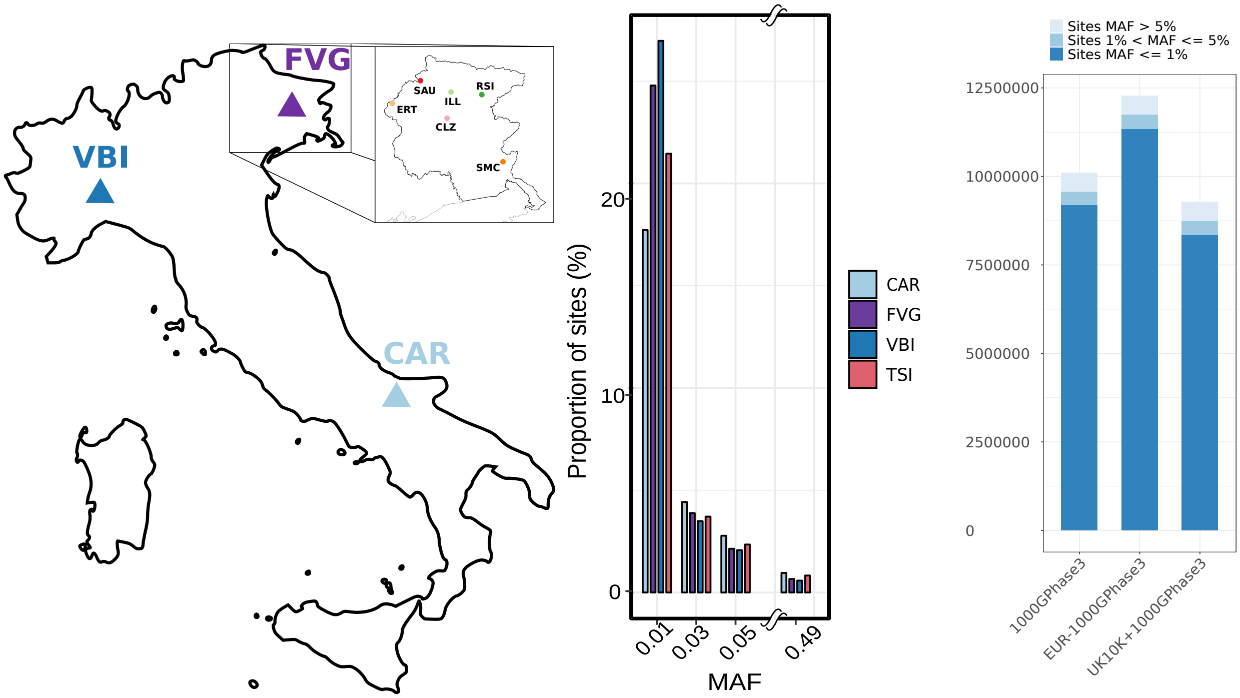
A bird’s-eye view of Italian genomic variation through whole-genome sequencing



(a) (b) (c)

Figure (1) a) Geographic localization of the three study cohorts. b) Minor allele frequency spectrum of the final INGI data set. For comparison, the Minor allele frequency spectrum of the TSI cohort from 1000GPhase 3 data has been added. c) Novel sites identified in the whole INGI dataset, compared to available resources. The majority of the private INGI sites are in the range of the rare variants (singletons sites are included).

