Supplemental Methods

Simulation Strategy for Clustering Pattern

Every set of configurations includes three parameters: sample size, core variant odds ratio, and percentage of influential variants. For each set of configurations, we designed a simulation pipeline that generated the odds ratio for each variant, genotype, and phenotype profiles for samples and ran multiple independent tests to obtain the empirical power (Supplemental Fig S1C).

To construct an influential clustering in the protein PDB:2OGV, we sample the odds ratio for all variants mapped to the protein (Supplemental Fig S1A) with (1). The core variant has the largest log odds ratio, and the rest have a log odds ratio dependent on their distance from the core variant. t is the parameter controlling the radiant effect from the core variant, which defaults to 7Å. r_i is the distance of a surrounding variant to the core variant in angstroms (Å) within the protein. We set the cut-off radius as 14Å. Any variants beyond 14Å from the core variants are considered neutral and assigned an odds ratio of 1.

$$\beta(v_i) = \beta_0 e^{-\left(\frac{r_i^2}{2t^2}\right)} \quad (1)$$

We generated genotypes with the following strategies. We randomly selected 50 variants from the protein PDB:2OGV. The log minor allele frequencies for all the variants are randomly sampled from a uniform distribution with an interval (-4, -2.3). When the population size was determined, we used a binomial distribution to obtain the overall minor allele count in the population. The minor alleles would be randomly assigned to subjects. Given the genotypes sampled and odds ratio determined by the clustering pattern, we simulated the phenotype based on the logistic regression. The probability of individual *j* being a case $y_j = Pr(1|G_j)$ is shown in (2). Then the binary phenotype of case and control would be sampled from a Bernoulli distribution with the probability equivalent to y_j. G_j is the genotype vector for individual *i*. β_0 is associated with the population prevalence and is commonly estimated by $\sum_j Pr(1|G_j) - \beta_0$. However, since we only

simulate rare variants, the difference between $\sum_j Pr(1|G_j)$ and β_0 is less than 10% of the β_0. Therefore, we approximate the population prevalence to β_0.

$$log\left(\frac{y_j}{1-y_j}\right) = \beta_0 + \beta_j G_j^T \quad (2)$$

We used empirical power to determine the performance of POKEMON and other methods. The significance level was 0.05. We derived the empirical power by running 100 independent tests and calculating the percentage of tests with a p-value within the significance level. For each independent test, the odds ratios of the variants were fixed, while the variants being sampled were randomly generated. For POINT, a successful test is where any variant that passes a Bonferroni threshold is influential (i.e., the simulated odds ratio for the variant is larger than 1).

Simulation Strategy for Dispersive Pattern

We carried out the same strategy to simulate dispersive patterns for each set of configurations. The odds ratio for all variants was assigned with the same value (e.g., 1.1) (Supplemental Fig S1B). We randomly selected 30 variants from the protein PDB:2OGV. The log minor allele frequencies for all the variants were randomly sampled from a uniform distribution with an interval (-4, -2.3). When the population size was determined, we used a binomial distribution to sample the overall minor allele count and randomly assigned the allele copies to simulated subjects. We simulated the binary phenotype case and control for all subjects with the same strategy used in simulating the clustering pattern. The empirical power was obtained by the percentage of successful tests out of 100 independent tests.

Cluster identification

POKEMON first classified the variant by the percentage of case subjects that carry it. Variants carried by more than 50% of the case subjects were classified as case variants; otherwise, they were classified as control variants. Next, POKEMON clustered on case variants and control variants, respectively. DBSCAN clustering algorithm was adopted here with the maximum distance of 14Å and the minimum number of neighborhoods as 5. Clusters with case variants

were classified as case clusters, and clusters with control variants were classified as control clusters.