**Frequently Asked Questions (FAQ) for the**

**Global Lipids Genetics/GIANT Consortium HRC and 1KG phase3 Imputation and Analysis plan version 2 or 2a, April 2017**

**A) What do I do about a warning or error regarding missing header lines?**

1) Ensure you are using bcftools 1.3.1 and tabix 0.2.6

If you only have a warning about a Genotype headers missing, you can ignore it.

If you have an error please add the following 2 lines to a file and use bcftools to update your vcf files with this command.

bcftools annotate -h addLines.txt chrX.no.auto\_male.dose.vcf.gz -O z > new\_chrX.no.auto\_male.dose.vcf.gz

##FILTER=<ID=GENOTYPED,Description="Site was genotyped">

##FILTER=<ID=GENOTYPED\_ONLY,Description="Site was genotyped only”>

**B) The frequencies are off for ChrX in the post-imputation qc report.html – did imputation work correctly?**

Yes. There appears to be an error in the allele frequency calculations for HRC chrX. Allele frequencies for the reference data set are shifted ~25%, so that variants at ~0% in an imputed sample appear to be at 25% in the reference panel and following the diagonal from there. This is a known issue with the reference panel values that has yet to be fixed. So long as your files are following a diagonal, then imputation was likely successful. If you have any concern, please contact us.

**C) Covariance matrices are really large – is that OK?**

Many cohort are seeing covariance matrices >100GB.

Do try to be sure that you have followed these steps to ensure the matrix is as small as possible:

1) Run association only on polymorphic variants (with at least 1 hard called heterozygote).

2) Included the window size adjustment to decrease the default matrix size. Make sure that the window size is woking correctly by looking at your log files.

3) Generate as few covariance matrices as is feasible for your cohort. (Hopefully a max of 2.)

For now, so long as your matrix is <500GB, please upload. We will monitor disk space on the server and adjust as needed. If you have sufficient disk space on your local system, please try to save your covariance matrix. We very much do not want to have to implement alternative methods, but that may be necessary as we proceed.

**D) Was data upload successful?**

The sftp is write only. You will not have the ability to view files in the folders. Please just upload, and we will contact you if there appears to be a problem with your files. However, please do send us an email after you have uploaded, so that we know to check for your files.

**E) The analysis is slow – any suggestions for speeding things up?**

Make sure you are not generating covariance matrices for all analyses. For the analyses with the largest N (ideally no more than 2 per ancestry) you should generate and upload the covariance matrices, using --meta score,cov[windowSize=500000],  but for all the rest of the analyses only use the option "--meta score". The latter is much faster.

In addition, the sample subsetting in the monomorphs step '*–S ID\_order\_chrX.txt'*can slow things down in really large samples. As long as you make sure that the samples and their order is consistent between the autosomes and combined X chromosome, you can exclude this step. An example for the monomorphic snps would then be:

*bcftools view -q1.0:major  chr${i}.dose.vcf.gz –O u | bcftools query -f '%CHROM\t%POS\t%REF,%ALT\n' >> STUDY\_PANEL\_ANALYST\_imputed.monomorphic.txt*