos-IDs and vertically added the novel *SNPID - ChrPosID* combinations to the remapping file. Note that for this to be done effectively, it is important to have information on the specific array(s) used for genotyping.

3) Adding SNP-IDs from older builds (<b36). At third, we vertically added rows with outdated SNP-IDs. The NCBI ftp-site provides files called “RsMergeArch.bcp.gz” that can be used to backtrack changes in SNP-IDs between previous builds (first column of the table, *OldMarkerName*) and the current build (second column of the table, *NewMarkerName*)[2]. To refer to b36 as the current build, we applied the “RsMergeArch.bcp.gz” file from dbSNP build 130, which is available from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/b130_archive (*OldMarkerName*: outdated SNP-ID <b36; *NewMarkerName:* b36 SNP-ID). To assign b36 ChrPos-IDs to the outdated SNP-IDs, we merged column *NewMarkerName* of the “RsMergeArch.bcp.gz” file with column *SNPID* of the remapping file. We then vertically added the novel *OldMarkerName - ChrPosID* combinations to the remapping file.

4) *Adding SNP-IDs from newer builds (>b36).* At last, we added newer SNP-IDs, which were to be found in the “RsMergeArch.bcp.gz” file from dbSNP build 131 (refers to genome build 37) that is available from ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/b131_archive (*OldMarkerName*: SNP-ID <b37; *NewMarkerName:* b37 SNP-ID). To assign b36 ChrPos-IDs to the newer b37 SNP-IDs, we merged column *OldMarkerName* of the “RsMergeArch.bcp.gz” file with column *SNPID* of the remapping file and vertically added the novel *NewMarkerName - ChrPosID* combinations to the remapping file.

Suppl. Table 1. Description of EasyQC report variables (File-level QC). The table lists and describes variables in the EasyQC-report (rep-file), which is created by the EasyQC-script for file-level QC ('filelevel_qc. [gwa|mc].ecf'). Many of the QC variables are critical, meaning that their value is expected to be zero. If values > 0 are observed for such critical variables, the study analyst should be contacted for clarification and (if needed) be asked for re-upload. In such cases it can also be helpful to save the affected SNPs to a separate file, which can easily be done by adding the parameter '--blnWriteCleaned 1' to the respective CLEAN function step.

Variable Name	Description
fileInShortName	Name of the Input-file
numSNPsIn	# SNPs in Input-file
numSNPsOut	# SNPs in the cleaned Output-file
numDrop_Monomorph	# SNPs excluded due to EAF=0 or EAF=1
numDrop_Missing_EA*	# SNPs excluded due to missing Effect_allele (critical)
numDrop_Missing_OA*	# SNPs excluded due to missing Other_allele (critical)
numDrop_Missing_P*	# SNPs excluded due to missing P (critical)
numDrop_Missing_BETA*	# SNPs excluded due to missing BETA (critical)
numDrop_Missing_SE*	# SNPs excluded due to missing SE (critical)
numDrop_Missing_EAF*	# SNPs excluded due to missing EAF (critical)
numDrop_Missing_N*	# SNPs excluded due to missing N (critical)
numDrop_invalid_EA*	# SNPs excluded due to invalid Effect_allele (critical)
numDrop_invalid_OA*	# SNPs excluded due to invalid Other_allele (critical)
numDrop_invalid_P*	# SNPs excluded due to invalid P (critical)
numDrop_invalid_SE*	# SNPs excluded due to invalid SE (critical)
numDrop_invalid_BETA*	# SNPs excluded due to invalid BETA (critical)
numDrop_invalid_EAF*	# SNPs excluded due to invalid EAF (critical)
numDrop_Nlt30	# SNPs excluded due to N<30
numDrop_MAClet6	# SNPs excluded due to MAC<=6
numDrop_MissingInformationType[a]*	# SNPs excluded due to missing Information_type (critical)
numDrop_Genotyped_LowInformation[a]*	# genotyped SNPs excluded due to having imputation quality < 1 (critical)
numDrop_Imputed_MissingInformation[a]*	# imputed SNPs excluded due to missing imputation quality (critical)
numDrop_LowInformation[a]	# imputed SNPs excluded due to low imputation quality (thresholds stated in **Table 2**)
numDrop_MissingCallrate[b]*	# SNPs excluded due to missing Callrate (critical)
numDrop_MissingPHwe[b]*	# SNPs excluded due to missing Phwe (critical)
numDrop_InvalidCallrate[b]*	# SNPs excluded due to invalid Callrate (critical)
numDrop_InvalidPhwe[b]*	# SNPs excluded due to invalid Phwe (critical)
numDrop_LowCallrate[b]	# SNPs excluded due to Callrate<0.95
numDrop_LowHwe[b]	# SNPs excluded due to P-HWE<1e-6
numDropSNP_ChrXY	# sex-chromosomal SNPs excluded
numRenamedMatch	# SNPs successfully matched to Chr:Pos-SNP ID format
numDuplicates.ChrPosID	# duplicated SNPs being excluded

