SUPPLEMENTARY MATERIAL

Please note that notation is carried over and equation numbering is referenced from the main text.

Section 1. Statistical Test for Interaction

Under the null assumption that distal SNPs segregate independently in the population (and are therefore in linkage equilibrium), the expected $\vec{1}$-frequency for a pair of distal variables in cases can be estimated by $\mathrm{E}\left[P_{\vec{v}}(\vec{1})\right] = P_v P_{v'}$: where P_v and $P_{v'}$ are the empirical 1-frequencies of v and v' respectively. Positive LD in cases results in an increase in the frequency of $\vec{1}$-carriers ($P_{\vec{v}}$), where LD is measured as a difference between the observed and expected frequency ($D_{\vec{v}}^{case} = P_{\vec{v}} - P_v P_{v'}$). If we denote $\Delta_{\vec{v}}^{case} = N D_{\vec{v}}^{case}$ as the number of excess $\vec{1}$-carriers in cases, then we derive the statistic

$$\left(LD_{\vec{v}}^{\text{case}}\right)^2 = \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_vP_{v'}} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_v(1-P_{v'})} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{N(1-P_v)P_{v'}} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{N(1-P_v)(1-P_{v'})}$$

$$= \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_v(1-P_v)P_{v'}(1-P_{v'})} \sim \chi^2 \; with \; 1 \; d.o.f.$$

From which, we get

$$LD_{\vec{v}}^{case} = \frac{D_{\vec{v}}^{case}}{\sigma_{\vec{v}}^{case}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{case} = \sqrt{\frac{P_v(1-P_v)P_{v'}(1-P_{v'})}{N}}$$

eq. S1.1

A similar analysis for controls gives us

$$LD_{\vec{v}}^{control} = \frac{D_{\vec{v}}^{control}}{\sigma_{\vec{v}}^{control}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{control} = \sqrt{\frac{p_v(1-p_v)p_{v'}(1-p_{v'})}{n}}$$

eq. S1.2

Consequently, the LD in cases and controls can be contrasted to derive an LD-contrast statistic

$$LD_{\vec{v}}^{diff} = \frac{\left(D_{\vec{v}}^{case} - D_{\vec{v}}^{control}\right)}{\sigma_{\vec{v}}^{diff}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{diff} = \sqrt{\left(\sigma_{\vec{v}}^{case}\right)^2 + \left(\sigma_{\vec{v}}^{control}\right)^2}$$

eq. S1.3

Under our null-hypothesis, we would expect to see no difference in LD between cases and controls,

$$\mathbb{H}_0 : LD_{\vec{v}}^{diff} = 0$$

Significant variable-pairs $\vec{v}$ are those for which $LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}$, where the $z_{\mathcal{B}} = \Phi^{-1}(1 - \mathcal{B})$ represents the number of standard deviations of the standard normal distribution $\mathcal{N}(0,1)$ required to achieve a significance level of $\mathcal{B}$. In the interest of clarity, we use the Bonferroni significance level without loss of generality (any other multiple test correction approach can be plugged in as easily). For example, in a dataset of $M = 450{,}000$ SNPs, if we perform $\left(\binom{450{,}000}{2} \times 4\right)$ pairwise tests (4 models tested per SNP-pair as per our binary encoding) genome-wide, giving us a significance threshold of $p = 1.2 \times 10^{-13}$. The LD-contrast cutoff required to achieve this significance level is $z_{\mathcal{B}} \approx 7$.

Section 2. Stage-1 filtering step

Consider a common disease with prevalence τ in the population. The LD between any two variables $\vec{v} = (v, v')$ in the entire population can be considered a mixture of two distributions

$$D_{\vec{v}}^{pop} \sim \tau\, D_{\vec{v}}^{case} \;+\; (1-\tau) D_{\vec{v}}^{control}$$

eq. S2.1

Assuming that physically unlinked alleles are in population-wide linkage equilibrium – i.e. $D_{\vec{v}}^{pop} \approx 0$ - we estimate that $\mathbb{E}\left[D_{\vec{v}}^{control} | D_{\vec{v}}^{pop} \approx 0 \right] = \frac{-\tau}{(1-\tau)} \mathbb{E}\left[D_{\vec{v}}^{case} | D_{\vec{v}}^{pop} \approx 0 \right]$. If variable-pairs exceed a disequilibrium cutoff $LD_{\vec{v}}^{case} \geq z_B'$ in cases (i.e. if $D_{\vec{v}}^{case} \geq z_B' \sigma_{\vec{v}}^{case}$), then for the variables to remain in population-wide equilibrium, the expected reverse disequilibrium in control required to counter the imbalance created by these cases is $\mathbb{E}\left[D_{\vec{v}}^{control} \mid D_{\vec{v}}^{pop} \approx 0,\, D_{\vec{v}}^{case} \geq z_B' \sigma_{\vec{v}}^{case} \right] \leq \frac{-\tau}{(1-\tau)} z_B' \sigma_{\vec{v}}^{case}$. Substituting in eq.1 (main text), we get

$$\mathbb{E}\left[LD_{\vec{v}}^{diff} \mid D_{\vec{v}}^{pop} \approx 0 \right] \;\geq\; \frac{1}{(1-\tau)} \times \frac{\sigma_{\vec{v}}^{case}}{\sigma_{\vec{v}}^{diff}} \times z_B'$$

eq.S2.2

For significant pairs $\left(LD_{\vec{v}}^{diff} \geq z_B \right)$, by assuming the marginal frequencies of both variables are approximately equal in cases and controls (i. e. $P_v \approx p_v$ and $P_{v'} \approx p_{v'}$), we get

