GWAS for Compound Heterozygous Traits: Phenotypic Distance and Integer Linear Programming

Dan Gusfield, Rasmus Nielsen

December 11, 2016

GWAS

In Genome Wide Association Studies (GWAS) we try to locate mutations that are *causal* for a particular *trait* by comparing genomic information from individuals who are *believed* to have the trait (**cases**), with information from individuals who are believed to *not have* the trait (**controls**).

The goal is to find sites in the genome which statistically distinguish the cases from the controls. The method considers each site separately. This has worked well for simple traits caused by a single particular mutation at a single site (pure Mendelian traits).

Complex Traits

However, for many *complex* traits, involving more than one site, or mutations with small effect, or with large errors in determining who has the trait, GWAS have only recovered a small proportion of the variance in trait prevalence known to be caused by genetics. Ex. Diabetes II.

The most common explanation is the presence of *multiple* interacting or causal mutations that cannot be identified individually due to a lack of statistical power.

But, the mutations are often concentrated in a small number of genes, or loci. Many are thought to be *Compound Heterozygous* traits, a *subset* of the *recessive* traits.

Compound Heterozygous (CH) Trait

								CH
X_g (SNPs):	0	1	1	0	0	0	1	
$H_{1,1}$:	1	0	1	0	1	0	0	
$H_{1,2}$:	1	1	0	0	1	1	0	1
$H_{2,1}$:	1	0	0	1	1	0	0	
$H_{2,2}$:	0	1	1	0	1	0	1	0

Table : Vector X_g and two haplotype pairs. CH(1) is 1, and CH(2) is 0.

In this talk, we discuss an GWAS approach to finding loci that are causal for CH traits.

Formal Model of a CH-trait at a Causal Locus

Binary vector X_g denotes which of the m SNP sites are *causal* (i.e., contribute to the CH-trait), and which are not. H is the set of haplotype pairs for n individuals.

Given X_g and H, we define $CH(i)$:

$$CH(i) = [\bigvee_c (X_g(c) \wedge H_{i,1}(c))] \wedge [\bigvee_c (X_g(c) \wedge H_{i,2}(c))] \qquad (1)$$

In words, $CH(i)$ will have value 1 if and only if there is a SNP site c with $X_g(c) = 1$, where site c in haplotype $H_{i,1}$ also has value 1; and there is also a site c' (possibly c) with $X_g(c') = 1$, where site c' in haplotype $H_{i,2}$ also has value 1.

CH

We let CH denote the vector of length n, containing the values $CH(1), ..., CH(n)$.

Phenotypes The phenotypes of the individuals are recorded by the observable vector T. The *cases* have T-value of 1; the *controls* have T-value of 0.

$CH \ + \ $ noise $\rightarrow T$.

Without false positive or negatives, $T = CH$. But there are always some false positives or negatives, so T will differ to some extent from CH.

In the simulations I will discus later in the talk, about 30% of the cases (T value of 1) are false positives, and about 5% of the controls (T value of 0) are false negatives, at a causal gene.

1001100010001001001 True, but unknown SNPs

0000000000000000000 0
0000000000000000100

0000100000000000000 1
0010000100011001010

0001000000000000000 0
0000000000000000000

0000000000000000000 0
0000000001000000000

0000000000000000000 1
0000001000000000001

0001000000000000000 1
1000000010000100000

0100010000000000000 1
0000000000100000000 Low statistical power.

Hidden Phenotypic Distance

Definition: Given CH (which is a function of X_g and H), the *Hidden Phenotypic Distance* is the Hamming Distance (number of positions where the vectors differ) between CH and T. This is written *HPD(CH,T)*.

Without false positives or negatives, $HPD(CH,T) = 0$.

Compound Heterozygous (CH) Trait

								CH	phenotype T
X_g (SNPs):	0	1	1	1	0	0	1		
$H_{1,1}$:	1	0	1	0	1	0	0		
$H_{1,2}$:	1	1	0	0	1	1	0	1	1
$H_{2,1}$:	1	0	0	1	1	0	0		
$H_{2,2}$:	0	1	1	0	1	0	1	1	0

Table : Hidden Phenotypic Distance is the Hamming Distance between vectors CH and T. In this example, it is 1.

