SpiroMeta Haplotype Reference Consortium (HRC) Imputation Analysis Plan

**Draft:** 12/04/2017 **Contact:** Vicki Jackson ([vej4@le.ac.uk](mailto:vej4@le.ac.uk))

# Overview

With the full release of the UK Biobank data being made available imminently, we are planning to combine analyses of lung function using these data with analyses undertaken by SpiroMeta studies. SpiroMeta studies may contribute in two ways:

1. We shall utilise the already complete results used in the 1000 Genomes meta-analysis.
2. We are inviting studies to contribute new analysis undertaken using genotypes imputed to the new HRC panel. For these new analyses, we are primarily seeking contributions by studies who either did not contribute to the 1000 Genomes discovery analysis, or whose sample size has significantly increased since the 1000 Genomes effort.

Given we are combining these results with analysis undertaken in UK Biobank, we are asking that studies who wish to contribute analysis using HRC imputed data aim to upload these results by **Friday 26th May 2017.**

# Imputation and Quality Control

It is assumed that all contributing studies will have already completed standard quality control checks of the genotype data and imputation to the HRC imputation panel, using either the Michigan or Sanger imputation servers. It is also assumed that genome-wide genotypes correspond with GRCh37 / hg19, and are aligned to the forward strand. To facilitate downstream QC of results centrally, we ask studies to provide the following:

1. Please could studies complete the attached spreadsheet with details of the QC carried out, prior to imputation, and details of the imputation itself.
2. We request that studies run this post-imputation tool, developed by Will Rayner (available: <http://www.well.ox.ac.uk/~wrayner/tools/Post-Imputation.html>), and include the resulting output with their results upload.
3. We also require basic metrics regarding the quality of all SNPs. For studies who used the Michigan imputation server, please provide the info files and qcreport.html (if downloaded from the Michigan server). Alternatively, or for studies who used the Sanger server, please run qctool (<http://www.well.ox.ac.uk/~gav/qctool/#documentation>), which takes the vcf file as input, and using the –snp-stats option produce these summary statistics.

# Software

Studies may use any suitable software for association analysis. For large studies (N>5000) with related individuals, BOLT-LMM is recommended as an efficient algorithm for mixed model association testing; this software shall be used for the analysis of UK Biobank data. Studies with unrelated individuals may use standard linear models (eg using SNPTEST etc.), with appropriate adjustment for ancestry principal components. Where possible, we strongly encourage use of programs/options that make use of the posterior probabilities of the genotype calls rather than a simple threshold approach. Please contact Vicki ([vej4@le.ac.uk](mailto:vej4@le.ac.uk)) to discuss any software issues.

# Phenotypes and Association Analysis

We are asking for association analyses to be undertaken by all studies for FEV1, FVC and FEV1/FVC. Where studies have PEF available, we also ask studies also undertake analysis for this trait if possible; however please prioritise the other three traits. Similarly, we ask studies to prioritise analysis of the autosomal chromosomes, with analysis of the X chromosome if time allows.

### AUTOSOMES

Could **all studies** please undertake the following analyses:

1. Never-smokers only:

Restrict dataset to never-smokers only with complete data on both FEV1 **and** FVC. Undertake linear regression of age, age2, sex, height, on the trait. Studies of unrelated individuals should also adjust for ancestry principal components. Transform the residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for SNP association testing under an additive genetic model.

1. Ever-smokers only:

a.Analysis as for 1 above, but restrict to ever-smokers only.

b. Restrict dataset to ever-smokers only with no missing data for the **pack-years** variable and complete data on both FEV1 **and** FVC. Undertake linear regression of age, age2, sex, height, ancestry principal components (if appropriate) and **pack-years** on the trait. Transform residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for SNP association testing under an additive genetic model.

Could **studies with related individuals only** please also undertake the following analysis:

1. All individuals:

Restrict dataset to individuals with complete data on both FEV1 **and** FVC. Undertake linear regression of age, age2, sex, height, and **smoking status** (never versus ever smoker) on the trait. Transform the residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for SNP association testing under an additive genetic model.

### CHROMOSOME X

**Studies with unrelated individuals:** undertake analyses 1 & 2 above, separately in males and females. For males, please code SNPs as 0 for 0 copies of the coded allele and 2 for 1 copy of the coded allele, making sure that the same coded allele is used for females and males. Residuals calculated with males and females combined (as for the autosomal analysis) should be used for this analysis.

*Please note for chromosome X, the imputation server generates separate vcf files for males and females and individuals with discordant sex information are removed; as a result, the chrX vcfs may contain fewer individuals than the autosomal files, if these individuals were not removed prior to imputation. When creating phenotype files for chromosome X analyses, please ensure the ordering of individuals corresponds with the VCF file. Lists of samples included in each vcf may be extracted using bcftools (*[*https://samtools.github.io/bcftools/bcftools.html*](https://samtools.github.io/bcftools/bcftools.html)*) by running:*

bcftools query –l vcfname.vcf.gz > vcf\_ordered\_ids.txt

**Studies with related individuals:** undertake analyses 1, 2 & 3 above, separately in males and females. For males, please code SNPs as 0 for 0 copies of the coded allele and 2 for 1 copy of the coded allele, making sure that the same coded allele is used for females and males. Please also undertake the above analyses in males and females together, with sex as an additional covariate. Residuals calculated with males and females combined (as for the autosomal analysis) should be used for this analysis.