$$z_B' \;\geq\; (1-\tau) \times \sqrt{\frac{N+n}{n}} \times z_B$$

eq.S2.3

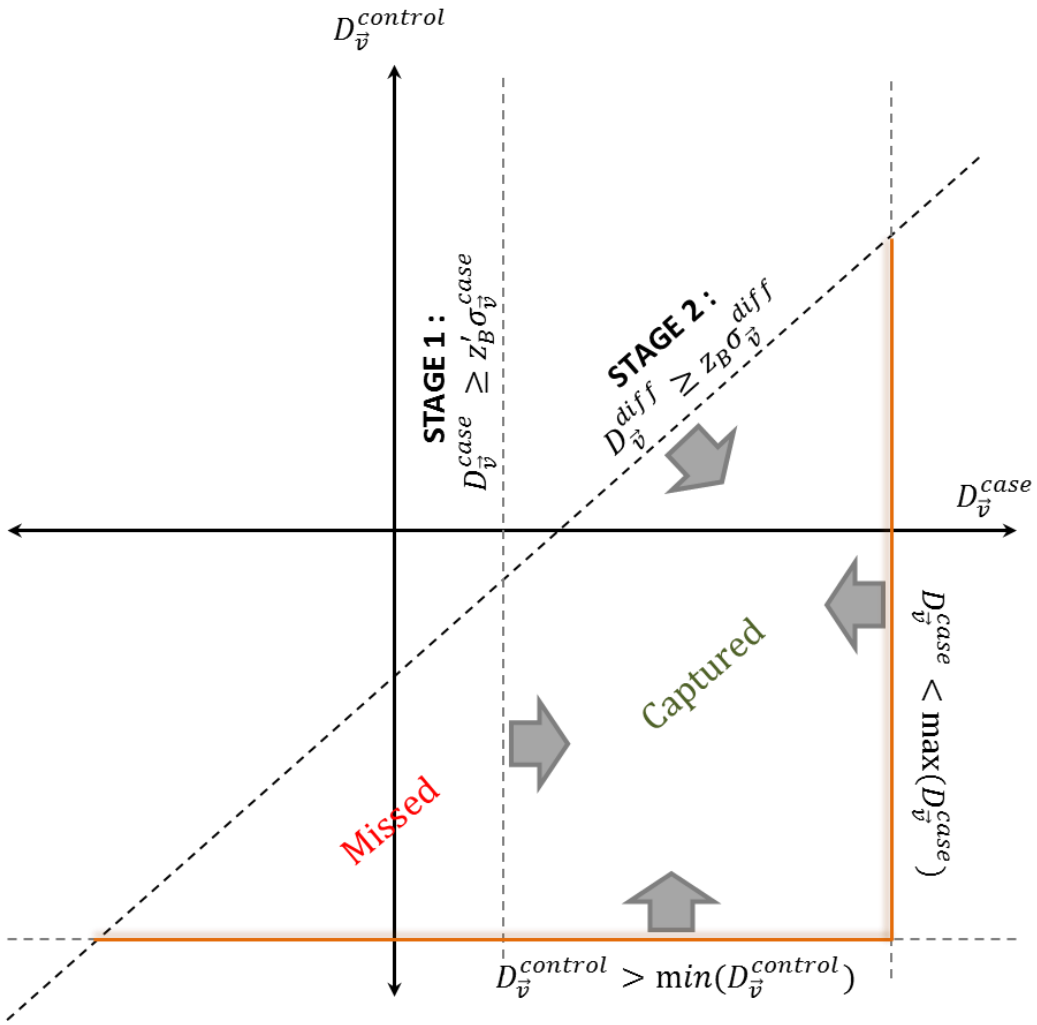
This result expresses the disequilibrium cutoff z_B' in cases as a function of the disequilibrium-contrast cutoff z_B in cases versus controls. To extend the example in Supplementary Section 1, if $z_B \approx 7$ is the LD-contrast Bonferroni cutoff for a common disease with population prevalence 5% in a dataset with an equal number of cases and controls, from eq.S2.3 we can estimate that in cases $z_B' \geq 9.4 \geq z_B$.

We say a stage 1 case-only analysis is *approximately complete* under this prescription, because while we have determined the expected value of the statistic for significant pairs, we have not characterized the full distribution.

This approximation depends on standard assumptions, particularly that most alleles have similar frequencies in cases and controls. For interactions between variables with large causal or protective marginal signals, we make the following observations: (i) For variables whose minor allele is enriched in cases (i.e. causal association) the approximation is actually violated in our favor: it underestimates our power to find interactions. (ii) For the converse case, where the minor allele is depleted in cases (i.e. protective association), the approximation does indeed falter. However, using an ultra-conservative approach in which we lower the stage 1 cutoff for candidates to be the same as the stage 2 cutoff (i.e. $z_B' = z_B$), is observed to be sufficient to accommodate such violations in practice. In other words, we can largely capture interactions between SNPs whose frequency is greater in cases as well as greater in controls (SNPs with main effects). Lastly and importantly, these loci have typically already been identified by single-locus association and are therefore accessible to a candidate gene based analysis.

Section 3. <u>The approximate nature of a stage-1 case-only analysis</u>

The area enclosed within the orange lines represents the region occupied by significant variable-pairs in linkage disequilibrium space (LD in cases on the X-axis, controls LD on the Y-axis). The three sides of the triangle are the maximum (i.e. positive) LD in cases, the minimum (i.e. negative) LD in controls and the LD-contrast threshold used to demarcate significant pairs. The dotted vertical line represents the stage 1 cutoff used for shortlisting tuples through a case only analysis. The number of significant pairs in the area labeled "Missed" and "Captured" depends upon the density distribution of pairs in this space. We show that for common diseases almost all statistically significant pairs lie in the captured region (see section on Approximately Complete Search).



Section 4. Applying group sampling to a genome-wide scan.

In the toy example described in the main text, we restricted our discussion to finding pairs of variables that occupy a narrow frequency window w. To generalize the approach to a genome-wide search for significant LD-contrasts, we first partition the entire spectrum of frequencies $[0,1]$ into R windows $W = \{w_0, \ldots, w_{r-1}\}$ of ranges $\mathrm{E} = \{\epsilon_0, \ldots, \epsilon_{r-1}\}$ respectively, where each $w_r = [\eta_r, \eta_{r+1})$; $\eta_{r+1} = \eta_r + \epsilon_r$ and $\eta_0 = 0, \eta_r = 1$. We then allocate each of the $2M$ variables genome-wide to their appropriate frequency windows (Supplementary Section 8). As before, the number of variables in a window w_r is denoted $V(w_r)$, and every variable is assigned to exactly one window: $\sum_r |V(w_r)| = 2M$. In practice, we find that using around 50 windows is adequate to cover the frequency spectrum even in large datasets of $\geq 10^{5-6}$ SNPs.