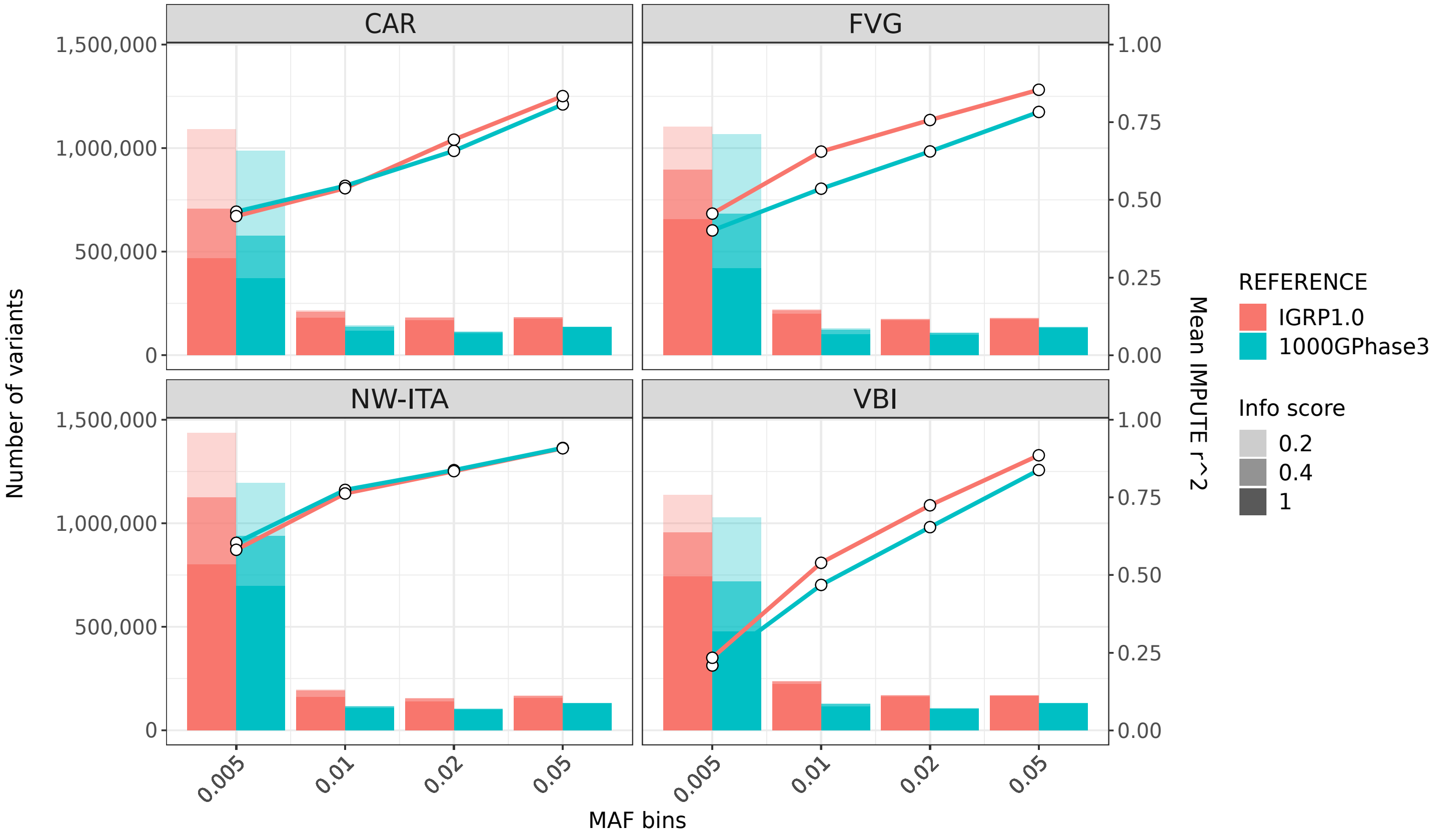
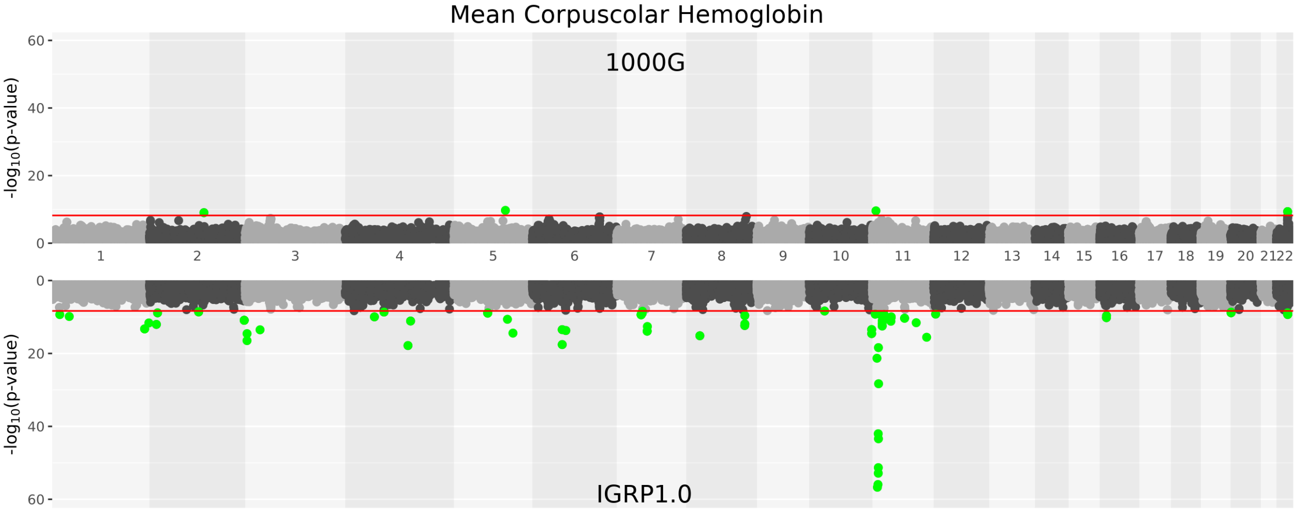