[a] only relevant for GWA data; [b] only relevant for typed MetaboChip data; * critical (expected to be zero)

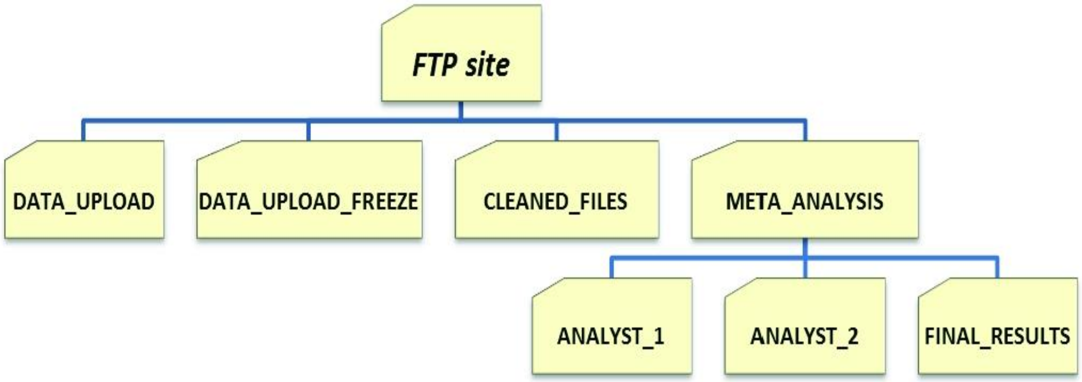
Suppl. Table 2. Description of EasyQC report variables (Meta-level QC). The table lists and describes variables in the EasyQC-report (rep-file), which is created by the EasyQC-script for meta-level QC (script ‘metalevel_qc.ecf’). In particular the AFCHECK variables can be helpful to investigate allele frequency issues. If high numbers of SNPs with invalid or mismatching alleles or outlying allele frequencies are observed (see variables AFCHECK.AlleleInInvalid, AFCHECK.AlleleMismatch or AFCHECK.numOutlier respectively), the source of the problem should be investigated. For this purpose, it can be helpful to examine the automatically written files ‘*AFCHECK.invalid.txt’, ‘*AFCHECK.mismatch.txt’ or ‘*AFCHECK.outliers.txt’, which contain the affected SNP subsets of the input data..

VariableName	Description
fileInShortName	Name of the Input-file
numSNPsIn	# SNPs in Input-file
numSNPsOut	# SNPs in the Output-file
Nmax	Maximum sample size
SEmedian	Median Standard Error
AFCHECK.NotInRef	# SNPs from the input that are not present in the allele-frequency reference
AFCHECK.NotInIn	# SNPs from the allele-frequency reference that are not present in the input
AFCHECK.Checked	# SNPs that are present in the allele-frequency reference and thus are being checked for Allele-frequency issues (i.e. drawn in the allele frequency graphs)
AFCHECK.StrandChange	# SNPs with switched strand (i.e. + in input, - in ref; or vice versa).
AFCHECK.AlleleMatch	# SNPs with matching alleles, same direction and same strand (e.g. +AC in input and +AC in reference)
AFCHECK.AlleleChange	# SNPs with matching alleles, switched direction and same strand (e.g. +AC in input and +CA in reference)
AFCHECK.n4AlleleMatch	# non-palindromic SNPs with matching alleles, same direction and switched strand (e.g. +AC in input and +TG in reference)
AFCHECK.n4AlleleChange	# non-palindromic SNPs with matching alleles, switched direction and switched strand (e.g. +AC in input and +GT in reference)
AFCHECK.AlleleMismatch	# SNPs with allele mismatch (e.g. +AG in input and +AC in reference)
AFCHECK.AlleleInMissing	# SNPs with missing input alleles
AFCHECK.AlleleInInvalid	# SNPs with invalid input alleles (other than A,C,G,T)
AFCHECK.StrandInInvalid	# SNPs with invalid input strand (other than +,-)
AFCHECK.AlleleRefMissing	# SNPs with missing reference alleles
AFCHECK.AlleleRefInvalid	# SNPs with invalid reference alleles (other than A,C,G,T)
AFCHECK.StrandRefInvalid	# SNPs with invalid reference strand (other than +,-)
AFCHECK.cor_Freq1.ref_EAF	Pearson Correlation between the input and the reference allele frequency after aligning all directions to the reference
AFCHECK.numOutlier	# outlying SNPs that differ > 0.2 in terms of allele frequency from the reference allele frequency
Lambda.P.GC	GC Lambda of the input file

