

Supplementary data

Figure S1 | Population outliers were identified by multi-dimensional scaling (MDS) analysis. We used 11,192 SNPs that had minor allele frequencies (MAF) >0.05 , and pruned for SNPs in linkage disequilibrium and shared between ImmunoChip and HapMap3 samples. A global analysis of our north Indian cohort together with CEU, CHB, YRI and GIH samples from HapMap, our samples were laid over the GIH (panels A and B). More local analysis showed that the cases and controls were well matched and there was a difference between the HapMap Indians and our cohort, mostly separated by component 2 (panels C and E). Subtle population substructures were observed (panels D and E), but a similar pattern was observed for both our cohort and the GIH samples, and was equally distributed between cases and controls, indicating that the observed clusters were a genetic mark of these ethnic populations, rather than bias in our cohort sampling.

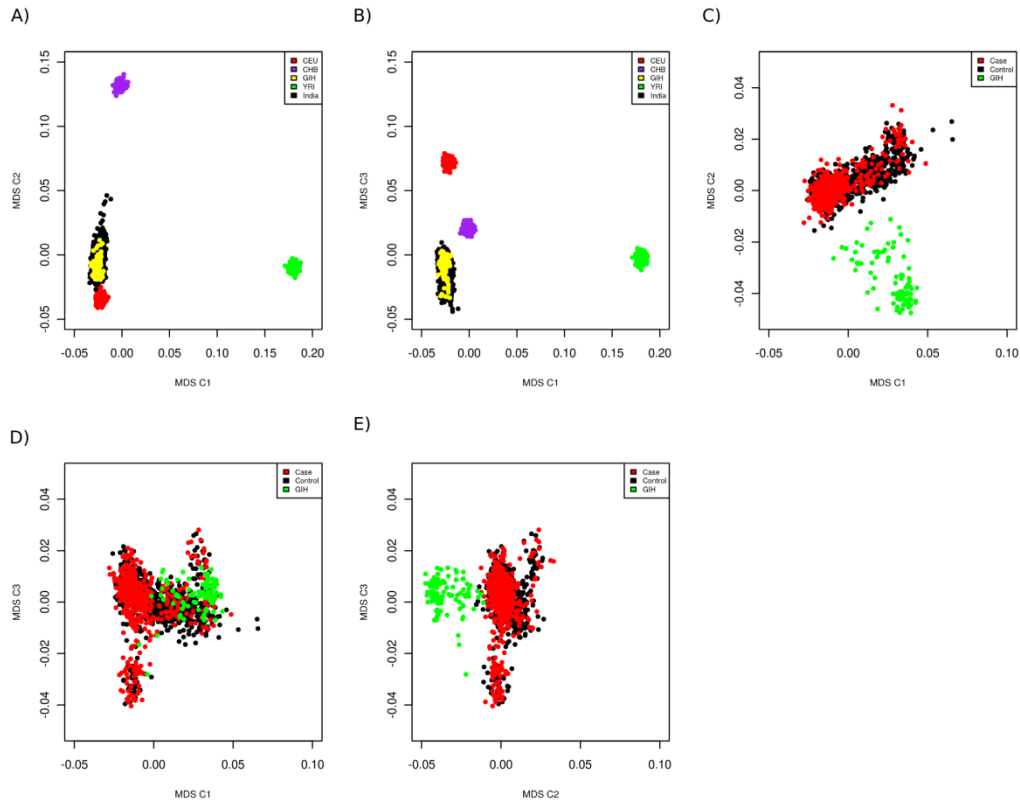


Figure S2 | Power calculations for detecting significant associations under various minor allele frequencies (MAFs) and odds ratios (OR) for a sample size of 497 cases and 736 controls.

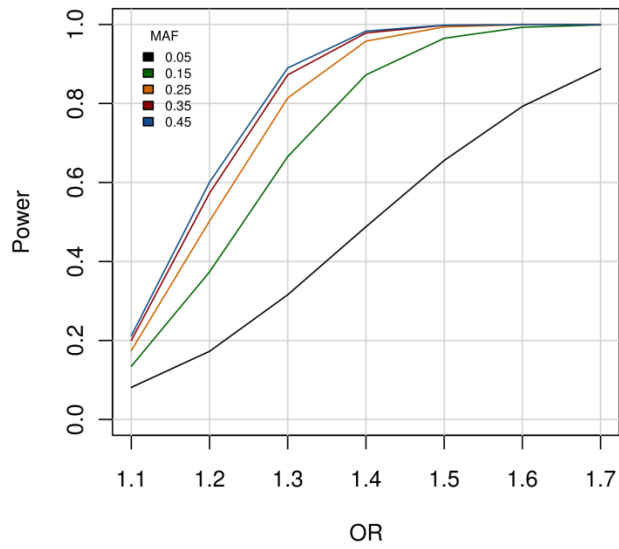


Figure S3 | Minor allele frequency (MAF) correlations between north Indians and the Dutch, and within the European controls (UK, Italian, Polish and Spanish). SNPs tagging loci which were transferred are marked in green (panel A). In the north Indian cohort, all coeliac disease SNPs were polymorphic and common (MAF>0.05), however, the frequency was very different to that in the Dutch cohort (panel A) or in the European samples (panels B-E).

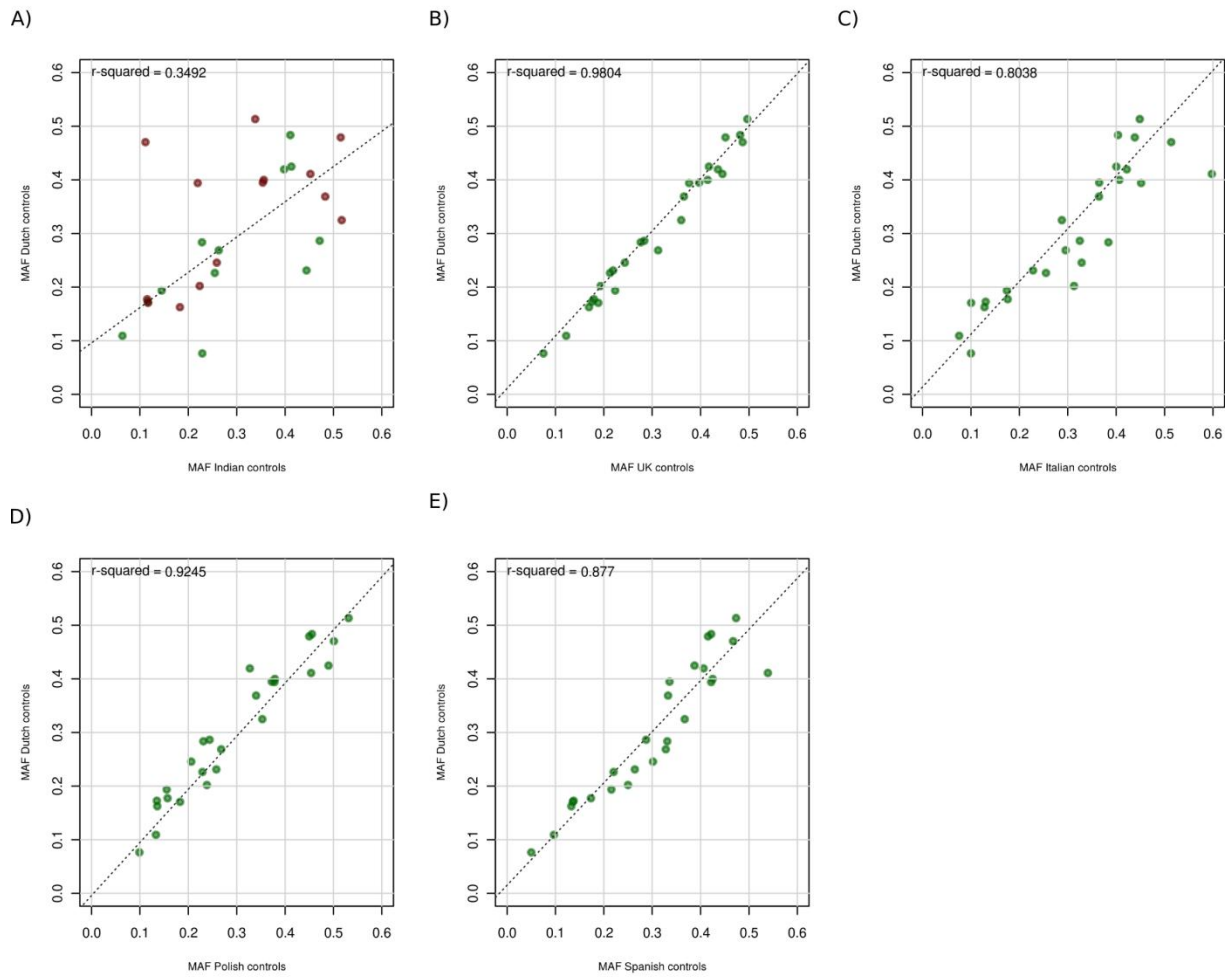


Figure S4 | Comparison of the directionality of associations in north Indians and Dutch. Five SNPs showed opposite directions.

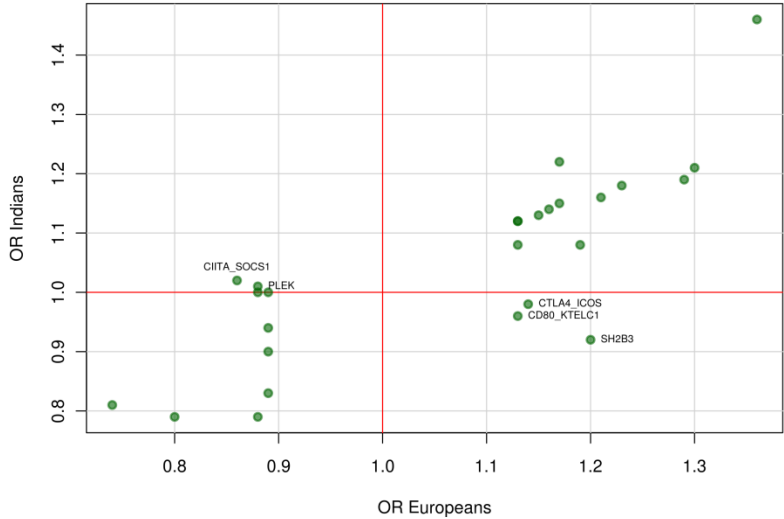


Figure S5 | Distribution of frequencies of variants unique for European populations sequenced in the 1000 Genomes Project but not genotyped on the ImmunoChip in our north Indian samples. The great majority of variants that were not covered by ImmunoChip are of very low frequency (MAF<0.05), and hence likely to be population-specific.

