The American Journal of Human Genetics, Volume *90*

Supplemental Data

Linkage-Disequilibrium-Based Binning

Affects the Interpretation of GWASs

Andrea Christoforou, Michael Dondrup, Morten Mattingsdal, Manuel Mattheisen, Sudheer Giddaluru, Markus M. Nöthen, Marcella Rietschel, Sven Cichon, Srdjan Djurovic, Ole A. Andreassen, Inge Jonassen, Vidar M.Steen, Pål Puntervoll, and Stéphanie Le Hellard

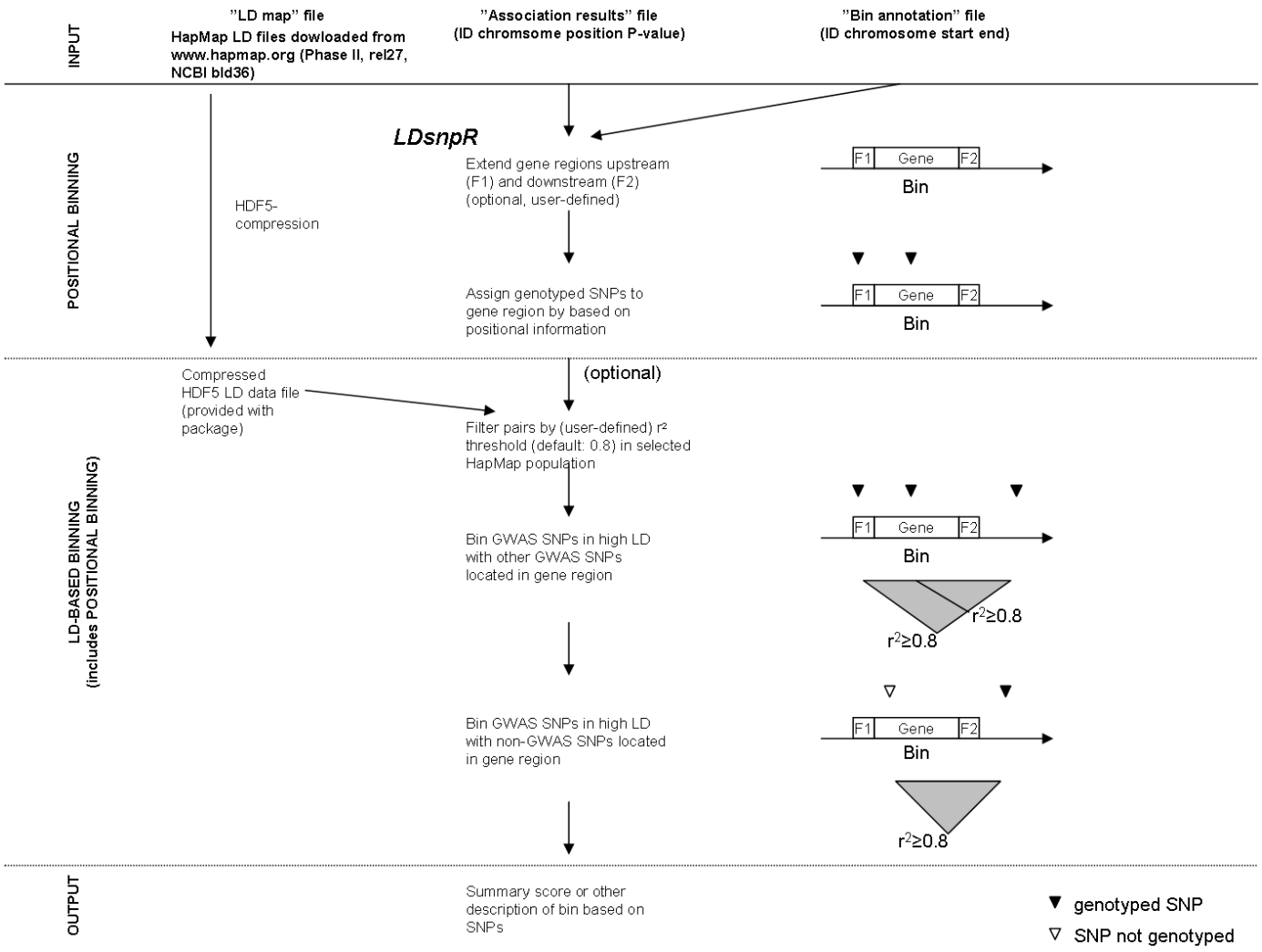

Figure S1. LDsnpR Workflow

The LDsnpR software, including details about its development and use, can be downloaded from http://services.cbu.uib.no/software/ldsnpr. Briefly, LDsnpR requires three input files. The first is a tab-delimited "Association Results" text file, containing the SNP identifiers, their chromosomal position and association result. The second is a tab-delimited "Bin Annotation" text file, containing the bin identifiers (e.g. EnsEMBL gene identifiers) and the start and stop coordinates of each bin. The third is a "LD map" file, containing the pre-calculated pairwise SNP LD information, which is specially compressed in HDF5 format (http://www.hdfgroup.org/HDF5) for computational efficiency. The binning process begins by extending the bin regions, which are defined in the "bin annotation" file, an amount specified by the user and then assigning SNPs to that region if they are physically located within that extended region – i.e. positional binning. Then, based on the user-defined LD threshold (default $r^2 \geq 0.8$), LDsnpR searches the "LD map" file and identifies the relevant SNP pairs in which a genotyped SNP is paired with a SNP (genotyped or not) physically located within a bin. It then assigns the former, genotyped SNP to the bin in which the latter SNP is physically located – i.e. the LD-based binning approach. The LD-based binning approach includes positional binning. Finally, LDsnpR produces a text file of all of the bins and a user-declared summary of the bins, such as the number of SNPs per bin, the minimum *P*-value of all SNPs in the bin or a list of the SNPs in the bin, which can then be used for subsequent analyses. In the current version of LDsnpR, positional information for all SNPs must correspond to NCBI Build 36 of the human genome. Both "Bin Annotation" and "LD map" files used in this study are available for download with the package.

Table S1. Comparison of Positional and LD-Based-Binning Results for Top SNP Hits in Each GWAS

Reported SNPs[a]	EnsEMBL Genes[b] - Positional Binning	EnsEMBL Genes[b] - LD-Based Binning
WTCCC BP		
rs420259	PALB2	PALB2
rs4027132	-	-
rs7570682	-	-
rs1375144	DPP10	DPP10
