Commands I Used to clean the data :

Unzip filed : gunzip -d <file name>

Checking parts of the file : (Sanity check) cat chr1.vcf | head -100 > example.vcf

Replace the file first column : perl -pe 's/^chr25/chrY/' chr25.vcf > chrY.vcf

Bcftools was used to get the total variant count :

bcftools index some.vcf.gz # create the tbi

bcftools index --nrecords some.vcf.gz # get total variant count

Chromosome 21:

Before Impuation : Total variant count : 87,142

After Imputation : Total variant count : 7,75,275

1.Extracted 100,00 variants from my vcf file

2. Calculate allele frequency : include sites that have only 2 alleles

3. Calculate mean depth

for chr in `bcftools view -h MadCaP.vcf.gz | perl -ne 'if (/^##contig=<ID=([^,]+)/) { if (length($1)<=2) {print "$1\n"} }'`; do bcftools view -Oz -r $chr MadCaP.vcf.gz > trial\_$chr.vcf.gz &

done

How to assess imputation quality?

1. Quality score vs MAF bins
2. SNP count vs quality score
3. Mask data from random chromosomes ->impute -> average genotype errors
4. Hellinger score
5. Concordance score

For chr 21 :

Before imputation number of snps = 16,218

Snps after imputation ( computed using bcftools) = 775275

Snps after filtering with R2 > 0.8 and MAF > 0.5 = 124700

bcftools view -i 'R2>.8 & MAF>.01' -Oz chr21.dose.vcf.gz > chr21\_0.01.filtered.vcf.gz

sed -n '/#/!p' chr21.filtered.vcf | cut -f 8 chr21.filtered.vcf > refined\_0.05.txt