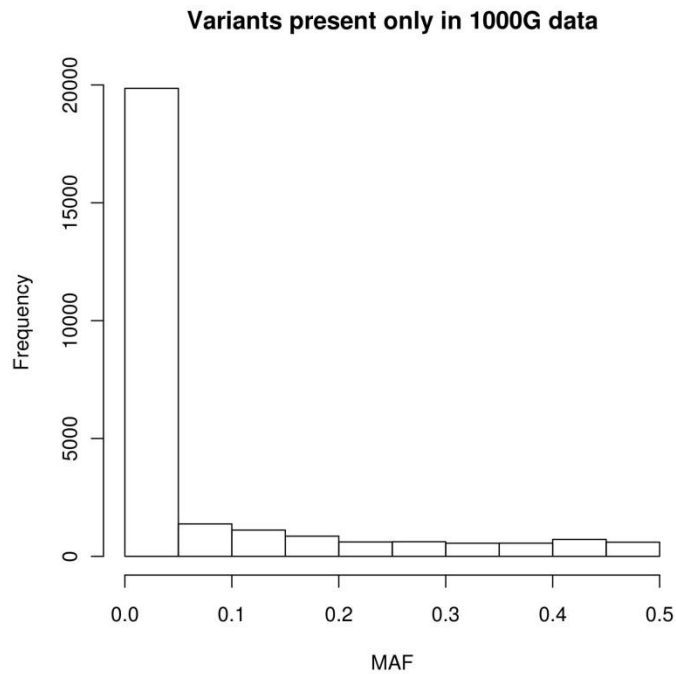


Figure S6 | Linkage disequilibrium (LD) correlation between the European index SNP and top transferable SNP (panel A), as well as between all the transferable SNPs (panel B) in north Indians and Dutch.

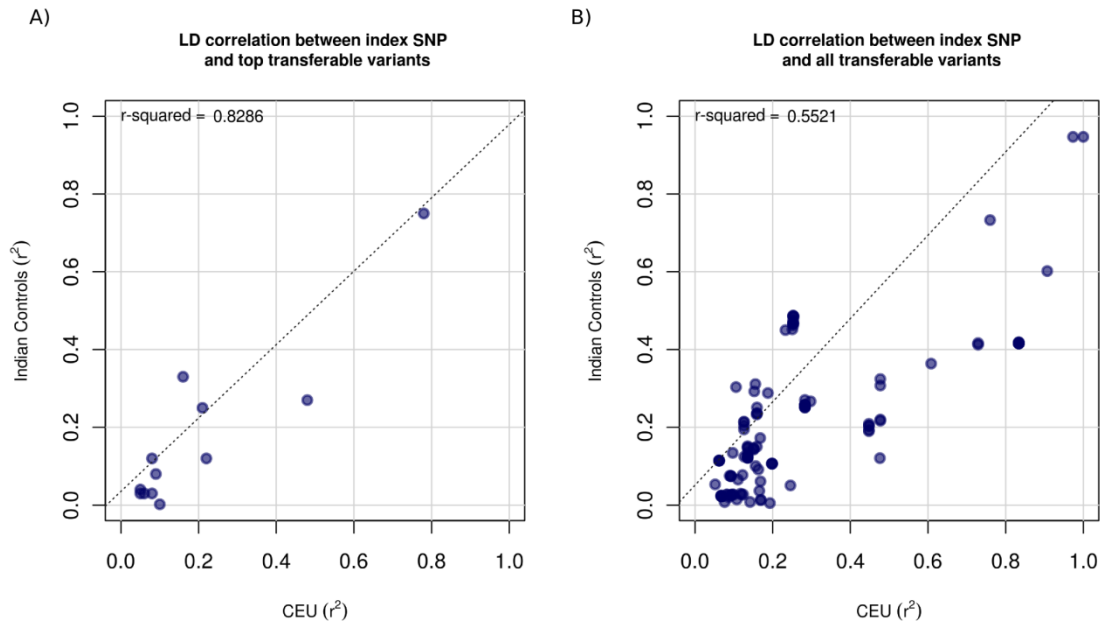


Figure S7 | (A) Comparison of the association signals between north Indians (dark blue) and Dutch (grey). The most significant transferable SNP is depicted in dark green and all the variants in strong linkage disequilibrium with it ($r^2 > 0.8$ based on north Indian controls) are shown in a lighter shade of green. All transferable SNPs ($p < 0.01$) are in yellow and the European index SNP is in red. In some cases colours may overlap. Comparison of the LD structure between Dutch controls (upper LD block) and north Indian controls (lower LD block), measured by D' (B) and r^2 (C).

At five loci (*IL18RAP*, *CCR1/5*, *IL12A*, *IL2/IL21* and *ETS1*), the pattern of association in north Indians overlapped with the Dutch, however for some loci, the association signal could be refined to a smaller region. Another five loci (*PLEK*, *UBE2E3*, *THEMIS/PTPRK*, *TAGAP* and *ICOSLG*) showed shift in the association patterns.

At 2p14 the Dutch signal spreads over the middle LD block and covers the *PLEK* gene, whereas the north Indian signal locates at 69 kb away from the European top SNP, downstream of this gene. The north Indian signal is located in a block of low LD and is poorly correlated with the European index signal.

Although in the 2q12.1 region (*IL18RAP* locus) the European and north Indian signals overlap, the LD is lower in the north Indians and the main LD block covering the association signals is 7 kb smaller than in the Dutch. This LD block covers *IL18R1*, *IL18RAP* and part of *SLC9A4* but clearly excludes *IL1RL2* and *IL1RL1*. The only transferable SNP mapped in the intron of *IL18R1*. Both Dutch and north Indian association signals show very high D' with the European Index marker (Table 1).

At 2q31.3 the north Indian association signal is stronger than that in the Dutch population. In the Dutch, this region is covered by three LD blocks, with the association signal mapping in the second block, an intergenic region downstream of the *UBE2E3* gene. In north Indians the LD is even more broken down and the association signal localizes in two clusters of SNPs: first, in a small block in close proximity to the top European SNP, and second, stronger signal, in a distal block partly corresponding to Dutch third LD region.

At the *CCR1/5* locus, both populations show a similar LD background and overlapping association signals, which narrows down the association signal to a smaller

region of ~250 kb, including the *CCR1*, *CCR2*, *CCR3*, *CCR5*, *CCRL2* and *LTF* genes. The north Indian top transferable SNP mapped in the exon 2 of *CCR5*.

At the *IL12A* locus, association signals localized in the intergenic region between *SCHIP1* and *IL12A* with the top transferable north Indian SNP mapping in a small LD block of 6 kb (161176264 bp – 161182066 bp) near the promoter of *IL12A*.

The coeliac disease region at chromosome 4q27 is characterized by very strong LD, of four genes, including two plausible, immune-function candidate genes, *IL2* and *IL21*. In our north Indian cohort the LD was broken down due to the larger number of low frequency SNPs (MAF < 0.1), resulting in two smaller LD blocks in north Indians, compared with one large block in the Dutch. The association signal was spread along the whole region, although an outstanding cluster of most strongly associated SNPs was located in the small 21 kb LD block (123246379 bp - 123267309 bp) adjacent to the large LD block that covers the *KIAA1109* gene and top European SNP.

At the 6q22.33 locus we observed a cluster of correlated variants in the intronic region of the *THEMIS* gene. However, the Dutch signal at this locus maps 124 kb upstream, clearly pointing towards the 3'UTR of the *PTPRK* gene rather than *THEMIS*.

At the *TNFAIP3* locus only a single SNP was transferable, therefore it was not possible to deduce the association patterns.

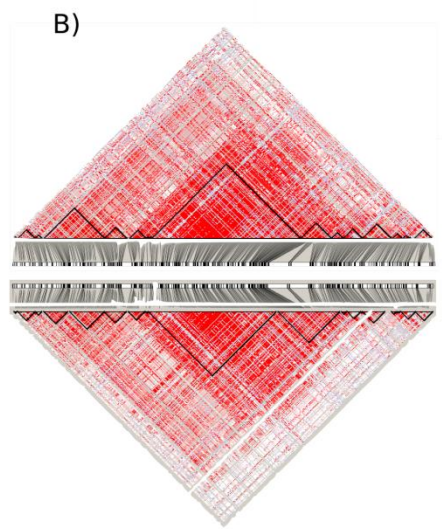
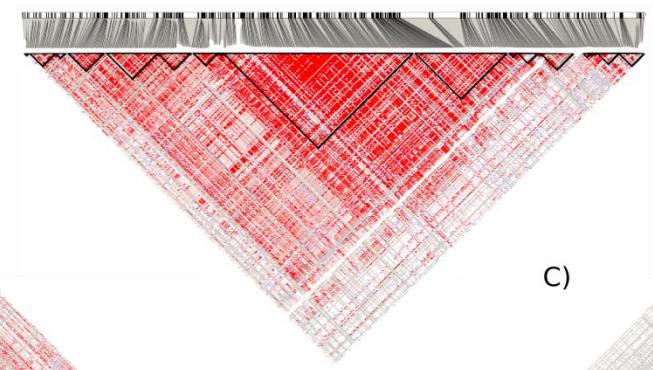
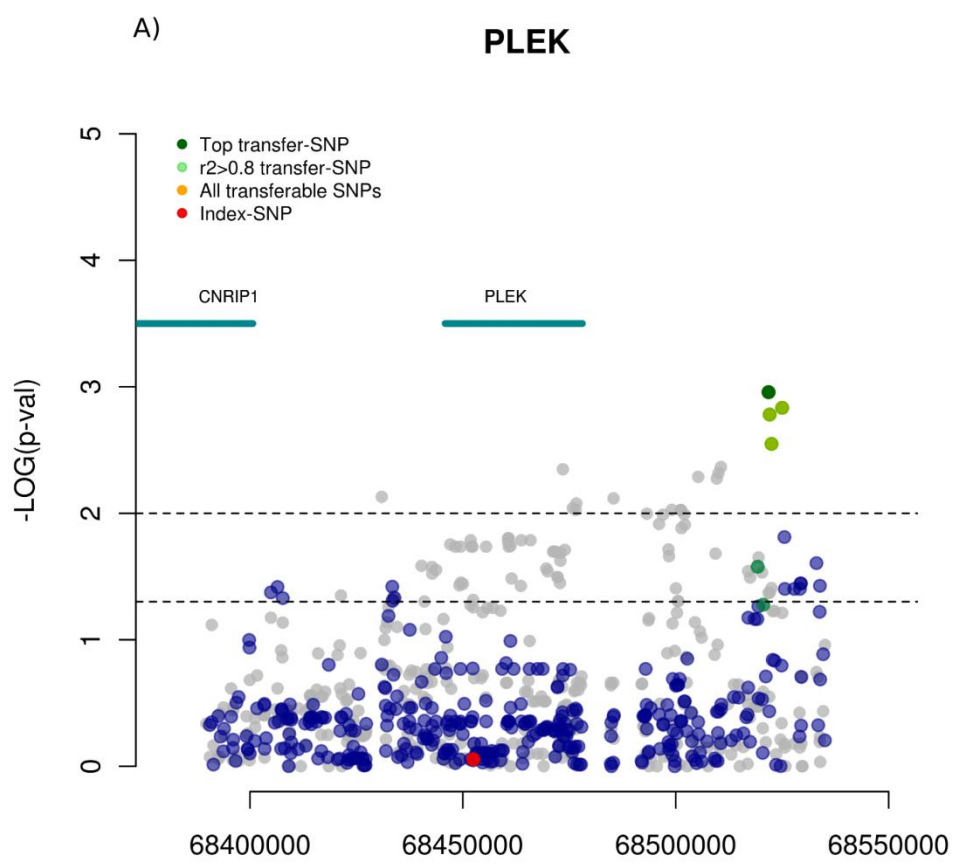
At the 6q25.3 locus, the European top SNP is located in the first exon of the *TAGAP* gene, whereas the north Indian signal is located in the promoter region, which partly also overlaps with the Dutch signal.