Consider any pair of windows $\{w_A, w_B\}$. There are $\binom{W}{2} + W$ such window pairs (including the possibility that A=B). For all $\vec{v} = (v_A, v_B)$ comprising of one variable $v_A \in V(w_A)$ and the other $v_B \in V(w_B)$, the minimum and maximum expected $\vec{1}$-frequencies can be derived as per eq.5 (main text). If $\mathbb{H}_0'$ holds for $\vec{v}$, then $(P_{\vec{v}})^k \leq (\eta_{A+1}\eta_{B+1})^{2k}$. If $\mathbb{H}_0'$ is rejected by $\vec{v}$ then $(P_{\vec{v}})^k \geq (\eta_A\eta_B + \delta_{A\times B})^{2k}$, where $\delta_{A\times B}$ is derived as

$$\delta_{A\times B} = \sqrt{\frac{\eta_A(1-\eta_A)\eta_B(1-\eta_B)}{N}}\, \mathrm{z}_{\mathcal{B}}$$

eq.S4.1

We are required to find the group-sampling parameter values $(k_{A\times B}, t_{A\times B})$ for this window pair, at which we can guarantee that all significant-pairs which reject $\mathbb{H}_0'$ are observed in at least one group with probability greater than the user-specified threshold $(1-\beta)$. Furthermore, our solution $(k_{A\times B}, t_{A\times B})$ has to be "optimal" in two ways : (i) the false positive rate should be low (which requires number of individuals per draw - $k_{A\times B}$ - to be large) because these candidates will have to be kept in memory, only to be screened out later by stage 2, and (ii) the solution should not consume too many compute cycles (large $k_{A\times B}$ requires a large number of random draws $t_{A\times B}$ to achieve the desired power, which in turn drives up the number of compute cycles). Details of the optimization procedure we employed to find the best $k_{A\times B}$ and $t_{A\times B}$ are provided below.

To summarize, group-sampling lets us restrict our test to an exponentially (in $k_{A\times B}$) small fraction $f_{A\times B} \leq t_{A\times B}.(\eta_{A+1}\eta_{B+1})^{2k_{A\times B}}$ of the pairs of variables in $V(w_A) \times V(w_B)$ at minimum computational

"cost" (as discussed in the optimization section below), and simultaneously guarantees that all significant variable pairs will be captured in this fraction of the universe with power $\geq (1-\beta)$. This makes stage 1 of our search experiment extremely rapid.

Although the sheer size of the universe of combinations $U_{AB} = |V(w_A) \times V(w_B)|$ can suggest a large number of false-positives αU_{AB} in stage 1 overall. We make three observations to alleviate this concern: (i) This constitutes an upper bound which (by definition) is rarely encountered in empirical data, (ii) Most false-positive pairs are observed in more than one sampled group. However, these are stored in memory using a hash-table the very first time they are encountered, and have to be tested only once by the stage 2 analysis. The upper bound appears large because it does not account for such "over counting", and finally (iii) A poor stage 1 false-positive rate comes at a computational cost, but does not affect the accuracy of the algorithm. False-positive candidates like these are screened out in stage 2.

Details of the optimization procedure:

Translating eq.5 (main text) to our current setting, and expressing k as a function of t gives,

$$k_{A\times B}(t) \leq \frac{\log(1-\beta^{1/t})}{\log(\eta_A \eta_B + \delta_{A\times B})}$$

eq.S4.2

There are a total of $\binom{|V(w_A)|+|V(w_B)|}{2}$ potential variable-pairs between these windows. The number of pairs that emerge purely by chance over $t_{A,B}$ random draws is estimated as

$$Chance_{A\times B}(t) \approx \binom{|V(w_A)| + |V(w_B)|}{2} . t_{A\times B} . (P_{\vec{v}})^{k_{A\times B}}$$

And since $(P_{\vec{v}})^{k_{A\times B}} \leq (\eta_{A+1}\eta_{B+1})^{2k}$ for these chance pairs, we get

$$Chance_{A\times B}(t) \leq \binom{|V(w_A)| + |V(w_B)|}{2} . t_{A\times B} . (\eta_{A+1}\eta_{B+1})^{2k_{A\times B}}$$

eq.S4.3

This gives us an upper bound on the number of pairs that turn up purely by chance. We confirmed this bound in practice: since most co-occurring variables are encountered in several random draws, they need not be investigated more than once if we record them in a hash-table in memory. We can now find the optimal parameter values $k_{A\times B}$ and $t_{A\times B}$ that satisfy eq.S4.2 while minimizing the overall cost function,

$$k_{A\times B}, t_{A\times B} \longleftarrow \mathrm{argmin}_t(\lambda_1 Chance_{A\times B}(t) + \lambda_2 t + \lambda_3 \mathbb{I}(k < 2))$$

eq.S4.4

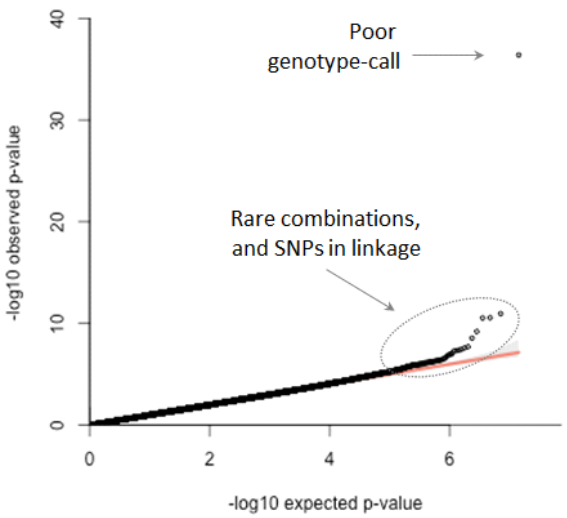
where λ_1, λ_2 and λ_3 are the cost of shortlisting and validating a chance pair, the cost of a random draw, and a Lagrange multiplier to avoid degenerate values of k respectively, while $\mathbb{I}(.)$ is the indicator function. These costs depend on the particular software implementation and data-structures used.

