Supplemental Material for

*JBASE: Joint Bayesian Analysis of Subphenotypes and Epistasis*

September 29, 2015
Contents

1 Effects of phenotype heterogeneity on association studies 3

2 Statistical Model 3
   2.1 Derivation of the Posterior Distribution .......................... 5
   2.2 Inference ......................................................... 7
   2.3 Post-processing ................................................ 10
   2.4 Parameter Choices .............................................. 12

3 Simulation Studies 13
   3.1 Data Generation .................................................. 13
   3.2 Additional Performance Results on Simulations ................. 16
      3.2.1 Threshold Analysis ....................................... 16
      3.2.2 Effect of MAF Distribution .............................. 20

4 Type-2 Diabetes Experiments 21
   4.1 Experimental Setup ............................................. 21
   4.2 Analysis of the Candidate Markers ............................ 23
   4.3 Genotyping Quality, LD and Stratification Analysis of subphenotypes 25
   4.4 Genotyping Quality Analysis of the Replication Markers .... 31
1 Effects of phenotype heterogeneity on association studies

At the center of our method is the hypothesis that a complex phenotype can be thought of as a mixture of simpler phenotypes with shared and subphenotype-specific genetic variants that drive similarities and distinctions between these subphenotypes. Figure 1 of the main text illustrates a toy example containing two subphenotypes: Case-1 and Case-2. These two types of cases share some associated genetic variants (SNP3) but are also driven by variants specific to each of them (SNP1 and SNP5 for Case-1 and SNP4 for Case-2). Assume we have 10 controls, and 10 cases consisting of 5 Case-1 and 5 Case-2 patients. In this setting, the classical GWAS case-control approach would only discover SNP3 as it has a marginal association P-value < 6e-05. SNP1, SNP5 and SNP4 will go undetected (Fig. 1A of the main text). This problem is caused by the heterogeneity of the patient population that is not known a priori and cannot be accounted for by the standard association detection algorithms. To detect SNP1, which is specific to Case-1 subphenotype, Case-2 subpopulation should be pooled with controls and compared to Case-1 cohort as is done in part B. This is because SNP1 has no effect on the phenotypes of Case-2 cohort. In brief, comparing Case-1 subpopulation against the rest of the population recovers SNP2, SNP3 and SNP5. Similarly, comparing Case-2 subpopulation reveals SNP3 and SNP4 as associated markers. Overall, subphenotyping is able to recover all markers correctly (Fig. 1B of the main text) including the shared one, i.e. SNP3.

In addition to causing false negatives, not accounting for subphenotypes can result in other problems. For example, assume that in our toy example above, a two-level GWAS is employed: associations discovered in the first study were to be replicated in another cohort in a second GWAS. Further assume that due to some confounding factor, all the case population in the first study comes from Case-1 and the second one from Case-2. In such a setting, SNP2, SNP3, and SNP5 would be detected as potential associations to be validated in the second step. However, in the second step, only SNP3, which is a shared causal variant of both subphenotypes, would be discovered, resulting in SNP2 and SNP5 associations deemed spurious even though they are not. Hence, we believe that disease heterogeneity is one of the main underlying reason for failure in GWAS replications.

2 Statistical Model

In this section, we will derive the exact form of the posterior distribution required for the inference algorithm. Recall that given \( N \) individuals that are genotyped over \( M \) markers and are profiled on \( F \) phenotype features, JBASE aims to find a partitioning of the population into \( K \) subphenotypes, each of which has a homogeneous phenotype distribution whose variation is associated with genotypic variation at the identified marginal and epistatic causal markers (See Fig. S1). We start by introducing some notation:

- The data \( \mathcal{D} \) is represented in an \( N \times (M+F) \) matrix, where each individual \( i \) (\( i^{th} \) row) consists of \( g_i \)
Figure S1: **The JBASE Model:** Given genotype and phenotype matrices as input, JBASE outputs a partitioning of the population into $K$ subphenotypes, each of which has a more homogeneous phenotype distribution as well as group-specific allele distributions at the identified marginal and epistatic causal markers.

- We assume that each individual $I_i=(g_i, z_i)$ belongs to a phenotypic subpopulation $k \in \{1, \ldots, K\}$. This information is captured by the latent variable $C_i \in \{1, \ldots, K\}$, parameterized by the mixing co-efficient vector $\pi$, that needs to be inferred.

- We model each subphenotype’s genotype submatrix as a collection of three genotype components: the null, the marginal, and the epistatic component, denoted with $G_{0}^k = \{g_{i,0} : \forall i \text{ s.t. } C_i = k\}$, $G_{1}^k = \{g_{i,1} : \forall i \text{ s.t. } C_i = k\}$ and $G_{2}^k = \{g_{i,2} : \forall i \text{ s.t. } C_i = k\}$ respectively, where $g_{i,0}$, $g_{i,1}$ and $g_{i,2}$ represent the genotype profile of individual $i$ over the null, marginal, and epistatic markers. Note that the underlying genotype distribution in the null component is the same across all subphenotypes (where the marker belongs to the null component), thus we use $G_{0}$, i.e. without the $k$ superscript, to represent the union of null components across all subphenotypes.

- Similarly, we use $Z^k$ to denote $k^{th}$ subphenotypes’ phenotype submatrix.

- As marker-to-genotype-component assignments are unknown, we introduce $S_{km} \in \{0, 1, 2\} : m \in \{1, \ldots, M\}$, $k \in \{1, \ldots, K\}$ latent variable to represent the assignment of marker $m$ to one of the three marker com-
ponents of population $k$. Hence, $S_k$ denotes the marker-to-genotype component assignment vector for subphenotype $k$.

he priors are listed in Table S1 (See Subsection 2.4 for recommendations on setting the hyper-parameters).

| Model Component                        | Parameter | Prior             |
|----------------------------------------|-----------|-------------------|
| Null marker $m$                        | $\psi_m$  | Dirichlet($\beta_{m1}, \beta_{m2}, \beta_{m3}$) |
| Marginal marker of subphenotype $k$    | $\theta_{km}$ | Dirichlet($\gamma_{km1}, \gamma_{km2}, \gamma_{km3}$) |
| Epistatic component subphenotype $k$  | $\phi_k$  | Dirichlet($\delta_k1, \ldots, \delta_{ke_k}$) |
| Phenotype of subphenotype $k$          | $\omega_k$ | Dirichlet($\omega_{k1}, \ldots, \omega_{kT}$) |
| Mixing proportions for markers of subphenotype $k$ | $\alpha_k$ | Dirichlet($\lambda_0, \lambda_1, \lambda_2$) |
| Mixing proportions for subphenotypes  | $\pi$     | Dirichlet($\rho_1, \ldots, \rho_K$) |

Table S1: Priors Distributions for Model Parameters: JBASE uses conjugate Dirichlet priors as the prior distribution for parameters governing its multinomial distribution components.

