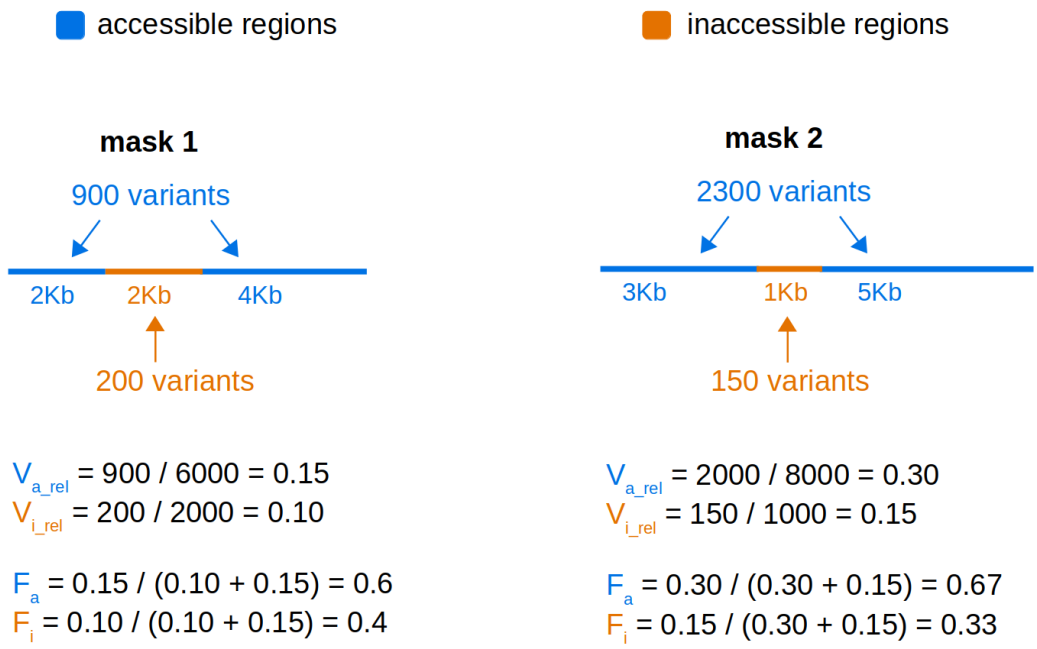
Supplementary Information for the Article

Impact of the inaccessible genome on genotype imputation and genome-wide association studies

Eva König[1], Jonathan Stewart Mitchell[1], Michele Filosi[1], Christian Fuchsberger[1]

[1] Institute for Biomedicine (affiliated to the University of Lübeck), Eurac Research, Via Volta 21, 39100 Bolzano, Italy



