Pre-phasing and Imputation Report for NEAM

Reference Panel Choice

Between the easily available data released by Impute2, we have chosen to use the latest Phase I release of 1000 genomes, found [here](https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated_SHAPEIT2_16-06-14.html). The latest version has been shown to have serious problems, missing approximately 2.6M SNPs, and therefore until the reason for this discrepancy is established, we will use the slightly older reference data.

Pre-phasing

NEAM genotypes were split by chromosome to facilitate easy parallelization on the HPCVL cluster.

app/[split\_by\_chr.sh](https://github.com/Chris1221/impute/blob/master/app/split_by_chr.sh)

app/[m\_split\_by\_chr.sh](https://github.com/Chris1221/impute/blob/master/app/m_align_check.sh)

* bash app/m\_split\_by\_chr.sh

After splitting by chromosome, we are required to ensure that the strand alignment of our files matches that of the reference panel.

app/[align\_check.sh](https://github.com/Chris1221/impute/blob/master/app/align_check.sh)

app/[m\_align\_check.sh](https://github.com/Chris1221/impute/blob/master/app/m_align_check.sh)

* bash app/m\_align\_check.sh

This returned errors such as the following:

Mode : Summarizer

Options : -B /scratch/hpc2862/CAMH/jen/NEAM/NEAM\_QC\_complete\_150623\_chr10 -M /scratch/hpc2862/CAMH/jen/ALL.integrated\_phase1\_SHAPEIT\_16-06-14.nomono/genetic\_map\_chr10\_combined\_b37.txt --input-ref /scratch/hpc2862/CAMH/jen/ALL.integrated\_phase1\_SHAPEIT\_16-06-14.nomono/ALL.chr10.integrated\_phase1\_v3.20101123.snps\_indels\_svs.genotypes.nomono.haplotypes.gz /scratch/hpc2862/CAMH/jen/ALL.integrated\_phase1\_SHAPEIT\_16-06-14.nomono/ALL.chr10.integrated\_phase1\_v3.20101123.snps\_indels\_svs.genotypes.nomono.legend.gz /scratch/hpc2862/CAMH/jen/ALL.integrated\_phase1\_SHAPEIT\_16-06-14.nomono/ALL.integrated\_phase1\_v3.20101123.snps\_indels\_svs.genotypes.sample --output-log /scratch/hpc2862/CAMH/jen/NEAM/neam\_alignment\_chr10

Version : v2.r790

Date : 19/08/2015 11:27:00

MODE -summarise : GENERATING SUMMARY STATISTICS OF THE INPUT DATA

\* Autosome (chr1 ... chr22)

\* Reference panel of haplotypes used

Parameters :

\* Seed : 1439998020

\* Parallelisation: 1 threads

\* Ref allele is NOT aligned on the reference genome

Reading site list in [/scratch/hpc2862/CAMH/jen/NEAM/NEAM\_QC\_complete\_150623\_chr10.bim]

ERROR: Duplicate site pos=431161 ref=C alt=T

Indicating that duplicated sites are interfering with the process. I was pretty sure that the EXM sites are interfering with this, so I want to remove all duplicates, but instead of just removing either, I want to preferentially eliminate the EXM sites which are duplicated. Want to lose as few rs#s as possible.

Please see app/[remove\_dup.R](https://github.com/Chris1221/impute/blob/master/app/remove_dup.R) for the procedure of checking for EXM duplicates, and then doing a second scan to ensure no duplicates are found.

Once the duplicates are removed, we repeat the splitting process as above except excluding these duplicates.

We then do the check again to get a list of positions that are flipped, using the above mentioned code.

Once this runs through successfully, we receive two files per chromosome.

neam\_alignment\_chr5.snp.strand

neam\_alignment\_chr5.snp.strand.exclude

The first lists all problems with the file, and the second gives an easy to use exclusion list. However, we are only interested in the strand flip errors (Case 1 error on the Shapeit website). We will use the other error cases later to produce a new exclusion list as well.

To obtain the sites which require flipping, we isolate the strand flip issues and pipe to a new file.

cat neam\_alignment\_chr${CHR}.snp.strand | grep "strand" | awk '{ print $2 }' > flip.chr${CHR}

To obtain the sites which have other problems, we do the same thing and keep these in a similar file.