Section 5: QQ plots for LD-contrast test (sub genome-wide)

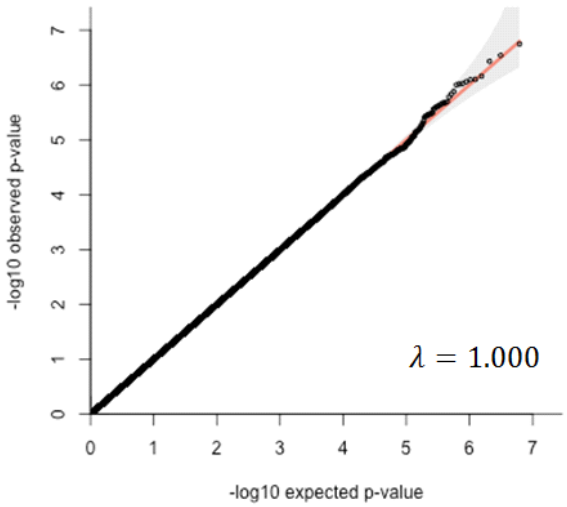
We drew 15 million random pairs of binary variables $\vec{v} = (v, v')$ from the cleaned WTCCC dataset, and contrasted their LD between bipolar cases and joint controls. We consider this a representative sample of the full space of 400 billion pairs which is computationally difficult to test. We observed over-dispersion that suggested a deviation from the null-hypothesis. We filtered out pairs with an extremely low expected number of minor allele co-carriers (i.e. remove pairs with $NP_vP_{v'} < 4$ or $np_vp_{v'} < 4$ in cases or controls respectively) because these might inflate the statistic due to unstable variance estimates (denominators in equations. S1.1, S1.2 and S1.3). Further, we filtered out pairs comprising of SNPs in genetic linkage (<5cM apart) which cannot be treated as independent random variables when calculating co-carrier expectations of the 2×2 table : we observed an over-dispersion of LD-contrast p-values on random pairs of such physically linked SNPs. However, since interactions between nearby markers (e.g. neighboring genes or markers within a gene) quite possibly comprise a significant portion of the interaction space, modifications to the test that can adjust for this variance inflation due to multi-collinearity are the subject of future work. Lastly, in addition to WTCCC prescription, we removed SNPs whose CHIAMO genotype-calling confidence was <95% in >1% of the individuals of the dataset. Conservative filters help us avoid false positives when we report pairs in the genome-wide significance range ($p < 10^{-12 \text{ to } 13}$). The resulting QQ plots show that for pairs that pass our filters, the LD-contrast test operates on a robust null-hypothesis and does not suffer from any residual over dispersion in BD data ($\lambda = 1.000$).

We note here that each QQ plot only presents a subset of randomly chosen variable-pairs out of the potential 4.15×10^{11} pairs that exist genome-wide in this dataset. It is computationally prohibitive to test (on the order of) a trillion pairs of variables, sort their p-values and plot as many points without specialized software and computational infrastructure (indeed, avoiding this is the primary motivation of our work) but genomic over-dispersion (inflation of the median value) can be estimated robustly with a representative sample of the universe of pairs. Additionally, for SNP-pairs that do pass the Bonferroni cut-off genome-wide, we also perform a permutation analysis to verify their significance.

Before Cleaning



After Cleaning



Section 6: <u>Synthetic dataset construction</u>

We tested the accuracy of the randomization algorithm under various simulated scenarios. In particular, we were interested in whether SIXPAC always finds SNP-pairs with a significant LD-contrast level at (or above) the computational power requested by user.

Using the original WTCCC BD case-control cohort, we simulated 3 datasets to contain SNP-pairs with significant LD-contrast. These datasets capture a range of different scenarios concerning disease prevalence levels in the population (1% to 25%), minor allele frequencies of interacting SNPs (5% to 40%), as well as mode of interaction (recessiveness and dominance). The datasets were synthesized through a technique called *chromosomal shuffling*, as follows:

i. First choose one SNP on each chromosome (All MAF 40% for dataset 1, MAF 30% and MAF 10% on alternate chromosomes for dataset2, and all MAF 5% for dataset3).

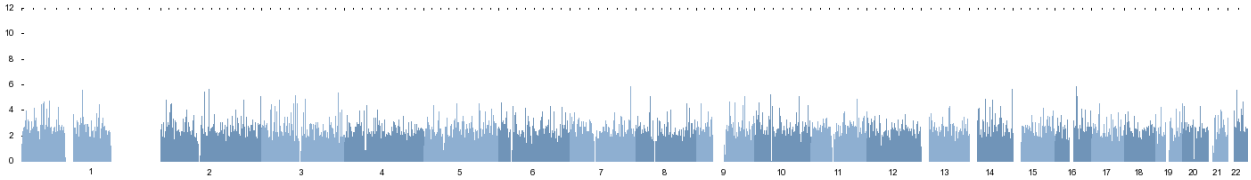
ii. Ascertain that these SNPs presented no discernible marginal significance (p >0.1) according to five standard single-locus association tests (allelic, genotypic, trend, dominance and recessive) offered by Plink (Purcell et al. 2007).

iii. Next, we introduce LD-contrasts into the dataset without changing the marginal frequencies of the SNPs either in cases on controls. To do this, we swap entire chromosomes among cases (controls) so as to increase (decrease) the number of co-carriers of minor alleles in these cohorts. The ratio of cases/controls to shuffle during each iteration is determined by the prevalence estimate.
For example, we create one additional recessive-recessive co-carrier in the case dataset (without affecting the marginal signal) we follow these steps.
Let case 2 be recessive at SNP A (chr21) and case 3 be recessive at SNP B (chr22), then:
(i) swap chromosome 21 of case 1 and case 2
(ii) swap chromosome 22 of case 1 and case 3.
Case 1 is now a carrier of the recessive-recessive pair at SNPs A and B. The controls can have the number of co-carriers depleted by analogous shuffling.