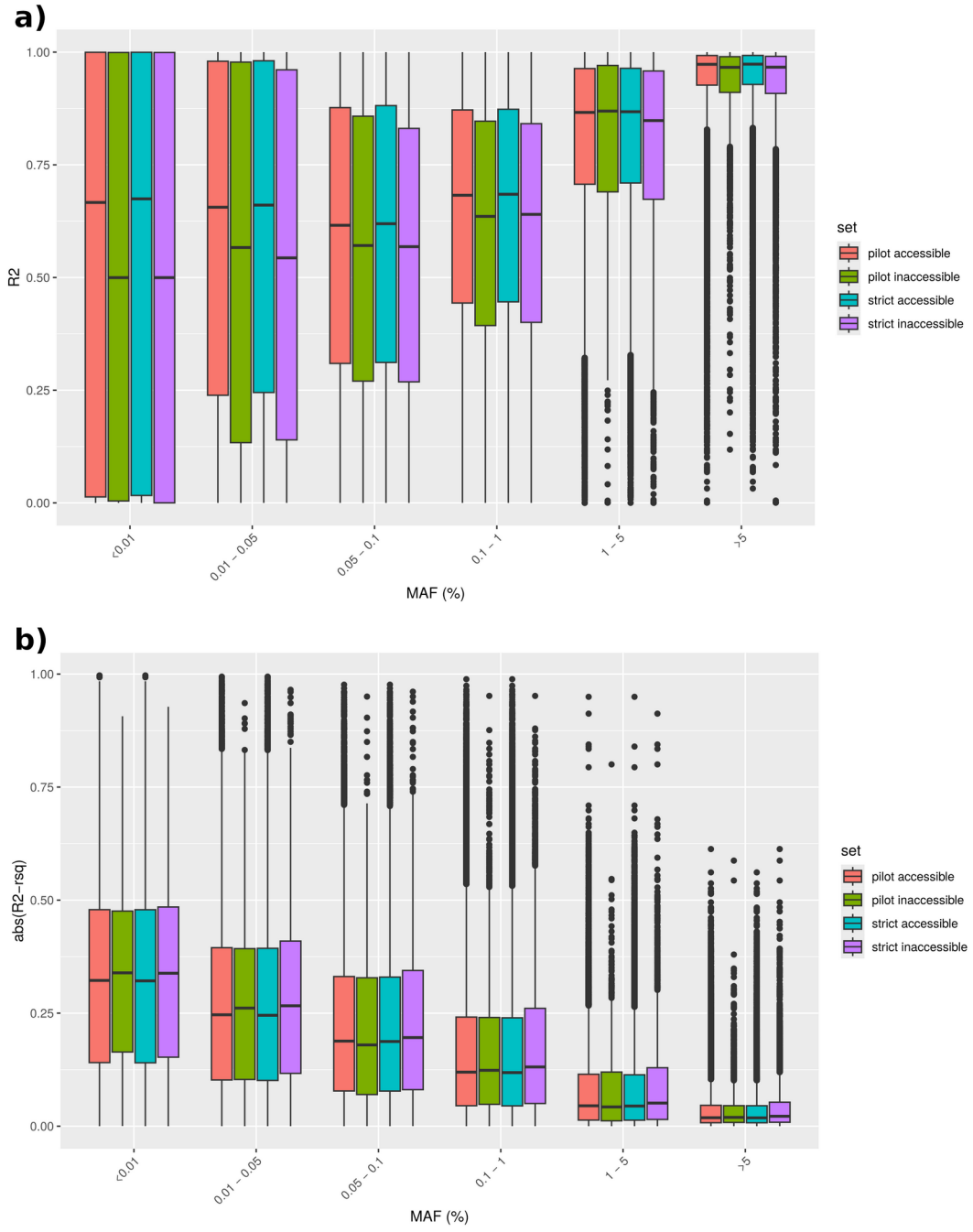
V_{a_rel} = 900 / 6000 = 0.15
V_{i_rel} = 200 / 2000 = 0.10

F_a = 0.15 / (0.10 + 0.15) = 0.6
F_i = 0.10 / (0.10 + 0.15) = 0.4

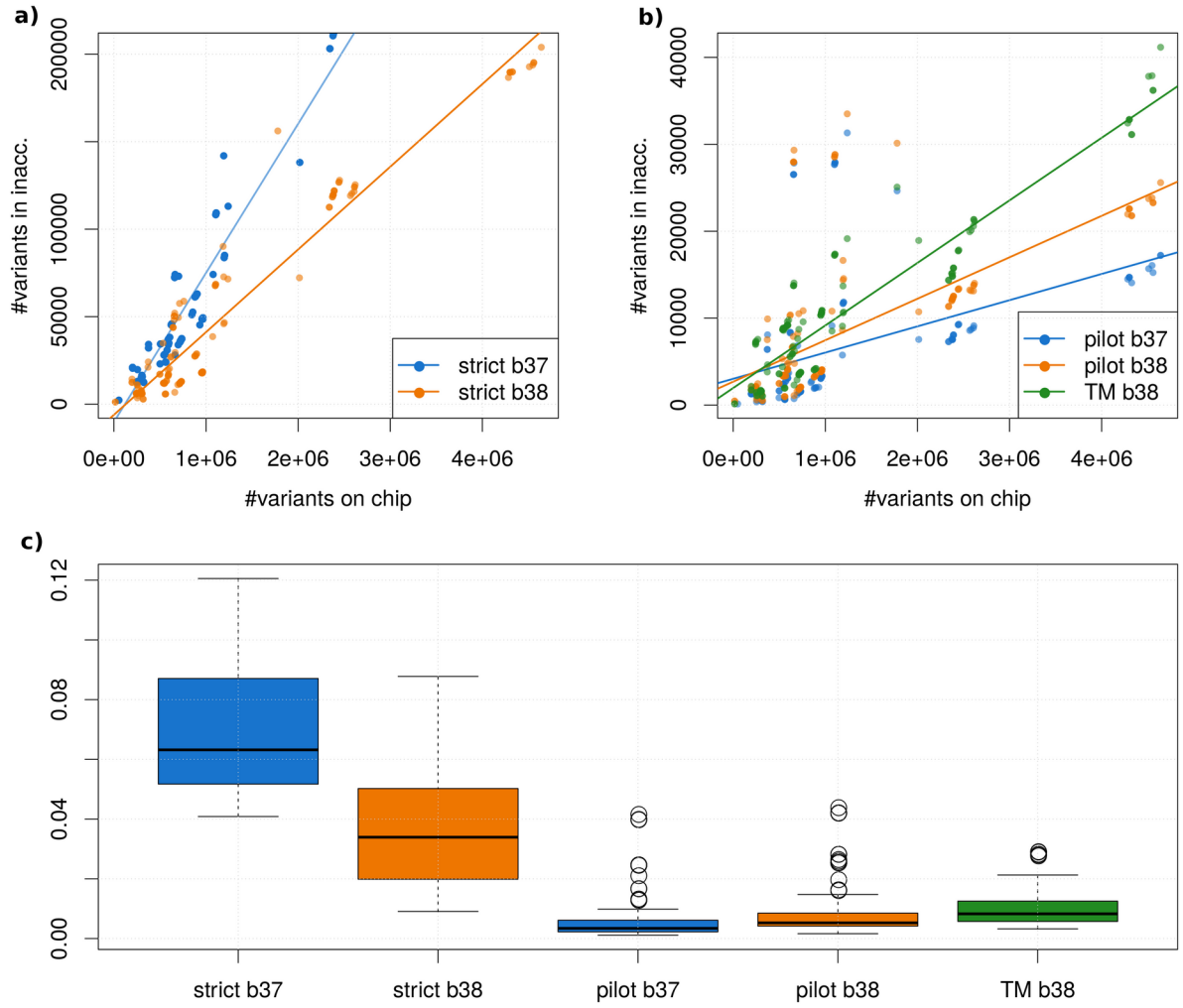
V_{a_rel} = 2000 / 8000 = 0.30
V_{i_rel} = 150 / 1000 = 0.15

