1 Supplementary Data

1.1 Long-range LD regions

	Chromosome	Start (Mb)	Stop (Mb)
1	2	134.7 (134.5)	137.3 (138)
2	6	27.5 (25.5)	33.1 (33.5)
3	8	6.6 (8)	13.2 (12)

Table S1: Regions found by snp_autoSVD for the POPRES dataset. Numbers in parentheses correspond to regions referenced in [Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PC1	100.0	-0.1	-0.0	0.1	-0.1	0.0	0.0	0.0	-0.0	-0.0
PC2	0.1	100.0	-0.0	0.1	-0.0	-0.0	-0.0	0.2	-0.1	-0.0
PC3	0.0	-0.0	99.9	0.9	0.1	-0.1	-0.3	0.2	0.4	0.1
PC4	-0.1	-0.1	-0.9	99.7	-1.0	0.7	0.6	0.2	0.3	0.9
PC5	0.1	0.0	-0.1	1.1	99.3	1.3	-0.8	1.3	-4.2	-2.4
PC6	-0.0	0.0	0.1	-0.7	-1.0	97.7	-3.5	6.1	7.9	-6.2
PC7	-0.0	-0.1	0.2	-0.3	-1.7	0.3	58.3	73.2	-25.9	9.1
PC8	0.1	-0.1	-0.3	0.4	-0.5	-5.3	-73.5	59.5	15.8	13.2
PC9	0.0	0.1	-0.4	-0.8	5.0	-7.6	27.8	11.0	91.9	9.0
PC10	0.1	0.0	0.0	-0.9	1.6	10.2	3.9	-19.6	-6.3	89.2

Table S2: Correlation between scores of PCA for the POPRES dataset when automatically removing long-range LD regions and when removing them based on a predefined table.

	Chromosome	Start (Mb)	Stop (Mb)
1	2	134.4 (134.5)	138.1 (138)
2	6	23.8 (25.5)	35.8 (33.5)
3	8	6.3 (8)	13.5 (12)
4	3	163.1 (n/a)	164.9 (n/a)
5	14	46.6 (n/a)	47.5 (n/a)

Table S3: Regions found by snp_autoSVD for the celiac dataset. Numbers in parentheses correspond to regions referenced in [Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PC1	100.0	-0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PC2	0.1	100.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0
PC3	0.1	-0.0	99.9	0.2	-0.0	0.1	0.1	0.1	0.0	-0.1
PC4	-0.0	-0.0	-0.3	99.9	-0.1	0.1	-0.1	0.0	0.1	0.1
PC5	0.0	0.0	0.0	0.1	99.7	0.9	-0.3	0.1	-0.8	-0.6
PC6	-0.0	0.0	-0.1	-0.2	-0.8	99.6	0.5	-0.5	-0.2	-0.4
PC7	-0.0	0.0	-0.1	0.0	0.5	-0.4	98.9	3.1	0.7	1.6
PC8	0.0	0.0	-0.2	-0.0	-0.2	0.5	-3.2	98.4	-4.5	-1.5
PC9	-0.0	-0.0	-0.0	0.0	0.6	0.1	-0.7	4.6	96.9	-10.7
PC10	-0.0	-0.0	0.1	-0.1	0.3	0.1	-1.2	1.5	8.6	92.7

Table S4: Correlation between scores of PCA for the Celiac dataset when automatically removing long-range LD regions and when removing them based on a predefined table.

