Phasing and imputation

Perform imputation of a small region for multi-sample vcf with BEAGLE using 1000 genomes GRCh38 reference panel. Before imputation of vcf, BEAGLE performs phasing of vcf. Let's proceed.

Input vcf: `/srv/common/imputation/data/samples.vcf.gz`
Ground-truth vcf: `/srv/common/imputation/data/ground_truth.vcf.gz`
Reference panel: `/srv/common/imputation/reference_panel/1kg_subset.vcf.gz`

1. Perform filtration of vcf that will be imputed. Use AF > 1%. How many positions are left? How many positions are filtered out? Why do we do filtration by AF before imputation?
2. Use script `/srv/common/imputation/perform_imputation.sh` to impute vcf. Figure out what values should be passed to the parameters.
3. How long did imputation take? What parameters of BEAGLE can be adjusted to speed up the process of imputation?
4. How many positions are in the reference panel? Does the size (number of positions, samples) of input vcf and reference panel matter for the speed of imputation process?
5. How many positions in the imputed vcf? What do you notice?
6. Calculate **genotype** concordance between imputed and ground truth vcf.