F_a = 0.30 / (0.30 + 0.15) = 0.67
F_i = 0.15 / (0.30 + 0.15) = 0.33

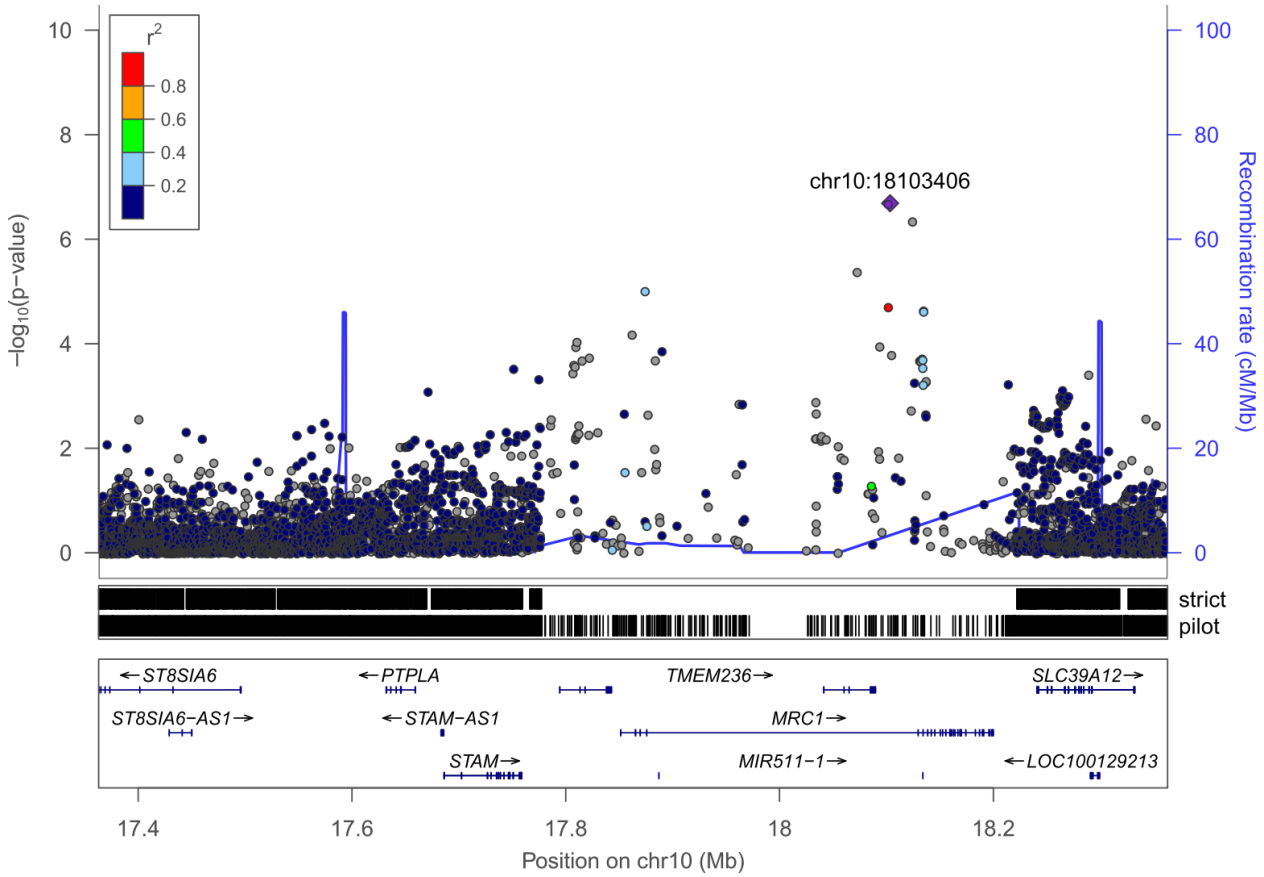
Supplementary Figure 1. Schematic representation of the computation of the relative proportions of variants in accessible and inaccessible regions displayed in Figure 1 using an example.