Similarly to *TNFAIP3* at the *ZMIZ1* locus only a single SNP was transferable. The top transferable north Indian SNP is localized in proximity to the European index SNP in the intronic part of the *ZMIZ1* gene.

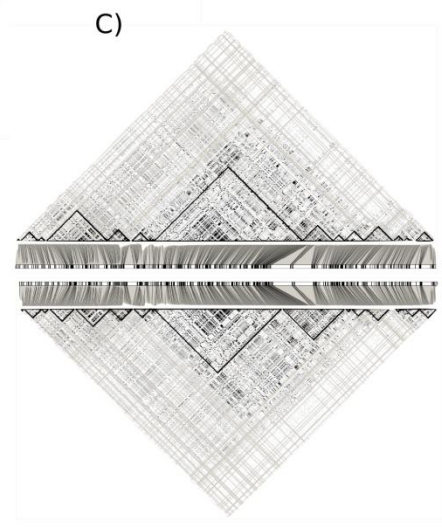
At the *ETS1* locus, the LD architecture is very similar between the Dutch and north Indians and the cross-population signals overlap but the north Indian signal is much more tightly clustered around the top European SNP (21 kb block, from 127882690 bp to 127904148 bp), whereas the Dutch signal is widely spread (a 103 kb region from 127882690 bp to 127985506 bp).

At the *ICOSLG* locus, the signal overlapped with the European one to some degree, but we noticed a tight cluster of SNPs further upstream of the *ICOSLG* gene,

suggesting that the causal variant could be captured within the 17 kb LD block (from 44435321 bp to 44452009 bp).

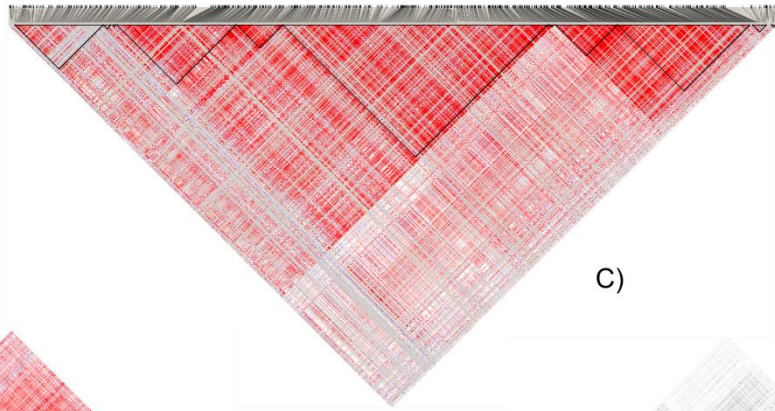
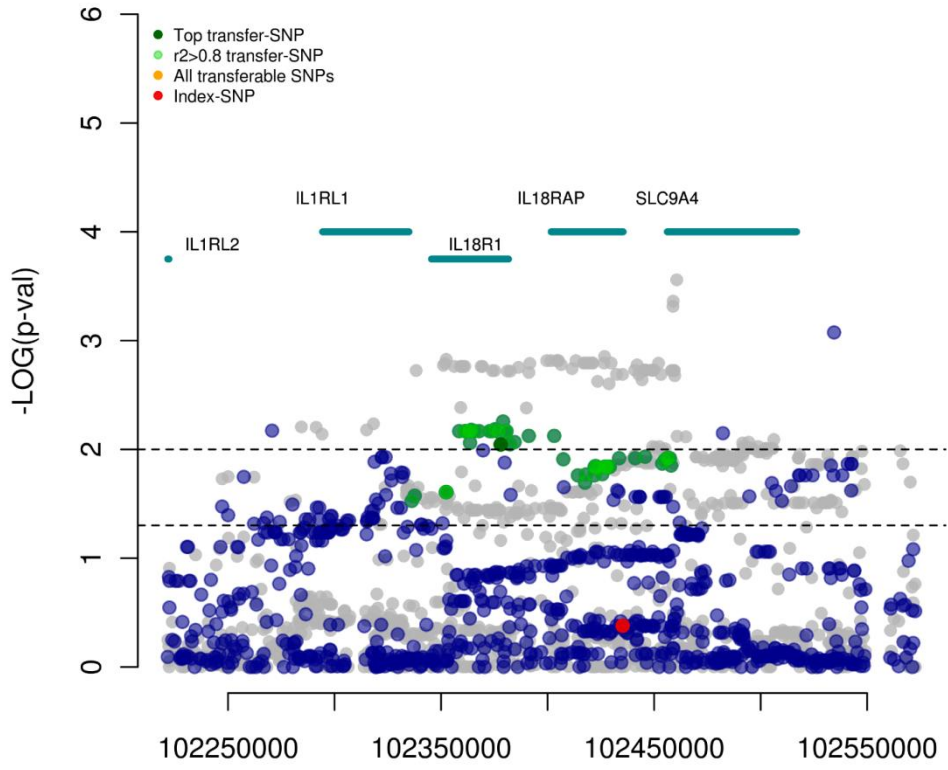


Netherlands
India

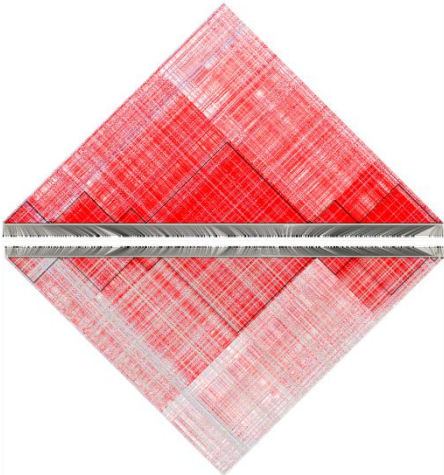


A)

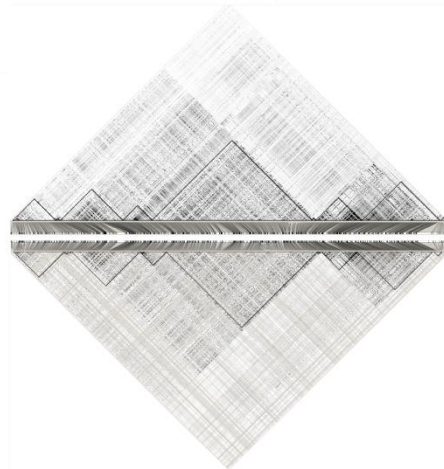
IL18RAP



B)



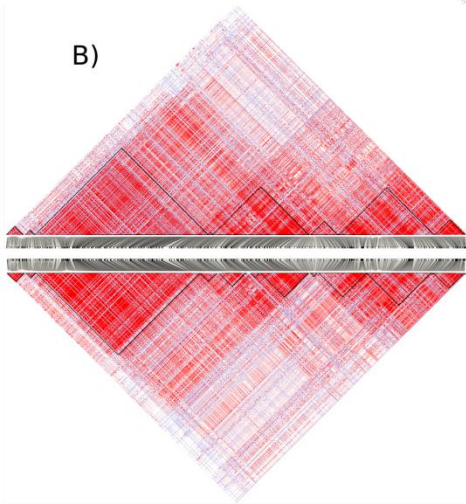
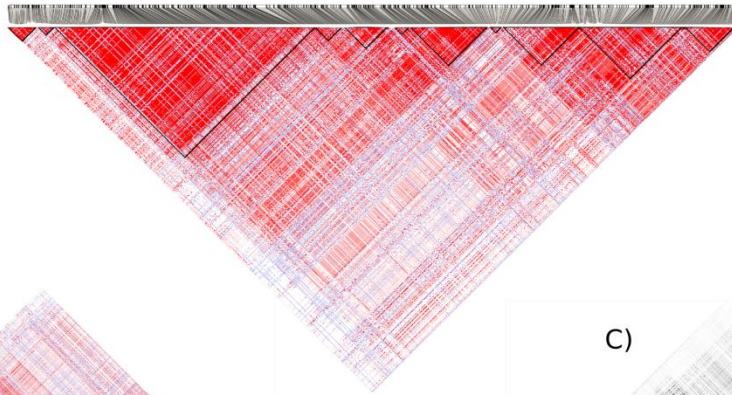
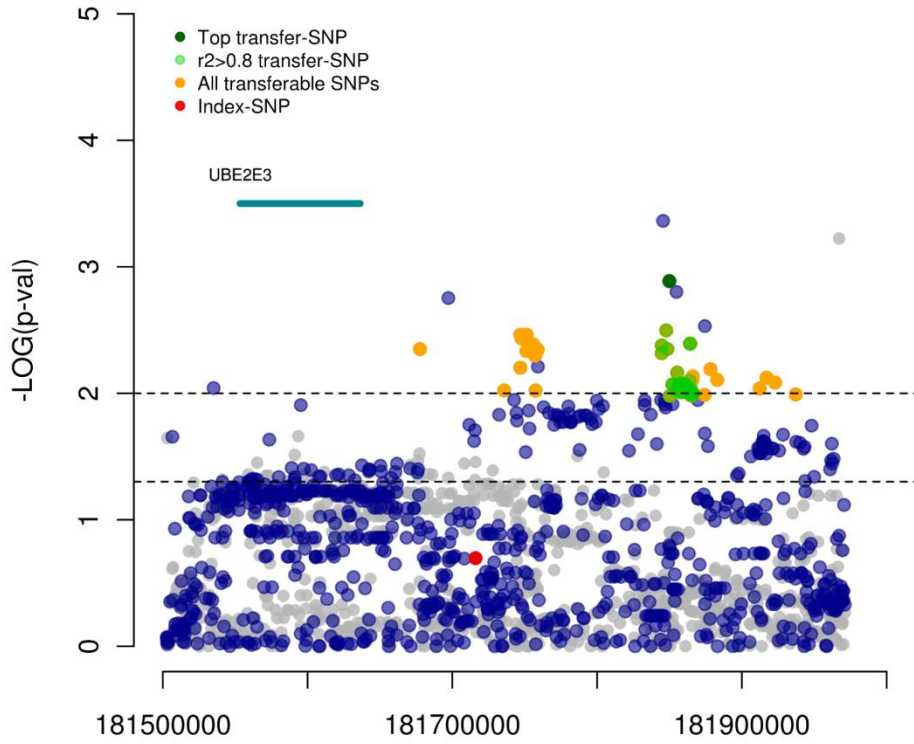
C)



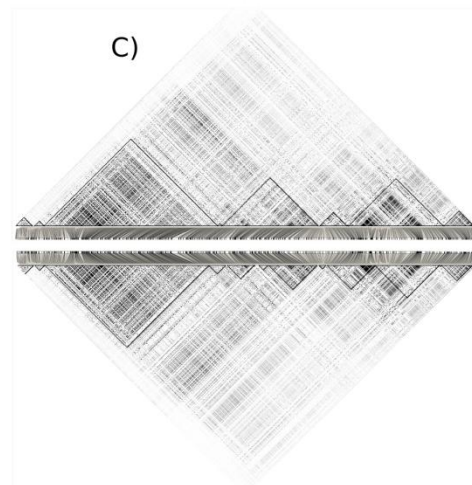
Netherlands
India

A)