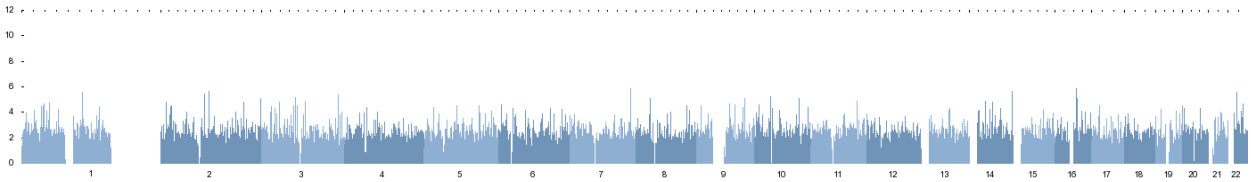
iv. By shuffling case and control chromosomes in this manner, we simulated 11 interactions (between SNPs on 22 autosomes) in each of the 3 datasets – each interaction at different levels of LD-contrast significance from $p = 10^{-2}, 10^{-4}, \ldots, 10^{-22}$ (decrements of $\approx 10^{-2}$). Note that because of the

discrete nature of the shuffle, it is not always possible to achieve accurate LD-contrast p-values in the synthetic data (e.g. $p = 1.0 \times 10^{-2}, 1.0 \times 10^{-4}, \ldots, 1.0 \times 10^{-22}$). Instead, after each swap we perform an LD-contrast test between this SNP-pair to check if the co-carrier imbalance introduced between cases and controls is sufficient to provide the required level of significance. We stop when we cross this level.

Chromosomal shuffling allows us to effectively manipulate LD between SNPs in cases (and controls) without changing the marginal association signal at all. This can be verified by checking that the Manhattan plot (single-locus, allelic model) ***before*** chromosomal shuffling in cases



And the corresponding Manhattan plot ***after*** simulating recessive-recessive co-carriers in the cases are exactly identical.



Although differentiating LD in the case and control datasets is not intended to directly confer statistical epistasis, we can also analyze these simulated SNP-pairs using the traditional model for interaction in a case-control study. This involves applying logistic regression, which tests whether the two loci - when considered in conjunction - result in a deviation from multiplicative odds.

First we test whether the SNPs in the synthetic datasets have any main effects – by testing the term β_1 or β_2 term in a logistic regression $\ln(\frac{P}{1-P}) = \beta_1 \mathrm{G}_1 + \beta_2 G_2$, where G_1 and G_2 are binary predictor variables that encode :

i. recessive carrier status for interacting SNPs in dataset1
i. recessive and dominant carrier status respectively for interacting SNPs in dataset2, or
ii. dominance carrier status for interacting SNPs in dataset3.

As the tables below confirm, in all 3 scenarios, LD-contrast does not inflate main-effect estimates. Next we test for the multiplicative interaction effects – by testing for significance of the β_{12} term in a logistic regression using the full model: $\ln\left(\frac{P}{1-P}\right) = \beta_1 \mathrm{G}_1 + \beta_2 G_2 + \beta_{12} G_1 G_2$. We note that increasing LD-contrast is strongly indicative of increasing statistical epistasis on this scale (see 3 dataset tables below), although the correlation between the 2 tests is not perfect (see discussion elsewhere (Purcell; Kam-Thong et al. 2010)). Fully elucidating the wide range of models and alternate parameterizations that may be visible through such LD-contrasts is the subject of future work.

Dataset 1. Common × Common Interaction

LD-contrast simulated between a 40% MAF SNP (in recessive mode) with a 40% MAF SNP (in recessive mode), disease prevalence 25%.

Simulated interactions (approximate significance)	Main effects models		Full model	Empirical significance (p-value)	
	odds ratio (e^{β_1})	odds ratio (e^{β_2})	odds ratio ($e^{\beta_{12}}$)	β_{12} term (epistasis)	LD-contrast
chr 1 – chr 2 (10^{-2})	1.0	1.01	1.6	0.01	8.2E-03
chr 3 – chr 4 (10^{-4})	0.99	1.0	2.0	2.0E-04	6.8E-05
chr 5 – chr 6 (10^{-6})	1.0	1.0	2.6	3.7E-06	6.0E-07
chr 7 – chr 8 (10^{-8})	1.0	1.0	2.9	1.5E-07	8.2E-09
chr 9 – chr 10 (10^{-10})	1.0	1.0	3.5	1.7E-09	4.2E-11
chr 11 – chr 12 (10^{-12})	1.0	0.99	4.0	1.2E-11	8.9E-13
chr 13 – chr 14 (10^{-14})	1.0	0.99	4.2	2.6E-12	7.2E-15
chr 15 – chr 16 (10^{-16})	0.99	1.0	4.5	3.6E-14	5.5E-17
chr 17 – chr 18 (10^{-18})	1.0	1.0	5.3	3.2E-16	8.2E-19
chr 19 – chr 20 (10^{-20})	1.0	0.99	6.3	<2E-16	5.7E-21
chr 21 – chr 22 (10^{-22})	1.0	1.0	6.5	<2E-16	3.3E-23

Dataset 2: Rare × Common Interaction.

LD-contrast simulated between a 10% MAF SNP (in dominant mode) with a 30% MAF SNP (in recessive mode), disease prevalence 10%.

Simulated interactions (approximate significance)	Main effects models		Full model	Interaction significance (p-value)	
	odds ratio (e^{β_1})	odds ratio (e^{β_2})	odds ratio ($e^{\beta_{12}}$)	β_{12} term (epistasis)	LD-contrast
chr 1 – chr 2 (10^{-2})	1.01	0.99	1.7	0.01	8.8E-03
chr 3 – chr 4 (10^{-4})	1.01	1.0	2.3	4.0E-04	7.9E-05
chr 5 – chr 6 (10^{-6})	0.99	0.99	2.8	4.7E-06	5.9E-07
chr 7 – chr 8 (10^{-8})	1.01	1.0	3.4	2.4E-07	6.9E-09
chr 9 – chr 10 (10^{-10})	1.0	0.99	3.8	2.7E-09	3.5E-11
chr 11 – chr 12 (10^{-12})	1.0	1.0	4.2	2.6E-10	6.9E-13
chr 13 – chr 14 (10^{-14})	1.01	0.99	5.4	1.6E-12	6.7E-15
chr 15 – chr 16 (10^{-16})	0.99	0.99	5.7	5.7E-14	5.8E-17
chr 17 – chr 18 (10^{-18})	1.0	1.0	5.6	1.3E-14	5.3E-19
chr 19 – chr 20 (10^{-20})	1.0	1.0	6.0	7.6E-16	9.9E-21
chr 21 – chr 22 (10^{-22})	0.99	1.0	7.1	<2E-16	3.0E-23

Dataset 3: Rare × Rare Interaction.