2.1 Derivation of the Posterior Distribution

Recall that in Section 2 of the main text, we derived the likelihood function as follows:

$$P(D=(G, Z)|C, S, Q) = P(G_0|\psi, C, S) \times \left\{ \prod_k P(G_k^1|\theta_k, C, S_k) P(G_k^2|\phi_k, C, S_k) P(Z^k|\omega_k, C) \right\}$$

, where we used $Q=\{\psi, \theta, \phi, \omega, \pi, \alpha\}$ for notational convenience. For the same reason, we use $H=\{\rho, \beta, \lambda, \gamma, \tau, \delta\}$ to represent the hyper-parameters of the model. Let us also introduce $V=\{\psi, \theta, \phi, \omega\}$ to denote $Q\setminus\{\pi, \alpha\}$.

Given these the posterior can be decomposed as follows:

$$P(S, C, Q|D, H) \propto P(D|C, S, Q) \times P(S, C, Q|H)$$

$$\propto P(D|C, S, Q) \times P(S|C, Q|H)$$

$$\propto P(D|C, S, Q) \times P(S, C|Q, H) \times P(Q|H)$$

$$\propto P(D|C, S, V) \times P(S, C|\alpha, \pi) \times P(V|H) \quad (1)$$

It is essential to recall that we are interested in inferring only the latent variables $S$ and $C$. We would like to achieve this without the need to infer the other model parameters $Q = \{\psi, \theta, \phi, \omega, \pi, \alpha\}$. This is because it is easier and more efficient to work with $P(S, C|D, H)$ than to work with $P(S, C, Q|D, H)$. As such we use a common trick often used in conjugate-based latent Bayesian models: we simply integrate out parameters that we are not interested in:
\[ P(S, C | D, H) \propto \int P(S, C, Q | D, H) dQ \]
\[ \propto \int P(D | C, S, V) P(V | H) dV \times \int P(C | \pi) P(\pi | \rho) d\pi \times \int P(S | \alpha) P(\alpha | \lambda) d\alpha \] (2)

The following key observations enable us to carry out the above integrations analytically.

1. Using conjugate priors, we obtain:
\[
\int P(C | \pi) P(\pi | \rho) d\pi = \frac{\Gamma(||\rho||_1)}{\Gamma(N + ||\rho||_1)} \prod_k \frac{\Gamma(n_k + \rho_k)}{\Gamma(\rho_k)}
\] (3)

and
\[
\int P(S | \alpha) P(\alpha | \lambda) d\alpha = \prod_k \int P(S_k | \alpha_k) P(\alpha_k | \lambda) d\alpha_k = \prod_k \frac{\Gamma(||\lambda||_1)}{\Gamma(M + ||\lambda||_1)} \left\{ \prod_j \frac{\Gamma(n_{kj}^m + \lambda_j)}{\Gamma(\lambda_j)} \right\}
\] (4)

2. Given \( C \), the genotype and phenotype likelihoods are independent:
\[
\int P(D | C, S, V) P(V | H) dV = \int P(\mathcal{G} | C, S, V) P(\mathcal{Z} | C, S, V) P(V | H) dV
\] (5)

3. Given \( S \) and \( C \), the likelihood decomposes over the \( k \) subpopulations as follows:
\[
\int P(D | C, S, V) P(V | H) dV = \int P(\mathcal{G}_0 | \psi, S, C) P(\psi | \beta) d\psi
\]
\[
\times \prod_k \left\{ \int P(\mathcal{G}_k^1 | C, S_k, V_k) P(\mathcal{G}_k^2 | C, S_k, V_k) P(\mathcal{Z} | C, S_k, V_k) P(V_k | H) dV_k \right\}
\] (6)

4. Given \( H \), the prior probability decomposes as follows:
\[
P(\mathcal{Q} | H) = \prod_m P(\psi_m | \beta_m) \times \prod_{m,k} P(\theta_{km} | \gamma_{k,m}) \times \prod_k P(\phi_k | \delta_k) \times \prod_k P(\omega_k | \tau_k)
\] (7)

Given Eqs. 3-7, we are ready to put the full posterior together:
\[
P(S, C|D, H) \propto \int P(D|C, V)P(V|H)d_V \times \int P(C|\pi)P(\pi|\rho)d_\pi \times \int P(S|\alpha)P(\alpha|\lambda)d_\alpha
\]

\[
\propto \prod_k \left\{ \int P(G^k|C, S_k, \theta_k)P(\theta_k|\gamma_k)d_{\theta_k} \right\} \times \left\{ \int P(G^2|C, S_k, \phi_k)P(\phi_k|\delta_k)d_{\phi_k} \right\}
\]

\[
\times \left\{ \int P(Z^k|C, S_k, \omega_k)P(\omega_k|\tau_k)d_{\omega_k} \right\} \times \left\{ \int P(G_0|\psi, S, C)P(\psi|\beta)d_{\psi} \right\}
\]

\[
\times \left\{ \int P(C|\pi)P(\pi|\rho)d_{\pi} \right\} \times \prod_k \left\{ \int P(S_k|\alpha_k)P(\alpha_k|\lambda)d_{\alpha_k} \right\}
\]

\[\text{marginal component} \quad \text{epistatic component} \quad \text{phenotype component} \quad \text{null component} \quad \text{priors of latent components}
\]

We are now ready to carry out the integrations analytically thanks to the conjugate priors:

\[
P(S, C|D, H) \propto \left\{ \frac{\Gamma(||\rho||_1)}{\Gamma(N + ||\rho||_1)} \prod_k \frac{\Gamma(\rho_k + n_{k})}{\Gamma(\rho_k)} \right\} \times \left\{ \frac{\Gamma(||\lambda||_1)}{\Gamma(M + ||\lambda||_1)} \prod_j \frac{3 \Gamma(n_{k}^{\text{pen}} + \lambda_j)}{\Gamma(\lambda_j)} \right\}
\]

\[
\times \prod_m \left\{ \frac{\Gamma(||\beta_m||_1)}{\Gamma(N_m + ||\beta_m||_1)} \prod_j \frac{3 \Gamma(\beta_{mj} + n_{mj}^{\text{null}})}{\Gamma(\beta_{mj})} \right\}
\]

\[
\times \prod_{k,m,S_{k,m}=1} \left\{ \frac{\Gamma(||\gamma_{km}||_1)}{\Gamma(N_{k} + ||\gamma_{km}||_1)} \prod_j \frac{3 \Gamma(\gamma_{kmj} + n_{kmj}^{\text{mar}})}{\Gamma(\gamma_{kmj})} \right\}
\]

\[
\times \prod_k \left\{ \frac{\Gamma(||\delta_k||_1)}{\Gamma(N_{k} + ||\delta_k||_1)} \prod_j \frac{3 \Gamma(\delta_{kj} + n_{kj}^{\text{pen}})}{\Gamma(\delta_{kj})} \right\}
\]

\[
\times \prod_k \left\{ \frac{\Gamma(||\tau_k||_1)}{\Gamma(N_{k} + ||\tau_k||_1)} \prod_l \frac{3 \Gamma(\tau_{kl} + n_{kl}^{\text{pen}})}{\Gamma(\tau_{kl})} \right\}
\]

\[\text{null component} \quad \text{marginal components} \quad \text{epistatic components} \quad \text{phenotype}
\]

\[2.2 \text{ Inference}
\]

We infer the latent variables \( C \) and \( S \) with Markov Chain Monte Carlo (MCMC), specifically with the Metropolis-Hastings (MH) algorithm.