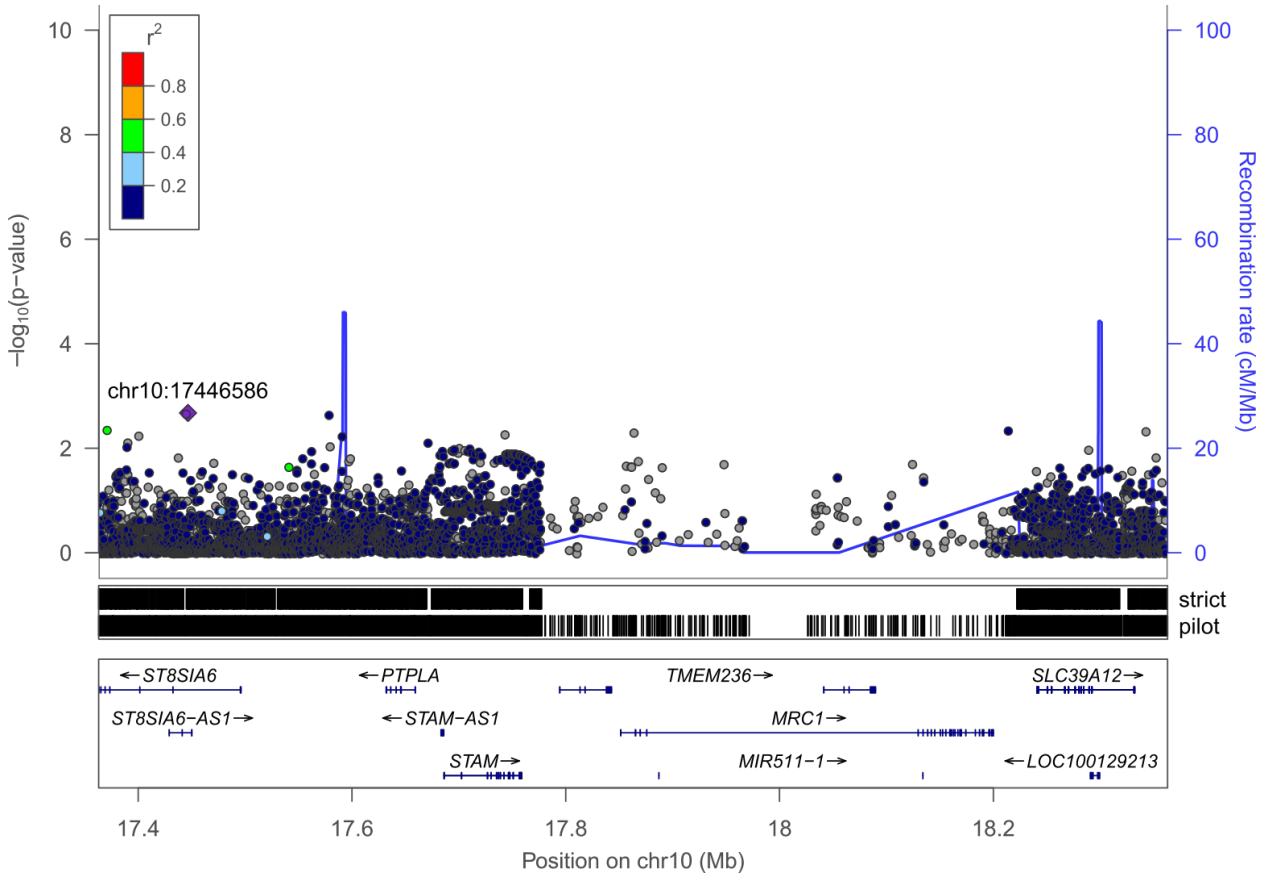
Supplementary Figure 2. Comparison of imputation quality in CHRIS HRC imputed data between accessible and inaccessible regions. Variants are stratified by minor allele frequency (MAF) in the HRC imputed dataset. **a)** Squared correlation of imputed dosages and WES hard calls (R2) for all variants that were both imputed and sequenced, but not genotyped. **b)** Absolute difference of R2 and the imputation quality statistic rsq estimated by the imputation software.



