BroadABC Quality Control and Meta-analysis Protocol

Steps to include in analysis pipeline:

- concatenate summary files if different files per chromosome so that 1 file per GWAS
- check order of columns
- check header

Format checks GWAS files

SNP column consists of -999, or specific SNP format (note that very many different formats are used across sets)
CHR column is -999 or integer 1-22
POS column is -999 or positive integer
EFF_ALL is -999, A, C, T or G or indel (different formats used for indels)
NONEFF_ALL is -999, A, C, T or G or indel (different formats used for indels)
BETA/OR is -999 or numeric
SE is -999 or numeric
P is -999 or numeric between 0 and 1
MAF is -999 or numeric between 0 and 1
HWE is -999 or numeric between 0 and 1
IMP is -999, 0 or 1
INFO is -999 or numeric between 0 and 1
INFO_TYPE is
N_EFF is NA or numeric >0

- Check if missing values; how many and why?
- Check number of duplicate SNPs, if more than 20, check by hand what is going on
- Check whether the reported alleles are in line with the reference set, if wrong for > 50, check by hand
- QQ-plot cleaned dataset, lambda cleaned dataset
- Manhattan plot - cleaned dataset
- Histogram of MAF - MAF > 0.01
- Histogram of INFO - MAF > 0.01
- Plot allele frequency effect allele versus frequency same allele in reference set
- Plot reported p-value versus Wald-test p-value
- Check whether n and MAF predict SE within cohort (without outliers)
- Box plot BETAs and SEs & estimated SDs across cohorts

Cleaning files

- QC include only SNPs for which
    - MAF > 0.01 (column 9 MAF)
    - INFO > 0.3 , (separate run on INFO > 0.6) (column 12 INFO)
    - call rate 98-99% from effective sample size?
    - HWE > 1*10-7 or 1*10-6
- What to do with ambiguous SNPs (G:C, T:A) – exclude?
    - METAL will handle differences in effect allele vs reference allele, but ambiguous SNPs need to be cleaned out or we need to know that they are on the same strand
- Make a column with direction of effect (from log-odds for the dichotomous GWAS)
- Calculate SNP availability in total sample . Use ≥ 80% as filter? (run meta-analysis both using this filter and without? See e.g. educational attainment GWAS: "Only SNPs with an availability of ≥80% in the total sample were selected"
- Exclude missing BETA/SE/PVAL
- Exclude SNPs with different alleles than reference set or strand issues
  Only keep unique SNPs (select the first one in case of duplicate SNPs. Do this as last step.)

Meta-analysis

Fixed-effects meta-analyses using METAL (at least men, women, combined)

- Z-scores weighted by effective sample size
- Genomic control correction? (I would assume yes)
- Variables needed:
    - SNP
    - EFF_ALL
    - NONEFF_ALL
    - Direction of effect
    - PVALUE
    - N
- Additional possible QC measures:
    - "Leave one out"
    - Leave one out and calculate genetic risk scores from the meta-analysis results and check association between this risk score and the phenotype in the sample that was left out.

Post meta-analysis checks

- QQ-plot meta-analysis results
- Manhattan plot meta-analysis results
- Consistency of sign BETA across cohorts for top SNPs (top 1000-10000 or so)

- Check for heterogeneity top signals (statistical test of heterogeneity)
- Forrest plot top hits (visual test of heterogeneity)
- Check frequency of the allele across cohorts for top hits, and frequency in all separate ethnicity panels in 1000G
- Check local LD and p-value and direction of beta for SNPs in LD with the top hits
- Check info score, HWE, call rate, n of genotyped and imputed SNPs for top hits

Plan:

- JT and AJ write scripts separately
- Data preparation and QC:
    - Choose 1 sample to check scripts with (e.g. QIMRB or the first one we have GWAS results for)? Or one quantitative and one dichotomous. (i.e. run with both scripts and compare results)
    - Then divide who runs the scripts on which of the remaining samples
    - Randomly choose another sample to check
- Meta-analyses: Both JT and AJ run the final meta-analyses (women, men, combined)

Replication:

- SNPs with p-values less than $1*10^{-6}$ for replication
- Genes with p-values less than $1*10^{-4}$ for replication

Combined discovery + replication meta-analysis:

- Conduct MA with all samples

Biological annotation:

- Gene-based analyses
- Pathway-analyses

Simulation

- Poisson distribution?
- http://www.stat.berkeley.edu/~vigre/activities/bootstrap/2006/wickham_R.pdf
- Johnson distribution
- http://cran.r-project.org/web/packages/JohnsonDistribution/JohnsonDistribution.pdf