At iteration \( l \), we update each \( C_i \) and \( S_{k,m} \) variables in turn as follows:
For each $C_i$:

- We uniformly randomly choose a new subphenotype $C'_i$, s.t. $C'_i \neq C_i$
- We next check the following constraints:
  
  * **Subphenotype size constraint**: If the new assignment causes any of two effected subpopulations to become smaller than the minimum allowed subpopulation size, then the proposal is rejected. In order to prevent over-fitting, we reject updates proposing to decrease subpopulation’s size below a given threshold, which we set to $\frac{N}{20}$ in our experiments.
  
  * **PCA constraint**: If the new assignment causes the mean value of the PCA values between the effected populations to differ beyond the allowed threshold in any of the PCA dimensions, the proposal is rejected. This is to prevent hidden confounding effects that cause subphenotypes. Given the PCA coordinates of each individual, as obtained from the PCA analysis of the genotype matrix, proposals that cause assignment of individuals to subpopulations to exceed a user-defined difference are rejected immediately. Unfortunately, this threshold is dependent on the data type as PCA coordinates are compared across different datasets. In our experiments, we set this threshold to 0.01 by analyzing the PCA main axis separation of discovered subphenotypes when PCA was not taken into account. Our experiments (See Section 4.3) and subsequent analysis of the PCA coordinates of the 10 most important dimensions, as well as linear mixed model based analysis, showed that this simple approach performs very well in practice.
  
  * **MAF constraint**: If a marker is assigned to a component as a marginal then its MAF is enforced to be higher in that component than in the rest of the population. This constraint is based on the assumption often used in GWAS studies that the minor allele of a gene associated with the disease, is more common in patients than in the controls. If the newly proposed $C_i$ invalidates this constraint, then the proposal is rejected.
  
  * **Hardy-Weinberg Equilibrium (HWE) constraint**: We require each marker to adhere to HWE over the subphenotypes in which it is assigned to the null component at $p<0.01$ significance level. If the proposed assignment violates this constraint, it is rejected.

- If the proposal is not rejected so far, $C_i$ is updated to the new state with probability

$$min\{1, \frac{P(C', S|D, H)}{P(C, S|D, H)}\}$$

For each $S_{km}$:

- We uniformly randomly choose a new genotype component $S'_{km} \neq S_{km}$
- We next check the following constraints:
* **Genotype size constraints:** If the new assignment causes the number of marginal and/or epistatic components to grow larger than the maximum size, then the proposal is rejected.

* **LD (Linkage Disequilibrium) constraint:** If the new assignment violates LD constraints specified by the distance threshold or the input LD pair list of markers, then the proposal is rejected. A suitable LD pairs list can be obtained by programs such as PLINK [1]. Alternatively, a distance threshold can be specified, below which two nearby markers are assumed to be in LD. Proposals that put more than one marker from an LD group of markers into an epistatic component are rejected as LD patterns induce false positive epistatic interactions and are immediately rejected.

* Proposals violating MAF and HWE constraints are rejected as well.
  - If the proposal is not rejected so far, $S_{km}$ is updated to the new state with probability
    \[
    \min\{1, \frac{P(C, S'|D, H)}{P(C, S|D, H)}\}
    \]
    These proposals satisfy the detailed balance condition and work well in practice. We monitor the convergence using trace plots, some of which are shown in Fig. S2. Although in our experiments we sampled 200,000 samples for each of the $S_{km}$ and $C_i$, more often than not, less than 100,000 was enough for obtaining good mixing.

Another important aspect of sampling is the acceptance/rejection rates of the proposal distribution. Due to several biological constraints described above, we obtained lower acceptance rates ($\approx 5\%$). Such low acceptance rates are expected given the low signal to noise ratio among the SNPs associated with the phenotype: it is expected that almost 99% of the SNPs belong to the null component whereas any one of them can be proposed to be marginal or epistatic. Analyzing the reasons for proposal rejections, we note that MAF and HWE constraints, which need to be checked for both $S_{km}$ and $C_i$, are responsible for $\approx 89\%$ of all proposal rejections. LD constraint, on the other hand, is responsible for the rejection of $\approx 5\%$ of the $S_{km}$ proposals. Similarly, subphenotype size constraint and PCA constraint are responsible for rejection of $\approx 1\%$ and $\approx 14\%$ of the remaining $C_i$ proposals respectively. Despite causing high rejection rates, these constraints ensure biologically valid states over $S_{km}$ and $C_i$. We efficiently calculate the constraints by tracking additional counters specific to each constraint without the need to go through the more costly operation of calculating the value of the posterior at the new state. These constraints are optional and can be turned on or off depending on the experiment. Finally, for computational efficiency, we apply constraints with higher rejection rates before the constraints with lower rejection rates. MAF constraint is an example of the higher rejection ratio constraint that is easy to calculate which prevents MCMC from exploring a large number of invalid states.

Finally, it is important to mention that, depending on the state of $S_k$, the size of the epistatic component
and hence the dimensionality of $\phi_k$, changes during sampling. However, since we integrate out $\phi_k$ and never perform inference on it, we do not need to resort to such algorithms as Reversible-jump MCMC [2].

![Figure S2: MCMC Convergence Analysis](image)

Figure S2: **MCMC Convergence Analysis**: Trace plot of (A) unnormalized $P(S,C|D,H)$ and (B) total number of epistatic and marginal markers and (C) ratio of the size of the larger subphenotype from a representative run of JBASE on the Mexico T1D dataset ($M=64$, $N=909$, runtime is $\approx 15$ hours).

### 2.3 Post-processing

JBASE has two goals: inferring the subphenotype assignments and inferring the causal markers (marginal and epistatic) that are responsible for the subphenotypes. We integrate out all model parameters except...
$\mathcal{C}$ and $\mathcal{S}$. Upon running MCMC, we obtain posterior distributions for these variables. Thus, we can now analyze how likely each patient is to belong to the same subphenotype as any other patient and what variants can explain genetic difference between the discovered subphenotypes. Depending on the final goal, the user of the method might be satisfied with such analysis of the posterior. We additionally illustrate a simple post-processing step to obtain clusters of patients. Such clustering maybe advantageous for translational purposes.