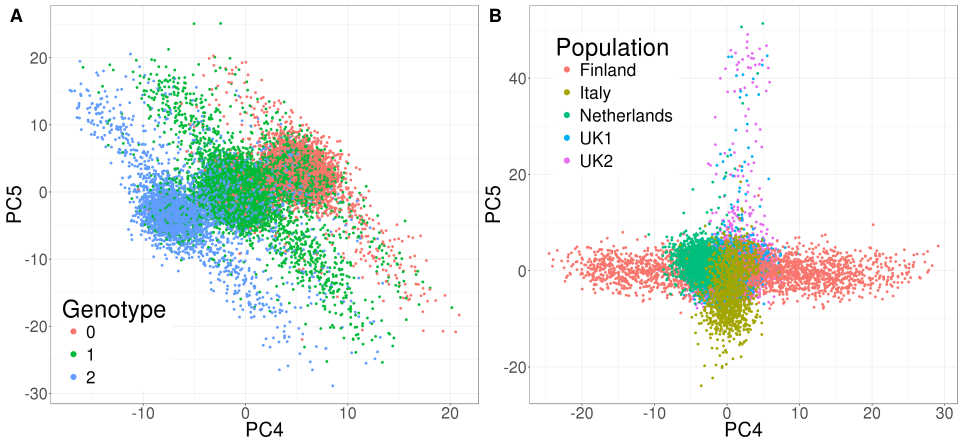

Figure S1: PC4 and PC5 of the celiac disease dataset. Left panel, PC scores obtained without removing any long range LD region (only clumping at $R^2 > 0.2$). Individuals are coloured according to their genotype at the SNP that has the highest loading for PC4. Right panel, PC scores obtained with the automatic detection and removal of long-range LD regions. Individuals are coloured according to their population of origin.

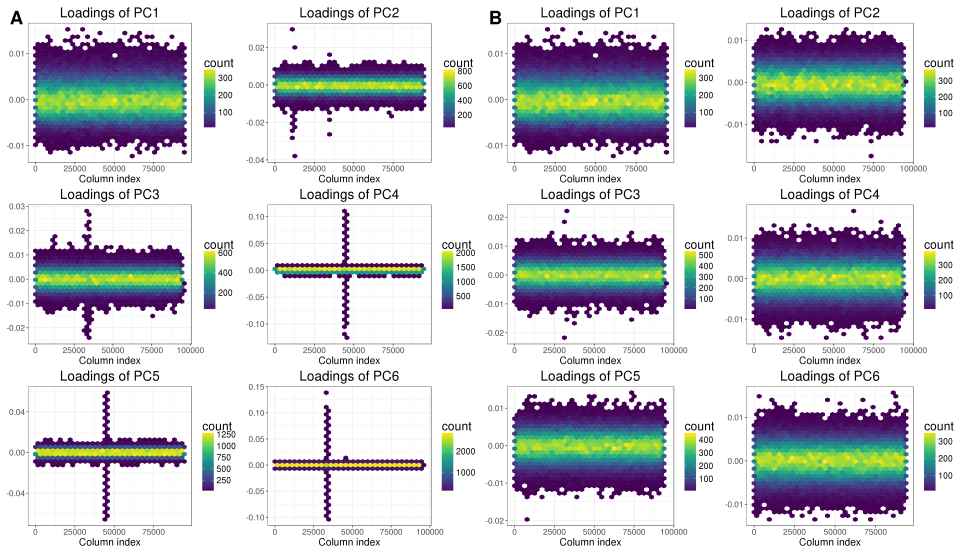

Figure S2: Loadings of first 6 PCs of the celiac disease dataset plotted as hexbins (2-D histogram with hexagonal cells). On the left, without removing any long range LD region (only clumping at $R^2 > 0.2$). On the right, with the automatic detection and removal of long-range LD regions.

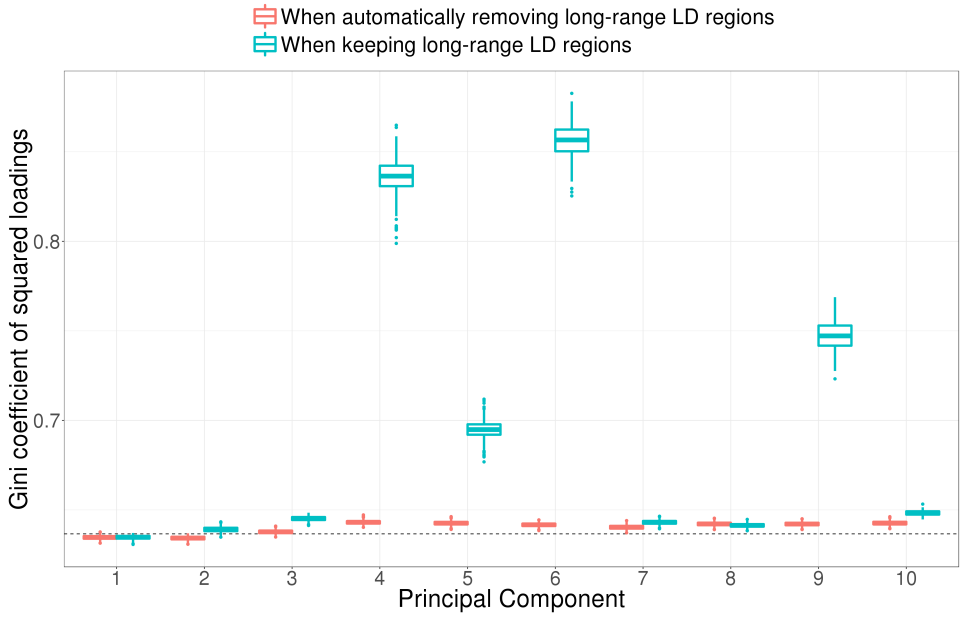

Figure S3: Boxplots of 1000 bootstrapped Gini coefficients (measure of statistical dispersion) of squared loadings without removing any long range LD region (only clumping at $R^2 > 0.2$) and with the automatic detection and removal of long-range LD regions. The dashed line corresponds to the theoretical value for gaussian loadings.

