Exome sequencing data processing and quality control

Exome sequence reads of 753 individuals were aligned to NCBI Build 37 of human reference genome using BWA [1]. SAMtools [2] was used for variant calling. Only variants with Phred-scale quality score ≥ 25 and read depth ≥ 8 were kept. In the case of no variant called at a site for an individual, we set the genotype as homozygous reference allele if read depth at that site ≥ 8, and as missing otherwise.

For NOD2 data, we focused on the +/- 500 Kb region around rs17221417 (the GWAS signal), which contained 608 SNPs. For this candidate region, we further applied the following quality control (QC) criteria: 1) missing rate per SNP ≤ 0.15; 2) missing rate per individual ≤ 0.1; 3) Hardy-Weinberg Equilibrium p value > 10^{-5}; and 4) MAF ≥ 0.2%. Finally, we used BEAGEL [3] (with default parameters) to impute the remaining sporadically missing genotypes and retained SNPs with allelic $R^2 ≥ 0.9$. After QC, 728 samples and 100 SNPs genotyped on each were kept for statistical analysis.

For ITPA data, we focused on the +/- 500 Kb region around rs6051702 (the GWAS signal), which contained 1353 SNPs. Applying the same QC as above, we obtained 715 samples and 338 SNPs genotyped on each.