There is no consistent assignment of patients to classes across iterations, what matters is which patients are consistently assigned to the same subphenotype. To obtain the clusters, we calculate the co-assignment matrix of size $\mathcal{L}_{N \times N} : \mathcal{L}_{i,i'} = \frac{\# \text{ of times } \mathcal{C}_i = \mathcal{C}_{i'}}{\text{of all iterations} - 2000}$ and cluster the $N$ individuals using the co-assignment matrix with any standard clustering algorithm, such as the hierarchical clustering, into $K$ groups.

Fig. S3: **Co-clustering Analysis**: Hierarchical clustering analysis of co-clustering frequencies of individuals from the Mexico-1 dataset reveals two clusters of individuals corresponding to the discovered phenotypes.

Fig. S3 illustrates the efficiency of this heuristic in practice on the Mexico-1 dataset. Two clearly separated clusters of tightly co-assigned individuals, which are visible to the human eye, start to emerge starting somewhere between 10,000 and 20,000 iterations, and these clusters pretty much stay the same for the rest of the iterations (not shown here), highlighting the robustness of the discovery of subphenotypes.

With $\mathcal{S}_{km}$ we do not have the *identifiability* problem, as each each assignment type (*null*, *marginal* and *epistatic*) is modeled with non-identical distributions. As such, we simply assign marker $m$ to the most
frequently assigned genotype category. One problem with this approach is that it prevents assigning the discovered components to specific subgroups. For example, in the discovered epistatic modules, we do not attribute the component to a final subphenotype. However, this does not mean such a goal is unattainable.

It is worth mentioning that when we sample from the posterior distribution, we keep every \( d \)th sample after the first 2000 burn-in iterations. \( d \) is adjustable and such a thinning strategy \((d > 1)\) is not a hard requirement. We prefer to thin the samples for efficiency purposes: storing every sampling iteration would mean having to store \( K \times M \times U \), where \( U \) is the number of iterations, for tracking \( S \) and \( M \times U \) for \( C \). This can cause prohibitively large memory consumption. Thinning can greatly decrease the required memory.

### 2.4 Parameter Choices

The user needs to set only the hyper-parameters of the model since we integrate out most of the model parameters except \( S \) and \( C \), which are inferred by the algorithm. Unless there is background knowledge about the data that needs to be incorporated into the model, we suggest setting hyper-parameters as follows: \( \lambda = \{0.99, 0.005, 0.05\}, \beta = \{0.64, 0.31, 0.05\} \) and \( \gamma = \{0.25, 0.50, 0.25\} \). These hyper-parameters are set according to our prior expectations on marker group ratios (null, marginal and epistatic) in the case of \( \lambda \), allele frequencies (AA,Aa,aa) of a moderately frequent marker in the case of \( \beta \) and allele frequencies of a marginal marker in the case of \( \gamma \). Indeed, instead of using a single \( \beta \), it is also possible to use the population-specific background genotype distribution of each marker. Apart from these, the user can also specify maximum number of marginal and epistatic markers.

Perhaps the most important parameters to set are the number of subphenotypes \( K \) and subphenotype size distribution controlled by \( \rho \). Setting \( K \) is not trivial, and should be done according to prior expectation of disease phenotype heterogeneity. Considering the fact that sample sizes for GWAS are usually on the order of hundreds (or thousands) of patients, we do not recommend setting \( K \) to more than 2 unless there is strong prior knowledge about the number of subphenotypes in the data. Using higher values will reduce statistical power. In our simulation experiments with the \((500, 0, 500)\) setting, we showed that even with an incorrect value set for \( K \), JBASE can usually find the true subphenotype groups. Some deterioration happens with incorrect \( K \): when setting \( K=2 \) when in reality \( K=1 \), JBASE still performs as well as BEAM and much better than others. The \( \rho \) parameter, on the other hand, parameterizes the skewness of the subphenotype sizes. Setting \( ||\rho||_1 \) to a high value would represent the prior belief of subphenotypes of uniform sizes whereas setting it to a low value, i.e. \( N/100 \) would allow for some subgroups to be significantly larger than others. As our simulation experiments show, the method is robust against a variety of parameter settings: runs with the true hyper-parameters that are used in the data generation have virtually the same performance (See Fig. 2 of the main text) as the runs where we used default and/or highly divergent values.
3 Simulation Studies

3.1 Data Generation

In this section we describe the input parameters and the generative model that we used to simulate data from the disease models shown in Fig. S4. We start with input parameters, which can be broadly categorized into two groups. The first category of the parameters control the size aspects of the simulation, including number of markers, cohort size, number of subpopulations, etc. The second category of parameters relate to the the simulated variants and the disease models. These parameters control the MAF spectrum (of only null markers), risk penetrance, odds ratio spectrum, etc. The complete catalog of parameters, along with their functions, is listed in Table S2.

In our simulations, we let $\Delta$, one of the two risk parameters (See Fig. S4), vary between 0.1 and 1 as these values generate the most concordant causal markers, as measured by MAF and odds ratio, to the published GWAS catalog markers [3]. (See Fig. S5 for MAF and odds ratio spectrum of GWAS markers as well as our simulated markers). The marOdds parameter ensures that a causal marker’s marginal association is not too strong. This constraint is due to the observation that complex, non-Mendelian traits are a byproduct of small contributions of many loci and it is hardly ever the case that a single marker has a very high odds ratio. However, with low probability, causal markers with high a odds ratio are allowed. With this scheme, we are able to simulate causal markers from diverse MAF and odds ratio combinations and capture the two main characteristics of common complex traits; namely, the moderate minor allele frequency and low odds ratios that make causal markers hard to detect.

| Input | Category       | Description                                                                 |
|-------|----------------|------------------------------------------------------------------------------|
| K     | Size           | Number of case subphenotypes, i.e. not including the control population.    |
| N     | Size           | Total cohort size including all disease subpopulations and the controls.    |
| M     | Size           | Number of markers.                                                          |
| D     | Size           | Number of phenotype traits.                                                 |
| C     | Size           | A vector of size $1 \times D$ containing number of distinct categories for each phenotype trait. |
| minMAF| Causal variants| Minimum allele frequency of a causal marker in the control population. Default is 0.05. |
| maxMAF| Causal variants| Maximum allele frequency of a causal marker in the control population. Default is 0.45. |
| $\theta$ | Causal variants | Parameter to control the amount of disease risk, i.e. penetrance Odds ratio upper bound for causal markers. Default value is 2. |

Table S2: **Parameters of the Simulation**: Two broad categories of parameters used for the disease model simulations.
We perform the following steps to simulate our data:

1. Sample $\pi$, i.e. the prior for probability distribution of subpopulation size ratios. For example, $\pi = [0.5, 0.3, 0.2]$ defines a prior probability of 0.5 for the first subpopulation, 0.3 for the second and 0.2 for the third. We always take the first subpopulation as the control population, into which no causal variant is embedded, i.e. all genotypes are sampled from the null genotype model and the size of the control population is kept around half of the total cohort size. Size of case subpopulations are allowed to vary between 0% to 50% of the total population.