ITGA4/UBE2E3

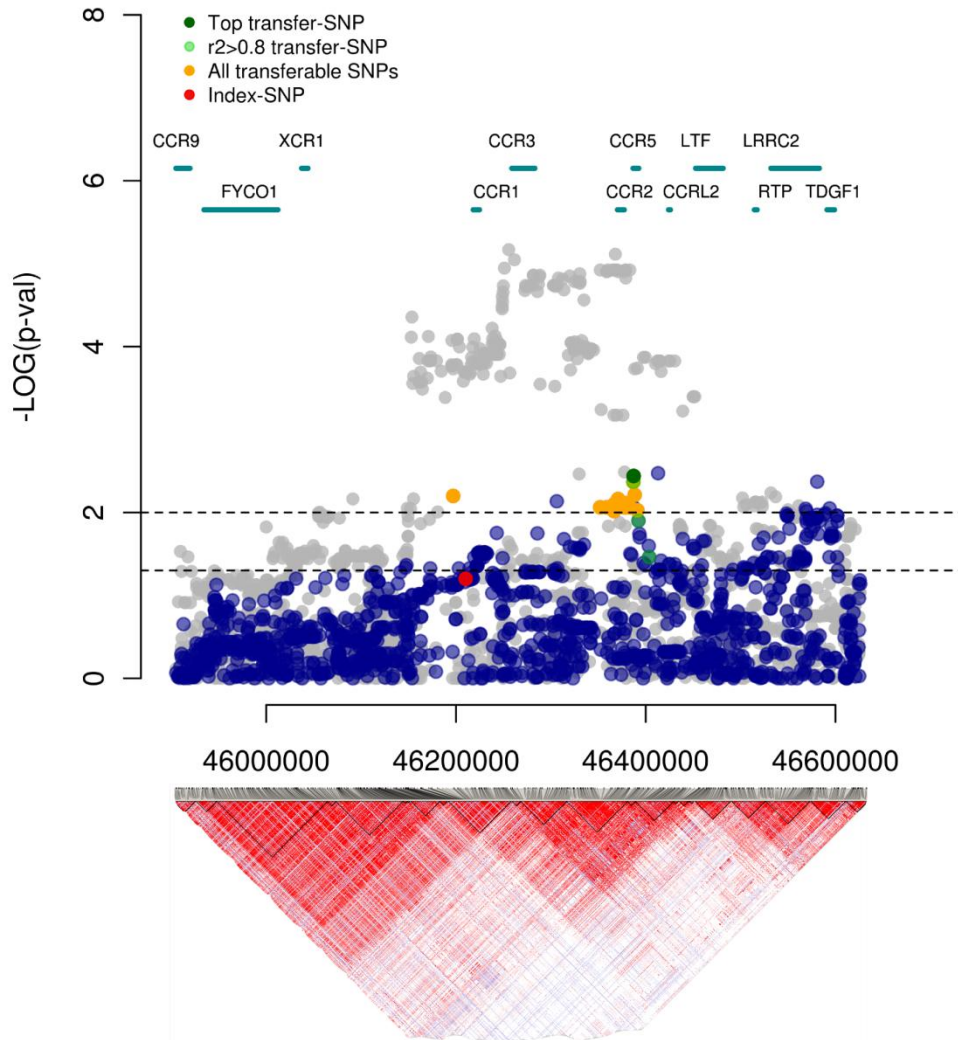


Netherlands
India

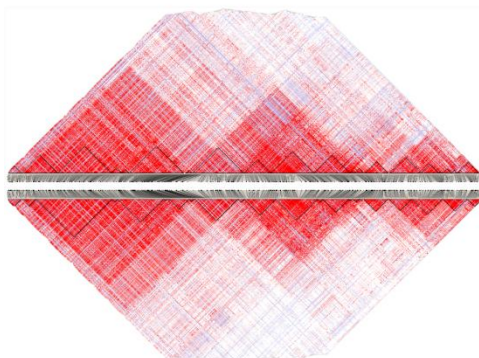


A)

CCRs

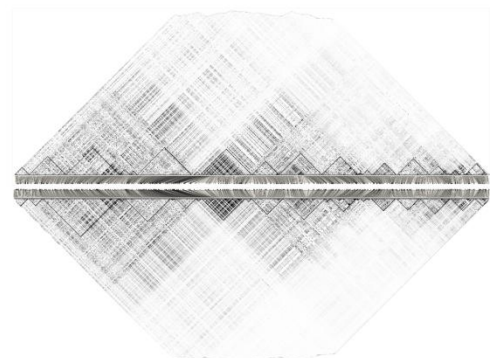


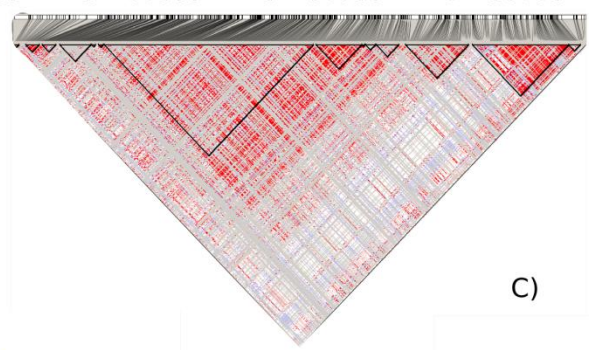
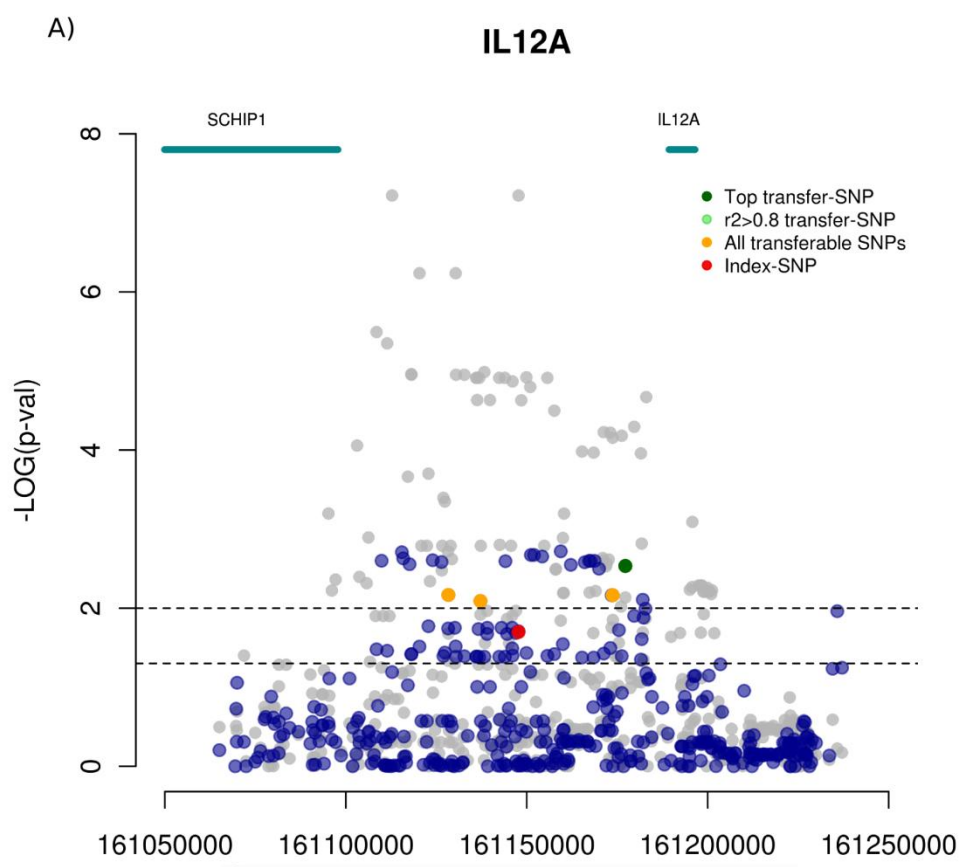
B)



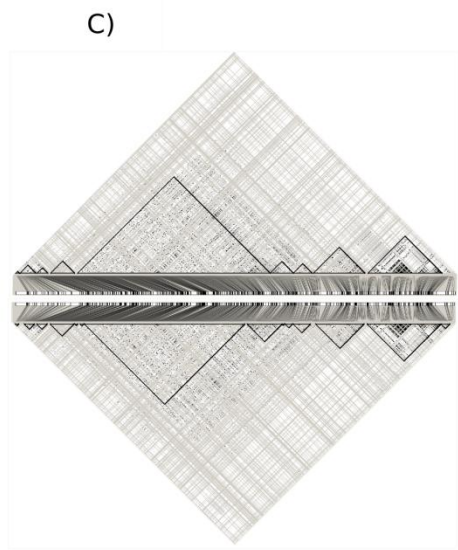
C)