*Please note, the male and female files may be merged together, using either bcftools (*merge *function) or alternatively using vcftools (*[*https://vcftools.github.io/perl\_module.html*](https://vcftools.github.io/perl_module.html)*) ‑*vcf‑merge *function). Again, please ensure the ordering of the samples in the phenotype file corresponds with the VCF file. For more guidance on this, please contact Vicki.*

# Data Uploads and File Naming Scheme

Please use the “Drop-off” facility on the University of Leicester filedrop to deposit all files: (<https://filedrop.le.ac.uk/>). Please combine and compress files, before uploading.

### MAIN RESULTS FILES

The following fields will be required for each SNP in the result files. It would be appreciated if the fields are named following the **bold** titles as below.

* **Markername**: as produced by the imputation
* **Chrom:** chromosome (integer)
* **Pos:** position (integer)
* **Bas\_all**: baseline allele - a single character: “A” “C” “G” “T” for SNPs. Note- the HRC panel includes only SNPs. If a merged panel (eg HRC + 1000 Genomes) has been used, please use coding “R”, “I” or “D” for INDELS and CNVs.
* **Cod\_all**: coded allele (effect allele) - a single character: “A” “C” “G” “T” for SNPs (“R”, “I” or “D” for INDELS and CNVs)
* **Freq**: allele frequency for **coded allele** (numeric data, 4 decimal places), **please ensure that the allele frequency reported corresponds to the subgroup analysed (ever-smokers or never-smokers)**
* **Beta**: effect size for each copy of the coded allele (numeric data, with at least 4 decimal places)
* **Se**: standard errors of beta (numeric data, with at least 6 decimal places)
* **Ntotal**: total number of individuals with phenotype and genotype data for the given marker (an integer)
* **Imp\_info**: r2\_hat or proper\_info for imputed SNPs (numeric data, 4 decimal places, matching the data type specified in the "Imputation\_Type"), if produced by association testing software.
* Integers should be supplied as a single integer with no decimal point.
* Please code missing values as a single dot character (“.”).
* Note that no quotes should be used around any data cells or headers.
* No row indices column or any other extra columns should be provided.
* Please upload results for **all** SNPs, without filtering for monomorphs or SNPs for which no association result is produced.

It would be helpful if the results of each analyses will be given in gzip compressed tab‐delimited txt‐format files, named as:

**cohortname\_phenotype\_analysis\_genome\_analyst\_version**.txt.gz

where:

**cohortname** will be an identifier for the specific cohort

**phenotype** will be one of “FEV1”,”FVC”, ”FF” (for the ratio FEV1/FVC), or “PEF”.

**analysis** will be one of ”smk”, “smkPY” (for the pack-years adjustment),”nonsmk” or “all” (studies with related individuals only)

**genome** will be “CHR{1-22}”, “CHRX\_females”, “CHRX\_males”, or “CHRX\_all”, for autosomal chromosomes {1-22}, or for X chromosome separately in females and males, or all individuals combined.

**Analyst** will be the initials of the analyst

**version** will be the date of the day of the uploading (ddmmyyyy)

For example a file name from the cohort BHS would be:

BHS\_FEV1\_smk\_CHR4\_vej\_28032017.txt.gz

### IMPUTATION AND SNP QC FILES

Please upload applicable files, as described in section B of this analysis plan, using naming as follows:

**cohortname\_**qc\_imputation\_details **\_analyst\_version**.xlsx– spreadsheet pertaining to pre-imputation QC checks and imputation details.

**cohortname\_**post\_imputation\_qc\_check **\_genome\_analyst\_version**.html – plots generated using Will Rayner’s post-imputation QC tool.

**cohortname\_** qcreport **\_genome\_analyst\_version**.html – qc report downloaded from the Michigan server.

**cohortname \_genome\_analyst\_version**.info.gz – info file downloaded from the Michigan server.

**cohortname \_genome\_analyst\_version**.snp-stats.gz – snp stats file, as produced by qctool.

where:

**cohortname** will be an identifier for the specific cohort

**analyst will be the analyst’s initials**

**version** will be the date of the day of the uploading (ddmmyy)

**genome** will be “CHR{1-22}”, “CHRX\_females”, “CHRX\_males”

### STUDY DESCRIPTIVES

Finally, we ask that studies complete the excel file circulated with this plan with descriptive information:

**cohortname\_**descriptives**\_analyst\_version.xlsx**

where:

**cohortname** will be an identifier for the specific cohort

**analyst will be the analyst’s initials**

**version** will be the date of the day of the uploading (ddmmyy)

Please also provide histograms of trait residuals (FEV1, FEV1/FVC, FVC and PEF) from linear regression after adjusting for covariates. Please provide these for males and females separately and for all individuals combined.