S1. Obtaining the ethnic proportions of cosmopolitan cohorts (Lee et al., 2015a)

Assume that individual genotypes in the study cohort is a mixture of random genotypes from (not admixed from) the *N* ethnic groups in the reference panel with weight vector $W = [w_i]_{N\times 1}$. Let *S* be the vector of GWAS SNP reference allele frequencies (RAFs). Let $P = [P_i]_{1\times N}$ be the RAF matrix of the reference population ethnicities for the measured SNPs, where P_i is the RAF vector of the i^{th} ethnicity of the reference panel. Under the assumption, the study cohort RAF vector can be expressed as a weighted sum of RAF vectors of reference population ethnicities with $W = [w_i]_{N\times 1}$: $S = \sum_{i=1}^{N} w_i P_i$. After straightforward algebraic manipulations:

$$Cov(P,S) = Cov(P, \sum_{i=1}^{N} w_i P_i) = Cov(P, PW) = Cov(P)W .$$

Linear/quadratic programming methods can be employed to estimate W subject to constraints $\sum_{i=1}^{N} w_i = 1$ and $0 \le w_i \le 1$. However, due to the large number of SNPs, even simply solving the linear system without constraints ($\hat{W} = Cov(P)^{-1}Cov(P,S)$) and substituting zero for the (very) few small negative proportions, can yield very accurate weights.

Alternatively, when GWAS RAFs are not available, users can pre-specify the weights based on their prior knowledge on ethnic composition of the study cohort of interest. This option should be most useful when i) fairly accurate proportion information about ethnicities involved in the cohort is available and ii) all ethnicities in the cohort have reasonably close proxies in the reference panels.

S2. Computing LD patterns for cosmopolitan cohorts (Lee et al., 2015a)

Based on estimated/pre-specified weights $\hat{W} = [\hat{w}_i]_{N\times 1}$, we estimate the cohort genotype correlation matrix Σ in a three step process. First, estimate the cohort genotype covariance matrix C of SNPs within a genomic region as

$$\hat{C} = \sum_{i=1}^{N} \hat{w}_i \hat{C}_i + \sum_{i=1}^{N} \hat{w}_i (\hat{\mu}_i - \hat{\mu})(\hat{\mu}_i - \hat{\mu})^T ,$$

where $\hat{\mu}_i$ and $\hat{C}_i$ are the estimated genotype mean (twice the RAF) vector and variance-covariance matrix for the i^{th} ethnic group respectively and $\hat{\mu}$ is the estimated cohort genotype mean vector computed as $\sum_{i=1}^{N} \hat{w}_i \hat{\mu}_i$. Second, normalize C to obtain the correlation matrix, Σ, by

dividing each covariance by the product of the corresponding SNP genotype standard deviations. To avoid ill-conditioned mixture correlation matrix due to the highly correlated LD structure, we add a ridge adjustment, heuristically set to $\lambda = 2/\sqrt{n}$ (where n is the sample size of the reference population), to the diagonal elements of Σ (Lee et al., 2015b; Pasaniuc et al., 2014; Pickrell, 2014). The estimated mixture LD matrix Σ is subsequently used in our imputation (DIST, Method S3) and gene-level joint testing procedures (JEPEG, Method S4).

S3. Imputing summary statistics of unmeasured functional SNPs (Lee et al., 2015b)

Let Z_u be the vector of Z-scores of unmeasured functional variants in the non-overlapping prediction window with a fixed length (0.1 mega base pairs (Mb) by default). Denote as Z_m the vector of Z-scores of all measured variants (including non-annotated measured variants) within the extended window [i.e. the prediction window with two fixed-length flanking regions (0.2 Mb by default)]. Let $\sum_{u,m}$ be the LD correlation matrix between the unmeasured and measured variants and $\sum_{m,m}$ be the LD correlation matrix among the measured variants, as estimated from a reference panel. $\sum_{u,m}$ and $\sum_{m,m}$ are estimated as presented in Text S2. By using the classical conditional mean formula (Lee et al., 2013), Z_u can be imputed as

$$Z_u = \sum_{u,m}(\sum_{m,m})^{-1} Z_m.$$

The variance-covariance matrix (proxy imputation information measure) of Z_u can be subsequently estimated as

$$I_u = \sum_{u,m}(\sum_{m,m})^{-1}\sum_{u,m}{}^{\mathrm{T}}.$$

To obtain imputation Z-scores with a variance of one, we normalize Z_u using the square root of I_u (Pasaniuc et al., 2014).