Netherlands
India



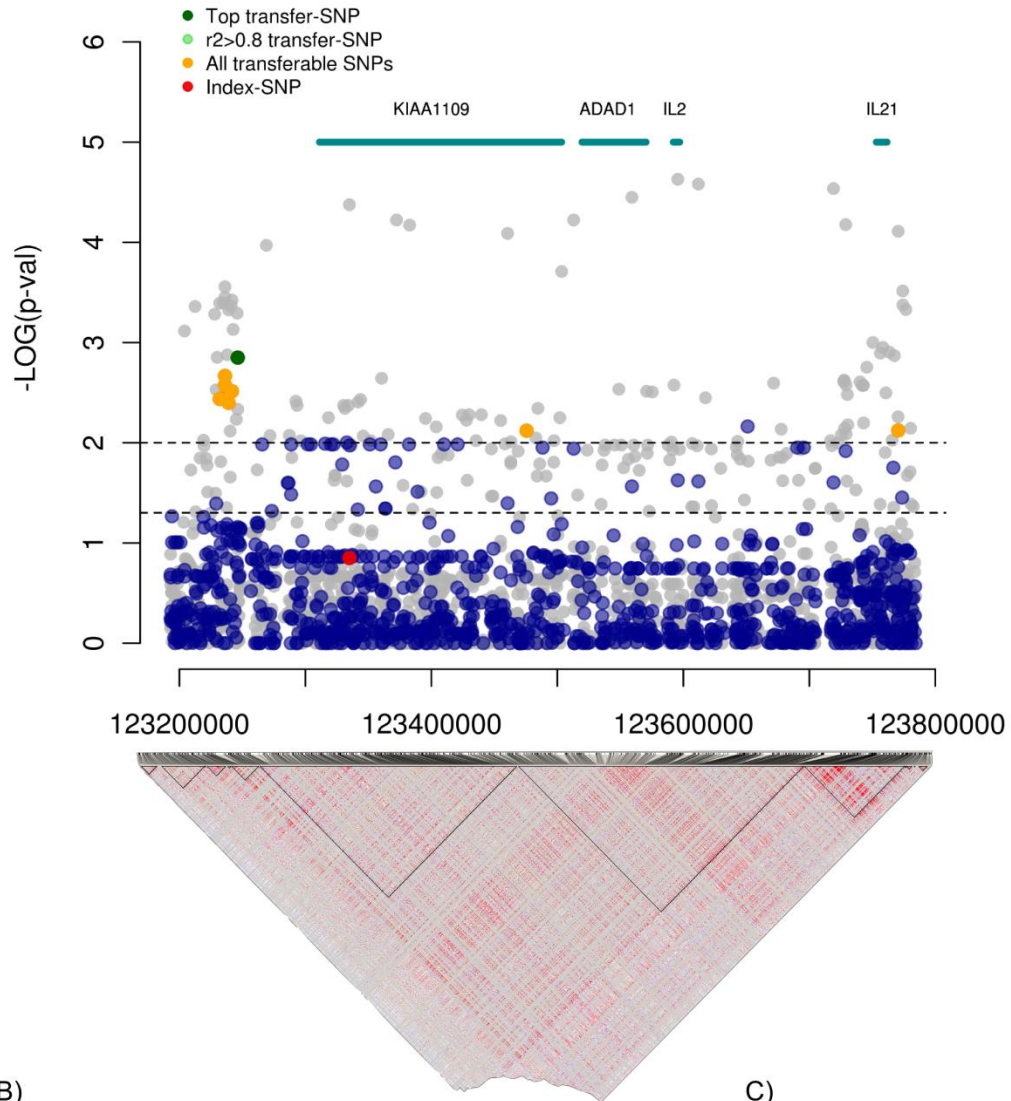


Netherlands
India

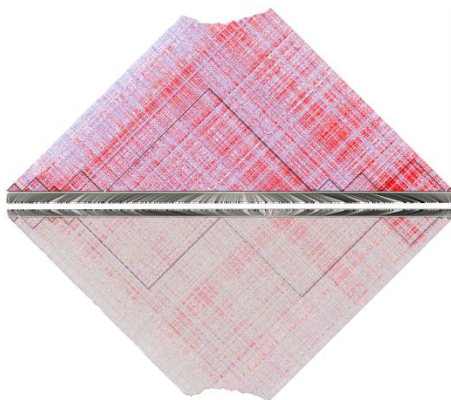


A)

IL2/IL21

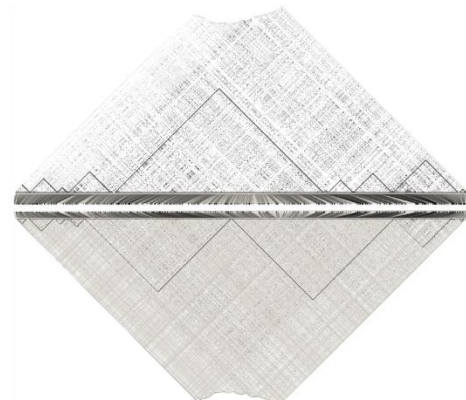


B)



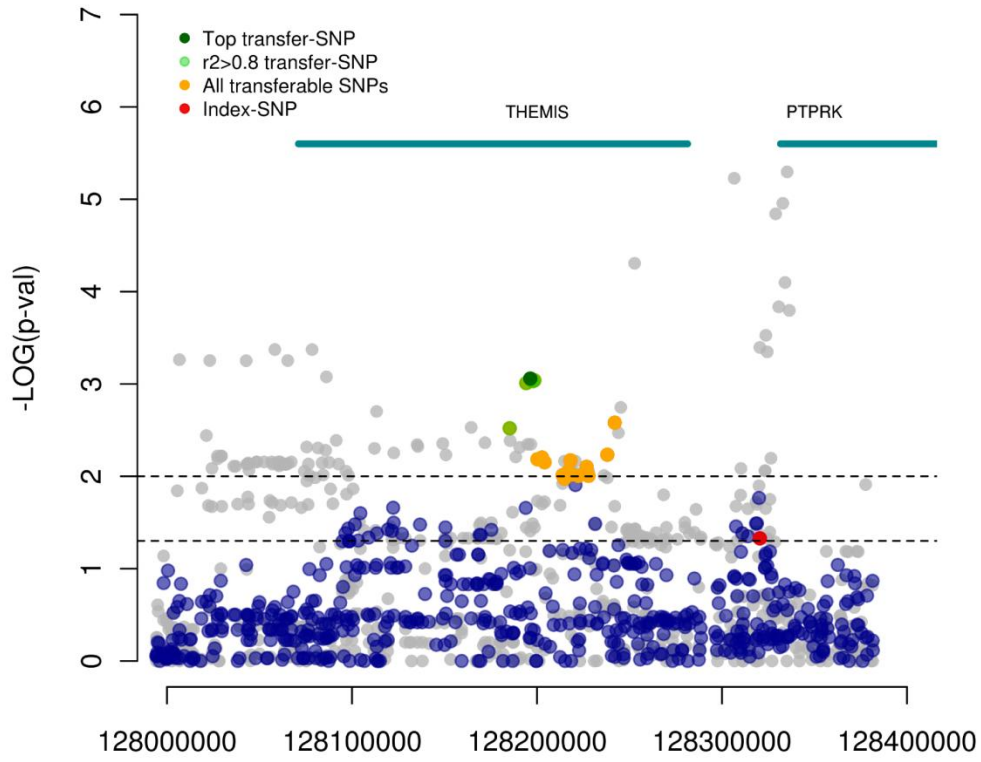
C)

Netherlands
India

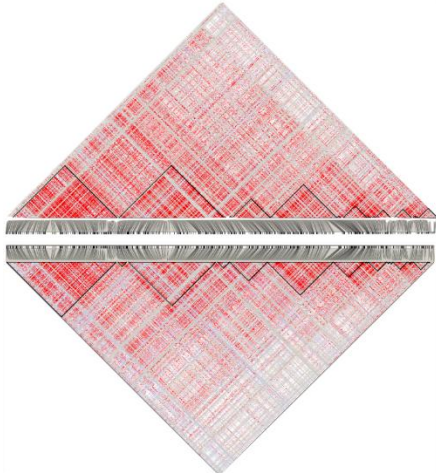


A)

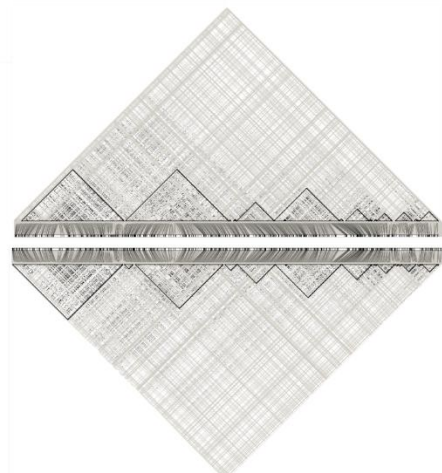
PTPRK/THEMIS



B)



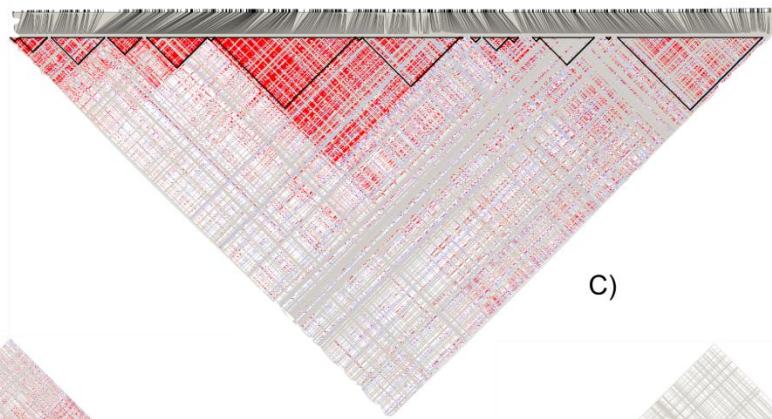
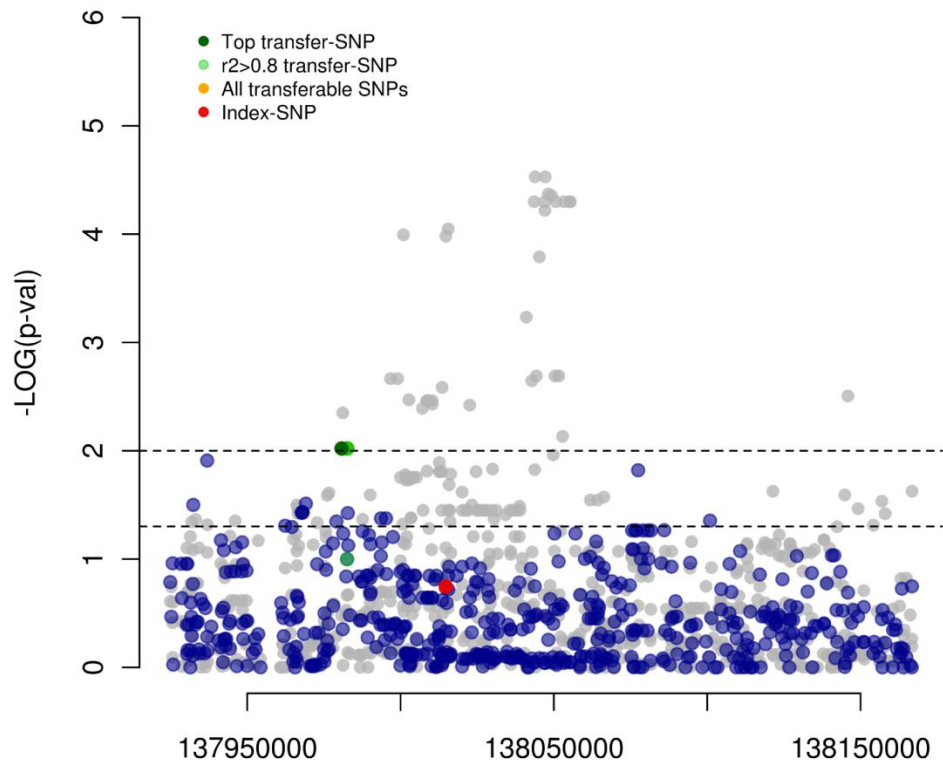
C)



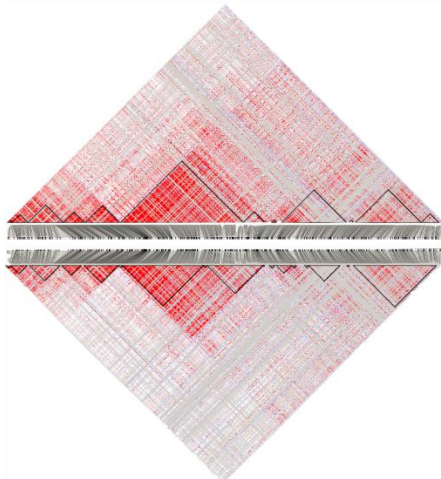
Netherlands
India

A)