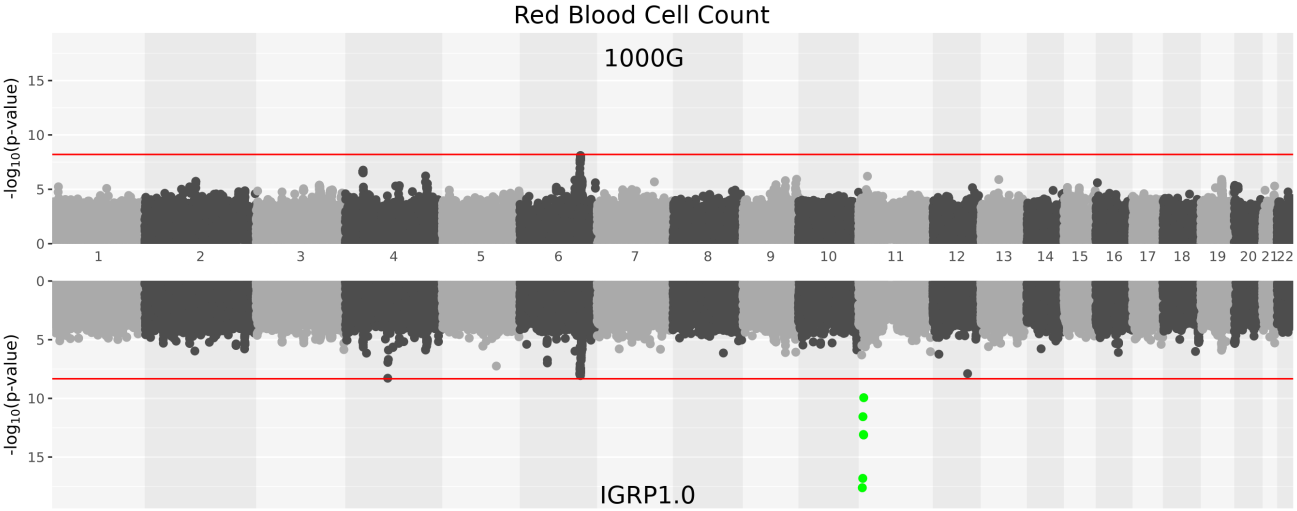