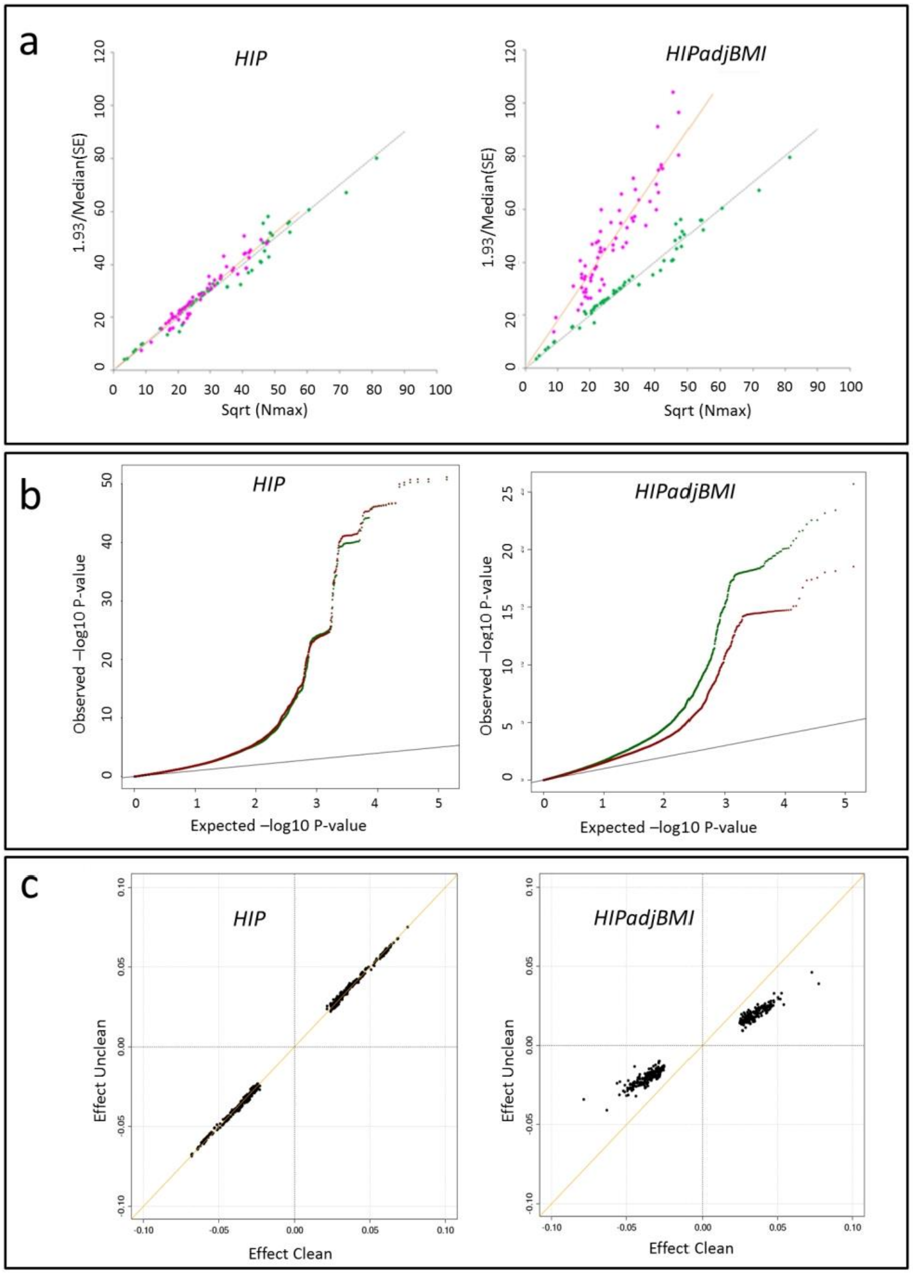
Suppl. Table 3. Description of EasyQC report variables (Meta-analysis QC). The table describes variables in the EasyQC-reports ("rep"-files), which are created by the EasyQC-script for meta-analysis QC (script 'metaanalysis_qc.ecf'): (i) 'metaanalysis_qc.rep' is created before the merging of the two meta-analysis results (two rows), and (ii) 'metaanalysis_qc.merged.rep' is created after the merging and provides comparison statistics (one row).