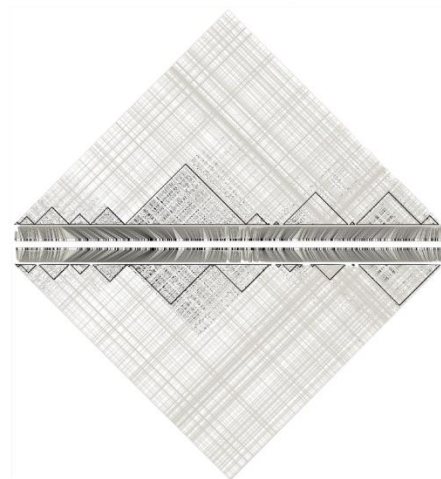
TNFAIP3



B)



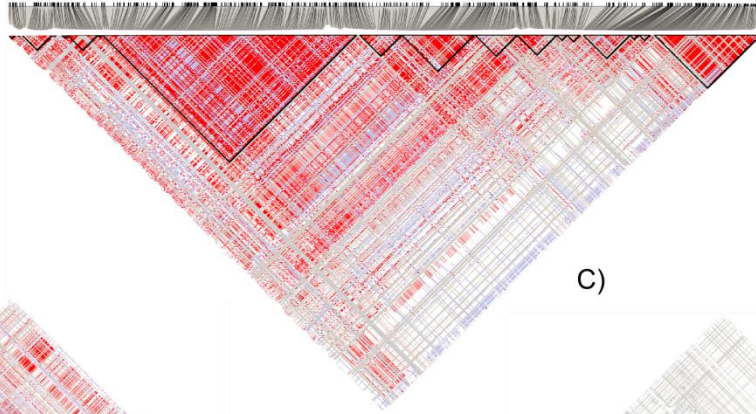
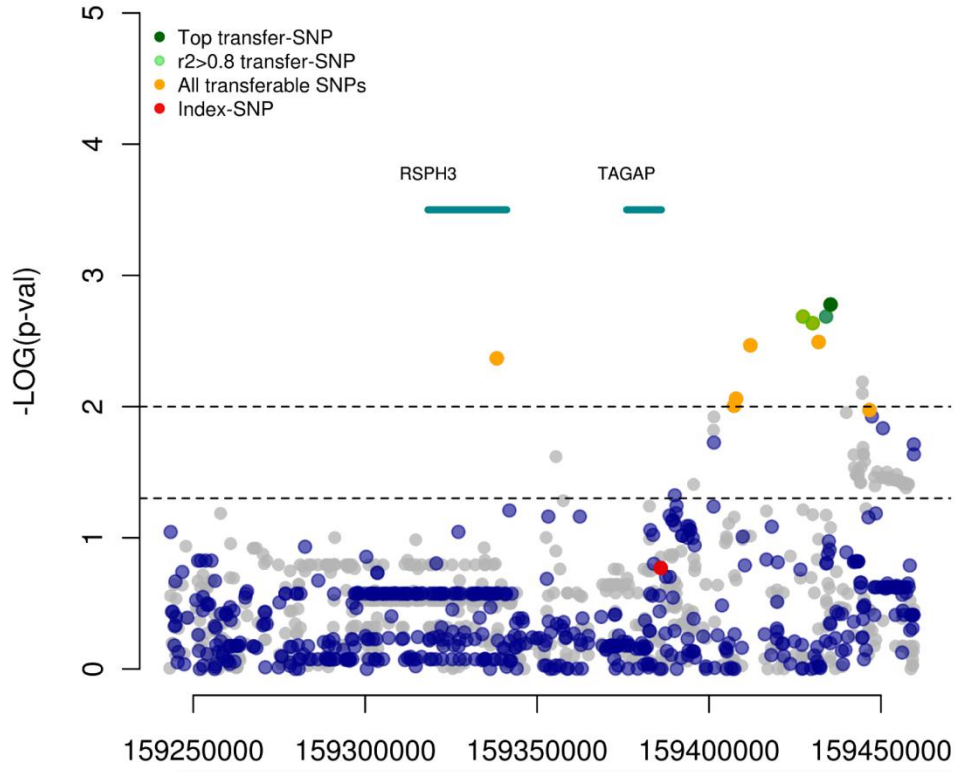
C)



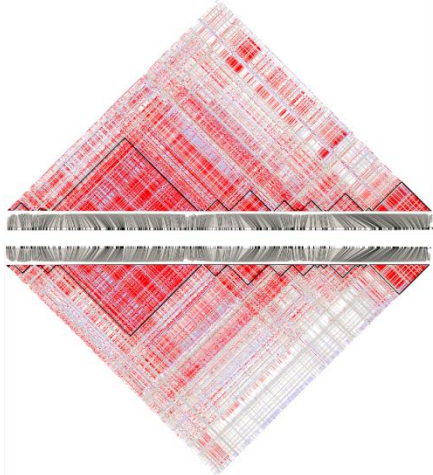
Netherlands
India

A)

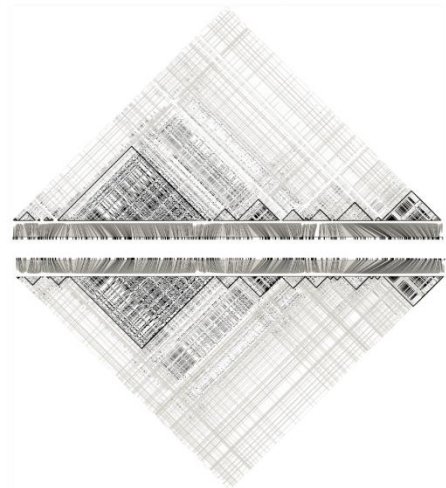
TAGAP



B)



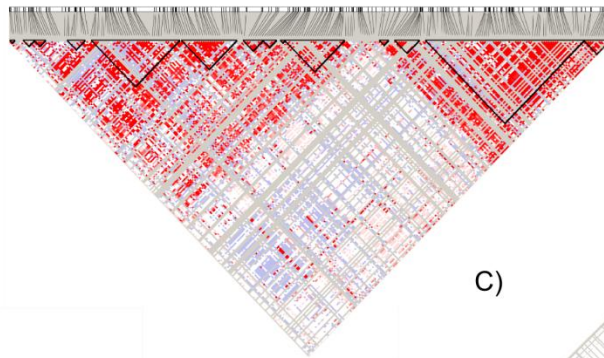
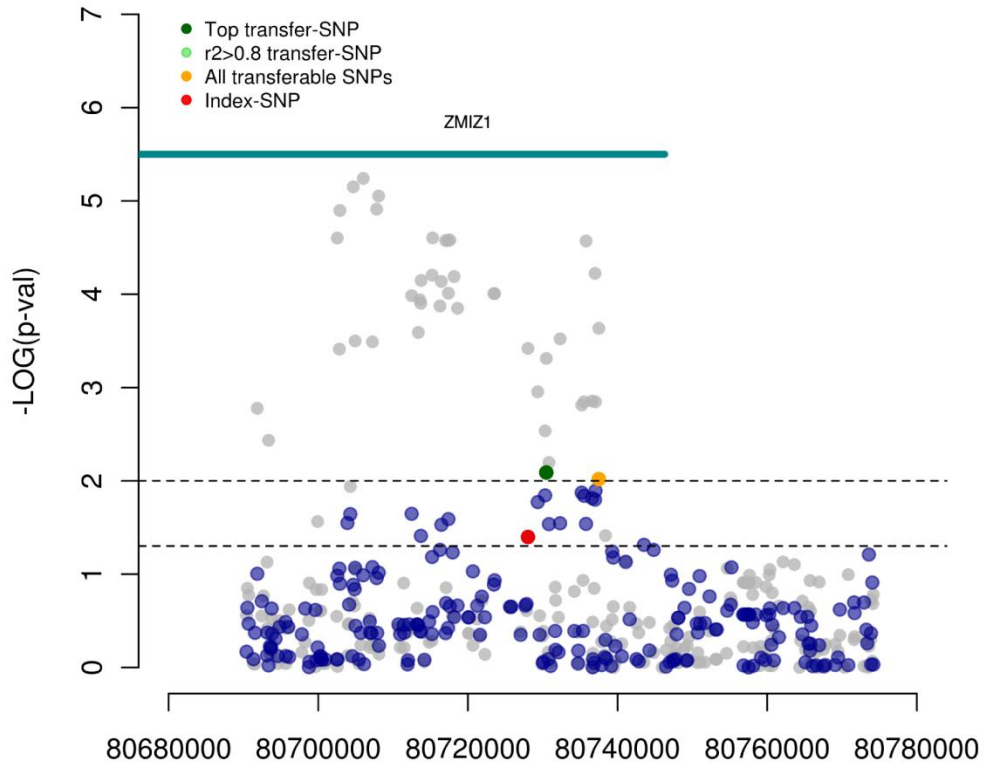
C)



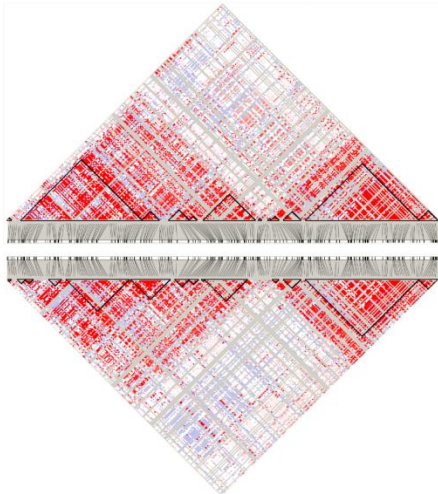
Netherlands
India

A)

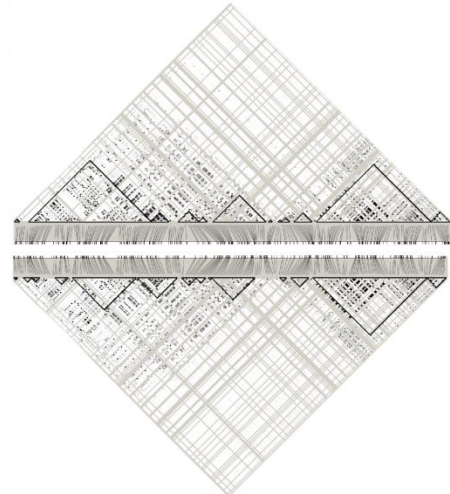
ZMIZ1



B)