S4. Testing for the joint effect of eQTL/functional SNPs (Lee et al., 2015b)

To test for the joint effect of eQTL/functional SNPs known to affect the expression of a gene, JEPEG was designed to rely solely on the measured and imputed association summary statistics. Based on the database-derived functional category information, JEPEG first groups

eQTL/functional SNPs affecting the same gene into the above mentioned six categories: 1) SNPs directly affecting protein function/structure encoded by a gene, i.e. protein function/structure (PFS) (e.g. stop codons), 2) SNPs affecting expression of a gene by disrupting its transcription factor binding sites (TFBS), 3) SNPs affecting the gene function by interrupting biogenesis of an miRNA (miRNA Structure), 4) SNPs affecting miRNA–mRNA target interaction (miRNA Target) and non-categorized/empirically derived 5) cis- and 6) trans-eQTLs. These functional SNPs can belong to one or more categories/genes simultaneously. To avoid a large number of degrees of freedom for the resulting test statistic (while simultaneously assessing the contribution of each functional category to the overall signal), we pool together statistics of all SNPs from the same functional category, in a single synthetic category score. This score is a weighted sum of the Z-scores associated with the SNPs in the functional category. The weighted sums of all functional categories influencing a gene are subsequently combined in a gene-level statistic by using a Mahalanobis-type statistic, which takes into account their multivariate correlation (as estimated from a relevant reference panel).

In more detail, let Z be the vector of measured and imputed Z-scores for m SNPs functionally associated with the gene under investigation, Y be the diagonal matrix of the square root of imputation information for the m functional SNPs, S be the weight matrix, as derived from the SNP annotation database, for the m functional SNPs belonging to the k functional categories. S consists of m column vectors representing weight scores of the k functional categories per SNP, which are pre-calculated on the basis of the consensus of results from diverse prediction methods and stored in the JEPEG annotation database (see Text 1 in the supplementary data of (Lee et al., 2015b). To down-weight SNPs with low imputation information, based on Y and S, we compute the adjusted weight matrix by accounting for the imputation information of the SNPs: $W = SY$. Let Σ_G be the correlation matrix of SNP genotypes, e.g. as estimated from a reference panel, U be the vector of weighted sum of Z-scores by category (i.e. the synthetic scores) and Σ_U be the variance-covariance/correlation matrix of U. Then, in mathematical notation:

$$U = WZ \text{ and } \Sigma_U = W\Sigma_Z W^T,$$

where Σ_Z is the covariance/correlation matrix of Z. Given that, under the null hypothesis of no association between genotype and trait (H_0), Z is asymptotically distributed as a multivariate normal with a zero mean vector and covariance matrix Σ_G, it follows that:

$$\Sigma_U = W\Sigma_G W^T.$$

Σ_G is estimated as shown in Text S2. Due to linkage disequilibrium, Σ_G might be close to singular, which results in unstable estimation of the gene-based test statistic. Thus, to stabilize JEPEG statistic, we add the DIST ridge adjustment to the diagonal elements of Σ_G. Based on the synthetic scores of all functional categories affecting the gene and their correlation structure, JEPEG computes an omnibus gene-level test as

$$T = U^T \sum_U^{-1} U ,$$

which, under H_0, is asymptotically distributed as a central χ^2 statistic with k df. The two-tailed p-values associated with the normalized U can be used as a post-hoc measure to evaluate the contribution of each functional category to the omnibus gene signal.

S5. Assessing the Type I error rate

To compare the Type I error rate of the proposed method with that of JEPEG, we considered two different cosmopolitan cohort scenarios based on the 1000 Genomes haplotypic data: 1) 30% CEU + 25% CHS + 5% PUR + 40% YRI (Cohort 1) and 2) 10% ASW + 15% CEU + 15% CHB + 12.5% CHS + 15% GBR + 10% MXL + 2.5% PUR + 20% YRI (Cohort 2). For each scenario, we simulated 100 genotypic data of 10,000 subjects for Ilumina 1M autosomal SNPs. Each simulation consists of 5,000 cases and 5,000 controls, for which the case-control status was randomly assigned (i.e. under the null hypothesis of no association). Summary statistics were obtained by regressing the simulated SNP genotypes on the case control status. Using the 100 summary data sets of each cohort, we estimated empirical Type I error rates for JEPEGMIX and JEPEG across different nominal levels. To compute JEPEG Type I error, we simply used 1000 Genomes Phase I Europeans as a reference population.

Fig. S1. Relative Type I error rate (the empirical Type I error rate divided by the nominal Type I error rate) as a function of the study cohort scenario and method used.