2. For each marker, sample $p_m$ representing MAF of marker $m$ in control population. We sample $p_m$ from the empirical MAF distribution of common variants used in GWAS studies, such as the Affymetrix 6 array’s SNPs’ MAF distribution. (See Fig. S5a).

3. Fill in the genotype matrix by sampling from the multinomial distribution parametrized by $(p_m^2, 2p_m(1-p_m))$.
4. Populate phenotype and the causal markers as follows:

(a) For each subpopulation:
   i. Choose a disease model and respective causal markers. This step is skipped for control population. As we have two case subpopulations, we have $5 \times 5 = 25$ disease risk model combinations to choose from for any given dataset.
   ii. For each marker of the current subpopulation, if the current marker is a causal one, overwrite the genotype of every individual of this subpopulation at this marker according to the disease model. This involves sampling genotypes from a multinomial distribution whose weight parameter is determined by the chosen disease model.
   iii. If the phenotype is categorical, sample $\omega_k$, which parametrizes the multinomial distribution of the phenotype. Otherwise, sample phenotype for each individual assigned to current subpopulation from the Normal distribution parameterized by the pre-determined hyper-parameters $\mu_k$ and $\sigma_k$. In our simulations we used $\mu_k \in (1, 2, 3)$ and $\sigma^2 = 0.5$.
   iv. If phenotype data is continuous, as is the case in our simulations, discretize it by uniformly binning across the range of the phenotype data.

(b) Check that the following integrity constraints are satisfied:
   - All markers are in Hardy-Weinberg Equilibrium (HWE) across the whole population. Causal markers are in HWE in the control population.
   - Odds ratios of causal markers are within the feasible ranges.

(c) Continue sampling until all constraints are satisfied.
Figure S5: **MAF and Odds Ratio Distributions:** (A) MAF spectrum of the Affymetrix-6 array used in the Mexico T2D experiments (B) Odds ratios distribution of the causal markers from our simulations (C) MAF distribution of the causal markers from our simulations (D) MAF distribution of the disease associated markers discovered by GWAS (January 2013) (E) MAF distribution of the disease associated markers discovered by GWAS (March 2015).

### 3.2 Additional Performance Results on Simulations

#### 3.2.1 Threshold Analysis

In Fig. 2 of the main text, we demonstrated each algorithms’ performance using the default thresholds when calling causal markers. For example BEAM uses a novel statistic called B-statistic, which was proposed as part of the algorithm, for calling markers rather than using posterior probabilities. For $\chi^2$-test based methods, we used multiple hypothesis corrected p-values, i.e. $p<0.01/(M=1000)$. In order to account for differences in the ways each algorithm assigns markers to the causal group(s), we also demonstrate results where we used comparable thresholds (See Fig. S6 and Fig. S7). For each algorithm, we found thresholds where every algorithm has average **Type-1 Error** of $5e−3$ and $2.5e−3$ as calculated over the 1000 datasets. Table S3 lists the thresholds we used for each of the algorithms. We chose $5e−3$ and $2.5e−3$ as target average...
Type-1 Error rates as these values fall into the range of Type-1 Error rates achieved by all algorithms.

When compared to performance results shown in Fig. 2 of the main text, the chosen thresholds cause JBASE and BEAM to have higher Type-1 Error and Power as we have to use much relaxed posterior probability thresholds to achieve the target Type-1 Error. On the other hand, algorithms which had relatively higher Type-1 Error in Fig. 2 of the main text, such as Multinom, have lower Power when stricter thresholds are used to obtain target Type-1 Error levels. Overall, we see that all algorithms’ Power performance remain qualitatively the same as Fig. 2. BEAM and JBASE perform the best across almost all cases. However, compared to Fig. 2, the Power gap between JBASE and BEAM shrinks, which indicates that BEAM’s performance can improve if one is willing to accept higher Type-1 Error.

| Algorithm    | Threshold Type      | Value for $5e^{-3}$ | Value for $2.5e^{-3}$ |
|--------------|---------------------|---------------------|-----------------------|
| BEAM         | Posterior Probability | 0.01                | 0.0059                |
| JBASE-Param1 | Posterior Probability | 0.023               | 0.0117                |
| JBASE-Param1 | Posterior Probability | 0.023               | 0.0116                |
| ChiSquare-CC | P-Value (Bonferroni) | 0.0352              | 2.75                  |
| ChiSquare-Multiway | P-Value (Bonferroni) | 1.75                | 4.07                  |
| Multinom     | P-Value             | 1e-6                | 0.014                 |
| OSACC        | P-Value             | 0.02                | 0.01                  |

Table S3: **Threshold Settings**: Types and value of the thresholds used for marker calling when fixing Type-1 Error rate of each algorithm.
Figure S6: Simulation Results for an average Type-1 Error=2.5e−3: Performance of all algorithms across four dimensions: subpopulation size (A,B), odds ratios (C,D), MAF (E,F) and disease model combinations (G,H) when using thresholds that achieves average Type-1 Error=2.5e−3.
Figure S7: Simulation Results for Average Type-1 Error = 5e−3: Performance of all algorithms across four dimensions: subpopulation size (A,B), odds ratios (C,D), MAF (E,F) and disease model combinations (G,H) when using thresholds that achieves average Type-1 Error = 5e−3.
3.2.2 Effect of MAF Distribution

In our simulations, disease models are parametrized by risk-related parameters $\alpha$ and $\Delta$ (See Fig. S4). Hence, we do not have direct control on MAF distribution of the embedded causal markers. We have originally performed the analysis under the MAF distribution shown in Fig. S5D with the corresponding power and Type-1 Error analysis shown in Figs. S8 A,B. Our results show that JBASE has the highest power, followed by BEAM. Type-1 Error is the best for BEAM, closely followed by JBASE and ChiSquare-Multiway. All three methods have negligible Type-1 Error.

We have consequently analyzed the performance of all methods under a more recent GWAS catalog MAF distribution (Fig. S5E). To that extent, we adaptively filtered some of the embedded markers such that the resulting MAF distribution of the embedded markers is close to Fig. S4D. We performed this random subsampling 100 times and report the average Power and Type-1 Error results in Fig. S8. In essence, this sampling approach increases the ratio of markers with low ($<0.1$) and high ($>0.4$) values. We see that all methods suffer from some decrease in power and increase in Type-1 Error. JBASE still achieves the best power, whereas BEAM maintains its lead on Type-1 Error.
Figure S8: Modified MAF Results: (A) Power and (B) Type-1 Error averaged over all datasets compared to (C) Power and (D) Type-1 Error where marker MAF distribution is modified to be close to real GWAS marker MAF distribution.

