**Email from UCSF**

I started with the EVE datafreeze in PLINK format.  The dataset was lifted over from hg18 to hg19. SNPs that didn’t map were removed.  The dataset was checked for snps mapping to the same position, samples or snps with missing% > 10%.  SNPs with MAF <1% were removed for these runs.  Alleles were checked against the 1KGP reference for strand and removed if not able to match.

Data was phased using SHAPEIT and imputed using IMPUTE2 with 1000Genomes Phase 1 integrated variant v3 phased reference (april 2012).

For each dataset, there is a corresponding shapeit ".sample" file in the directory "sampleOrder\_EVE" which lists the sampleIDs in the order they are found in the impute2 output files.  Please note, the first 2 rows are 0 and do not correspond to a sample.

Chromosomes are separated into 5mb chunks, and numbered where for each chunk n:

Endpoint of that 5mb region = (5mb x n).

Start = (endpoint - 5mb + 1).

For example, chr1.4.chs-hw-1kgv3.phased.impute2.gz represents the 4th segment imputed on chr1, with starting point bp 15,000,001 and ending point 20,000,000.

Some regions around the centromere and ends of certain chrs may cover more or less than 5mb, in order to respect centromere boundaries and prevent overly small regions at the ends.  Some values of n will not have a corresponding file if there were no snps in that region to be imputed, so simply looping from 1 to end n for each chr may pass over some purposely missing files.

Files provided are directly from IMPUTE2 output.  You may want to filter on imputation quality using the INFO measure found in the .impute2\_info files that were also provided.

**Files used by Patrick Muchmore to filter**

chs-hw.sample

chs-nhw.sample

USC\_HW\_genotypedSNPs\_hg19\_phased\_chr1.sample

IMPUTE\_output\_CHS-HW-1kg-p1v3/chr$1.\*.gz