1.2 Imputation

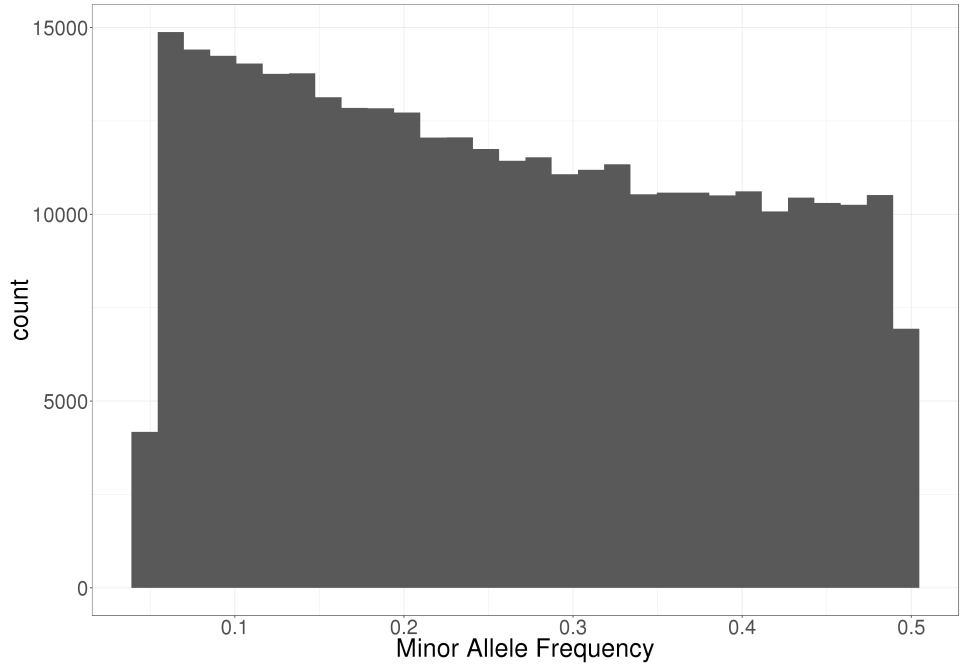

Figure S4: Histogram of the minor allele frequencies of the POPRES dataset used for comparing imputation methods.

References

[Chen and Guestrin(2016)Chen and Guestrin] Chen, T. and Guestrin, C. (2016). XGBoost : Reliable Large-scale Tree Boosting System. *arXiv*, pages 1–6.

[Price *et al.*(2008)Price, Weale, Patterson, Myers, Need, Shianna, Ge, Rotter, Torres, Taylor, Goldst Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D. D., Goldstein, D. B., and Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations.