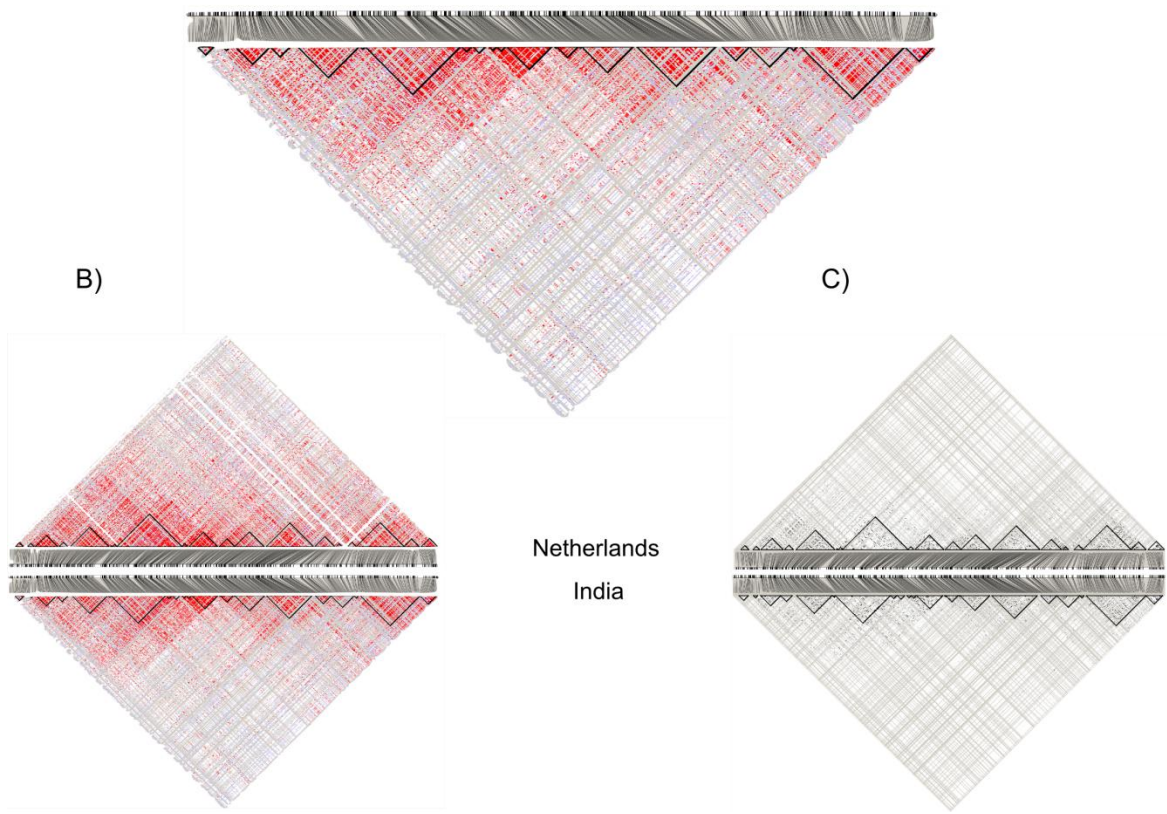
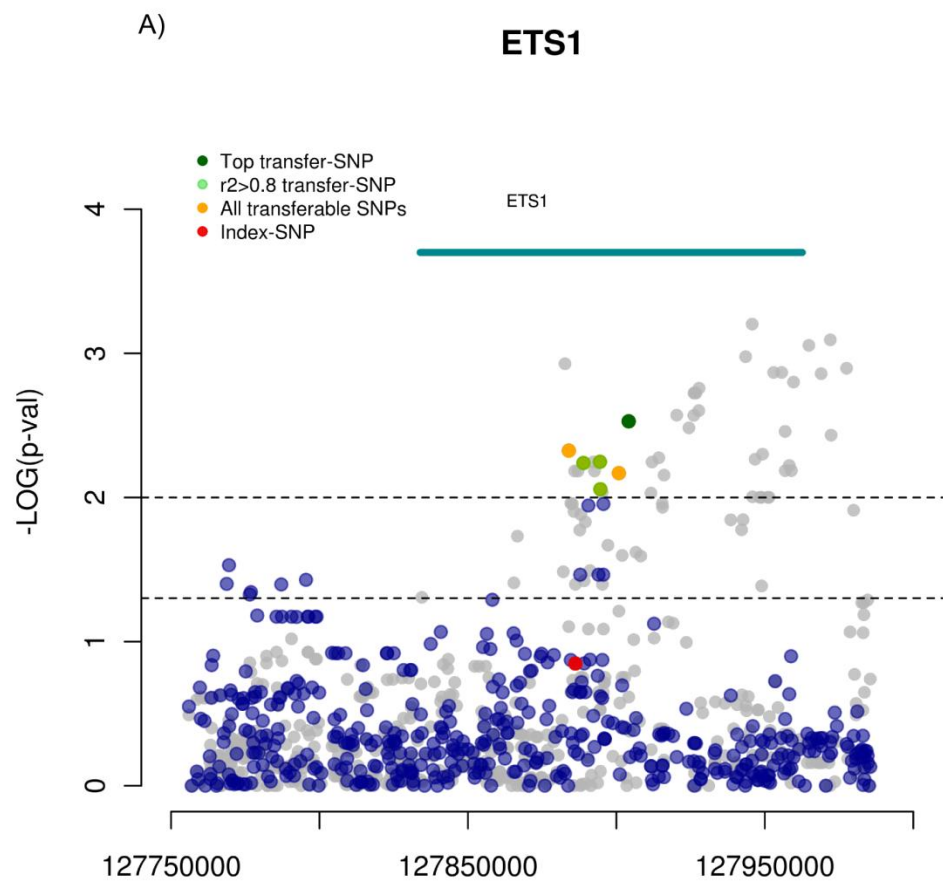


C)

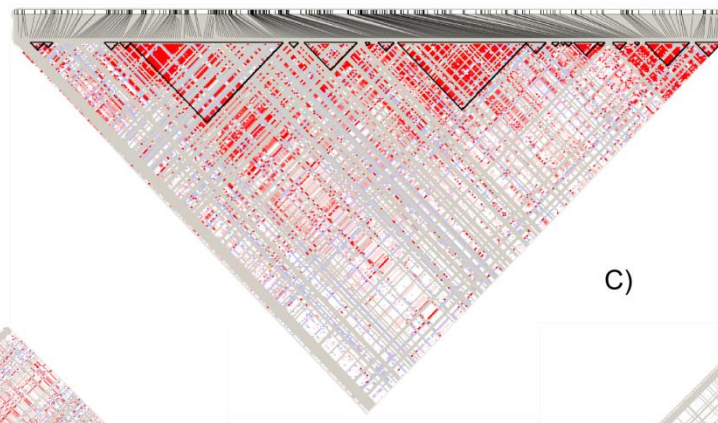
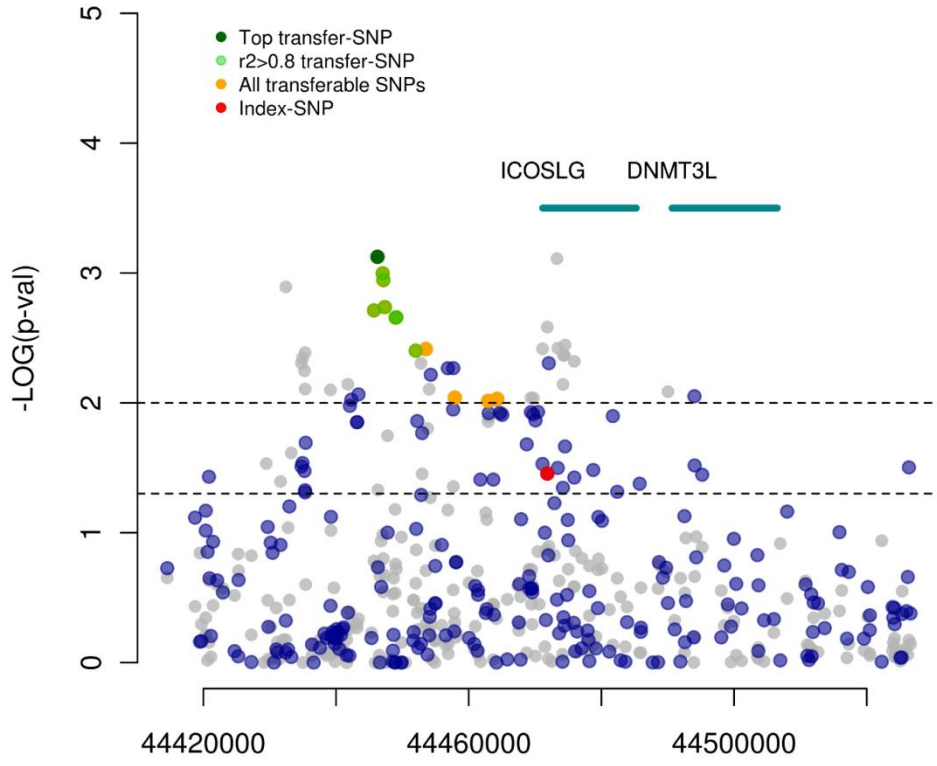


Netherlands

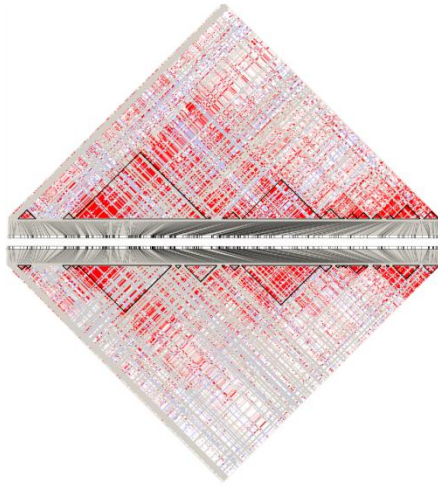
India



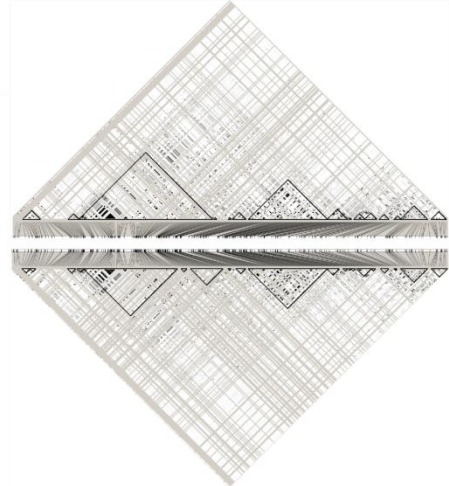
A) **ICOSLG**



B)



C)



Netherlands
India

Figure S8 | Correlation of the minor allele frequency (MAF) of SNPs in the *IL2-IL21* locus in north Indians and Dutch (A) and the distribution of allele frequencies between these two populations (B-C), showing a clear shift towards a higher proportion of low frequency markers (MAF<0.1) in the north Indian samples. Comparison of the LD structure at *IL2-IL21* after exclusion of variants with MAF<0.1, in the Dutch (upper blocks) and north Indians (lower blocks), measure by D' (red) and r^2 (grey).