LD-contrast simulated between a 5% MAF SNP (in dominant mode) with a 5% MAF SNP (in dominant mode), disease prevalence 1%.

Simulated interactions (approximate significance)	Main effects models		Full model	Interaction significance (p-value)	
	odds ratio (e^{β_1})	odds ratio (e^{β_2})	odds ratio ($e^{\beta_{12}}$)	β_{12} term (epistasis)	LD-contrast
chr 1 – chr 2 (10^{-2})	1.0	1.0	1.9	0.017	9.4E-03
chr 3 – chr 4 (10^{-4})	1.01	1.04	2.8	4.4E-04	5.2E-05
chr 5 – chr 6 (10^{-6})	1.0	0.99	3.1	1.8E-05	6.4E-07
chr 7 – chr 8 (10^{-8})	1.0	1.0	3.7	7.9E-07	4.2E-09
chr 9 – chr 10 (10^{-10})	0.99	0.99	4.0	9.7E-08	8.5E-11
chr 11 – chr 12 (10^{-12})	1.0	1.01	5.3	4.6E-10	2.9E-13
chr 13 – chr 14 (10^{-14})	0.99	1.01	5.1	8.2E-11	7.5E-15
chr 15 – chr 16 (10^{-16})	1.0	1.0	6.2	3.0E-12	8.1E-17
chr 17 – chr 18 (10^{-18})	1.0	1.0	6.3	1.7E-13	1.9E-19
chr 19 – chr 20 (10^{-20})	1.0	1.0	6.5	7.9E-14	3.4E-21
chr 21 – chr 22 (10^{-22})	1.0	0.99	7.9	5.1E-16	4.0E-23

Section 7: <u>Power of Algorithm</u>

To confirm that theoretical estimates of algorithm power were matched or exceeded by our implementation, we tested SIXPAC on the three simulated datasets, each containing 11 pairwise SNP-SNP interactions (LD-contrast) at different levels of significance as described in Supplementary Section 3.

SIXPAC accepts two critical inputs from the user, based on which it calculates search parameters

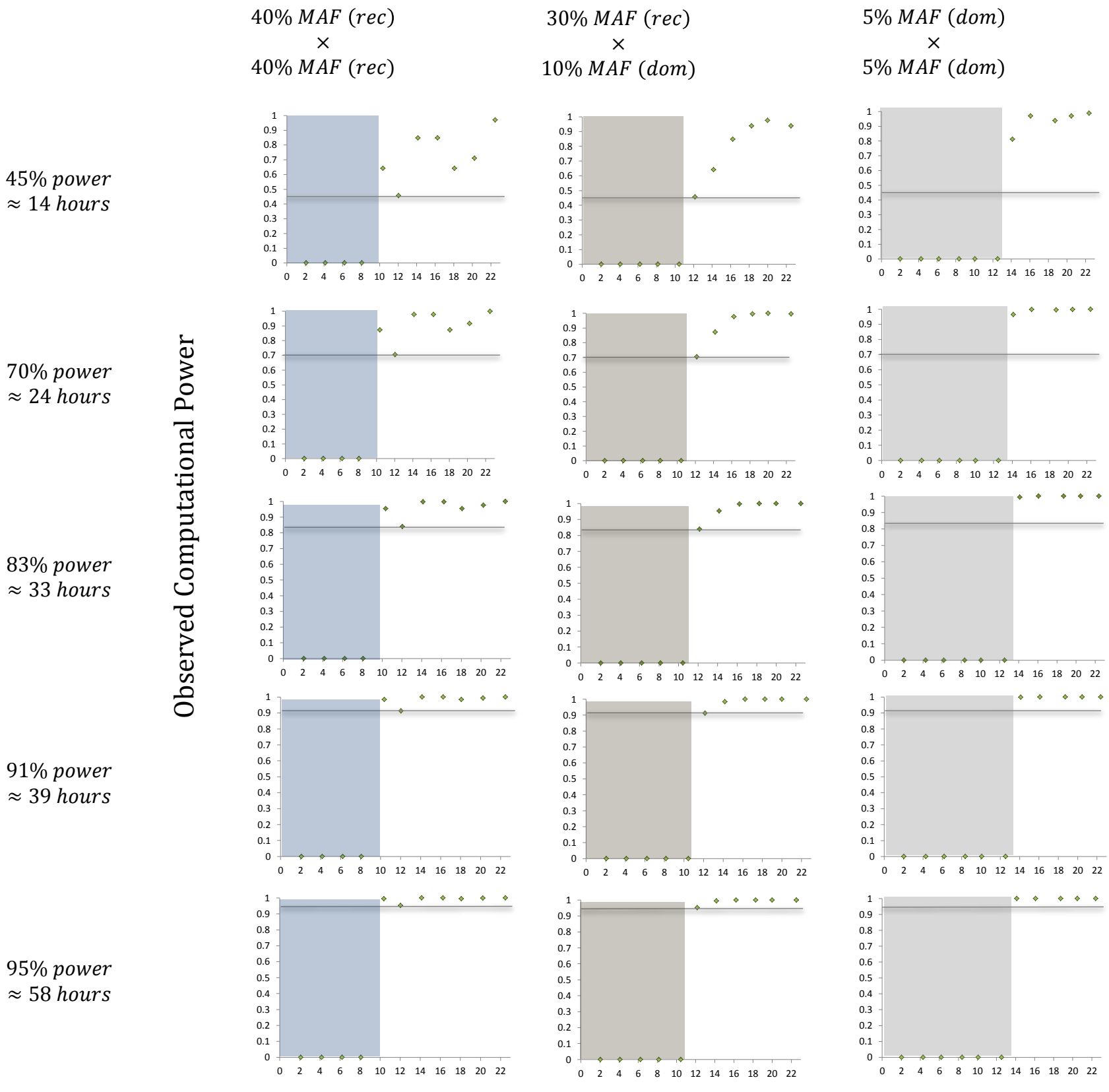
1. Significance cutoff as a p-value – all LD-contrasts above this cutoff must be reported.
2. Power (probability) to find these significant pairs, demanded by the user.