4 Type-2 Diabetes Experiments

4.1 Experimental Setup

The dataset, which we call the Mexico-1 dataset, consists of 983 T2D patients coming from mixed genetic backgrounds. Samples were genotyped for 315,658 common markers using the Affymetrix 5.0 platform and information on body mass index (BMI), waist, waist-to-hip ratio WHR, sex, age at diagnosis and at study time were collected. After careful preprocessing according to [5], we excluded the age phenotype as the study deliberately focused on the younger T2D patients and did not contain enough variation to detect age-related
subphenotyping effects. We focused on BMI and WHR only. We also filtered out 2547 ancestry markers reported in [5] as ancestry is strongly associated with diabetes and diabetes-related phenotypes in Mexican population. After preprocessing, P-value filtering and LD based clumping, we obtained 13,547 markers for JBASE analysis. Information on PCA coordinates for all samples was also available in order to test for confounding by population stratification. Fig. S9 illustrates our experimental setup and the details of the pipeline are as follows:

1. **Quality checks and data filtering:** In addition to preprocessing and quality controls performed in [5] (See Methods section), we undertook the following filtering and checks. We first removed individuals with missing phenotype data (n=74) and also markers with $MAF < 0.05$ or with Hardy-Weinberg deviation (P-value < 0.01). The filtering step resulted in n=909 individuals genotyped at 307,348 markers. The Mexico-2 dataset was processed in the same way.

2. **Data preprocessing:** On the phenotype side, we discretized the BMI and WHR values using equi-distant bins of 5 and 0.15, respectively, which resulted in seven categories for BMI and four for WHR. On the genotype side, we selected loci with P-value < 0.1 to obtain the set of 27,483 SNPs potentially enriched with true associations. For computational reasons, we further decreased this set to 13,547 using LD ($r^2 = 0.5$) and PLINK’s `clump` option (distance threshold: 250kb). We also acquired the scores of a PCA analysis to control for population stratification. Finally, we generated a list of markers that are in LD using PLINK’s `r^2` option with default distance option.

3. **Initial scan:** As the number of markers was large to perform detailed Bayesian analysis, we first applied JBASE to each pair of chromosomes independently to generate an initial set of candidate markers. We ran JBASE on each pair over 10 replicates each lasting 200,000 sampling iterations. Markers that were classified as marginal or epistatic in at least 10 runs were selected as candidate markers. From this step, we obtained 64 candidate SNPs, which we characterize below.

4. **Pooled analysis:** We finally pooled the 64 candidates and ran 100 concurrent MCMC chains, each with 200,000 iterations, on the pooled set. We then manually analyzed each run to discover the convergence pattern. We report the most frequently found pattern consisting of 2 epistatic modules (characterized below). The motivation for this step is that our model allows for only one epistatic and one marginal module per subphenotype and pooling all candidate markers into a joint run allows us to find the most significant association.
4.2 Analysis of the Candidate Markers

Running JBASE on \( \frac{22 \times (22-1)}{2} = 231 \) possible pairs of non-sex chromosomes resulted in 64 markers that were called *epistatic* or *marginal* in at least 10 runs. Before advancing to final marker analysis, we first analyzed this subphenotyping-based SNP set (SPS) and compared it to the complete set (CS) of 307,348 markers included on the SNP chip and against the tagSNP set (TS) containing 13,547 markers selected based on LD and *marginal* P-value analysis. We aimed to find whether the discovered candidates showed any overlap to previous GWAS findings, and if they did, what type of associations they were reported in. In addition, we compared these two sets of markers in terms of association enrichment to demonstrate that our candidate set of subphenotyping markers set (SPS) inferred by JBASE does not contain any bias due to
To quantify the chances of having the observed number of associations by chance and to rule out any confounding effects, we performed association enrichment analysis. Our selection of tagSNPs. As such, we performed association enrichment analysis to quantify the chances of having the observed number of associations by chance and to rule out any confounding effects.

| Phenotype                       | P-Value (SPS vs TS) | P-Value for (SPS vs CS) |
|---------------------------------|---------------------|------------------------|
| Aorta Abdominal                 | 0.00e+00*           | 0.00e+00*              |
| Diabetic Nephropathies          | 7.81e-09*           | 0.00e+00*              |
| Coronary Disease                | 0.00e+00*           | 0.00e+00*              |
| Adiposity                       | 7.39e-07*           | 0.00e+00*              |
| Insulin Resistance              | 0.00e+00*           | 0.00e+00*              |
| Cornea                          | 1.16e-05*           | 0.00e+00*              |
| Lipoproteins HDL                | 3.18e-05*           | 0.00e+00*              |
| Lipoproteins                    | 2.80e-04*           | 0.00e+00*              |
| Obesity                         | 5.41e-09*           | 0.00e+00*              |
| Lipids                          | 0.00e+00*           | 0.00e+00*              |
| Aorta                           | 2.16e-06*           | 0.00e+00*              |
| P-Selectin                      | 8.40e-03*           | 2.45e-14*              |
| Carotid Artery Diseases         | 1.17e-05*           | 0.00e+00*              |
| Diabetic Retinopathy            | 1.22e-02*           | 7.62e-14*              |
| Hip                             | 0.00e+00*           | 0.00e+00*              |
| Coronary Artery Disease         | 8.31e-04*           | 0.00e+00*              |
| Creatinine                      | 2.48e-06*           | 0.00e+00*              |
| Triglycerides                   | 0.00e+00*           | 0.00e+00*              |
| Kidney Diseases                 | 3.99e-02*           | 1.12e-09*              |
| Blood Pressure                  | 1.82e-14*           | 0.00e+00*              |
| Nonalcoholic Fatty Liver Disease| 7.15e-02*           | 4.47e-08*              |
| Cholesterol LDL                 | 9.40e-11*           | 0.00e+00*              |
| Body Mass Index                 | 6.60e-08*           | 0.00e+00*              |
| Body Weight                     | 7.04e-05*           | 0.00e+00*              |
| Cholesterol HDL                 | 2.75e-07*           | 0.00e+00*              |
| Cholesterol                     | 7.18e-04*           | 0.00e+00*              |
| Body Height                     | 3.94e-03*           | 0.00e+00*              |
| Diabetes Mellitus Type 2        | 2.48e-01*           | 1.27e-07*              |

Table S4: Association Enrichment Analysis: χ² test-based enrichment analysis of marker groups based on comparison of rate of published associations within 500kb neighbourhood of JBASE identified markers to that of (column-1) tagSNPs marker set (TS, m=13,547) and (column-2) the full marker set (CS, m=307348). Asterisk indicates significant enrichment at 0.01 level.