Supplementary Figure 3. Scatterplots and boxplot of number and proportion of variants on common genotyping chips. **a-b)** For each of the common genotyping chips, the total number of variants on the chip (x-axis) is plotted against the number of chip variants that are inaccessible according to the five masks. The solid line represents the linear regression model (y~x). **c)** Boxplot of the proportion of genotyping chip variants located in the inaccessible regions of the five masks for the common genotyping chips. Abbreviations: strict = 1000 Genomes phase 3 strict mask, pilot = 1000 Genomes phase 3 pilot mask, TM = TOPMed mask, b37 = GRCh37, b38 = GRCh38, # = number.



Supplementary Figure 4. A regional association plot (locusZoom) of the locus 10p12.33 where the previously genotyped rs2477642 has now been imputed into the 1000 Genomes phase 3 reference panel in GRCh37. The strict and pilot masks show in black regions of the genome defined as callable with different thresholds in the 1000 genomes phase 3 project.



Supplementary Figure 5. A regional association plot (locusZoom) of the genome-wide association results in the 1000Genomes phase 3 imputed data in GRCh37 at locus 10p12.33. Conditioning on rs2477642 shows that no additional signals in the region are associated with AST. The linkage disequilibrium between rs2477462 and all other variants is displayed as r^2 values calculated from the 1000 Genome Europeans.

Base Class	Pilot definition	Strict definition	% Bases Pilot GRCh37	% Bases Strict GRCh37	% Bases Pilot GRCh38	% Bases Strict GRCh38
N	base was N in reference genome	base was N in reference genome	6.8	6.8	5.3	5.3
L	depth of coverage was lower than 0.5 times the average	depth of coverage was lower than 0.5 times the average	1.1	1.1	1.4	1.4
H	depth of coverage was higher than 2 times the average	depth of coverage was higher than 1.5 times the average	0.2	0.5	0.6	1.0
Z	>20% of reads had mapping quality of zero	>0.1% of reads had mapping quality of zero	2.4	16.8	3.7	18.1
Q	average mapping quality was too low	average mapping quality < 56	0.0	3.1	0.0	0.0
P	Base passed all filters	Base passed all filters	89.4	71.7	89.0	74.1
0	An overlapping base was never observed in aligned reads	An overlapping base was never observed in aligned reads	0.0	0.0	0.0	0.0

Supplementary Table 1: Definitions of the seven base classes for the 1000 Genomes inaccessiblity masks and their prevalence in the reference genomes
(http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/README.accessible_genome_mask.20140520, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/README.accessible_genome_mask.20160622)

Inaccessibility mask	Reference genome	% of autosome [a]	% of variants in the respective imputation reference panel [b]	% of ClinVar variants (all/ pathogenic) [b]	% of chip variants [b]	% of EBI GWAS hits [b]	% of gene bodies [a]	% of exome [a]
1000G Pilot	GRCh37	4.4	2.6	2.8 / 1.3	0.5	1.3	3.1	4.8
1000G Pilot	GRCh38	4.6	1.4	3.0 / 1.5	0.6	1.8	2.7	4.4
1000G Strict	GRCh37	21.8	26.6	9.3 / 6.6	7.5	19.5	21.5	14.4
1000G Strict	GRCh38	18.9	20.0	6.7 / 4.4	3.9	14.2	17.8	11.8
TOPMed	GRCh38	3.2	0.9	4.0 / 2.3	0.8	1.2	1.2	3.3

Supplementary Table 2. Characteristics of inaccessible regions that are located outside of centromers and telomers.

a Percent of the autosome that is located in inaccessible regions outside centromers and telomers.

b Percent of the respective variant sets that are located in inaccessible regions outside centromers and telomers.