VariableName	Description
Report file: metaanalysis_qc.rep:	
fileInShortName	Name of the Input-file.
fileInTag	Tag of the input file (refers to the set --fileInTag).
numSNPsIn	Number of SNPs in Input-file.
numSNPsOut	Number of SNPs in the Output-file.
P.value_num	Number of P-values.
P.value_NA	Number of missing P-values.
P.value_min	Minimum P-value.
P.value_max	Maximum P-value.
P.value_median	Median P-value.
P.value_p25	25th percentile of the P-values.
P.value_p75	75th percentile of the P-values.
P.value_mean	Mean P-value.
P.value_sd	Standard deviation of the P-values.
N_num	Number of sample sizes.
N_NA	Number of missing sample sizes.
N_min	Minimum sample size.
N_max	Maximum sample size.
N_median	Median sample size.
N_p25	25th percentile of the sample sizes.
N_p75	75th percentile of the sample sizes.
N_mean	Mean sample size.
N_sd	Standard deviation of the sample sizes.
Lambda.P.value.GC	GC Lambda of the input file.
Report file: metaanalysis_qc.merged.rep:	
fileInShortName	This will be named by the script's name "metaanlysis_qc.1.merged".
fileInTag	This will be labeled according to the defined fileInTags: A1_A2
numSNPsIn	Number of SNPs in the merged data set. Since an inner join is being performed, this number reflects the number of overlapping SNPs between the two meta-analysis results.
numSNPsOut	Number of SNPs in the output (not meaningful in here)
corr_Pvals	Spearman correlation coefficient between P-values.

