

=====

README

=====

Scripts:

1) Phasing:

[S15_1_Phasing.py](#)

2) Imputation with GI panel:

[S15_2_Imputation.py](#)

3) Sensitivity Specificity assessment against UKB-WGS data

This analysis evaluates the accuracy of genotype imputation by comparing imputed genotypes against high-confidence WGS-derived genotypes from UK Biobank. The analysis is restricted to variants that are present in both the imputed dataset and the WGS data. The workflow proceeds chromosome-wise and aggregates results across all samples and variants to generate micro-averaged performance metrics.

Step1:

This step compares imputed genotypes against WGS genotypes, processing samples in 62 chunks (~122–125 samples at a time). It computes per-variant confusion matrix components for all overlapping variants.:

- TP (True Positives)
 - TP_HomR, TP_Het, TP_HomA
- FP (False Positives)
 - FP_HomR, FP_Het, FP_HomA
- TN (True Negatives)
 - TN_HomR, TN_Het, TN_HomA
- FN (False Negatives)
 - FN_HomR, FN_Het, FN_HomA

Script: [S15_3_0_Sensitivity_Specificity_Raw.py](#)

Step2:

These raw per-variant TP, FP, TN, FN counts for each chunk and chromosome are then aggregated by summing TP, FP, TN, FN component counts across all chunks ensuring coverage of all 7628 samples. This produces consolidated per-variant confusion counts per chromosome for all overlapping variants.

Script: [S15_3_1_Sensitivity_Specificity_Sum_Chunks.py](#)

Step3:

This step computes per-variant Sensitivity, Specificity and Precision using micro-averaging for overlapping variants, in all chromosomes. Final sensitivity, specificity, and precision metrics for all variants are calculated in a similar fashion, by aggregating confusion counts. ensuring that performance reflects imputation accuracy across all genotype classes rather than collapsing them into a binary framework.

Script: [S15_3_2_Sensitivity_Specificity_finalMetrics.py](#)