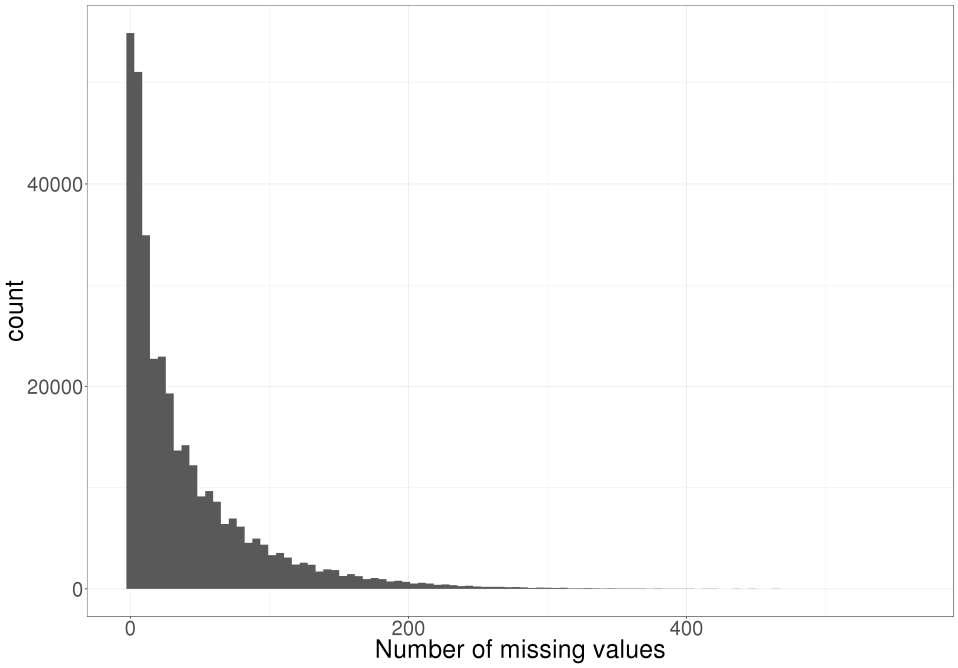

Figure S5: Histogram of the number of missing values by SNP. These numbers were generated using a Beta-binomial distribution.

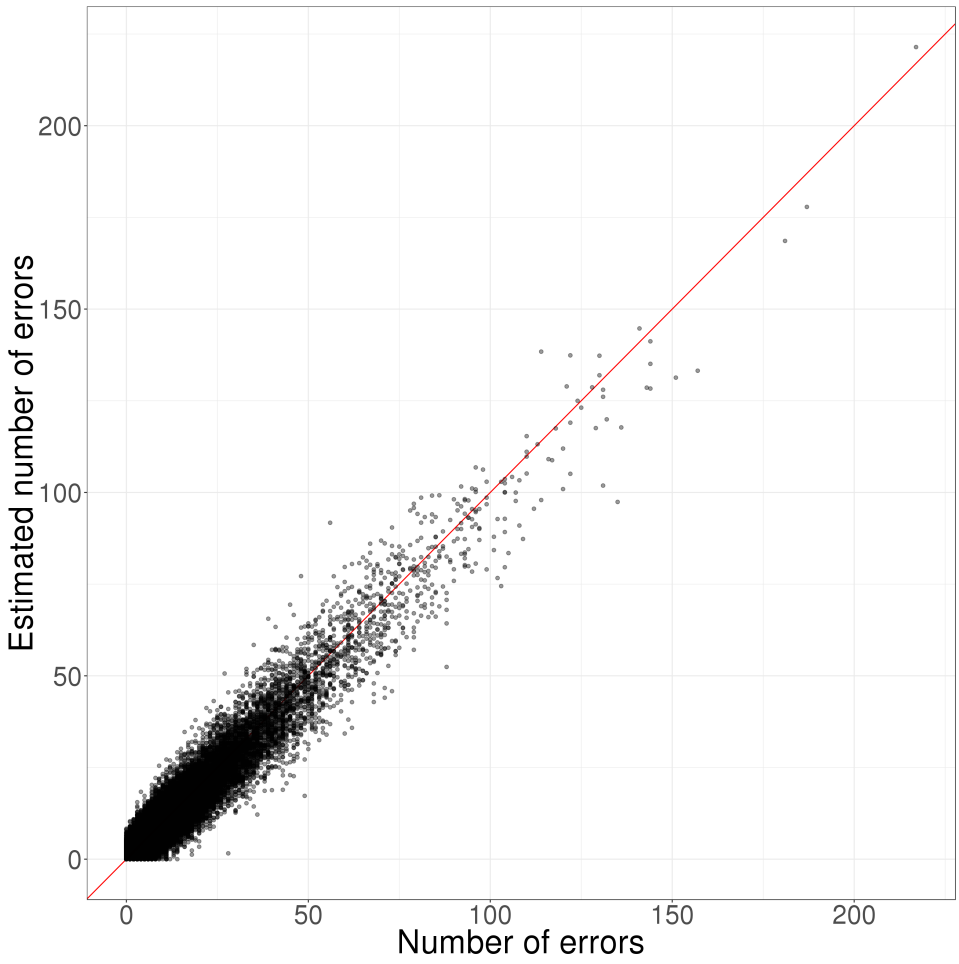

Figure S6: Number of imputation errors vs the estimated number of imputation errors by SNP. For each SNP with missing data, the number of imputation errors corresponds to the number of individuals for which imputation is incorrect. The estimated number of errors is a quantity that is returned when imputing with snp_fastimpute, which is based on XGBoost [Chen and Guestrin(2016)Chen and Guestrin,].