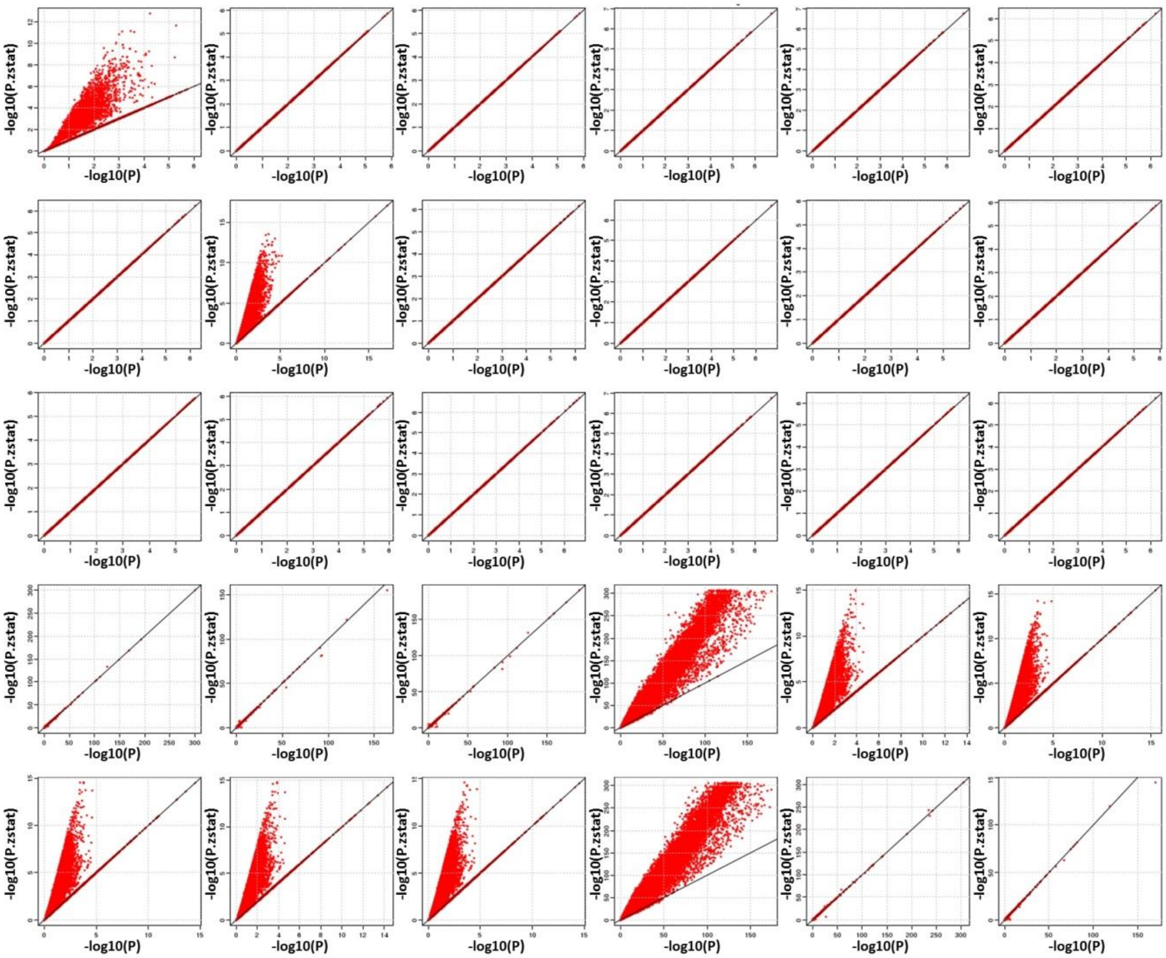
Supplementary Figure 1. Ftp-site directory structure. The DATA_UPLOAD directory is used for the collection of raw study-specific results, i.e. used by the collaborators to upload their results. Once all or at least files from >80% of studies have been collected, the DATA_UPLOAD folder should be frozen. The folder should be protected from further changes, be renamed to DATA_UPLOAD_FREEZE and a new DATA_UPLOAD folder should be created to collect any additional results. The CLEANED_FILES directory should be used for collection of cleaned files that passed the file-level QC routines. The META_ANALYSIS directory should be used to upload meta-analysis results and contains sub-folders, one for each meta-analyst (folders ANALYST_1 and ANALYST_2) and one to collect and freeze the final meta-analysis results (FINAL_RESULT).