Figure (2) Mean values of r^2 stratified by minor allele frequency (colored lines) and values of info scores stratified by minor allele frequency (bar plot) for Italian cohorts. An outbred cohort from North Italy (NW_ITA) was included for comparison.

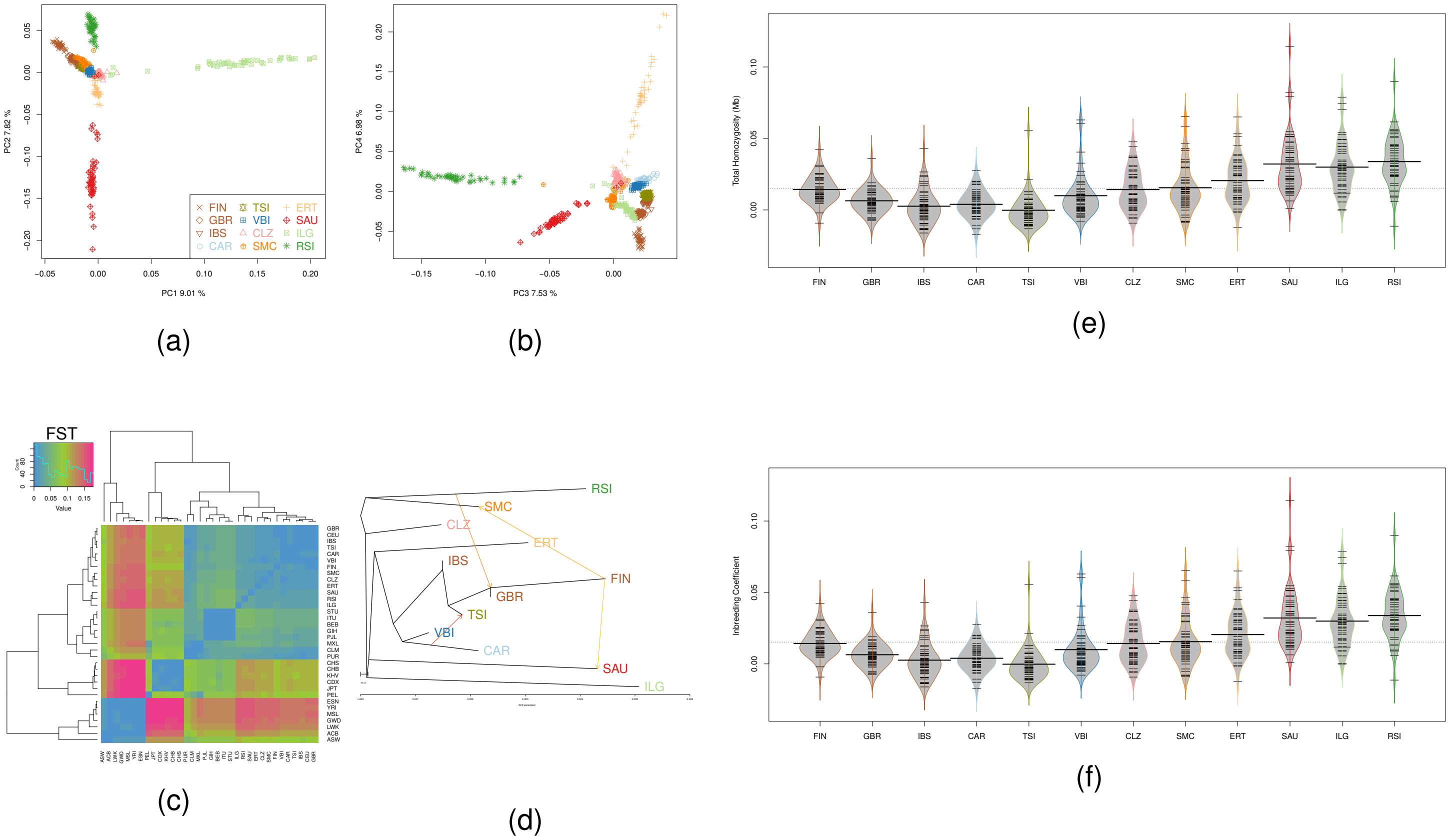


(a)