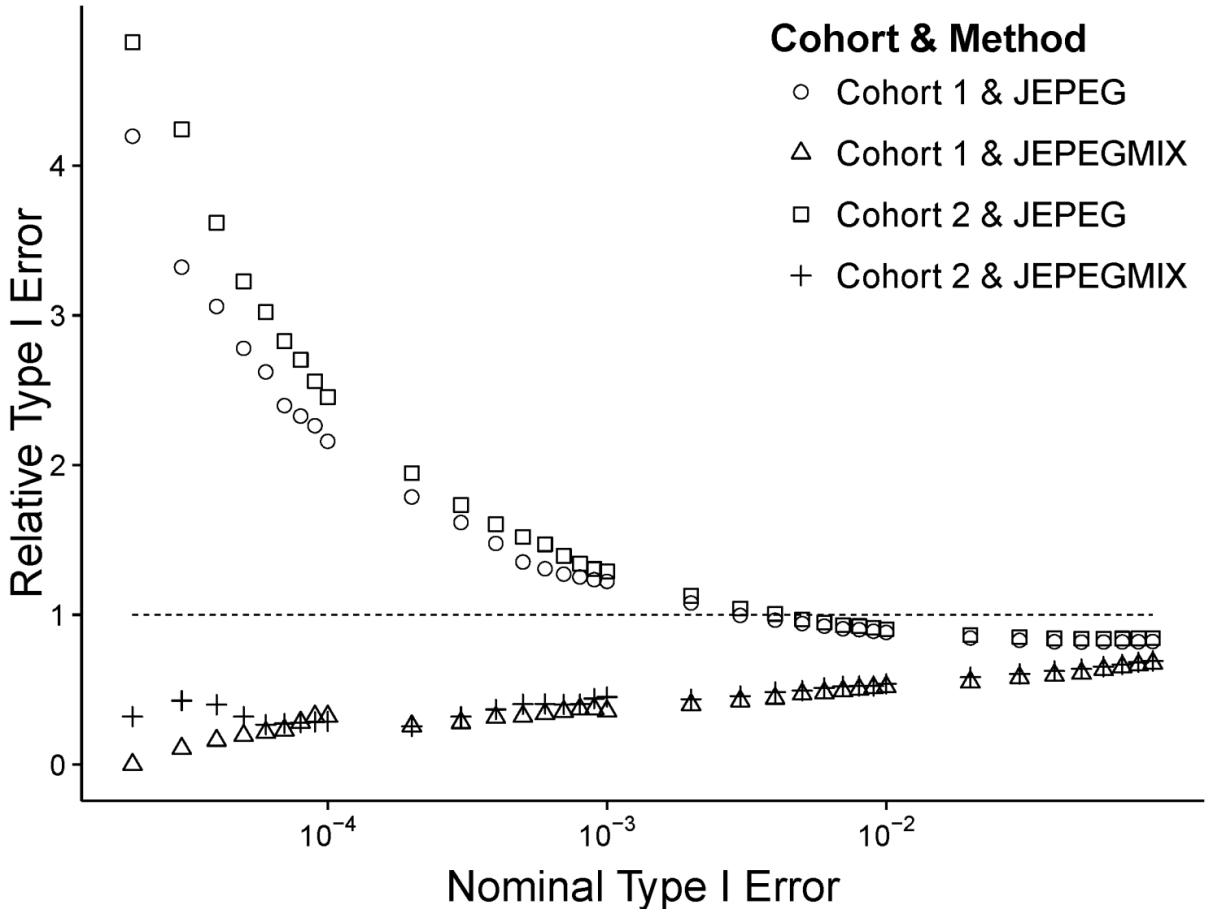

Table S1. Genes with JEPEGMIX p-values deemed significant at an FDR q-value < 0.05. Genes from the MHC region are colored in red.