For this, we obtained all inferred and published associations from dbSNP’s association track [6] and analyzed their neighborhoods in terms of proximity to members of the three marker groups described above. We marked a variant as true positive association (in the context of this analysis) if and only if it contains a published association with a given phenotype within 500kb. Under this scheme, we performed χ² test to compare the number of true positives associations to that expected by chance. Table S4 summarizes our findings. Basically, we observed significant enrichment for almost all phenotypes for which our candidate markers have a supporting proximal association, indicating that JBASE does a superior job of identifying...
markers just by subphenotyping T2D patients. These results are important for three reasons. First, it proves that the superior enrichment is not due to any bias in construction of the CS and SPS sets of markers. This is because of the fact that for almost all phenotypes, SPS is more enriched compared to CS. Second, the high enrichment achieved and the types of phenotypes for which enrichment is achieved demonstrate that our model is able to capture the heterogeneity of the T2D case cohort. Indeed, as we showed in the introduction section, classical GWAS can easily miss many causal markers in the case of heterogeneous disease groups. As no other method, to our best knowledge, analyzed this dataset with a subphenotyping approach, some non-overlap was expected. However, the high level of agreement with previous studies implies that JBASE does a superior job of picking up good signals. Note that most of the markers JBASE identified were not found to be significant in the original study [5].

4.3 Genotyping Quality, LD and Stratification Analysis of subphenotypes

To ensure that the discovered subphenotypes are not artifacts of data quality, LD and/or population stratification, we first investigated the genotype calling quality of the discovered markers. The separation in the genotype clusters indicates the error/bias-free genotyping of the four markers (See Figs. S15 and S16). Next we checked the haplotype structure in close and larger neighborhoods spanning the regions covered by the discovered markers using the Mexican population of HapMap-3 dataset [7, 8, 9]. Our analysis showed that for our SNPs, LD decays very rapidly $r^2 < 0.1$ within at most 500kb (See Fig. S12). Both regions are spanned by several small and seemingly independent haplotypes (See Figs. S13 and S14) suggesting the markers are not part of long-range haplotypes (LRH), which are usually hot-spots for recent positive selection. We also analyzed several studies, including HapMap-2 dataset, that published candidate LRHs both in Mexican and European populations and did not find any overlap, which provides further evidence against possible confounding due to LD [10, 11, 12, 13].

We also rigorously checked for potential confounding due to population stratification since the Mexico population has a high admixture rate. To do this, we first obtained the BMI and WHR of all samples as well as the PCA components for population stratification. Comparing the discovered subphenotypes across the first 10 principal axes showed that the partitioning is not due to population stratification (See Fig. S10). The insignificant P-values, using Wilcoxon rank sum test, prove that the discovered subpopulations do not differ from each other due to systematic genotype differences in the population, thereby proving JBASE’s efficacy for dealing with population stratification even in highly admixed populations. Moreover, analysis of the discovered subphenotypes against cryptic relatedness using the linear mixed model Fast-LMM showed that association p-values do not change after taking cryptic relatedness into account [14]. Here we calculated the kinship matrix with the full set of SNPs before clumping and association filtering. This analysis strongly suggests that the discovered subphenotypes are not confounded by hidden relatedness of the individuals in our study.
Finally, we used model-based clustering (mclust) [15] on phenotype data directly. We found that phenotype-only analysis failed to find stable subphenotypes or recover the subsets we found with JBASE. While analyzing markers’ association for the subphenotypes discovered with mclust we did not observe any statistically significant marginal or epistatic associations. These experiments suggest that our discovered subphenotypes are jointly driven by phenotypic and genotypic data.

Figure S10: Principal Component Analysis of the Mexico-1 Dataset: Analysis of the first 10 major PCA dimensions reveals no significant difference in the genetic background of the discovered subphenotypes Mexico-1 samples.
Figure S11: **Principal Component Analysis of the Mexico-2 Dataset:** Analysis of the first 10 major PCA dimensions reveals no significant difference in the genetic background of the discovered subphenotypes Mexico-2 samples.
Figure S12: LD Decay Patterns of the Epistatic Markers: Analysis of LD patterns of the four discovered epistatic markers reveals that LD drops to $r^2 < 0.1$ within immediate 500kb neighborhoods. The plots are generated with SNAP software [9] using Mexican population of the HapMap-3 [7] dataset.
Figure S13: **Haplotype Structure of Genomic Regions Covering Chr-13 Markers**: Haplotype structure analysis reveals that the region covering the members of the Chr-13 epistatic module contains several small haplotypes, providing further evidence for lack of an LRH (long-range haplotype). The plot is obtained from the HaploView software [8] using Mexican population of the HapMap-3 [7] dataset.
Figure S14: **Haplotype Structure of Genomic Regions Covering Chr-19 Markers:** Haplotype structure analysis reveals that the region covering the members of the Chr-19 *epistatic* module is covered by several haplotypes.
4.4 Genotyping Quality Analysis of the Replication Markers

We have plotted the cluster graphs obtained from Affymetrix’s Axiom array analysis platform for the four markers used in replication. Our aim is to make sure that there are no problems with the genotyping, such as inaccurate SNP calling and plate effects (See Figs. S15 and S16). The genotype clusters are well separated indicating that there were no problems while calling the genotypes for the SNPs in focus.
Figure S15: **Genotype Cluster Plot for the Mexico-1 dataset Markers**: Well separated genotype clusters indicate the error/bias-free genotyping of the four markers: (A) rs1159752 (B) rs4885712 (C) rs12461255 (D) rs8103847 used for replication.
Figure S16: **Genotype Cluster Plot for the Mexico-2 Dataset Markers**: Well separated genotype clusters indicate the error/bias-free genotyping of the four markers: (A) rs4805561 (B) rs4932867 (C) rs1929045 (D) rs4885712 used for replication.
Figure S17: **Genotype Frequency Analysis of Epistatic Modules**: Genotype frequencies of the rs1159752 and rs4885712 epistatic marker pair in the larger (n=631) obese (A) and the smaller (n=278) leaner T2D subphenotypes (B) and their difference (C). Genotype frequencies of the rs8103847 and rs12461255 epistatic marker pair in the larger (n=631) obese (D) and the smaller (n=278) leaner T2D subphenotypes (E) and their difference (F). P-values indicate the epistatic association p-values ($\chi^2$ test).
References

[1] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–575.

[2] Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82: 711–732.

[3] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. PNAS 106: 9362–9367.

[4] Zhang Y, Liu JS (2007) Bayesian inference of epistatic interactions in case-control studies. Nature Genetics 39: 1167–1173.

[5] Parra E, Below J, Krithika S, Valladares A, Barta J, et al. (2011) Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. Diabetologia 54: 2038-2046.

[6] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29: 308–311.

[7] The International HapMap Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

[8] Barrett JC, Fry B, Maller JB, Daly M (2005) Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics 21: 263-265.

[9] Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O’Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–2939.

[10] The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

[11] Koch E, Ristroph M, Kirpatrick M (2013) Long range linkage disequilibrium across the human genome. PLoS ONE 8: e80754.