For the purposes of this simulation experiment, we arbitrarily defined the LD-contrast significance cutoff at 3 different realistic values of $p < 10^{-10}, 10^{-11}$ and Bonferroni (1.2×10^{-13}) for datasets 1, 2 and 3 respectively. We note that any arbitrary cutoff value, lower or higher than these values, can be provided by the user. For the computational power parameter, we measured results over 5 different realistic values – 45%, 70%, 83%, 91% and 95% probability respectively. Here, power of the algorithm is defined as the probability of finding all SNP-pairs in the dataset with a significant LD-contrast. As we discussed in the main text, this is different from statistical power.

Each panel in the figure below represents the result of a SIXPAC run with a particular combination of power and significance cut-off. The shaded rectangle in each panel represents the significance cut-off : interactions below this threshold are not reported. The solid line represents the theoretical power required by the user – and guaranteed as per theoretical estimates. We wish to determine whether the interactions to the ***<u>right</u>*** of the shaded area are ***<u>above</u>*** the cutoff threshold line, as promised.

In the panels below, each interaction is represented by a green dot: the X-axis co-ordinate gives the – log(p) value of its LD-contrast, while its Y-axis co-ordinate gives the average observed probability of spotting the interaction by SIXPAC (100 runs) - under each particular power, cut-off setting. We can see that as per guarantees, pairs with an LD-contrast above the significant cut-off are always reported with probability greater than the user-prescribed baseline.

For each dataset, SIXPAC scanned approximately 400 billion pairwise tests (4 tests per SNP-pair). We report the times taken by each SIXPAC run on a Single Intel i7 processor (quad-core) with 8GB RAM alongside. We note that like any randomization algorithm, SIXPAC will require an infinite amount of compute time to reach 100% certainty of finding everything in a dataset, but can approach close to 100% with large compute savings.


40% MAF (rec) × 40% MAF (rec)
30% MAF (rec) × 10% MAF (dom)
5% MAF (dom) × 5% MAF (dom)
45% power ≈ 14 hours
70% power ≈ 24 hours
83% power ≈ 33 hours
91% power ≈ 39 hours
95% power ≈ 58 hours
Observed Computational Power
− log(pvalue) of synthetic interactions

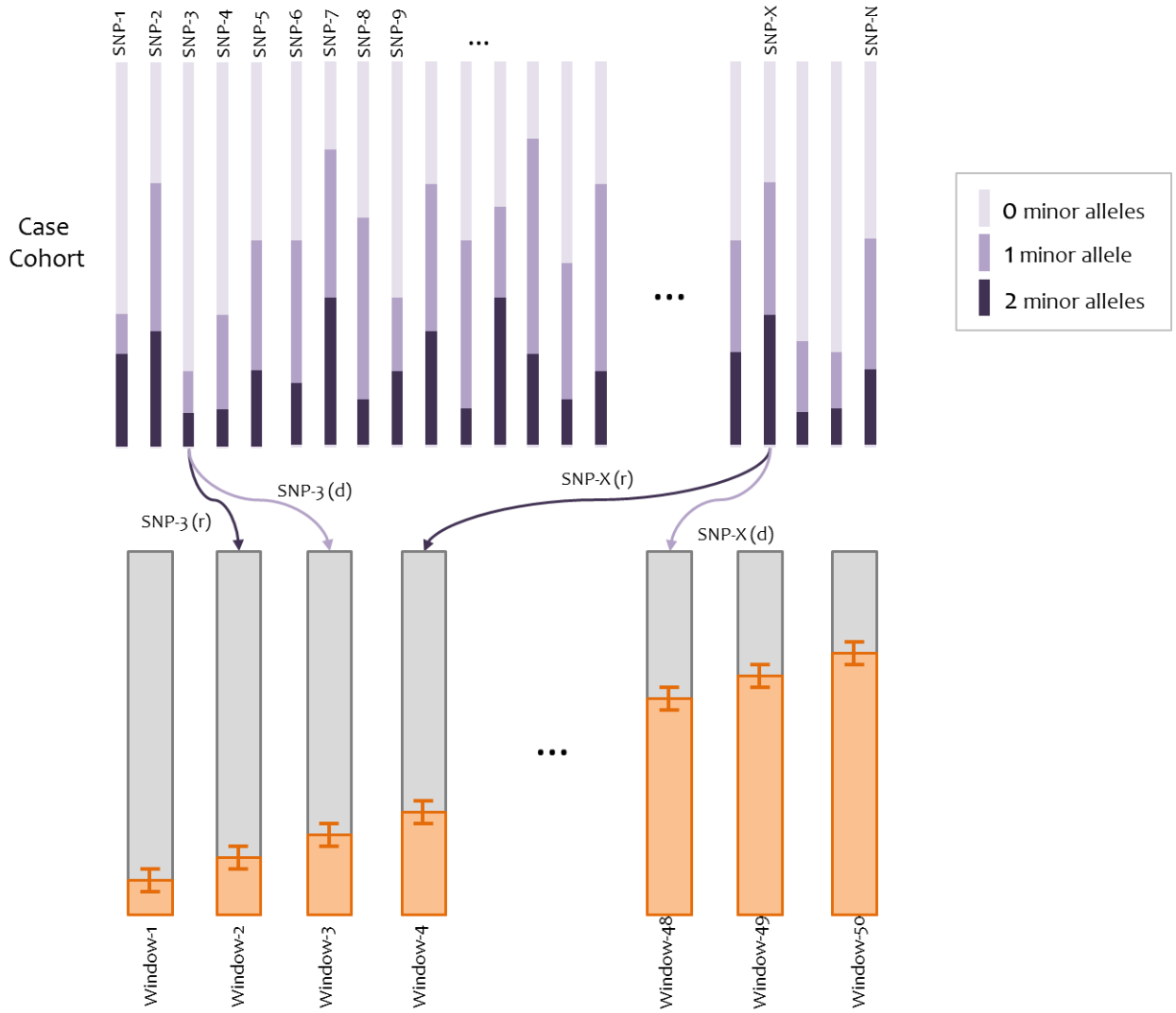
Section 8. <u>Frequency Binning</u>

For each SNP, we consider the empirical frequency of the 2 encoded binary variables in cases (recessive and dominance carrier status). Each variable is then assigned to a narrow frequency bin, as shown. Our algorithm operates by considering pairs of windows: since LD is a function of the frequency of two variables, we can conversely estimate how frequent a combination would need to be in order for the LD to be statistically significant. Group sampling exploits this difference in frequency between pairs with significant LD and pairs in equilibrium to rapidly shortlist interaction candidates.

