**Genetic risk score calculation for IGEMS using SBayesR**

**Revised protocol 2021-05-25 [PRS workgroup]**

Scripts, snp lists and documentation can be found here in Dropbox under:

\GE Interplay Documentation\WORKGROUPS\Genotypes\SBayesR\_steps

**General recommendations:**

* PRS files should be named as follows when shared with Patty: PHENOTYPE\_AuthorYear\_METHOD\_Cohort (e.g. EA3\_Lee2018\_SBayesR\_OATS, or AD\_Kunkle2021\_Clumping\_AO50)
* Variable names should be: TRAIT\_AuthorYear, i.e. AD\_Kunkle2021 or EA3\_Lee2018.
* A file with the Principal components should be named: PCs\_COHORT\_DATE (e.g. PCs\_SATSA\_June2021) and the variables are just PC\_1, PC\_2 etc.
* Please provide brief information for Patty with details what your file contains and how/when it has been derived (partly covered by the file naming), so that she can keep track of the different PRSs.
* Lead researcher needs to organize ‘Leave-All-Out’ sumstats if that is the preferred method. Default is Leave-One-Out (LOO) for each respective cohort where applicable, potentially resulting in small differences between sumstats per cohort while maximizing the discovery sample.
* The first group who rescales sumstats (without LOO) using GCTB should share the final rescaled sumstat file with the other cohorts so that the rescaling of sumstats does not need to be done separately on each site.

**Data needed:**

1. GWAS summary statistics - see PRS across IGEMS\_23.06.2020.xlsx for which discovery GWAS to be used for a given trait or use as agreed with the respective project members.

2. Sparse LD matrix generated using 50k random UKB samples by SBayesR authors (22GB file downloadable via the following link: 10.5281/zenodo.3350914)

3. Target genotype data for which to calculate PRSs

4. SNP list of ~0.95M HapMap3 SNPs available across all IGEMS data (without the MHC region): \WORKGROUPS\Genotypes\SBayesR\_steps\ SNPs\_IGEMS\_selected\_20210228.txt

**Retrieving and preparing the GWAS summary statistics (weights)**

1. **Download** summary statistics - get leave one out sumstats where target sample has also been part of discovery GWAS. [Leave all IGEMS out can be done on request with sumstats organized by project lead]
2. **Prepare**, filter and format summary statistics [format required for GCTB, i.e. GCTA-COJO format]
   1. Restrict to HapMap3 SNPs available across all IGEMS cohorts using the provided list [\WORKGROUPS\Genotypes\SBayesR\_steps\ SNPs\_IGEMS\_selected\_20210228.txt; list with rsID for all HapMap3 SNPS can be found here: \SBayesR\_steps\w\_hm3.noMHC.snplist, and list with chromosome and position (hg19) here: \SBayesR\_steps\HM3\_lists\_by\_chr\_with\_positions\_hg19\]
   2. Convert OR to log(OR) for binary traits so all weights are on a linear scale
   3. Filter out strand-ambiguous SNPs (C/G or A/T) and structural variants
   4. If details are provided, filter out low imputation quality SNPs (if imputation score available), INFO < 0.8 (IMPUTE2) or R2 < 0.8 (MACH), and filter out low MAF SNPs (MAF < 1%)
   5. Convert GWAS summary statistic to .ma format [for GCTB (GCTA-COJO format) columns need the following order: SNP identifier, effect allele, other allele, frequency of the effect allele, effect size, standard error, p-value and sample size]
      1. NOTE! If sample size is not available, use the reported sample size from the publication in the N column and the –impute-n option in the script (now commented out).
      2. See the details of input format from the tutorial in GCTB site (<https://cnsgenomics.com/software/gctb/#SummaryBayesianAlphabet>)
3. **Rescale** summary statistics using GCTB
   1. Get sparse LD matrix generated using 50k random UKB samples by SBayesR authors (22GB file downloadable via the following link: 10.5281/zenodo.3350914)
   2. Calculate the re-weighted effect size estimates by chromosome of your summary statistics using SBayesR with the sparse LD matrix [use and adapt example script \SBayesR\_steps\gctb\_sbayesr\_BMI\_chr1to22.sh.]
   3. Concatenate the output .snpRes files from each chromosome to one data set [use example script \SBayesR\_steps\concatenate\_snpRes.sh]
   4. Extract the columns "Name", "A1" and "A1Effect" from the file needed to calculate individual sum scores [use and adapt example script \SBayesR\_steps\get\_betas\_BMI.sh]

**Generating the scores (separately per site)**

1. Prepare the target genotype data for polygenic score calculation [several ways to do these steps so no tool specific details here]:
   1. Restrict SNPs to HapMap3 SNPs available across all IGEMS samples using the provided list \SBayesR\_steps\ SNPs\_IGEMS\_selected\_20210228.txt. Note that this list already has filtered out any out any rare (MAF<1%) and poorly imputed (INFO<0.8) variants from each chromosome across the IGEMS consortium and only includes snps available in all samples!!]
   2. Filter out any variant with duplicate position (the final scoring will not work if there are duplicate variants in the data)
   3. Merge all chromosomes to one dataset
2. **Calculate** the individual sum scores using PLINK --score function [use and adapt example script \SBayesR\_steps\plink\_score\_EXAMPLE\_BMI\_SBR.sh]