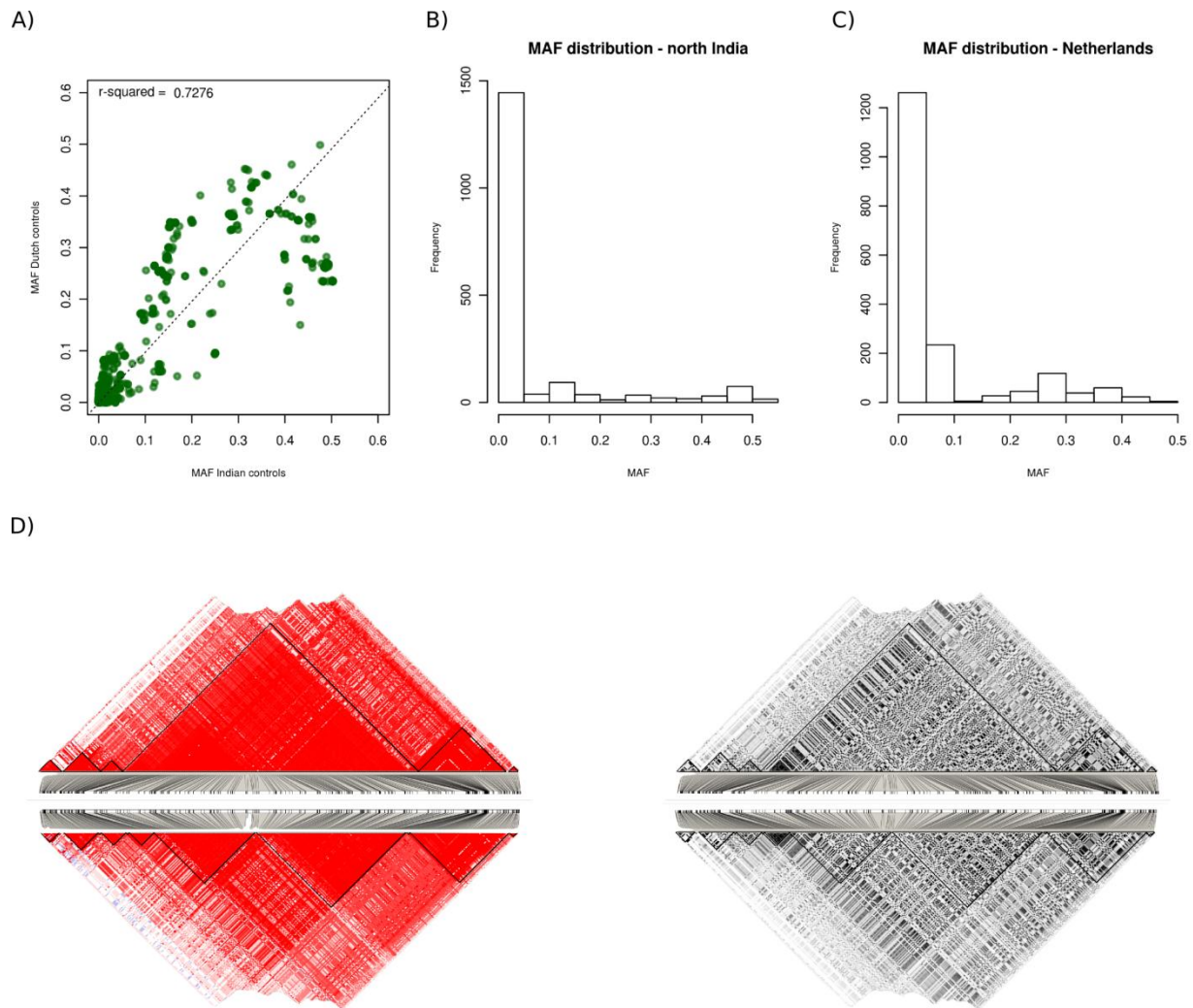
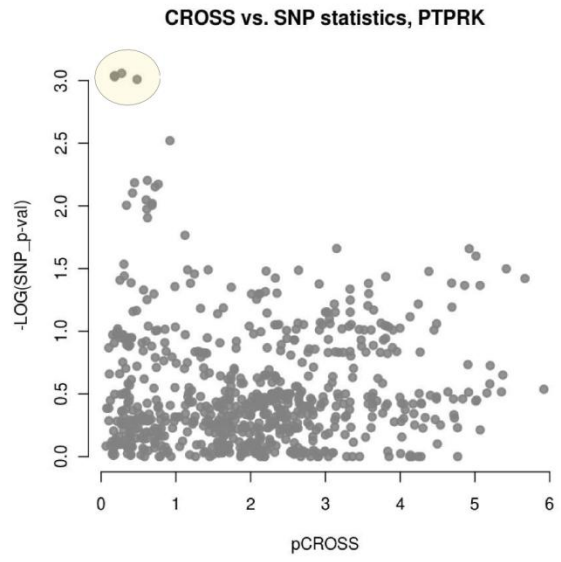
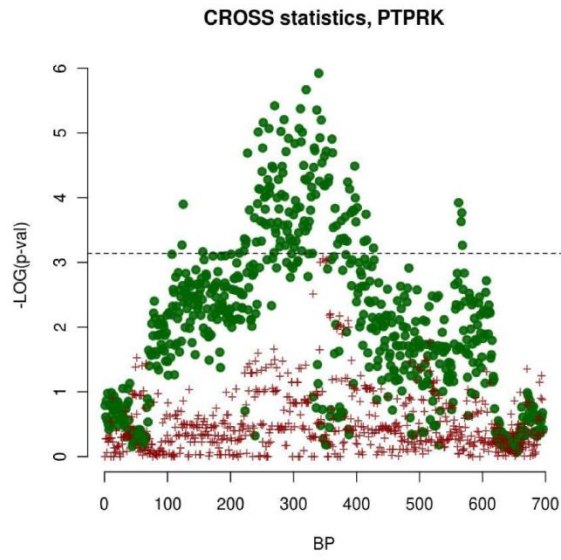


Figure S9 | Results of the CROSS test at *IL12A* and *THEMIS/PTPRK* loci. Left panels depict CROSS results (green dots) in combination with the single SNP test (red crosses). Right panels plot CROSS results against the single SNP association, SNPs clustering together (marked in the circles) were further taken for construction of haplotypes.

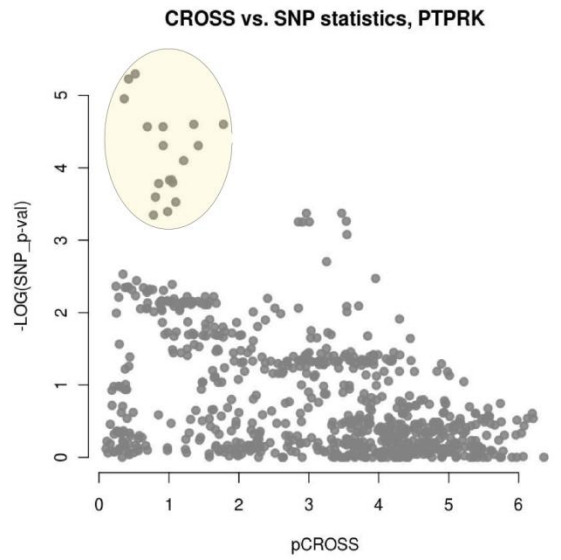
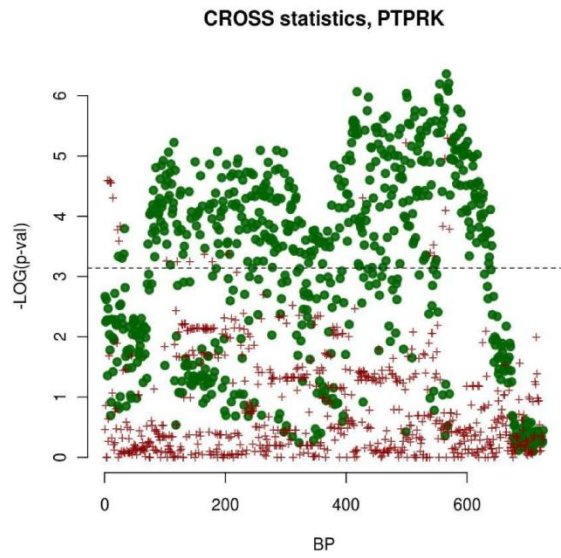
In the *THEMIS/PTPRK* locus we were able to distinguish one risk haplotype in north Indians and another risk haplotype in the Dutch (Table S5). Since none of the SNPs from the two population overlapped we constructed a “super-haplotype” which included all SNPs from north Indians and Dutch. Unlike the European index SNP *imm_6_128320491* (*rs802734*), the top transferable north Indian risk allele *imm_6_128196379**G (*rs4142030**G) was captured by the super-haplotype. In north Indians we observed eight and in the Dutch six different super-haplotypes (Table S5). The most associated risk super-haplotypes in both populations: #India_7 (OR=1.40, p-value=0.04 and #Dutch_5 (OR=1.49, pvalue = 8.68^{-06}) resembled the risk of the most associated haplotypes from Cross test: India_Cross: OR=1.35, p-value=0.0009 and Dutch_Cross: OR=1.51, p-value= 3.35^{-07} (Table S5). In phylogenetic analysis these haplotypes clustered together indicating they carry the same ancestor mutation. However lack of any overlap makes it difficult to draw any further conclusions (Table S5).

SNPs from two risk haplotypes at *IL12A* locus from both populations were merged to construct a single super-haplotype, which captures the top transferable north Indian risk allele *imm_3_161177252**T (*rs1498736**T). The risk was not recaptured by super-haplotype in Indians and increased from OR of 1.32 for the Indian_Cross haplotype to OR of 1.77 for #India_6 super-haplotype. The increase of the risk was driven by the Dutch_Cross 4 SNPs and as investigated in the Indians gave the risk measure of 1.71. In Dutch population the OR remained similar: Dutch_Cross OR=1.59, comparing to super-haplotype #Dutch_4 OR = 1.57. These two risk haplotypes also showed same ancestral origin in phylogenetic analysis.

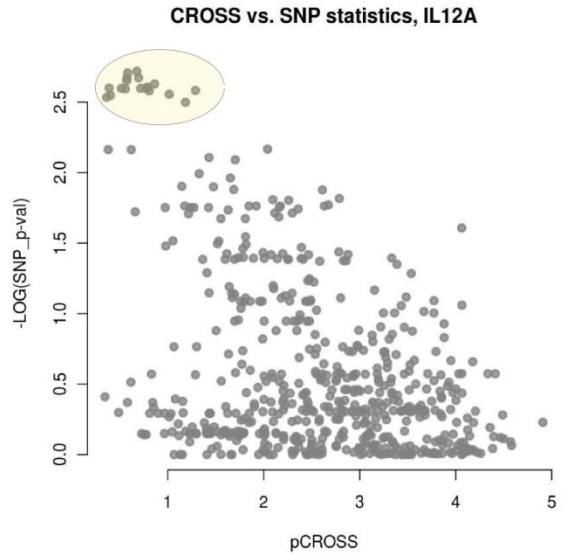
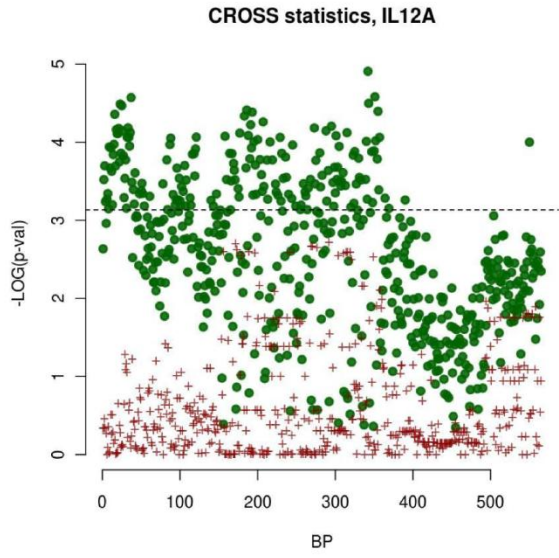
India



Netherlands



India



Netherlands