The optimal width of a frequency window is difficult to characterize analytically : it depends on the significance cut-off, statistical test being implemented, number of SNPs typed in the dataset as well as the number of samples.

A. On the one hand, having many windows with a narrow frequency range makes it easy to distinguish between statistically significant LD - $\left(\tilde{P}^2 + \delta_{w\times w}\right)$ - and a SNP-pair that is at the upper end of the frequency spectrum - $\left(\tilde{P} + \epsilon\right)^2$. On the positive side, this reduces the number of shortlisted candidates per window-pair. However, many narrow windows means a quadratic increase in the number of window-pairs that have to be considered – and each pair must go through millions of group-sampling iterations, which can be computationally expensive.
B. On the other hand, having fewer but wider frequency windows does not allow group sampling to distinguish between a pair with a statistically significant increase in frequency, and a pair that is at the upper end of the permitted frequency spectrum - $\left(\tilde{P}^2 + \delta_{w\times w}\right)$ and $\left(\tilde{P} + \epsilon\right)^2$ respectively, from the main text. This can result in a large false positive rate at the stage 1 shortlisting step.

For WTCCC size case-control datasets, using the LD-contrast test, at a significance level of $p < 10^{-12}$, we found that using 50 to 60 windows provided the best performance.


Case
Cohort
SNP-1
SNP-2
SNP-3
SNP-4
SNP-5
SNP-6
SNP-7
SNP-8
SNP-9
...
SNP-X
SNP-N
0 minor alleles
1 minor allele
2 minor alleles
SNP-3 (r)
SNP-3 (d)
SNP-X (r)
SNP-X (d)
Window-1
Window-2
Window-3
Window-4
Window-48
Window-49
Window-50

Section 9. Numerical Example: Detecting a strong joint effect between rare alleles in a large dataset

We describe a particular example of a joint-effect we pursue, in order to provide sense of the actual numbers involved. Consider a realistic GWAS dataset of 10,000 case and 40,000 control samples from a population. If the disease prevalence in this population is 4%, then cases are oversampled 5-fold by the ascertainment of this study.

Consider two unlinked SNPs of 5% MAF each (in HWE, same MAF in both cases and controls and thus no marginal signal at either). The dominant-variable of each SNP (which encodes whether an individual carries ≥ 1 minor alleles at the SNP) has a frequency of 9.75% in both datasets, and hence under the null hypothesis we expect 975 and 3,900 dominant carriers among the cases and controls respectively. Consequently, around ~95 cases and ~380 controls are expected to be "co-carriers" of these alleles when they are in perfect linkage equilibrium in both datasets ($LD_{\vec{v}}^{control} = LD_{\vec{v}}^{case} = 0$).

Now let us assume the specific alternative hypothesis under a certain interaction model (note: this may not be the only interaction that an LD-contrast captures, but is used here simply for illustrative purposes). Suppose that the disease penetrance for individuals carrying 1 or more minor alleles at both SNPs is 5%. If so, we expect to observe just ~19 fewer controls as co-carriers, leaving ~361 control co-carriers ($LD_{\vec{v}}^{control} = -1.08$). However, this small deflation in control co-carriers will be counterbalanced by an overabundance of ~95 co-carriers among cases due to the ascertainment bias (5-fold oversampling of cases), resulting in ~190 observed case co-carriers. This addition of 95 carriers to the background marginal count of 975 dominant carriers for each SNP, results in an observed marginal frequency of 10.7% in cases (up from 9.75%). Given these marginal frequencies, we would expect ~114.5 dominant co-carriers, which our observations exceed by ~75.5 ($LD_{\vec{v}}^{case} = 7.9,\ p \approx 1.4 \times 10^{-15}$).

Note that a signal of -19 (out of 40,000) vs. +75.5 (out of 10,000) co-carriers is highly significant ($D_{\vec{v}}^{control} = -0.000475,\ D_{\vec{v}}^{case} = +0.0075,\ LD_{\vec{v}}^{diff} = 7.6,\ p < 1.5 \times 10^{-14}$), and will pass the multiple testing burden of all pairs of variables in most experiments. In particular, we note here that the LD-case statistic was even more extreme and indicative of a significant LD-contrast, just as we had concluded in the section 6B on Approximately Complete Search.

Group sampling utilizes this difference in co-carrier frequencies as follows. If the dominant × dominant allelic combination was in perfect linkage equilibrium in both datasets, then by randomly sampling ($k = 4$) cases, the probability of all 4 cases being co-carriers by chance is $0.0095^4 = 8.1 \times 10^{-9}$. In the

alternative situation when there is penetrance of co-carriers, this probability is $0.0190^4 = 1.3 \times 10^{-7}$. If we draw 10^6 such groups of cases at random, then the probability that we will sample all co-carriers *at least once* is >12% if they are synergistic, while it is <0.7% if they are not. In this manner, group sampling makes it highly plausible that the joint-effects pair of variants will be observed under the alternative, not so under the null. Because group sampling utilizes Binary computer operations, even a million random draws can be accomplished in relatively insignificant amount of time.

References

Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, Erhardt-Lehmann A, Hemmer B, Rieckmann P, Daake M, Weber F, et al. 2010. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European journal of human genetics EJHG* **19**: 465-471. http://eprints.pascal-network.org/archive/00007993/.

Purcell S. Plink epistasis models documentation. http://pngu.mgh.harvard.edu/~purcell/plink/epidetails.shtml.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**: 559-575. http://pngu.mgh.harvard.edu/purcell/plink/.