rs2953145	***RNPEPL1***	***RNPEPL1; CAPN10***
rs4276227	CMTM8	CMTM8
rs683395	LAMP3	LAMP3
rs6458307	KIAA0240; ENSG00000181524	KIAA0240; ENSG00000181524
rs2609653	-	-
rs10982256	DFNB31	DFNB31
rs10134944	-	-
rs11622475	***TDRD9***	***C14orf2; TDRD9; ENSG00000211292***
rs1344484	-	-
rs3761218	***CDC25B; CENPB***	***C20orf27; SPEF1; CDC25B; CENPB***
TOP BP		
rs6679053	PLD5	PLD5
rs12052834	-	-
rs4257412	-	-
rs1529289	CNTNAP5	CNTNAP5
rs2420559	CNTNAP5	CNTNAP5
rs2699365	CNTNAP5	CNTNAP5
rs13216050	-	-
rs10901040	DPP6	DPP6
rs4960568	DPP6	DPP6
rs1510122	-	-
rs1567426	-	-
rs448245	-	-
rs264826	-	-
rs7844514	ENSG00000104324	ENSG00000104324
rs2578121	-	-
rs1435961	-	-
rs7100273	-	-
rs7081438	-	-
rs7119726	SHANK2	SHANK2
rs7310876	-	-
rs1750565	ENSG00000220558	ENSG00000220558
rs1798968	ENSG00000220558	ENSG00000220558
rs809846	ENSG00000220558	ENSG00000220558
rs1750567	ENSG00000220558	ENSG00000220558
rs11617400	GUCY1B2	GUCY1B2
rs9527256	-	-
rs8042197	-	-

rs1848053	-	-
rs931781	-	-
rs7183870	SLCO3A1	SLCO3A1
rs7184694	-	-
rs11652429	-	-
rs11080256	-	-
rs11658512	-	-
rs1292735	-	-
German BP		
rs1064395	***NCAN; HAPLN4***	***NCAN; HAPLN4; TM6SF2; SF4; KIAA0892; GATAD2A; MIRN640; ENSG00000213993***
rs11764590	MAD1L1	MAD1L1
rs10278591	***MAD1L1***	***MAD1L1; ENSG00000176349; ENSG00000215135***
rs6547829	BRE	BRE
rs985409	LHFPL3	LHFPL3
rs2209263	-	-
rs779279	-	-
rs9322993	***SIP1***	***SEC23A; SIP1; TRAPPC6B; PNN; CTAGE4***

SNPs extracted from WTCCC: Table 4 in ref. 18; TOP: Table 1 in ref. 19; and German: Table 2 in ref. 20. b. EnsEMBL 54 (May 2009), gene symbol, if known; otherwise EnsEMBL gene ID. Bold italicised rows indicate SNPs with different annotation after LD-based binning was performed.

Table S2. Effects of LD-Based Binning Compared to Resampling on the Ranks of SNPs and Genes in the WTCCC BP GWAS

Spearman Rank Correlation	**SNP Level**	**Gene Level (Positional Binning)**	**Gene Level (LD-Based Binning)**	**Gene Level (Positional / LD-Based Binning)**
MEAN	0.85	0.87	0.88	0.80
SD	0.0031	0.0034	0.0036	0.0018
Percent difference in top 2000 genes		**Gene level (Positional binning)**	**Gene level (LD based binning)**	**Gene level (Positional / LD based binning)**
MEAN	NA	25.6	25.6	30.7
SD	NA	0.76	0.98	0.62

Mean and SD (standard deviation) of pairwise measures (rank correlation or % difference) resulting from a re-sampling analysis in which 5% of the samples in the WTCCC dataset (genotype data from the European Genome-phenome Archive: www.ebi.ac.uk/ega) were randomly excluded (20 repetitions). SNP-level P-values (autosomes only) are based on the TREND test, as implemented in PLINK.