(b)

Figure (3) a) Manhattan plot of GWAS meta-analysis on MCH phenotype: results in the bottom panel are from newly imputed data while on the top panel we show GWAS results obtained using the 1000G reference panel for imputation. b) Manhattan plot of GWAS meta-analysis on RBC phenotype: results in the bottom panel are from newly imputed data while on the top panel we show GWAS results obtained using the 1000G reference panel for imputation.



(a)

(b)

(e)

(c)

(d)

(f)

Figure (4) Population structure: PCA of Italian samples. a) first two PC and b) the 3rd PC versus the 4th PC. Variance explained by the axis is reported. Each population from FVG has its own axis of variation. The first axis separates ILG from all other Italian populations. The second separates SAU from RSI; c) Pairwise matrix of FST; d) TREEMIX graph analyses with 3 migration edges, link between North Euopean population and some isolates are shown; e) Distribution of total amount of homozygosity in the all cohort minimum ROH length was set at 1 Mb; f) Beanplots of Inbreeding coefficient of 1000G European populations and Italian populations. All FVG population have higher inbreeding coefficient respect to other Italian and European population with the exception of FIN. The plot shows as in the INGI populations the distribution of the inbreeding values is more sparse respect to the reference Italian population TSI from 1000G.

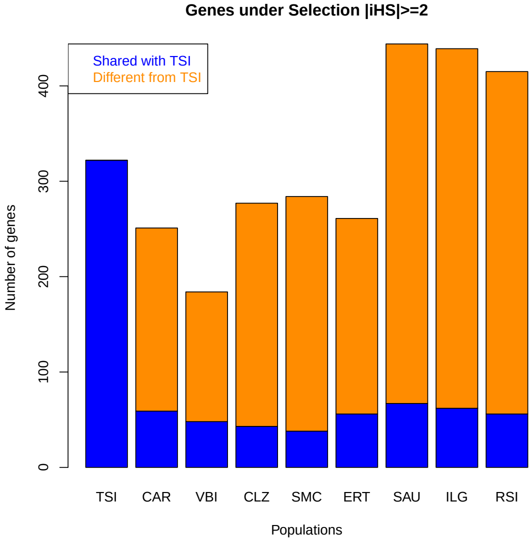


Figure (5) Proportion of genes with signatures of selection in INGI subpopulations respect to TSI. Blue colour represents the fraction of shared genes with TSI, orange colour represent the fraction of private genes respect to TSI.

