Supporting Information

Table S1: Predicted Effects of Tag SNPs and Variants in LD ($r^2 \geq 0.8$) on Transcription Factor Binding Motifs

The lead SNPs identified in this study are colored in red.

Name: dbSNP identifier of the GWAS hit.

Factor: name of the PWM of the proposed effect. PWMs from Factorbook[53] are subscripted _FACTOR (or lack a subscript, e.g. "UA8"); all others are from HOCOMOCO. The HOCOMOCO subscripts reflect models used to derive the PWMs as described.[54]

Match: the exact sequence of the genomic DNA around the SNP that is proposed to match the PWM.

SNV.pos (single nucleotide variant position): indicates the position of the SNP within the matched DNA sequence in the preceding column.

Strand: the + or - strand of the PWM match with respect to the reference genome.

Variant: the identity of the nucleotide variant reference and alternate alleles, presented as reference-->alternate.

Refpct: the percent of maximum PWM score assessed on the reference version of the genome given by the column "match".

Varpct: the percent of maximum PWM score assessed on the alternate SNP value of the genome if the Variant is substituted on the Match.

Refsnp.val: the frequency of the reference nucleotide within the PWM centered on the match sequence.

Varsnp.val: the frequency of the alternative nucleotide within the PWM centered on the match sequence. Note that Refsnp.val and Varsnp.val do not sum to 1 because a binary SNP represents only 2 of the 4 frequencies at each position. At each position of the PWM, frequencies of the four nucleotides (A, C, T, G) sum to 1.0.

Effect: indicates whether the alternate version improves upon or disrupts the reference match.

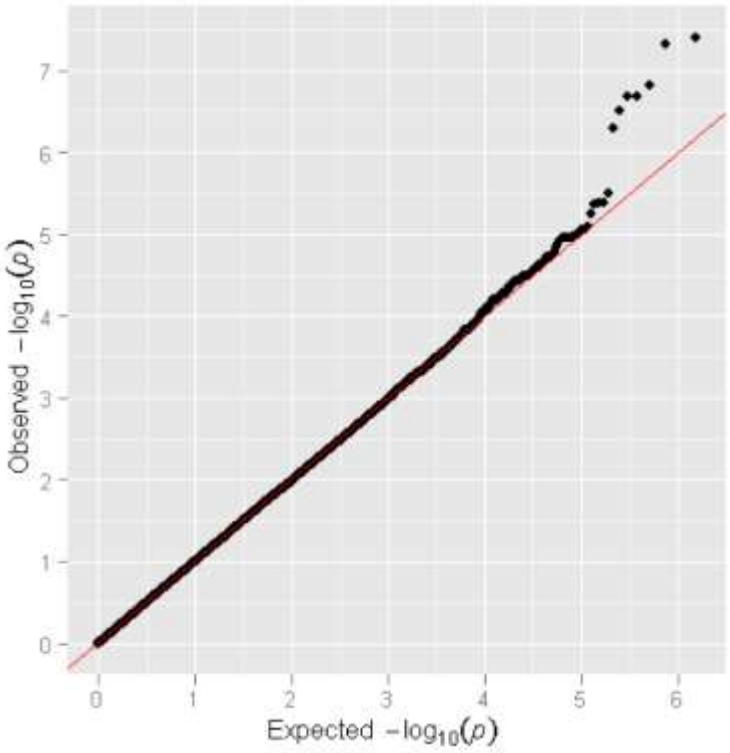

Figure S1: Q-Q plot

Q-Q plot showing the distribution of the observed P values from the logistic regression analysis for the GWAS scan against the expected distribution under the null hypothesis.