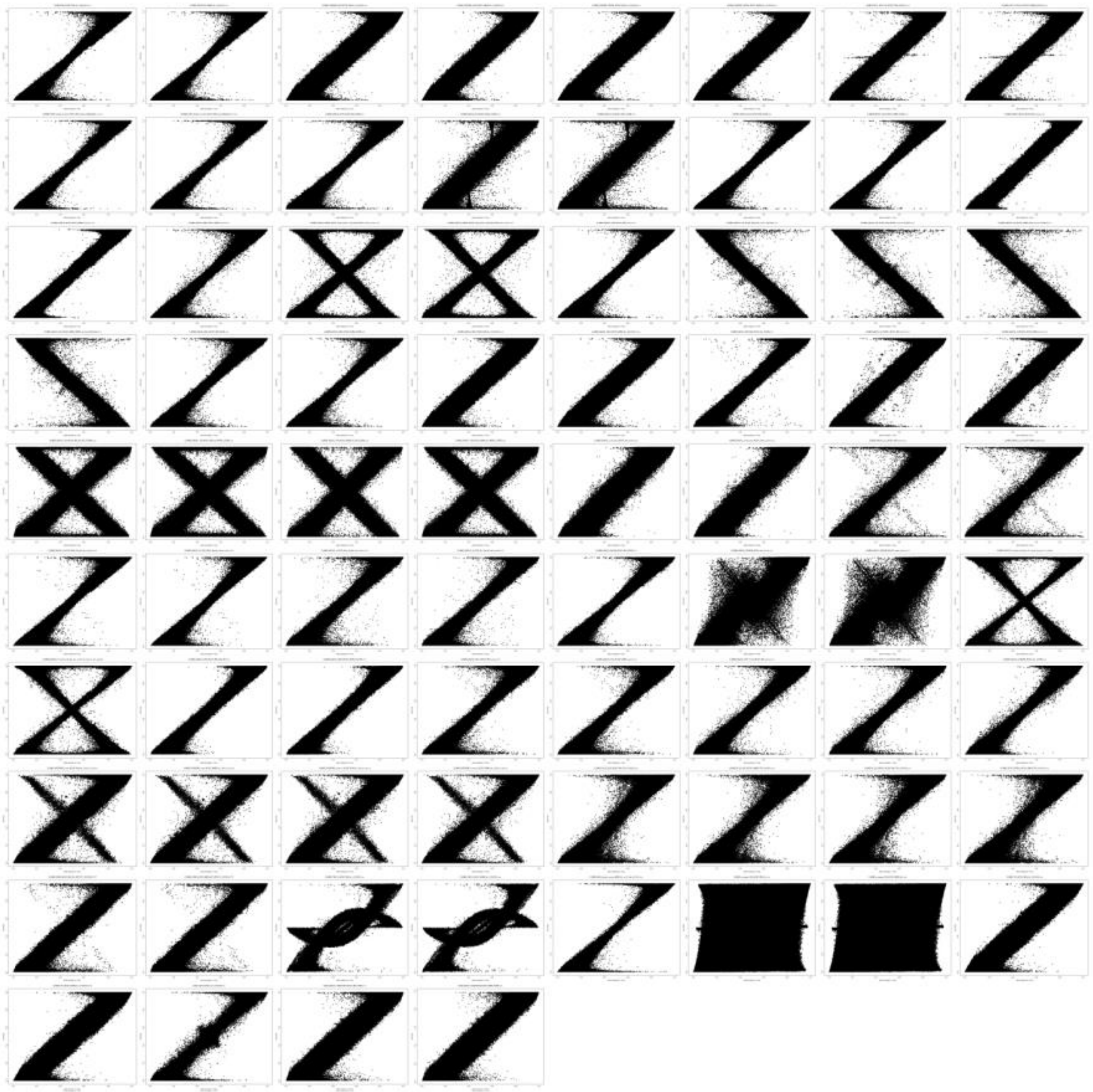
Supplementary Figure 2. Effect of the trait transformation issue. On the example of the phenotype hip circumference with and without adjustment for BMI (HIP, HIPadjBMI) in the GIANT Metabochip studies (81,000 subjects), it can be seen that (a) the trait transformation issue only affected the trait adjusted for BMI (SE-N plots; magenta: uncleaned studies affected by the issue; green: cleaned studies) ,(b) the uncleaned data had deteriorated power for the BMI-adjusted trait (QQ plot of association P-values from the Meta-analysis for all SNPs; red: meta-analysis on uncleaned data; green: meta-analysis on cleaned data) and (c) the uncleaned data yielded estimates biased towards the null for the BMI-adjusted trait (estimates from the Meta-analysis on uncleaned data on Y-axis and from cleaned data on X-axis).



Supplementary Figure 3. EasyQC panel of P-Z plots. Example EasyQC panel of plots to check whether reported P-Values (X-axis, on -log10 scale) match P-Values calculated from the Z-statistic using the reported beta estimates and standard errors (Y-axis, on –log10 scale) with one plot per file. Clearly, several files show deviations, which were due to deviating software specifications used by these studies, which were resolved with study analysts.



Supplementary Figure 4. EasyQC panel of EAF-plots. Example panel of plots to check issues with allele frequencies. Each plot contrasts the allele frequency of the input file (y-axis) with the allele frequency of the reference (x-axis). In this case the meta-analyzed GIANT height results have been used as reference to compare it to study-specific GWA results for height. Several issues can immediately be detected, which should be solved with the study analysts.



References

1. International HapMap, C. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861 (2007).
2. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29, 308-311 (2001).