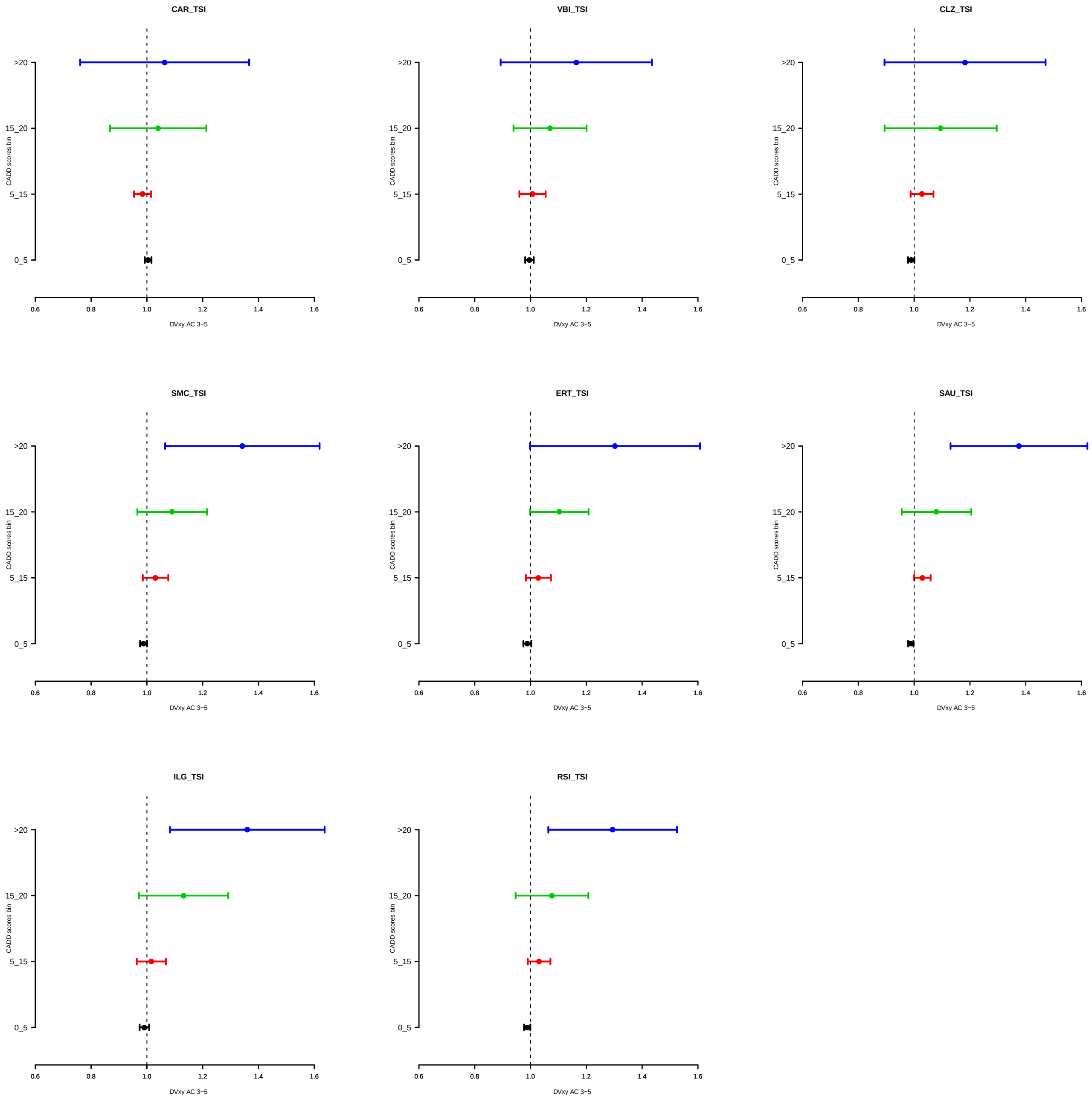


Figure (6) DVxy statistic for each INGI cohort using as reference the TSI population using variants between 3-5 allele count (AC) and binned for CADD score as show in the x axis. Confidence intervals were created using the distribution of all 22 chromosomes and represent 1 standard deviation. A value =1 means no enrichment, a value <1 means depletion in the INGI coohort respect to TSI and a value >1 means enrichment respect to the reference.

