Proceedings of the LXV SIGA Annual Congress
Piacenza, 6/9 September, 2022
ISBN: **978-88-944843-3-5**

Poster Communication Abstract – 4.11

LOW-COVERAGE IMPUTATION FOR HAPLOTYPE IDENTIFICATION FROM LOW-COVERAGE DATA

VINCENZI L.*, MARCOLUNGO L.*, DI VITTORI V.**, DEGLI ESPOSTI C.*, ALFANO M.*, SANTAMARINA C.**, BITOCCHI E.**, BELLUCCI E.**, NANNI L.**, ROSSATO M.*, PAPA R.**, DELLEDONNE M.*

*) Department of Biotechnology, University of Verona
**) Department of Agricultural, Food and Environmental , Marche Politechnic University

Phaseolus vulgaris, low-coverage sequencing, imputation, variant calling, haplotype

Reducing genome sequencing costs is one of the biggest challenges in the genomic field today. Low-coverage sequencing can significantly reduce investments; however, dedicated methods and bioinformatic pipelines should be evaluated to guarantee high confidence results. Imputation is one of these methods, which permits to infer missing information from low-coverage data based on a reference panel.

The purpose of the project was to compare the results of variant calling obtained using low-coverage data and imputation with the results produced using gold-standard genomic data, namely high coverage Whole Genome Sequencing (WGS). The final goal was to identify haplotypes in a large cohort of inbred individuals, exploiting the parental genotypes as reference panel for imputation.

As a case-study, in this work, we analysed the quantitative trait loci (QTLs) for pod-shattering in common bean (P.vulgaris), located onchromosome 05 and comprising a 1 Mb region. This locus was identified byGWAS analysing using an inbred population founded by two varieties ofP.vulgaris, MIDAS, an Andean variety, and G12873, a Mesoamerican one,showing opposite phenotypes in relation to the character of interest,namely non-shattering and shattering, respectively. The parental genomeswere assembled de-novo and their haplotypes were used to build a referencepanel for the imputation in the shattering locus.

We sequenced two introgression lines (ILs) with high coverage WGS (30x) by Illumina. These lines were chosen since they had opposite phenotypes, one shattering and non-shattering the other one. Sequencing data were then aligned on MIDAS and G12873 genomes and a variant calling was performed. To evaluate the imputation efficacy, variant calling has been performed on both ILs both using the full dataset or after downsampling at 5x. Missing sites derived from low-coverage datasets were imputed with Beagle using a reference panel generated on the basis of the parental lines' haplotypes.

The analysis could successfully identify the haplotype blocks in correspondence with the shattering or non-shattering region. In fact, one of the two lines showed the shattering haplotype in the region, inherited by G12873, while the other one showed the non-shattering block, close to the MIDAS haplotype. The comparison of results obtained with the variant calling at 30x and the imputed variants at 5x suggested that imputation can infer correct haplotypes very similar to those observed at high coverage, although some minor differences are found. Based on these results, the approach was applied to the genomic data generated from the inbred population to identify differences in the dimension of the region between the offspring individuals, possibly shortened by crossing-over events. In conclusion, this work demonstrated the potential of exploiting the imputation approach from low-coverage data as an alternative to high-coverage sequencing for the identification of haplotypes.