[12] Sabeti P, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

[13] Zhang C, Bailey DK, Awad T, Liu G, Xing G, et al. (2006) A whole genome long-range haplotype (wglrh) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122-2128.
[14] Listgarten J, Lippert C, Kadie C, Davidson R, Eskin E, et al. (2012) Improved Linear Mixed Models for Genome-Wide Association Studies. Nature Methods 9: 525-526.

[15] Fraley C, Raftery AE (1999) Mclust: Software for model-based cluster analysis. Journal of Classification : 297–306.
Table S5: Discovered markers: List of markers discovered from paired chromosome JBASE runs.

| SNP        | Chr | Location | Band | SNP        | Chr | Location | Band |
|------------|-----|----------|------|------------|-----|----------|------|
| rs4652099  | 1   | 173895830| q25.1| rs7538555  | 1   | 197529784| q32.1|
| rs10458590 | 1   | 203087879| q32.1| rs2320170  | 2   | 95603500 | q11.1|
| rs1160717  | 3   | 26389087 | p24.2| rs4682869  | 3   | 42898968 | p22.1|
| rs1527731  | 3   | 146764751| q24  | rs874147   | 4   | 90217281 | q22.1|
| rs2089990  | 4   | 93519212 | q22.1| rs11097407 | 4   | 95361588 | q22.2|
| rs7696248  | 4   | 101479821| q23  | rs994370   | 4   | 107143075| q24  |
| rs7664966  | 4   | 114277896| q26  | rs11098747 | 4   | 120224514| q26  |
| rs27412    | 5   | 111460574| q22.1| rs271128   | 6   | 133285984| q23.2|
| rs3107921  | 7   | 109204944| q31.1| rs7791565  | 7   | 114459956| q32  |
| rs1178439  | 8   | 15840128 | p22  | rs10867160 | 9   | 90991958 | q22.1|
| rs10283758 | 9   | 97040020 | q22.3| rs16919341 | 9   | 102520998| q31.1|
| rs7696959  | 9   | 121013663| q33.1| rs1041205  | 9   | 115464443| q32  |
| rs11010958 | 10  | 37283282 | p11.2| rs10776638 | 10  | 49590037 | q11.22|
| rs4112348  | 10  | 56318960 | q21.1| rs17816753 | 10  | 61931788 | q21.2|
| rs10997469 | 10  | 68416291 | q21.3| rs1080101  | 11  | 44000798 | p11.2|
| rs7929962  | 11  | 68742159 | q13.2| rs2851631  | 11  | 97308924 | q22.1|
| rs1948122  | 11  | 103391981| q22.3| rs2298511  | 11  | 10984756 | q22.3|
| rs1947234  | 12  | 37760758 | q12  | rs9527650  | 13  | 56860414 | q21.1|
| rs9540720  | 13  | 65820076 | q21.3| rs1324790  | 13  | 77639935 | q22.3|
| rs4885712  | 13  | 79737477 | q31.1| rs5096067  | 13  | 84292808 | q31.1|
| rs7332079  | 13  | 87969047 | q31.2| rs1159752  | 13  | 89977669 | q31.3|
| rs7982823  | 13  | 93632252 | q31.3| rs910790  | 15  | 38403706 | q15.1|
| rs2056493  | 15  | 43435564 | q21.1| rs2314277  | 15  | 52058085 | q21.3|
| rs7175117  | 15  | 59493888 | q22.2| rs12912276 | 15  | 66268636 | q23  |
| rs12445900 | 16  | 11312331 | p13.13| rs4782010  | 16  | 17316373 | p12.3|
| rs4356470  | 16  | 57087116 | q21  | rs9385666  | 16  | 64053820 | q21  |
| rs1055165  | 16  | 72012519 | q22.3| rs12604413 | 18  | 33879616 | q12.2|
| rs672862   | 18  | 39269881 | q12.3| rs11661691 | 18  | 45024184 | q21.1|
| rs12461255 | 19  | 23551796 | p12  | rs8103847  | 19  | 39501287 | q12  |
| rs12626370 | 21  | 25747994 | q21.2| rs2833117  | 21  | 31122843 | q22.11|
| rs8130115  | 21  | 34601958 | q22.12| rs2837348  | 21  | 40281706 | q22.2|
| rs16990991 | 22  | 42490017 | q13.2|
Table S6: **Association Enrichment Analysis:** χ² test-based enrichment analysis of marker groups based on comparison of rate of published associations within 500kb neighborhood of JBAS identified makers to that of (column-1) tagSNPs marker set (TS, m=13,547) and (column-2) the full marker set (CS, m=307,348). Asterisk indicates significant enrichment at 0.01 level.

| Phenotype                        | P-Value (SPS vs TS) | P-Value for (SPS vs CS) |
|----------------------------------|--------------------|------------------------|
| Aorta Abdominal                  | 0.00e+00*          | 0.00e+00*              |
| Diabetic Nephropathies           | 7.81e-09*          | 0.00e+00*              |
| Coronary Disease                 | 0.00e+00*          | 0.00e+00*              |
| Adiposity                        | 7.39e-07*          | 0.00e+00*              |
| Insulin Resistance               | 0.00e+00*          | 0.00e+00*              |
| Cornea                           | 1.16e-05*          | 0.00e+00*              |
| Lipoproteins HDL                 | 3.18e-05*          | 0.00e+00*              |
| Lipoproteins                     | 2.80e-04*          | 0.00e+00*              |
| Obesity                          | 5.41e-09*          | 0.00e+00*              |
| Lipids                           | 0.00e+00*          | 0.00e+00*              |
| Aorta                            | 2.16e-06*          | 0.00e+00*              |
| P-Selectin                       | 8.40e-03*          | 2.45e-14*              |
| Carotid Artery Diseases          | 1.17e-05*          | 0.00e+00*              |
| Diabetic Retinopathy             | 1.22e-02*          | 7.62e-14*              |
| Hip                              | 0.00e+00*          | 0.00e+00*              |
| Coronary Artery Disease          | 8.31e-04*          | 0.00e+00*              |
| Creatinine                       | 2.48e-06*          | 0.00e+00*              |
| Triglycerides                    | 0.00e+00*          | 0.00e+00*              |
| Kidney Diseases                  | 3.99e-02*          | 1.12e-09*              |
| Blood Pressure                   | 1.82e-14*          | 0.00e+00*              |
| Nonalcoholic Fatty Liver Disease | 7.15e-02           | 4.47e-08*              |
| Cholesterol LDL                  | 9.40e-11*          | 0.00e+00*              |
| Body Mass Index                  | 6.60e-08*          | 0.00e+00*              |
| Body Weight                      | 7.04e-05*          | 0.00e+00*              |
| Cholesterol HDL                  | 2.75e-07*          | 0.00e+00*              |
| Cholesterol                      | 7.18e-04*          | 0.00e+00*              |
| Body Height                      | 3.94e-03*          | 0.00e+00*              |
| Diabetes Mellitus Type 2         | 2.48e-01*          | 1.27e-07*              |