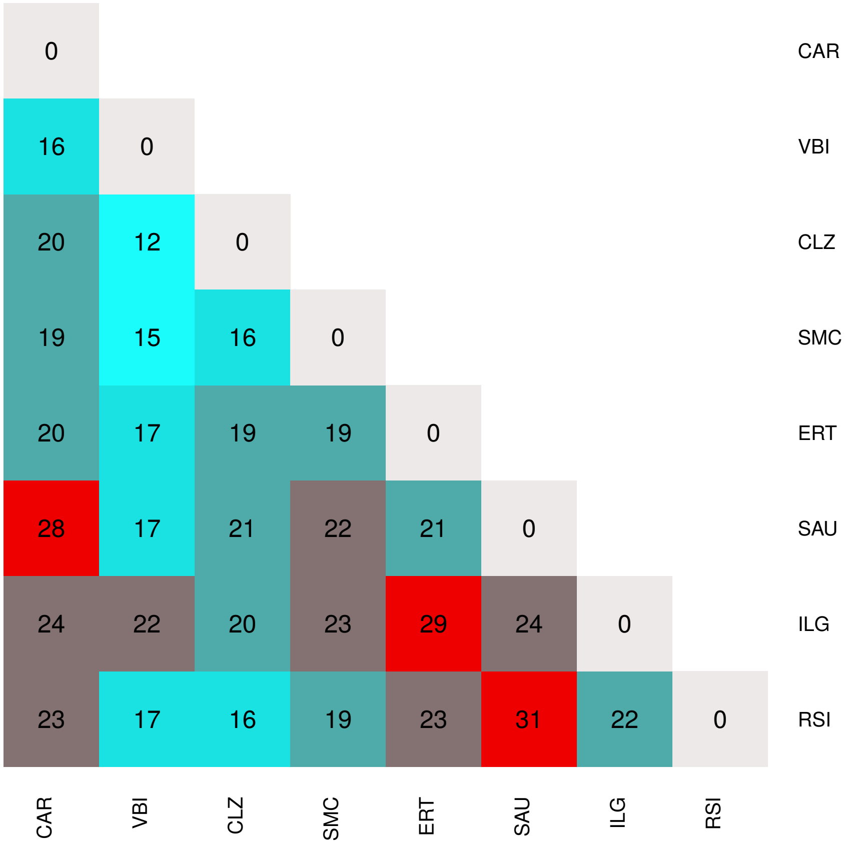


Figure (7) Heatmap of the ratio of DV variants respect to the Italian reference (3-5 AC ,CADD≥20, frequency fold enrichment≥3) that are different between pairs of INGI sub-populations with the DV variants that are shared between pairs.

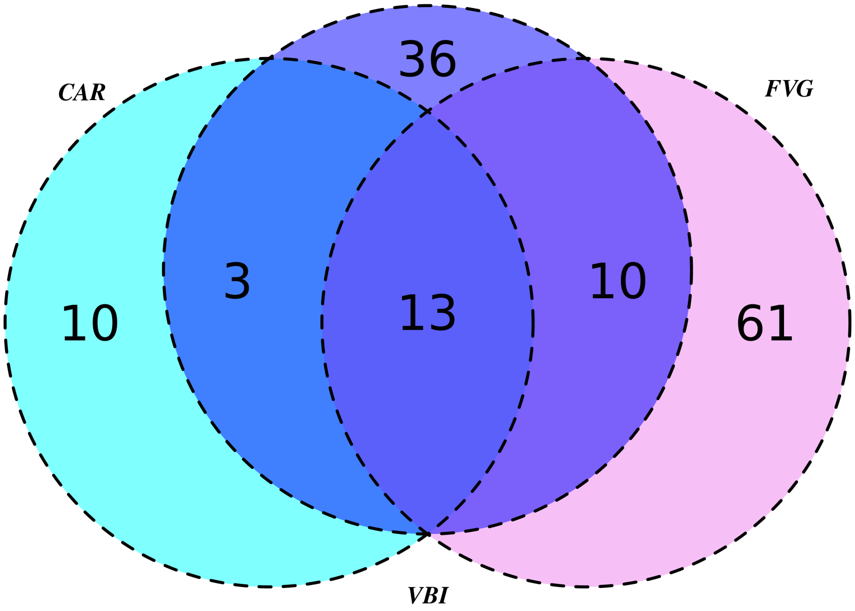


Figure (8) Venn diagram shows how the 133 genes harbouring HKO (with at least 1 homozygous individual) are distributed among INGI cohorts.