Gene Name	CHR	Start Position	End Position	Chi-square	Df	JEPEG p-value
BTN3A2	6	26,370,616	26,377,991	121.40	3	3.86E-26
BTN2A1	6	26,459,997	26,468,545	110.68	2	9.27E-25
OR12D3	6	29,274,486	29,408,720	94.13	2	3.63E-21
DPCR1	6	30,919,878	30,920,890	78.14	1	9.59E-19
KIAA1949	6	30,652,781	30,652,781	70.34	1	4.98E-17
HLA-DRB5	6	32,485,524	32,497,943	73.17	2	1.29E-16
HIST1H2BL	6	27,775,674	27,775,674	68.02	1	1.61E-16
OR2B2	6	27,879,200	27,879,982	67.45	1	2.16E-16
ZKSCAN4	6	28,219,661	28,219,695	66.73	1	3.12E-16
NKAPL	6	28,227,436	28,228,342	68.10	2	1.63E-15
VARS2	6	30,882,689	30,893,941	61.88	1	3.65E-15
BTN3A1	6	26,405,816	26,415,062	61.92	2	3.59E-14
SLC17A2	6	25,914,853	25,924,134	54.14	1	1.87E-13
SLC39A8	4	103,184,239	103,228,734	54.42	3	9.15E-12
HIST1H2AL	6	27,833,174	27,833,342	46.21	1	1.06E-11
ZNF323	6	28,294,550	28,297,313	52.60	3	2.24E-11
CYP17A1	10	104,596,981	104,596,981	41.09	1	1.46E-10
TTYH3	7	2,698,634	2,703,979	42.72	2	5.30E-10
SF3B1	2	198,283,305	198,283,305	38.31	1	6.02E-10
MAD1L1	7	1,976,457	1,976,457	38.78	2	3.80E-09
EP300	22	41,513,727	41,574,383	32.76	1	1.04E-08
ZKSCAN8	6	28,120,898	28,120,898	30.98	1	2.61E-08
ZKSCAN8	6	28,121,278	28,121,278	30.98	1	2.61E-08
NMB	15	85,198,606	85,200,520	34.84	2	2.73E-08
C10orf26	10	104,572,963	104,572,963	30.65	1	3.09E-08
PTPRF	1	44,045,573	44,083,507	34.44	2	3.31E-08
HIST1H1T	6	26,107,790	26,108,282	29.54	1	5.49E-08
FURIN	15	91,424,574	91,424,574	29.01	1	7.19E-08
GID4	17	17,948,475	17,948,475	28.85	1	7.82E-08
HLA-C	6	31,237,124	31,239,827	28.36	1	1.01E-07
ITIH4	3	52,852,538	52,861,211	27.21	1	1.83E-07
HIST1H2BPS1	6	25,732,302	25,732,302	26.91	1	2.13E-07
SYNGAP1	6	33,408,542	33,408,542	30.65	2	2.21E-07
DRG2	17	18,011,140	18,011,140	25.51	1	4.40E-07
SETD6	16	58,549,932	58,552,959	31.95	3	5.37E-07
LRRC48	17	17,896,205	17,896,205	24.97	1	5.83E-07
WBP2NL	22	42,416,056	42,423,110	24.71	1	6.66E-07
DOC2A	16	30,021,402	30,021,402	24.27	1	8.38E-07

MPHOSPH9	12	123,661,295	123,705,962	23.94	1	9.96E-07
CHRNA5	15	78,880,752	78,882,925	27.43	2	1.11E-06
DGKZ	11	46,387,868	46,387,868	23.56	1	1.21E-06
ITIH1	3	52,820,981	52,821,011	26.99	2	1.38E-06
SNX19	11	130,750,592	130,784,886	26.56	2	1.71E-06
CPEB1	15	83,215,251	83,215,251	22.48	1	2.12E-06
COBLL1	2	165,551,201	165,578,602	25.88	2	2.40E-06
PITPNM2	12	123,471,094	123,519,112	21.88	1	2.90E-06
PCDHA2	5	140,174,622	140,174,865	21.51	1	3.51E-06
VRK2	2	58,316,814	58,316,814	21.34	1	3.84E-06
NDUFA6	22	42,486,723	42,486,723	24.75	2	4.22E-06
DNAJA3	16	4,476,089	4,484,396	24.48	2	4.83E-06
GNL3	3	52,721,305	52,727,257	20.89	1	4.87E-06
MOV10	1	113,237,171	113,241,052	24.43	2	4.96E-06
ATXN7	3	63,898,497	63,982,082	24.17	2	5.66E-06
SBNO1	12	123,806,219	123,806,219	20.37	1	6.38E-06
SDCCAG8	1	243,493,907	243,493,907	20.14	1	7.18E-06
ALMS1	2	73,651,967	73,828,538	23.19	2	9.21E-06
PCNXL3	11	65,386,206	65,386,206	19.43	1	1.04E-05
IRF3	19	50,162,909	50,162,909	19.30	1	1.11E-05
DDHD2	8	38,095,662	38,095,662	18.88	1	1.39E-05
OPRD1	1	29,138,975	29,138,975	18.56	1	1.64E-05
RAI1	17	17,696,531	17,696,755	18.54	1	1.67E-05

Table S2. Canonical pathways for non-MHC genes which are enriched at a 0.05 significance level.

Pathway	p-value
Role of RIG1-like Receptors in Antiviral Innate Immunity	0.003
p53 Signaling	0.01
LXR/RXR Activation	0.02
Granzyme A Signaling	0.04
Glucocorticoid Biosynthesis	0.04
RhoGDI Signaling	0.04
NRF2-mediated Oxidative Stress Response	0.04
Calcium Signaling	0.05
Mitochondrial Dysfunction	0.05
ERK/MAPK Signaling	0.05
IL-15 Production	0.05
Androgen Biosynthesis	0.05
AMPK Signaling	0.05

Reference List

Lee,D. et al. (2015a) DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics*, doi:10.1093/bioinformatics/btv348.

Lee,D. et al. (2015b) JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, 31, 1176-1182.

Lee,D. et al. (2013) DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics*, 29, 2925-2927.

Pasaniuc,B. et al. (2014) Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30, 2906-2914.

Pickrell,J.K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, 94, 559-573.