The Phenotypic Distance Problem

Definition: Given only H and T, the *Phenotypic Distance Problem* is the problem of determining a vector $\widetilde{X_g}$, which induces a vector $\widetilde{CH}$ to *Minimize* the Hamming Distance between $\widetilde{CH}$ and T. This is written *PD(H,T)*.

Loosely, *PD(H,T)* measures how well the phenotypes fit the *CH* model, or, the deviation from what is expected (under the *CH*-model), and what is observed.

We want to compute Phenotypic Distance for up to $n = 4000$ haplotype pairs and $n = 250$ sites.

Computing PD(H,T)

Enumeration of 2^m possible $\widetilde{X_g}$ is infeasible.

But Integer Linear Programming (ILP) works efficiently in practice for this problem. Under 3 seconds for a causal gene, and under 1 minute for a non-causal gene, on a macbook pro (2.3 GH, 4 cpus, $1500 to buy).

Integer Linear Programming (ILP)

In ILP, we translate the problem of computing *PD(H,T)* into a set of *linear inequalities* on a set of *binary variables*, and a linear *objective function* on a subset of the variables.

The particulars of the inequalities is where the magic resides. For 4000 haplotype pairs and 250 sites, the ILP has about 10,000 inequalities and 8,000 variables (this is modest size). The translation from problem instance to ILP inequalities is written in (really) bad and simple Perl.

Then we use a commercial ILP solver that finds the optimal values of the variables. (GUROBI 6.0, free to academics and researchers)

An ILP Formulation for the Phenotypic Distance Problem

The ILP formulation is almost an immediate restatement of the problem requirements, with only a couple of subtlties.

Recall that an H_i pair with $T(i) = 1$ is called a *case*, even though, due to false-positives, indivdual i might not actually have the trait; similarly an H_i pair with $T(i) = 0$ is called a *control*.

The ILP Inequalities for Cases

For each case H_i, the ILP formulation for the Phenotypic Distance will have the following inequalities:

$$\widetilde{CH}(i) \leq \sum_{c:\ H_{i,1}(c)=1} \widetilde{X}(c)$$

$$\widetilde{CH}(i) \leq \sum_{c:\ H_{i,2}(c)=1} \widetilde{X}(c)$$

The first inequality ensures that for any case, $\widetilde{CH}(i)$ can be set to 1 *only if* some $\widetilde{X}(c)$ is set to 1 for a column c where $H_{i,1}(c) = 1$.

The second inequality says the similar thing for the second haplotype of a case, i.e., for $H_{i,2}(c)$. So, for any case H_i, $\widetilde{CH}(i)$ will be set to 1 *only if* the values of $\widetilde{X}$ and H_i satisfy equation 1.

Errors from Cases

It follows that in an ILP solution,

$$[(\text{the number of cases}) - \sum_{H_i \text{ a case}} \widetilde{CH}(i))]$$

is the number of cases (i.e., $T(i) = 1$), where $\widetilde{CH}(i)$ is set to 0. That is, it is the number of errors in the solution, contributed by the cases.

Next, we consider the inequalities for a control.

The ILP inequalities for Controls

Let f_i be the number of columns, c, in H_i where $H_{i,1}(c) = 1$, and let s_i be the number of columns, c, in H_i where $H_{i,2}(c) = 1$.

For each control H_i, the ILP formulation will have the three inequalities:

$$\sum_{c:\ H_{i,1}(c)=1} \widetilde{X}(c) \leq f_i \times Z_{i,1}$$

$$\sum_{c:\ H_{i,2}(c)=1} \widetilde{X}(c) \leq s_i \times Z_{i,2}$$

$$Z_{i,1} + Z_{i,2} - \widetilde{CH}(i) \leq 1$$

The first inequality ensures, for a control H_i, that $Z_{i,1}$ will be set to 1 *if* there is a column c where $\widetilde{X}(c)$ is set to 1 and $H_{i,1}(c) = 1$. The second inequality ensures, for a control H_i, that $Z_{i,2}$ will be set to 1 *if* there is a column c where $\widetilde{X}(c)$ is set to 1 and $H_{i,2}(c) = 1$. The third inequality ensures that $\widetilde{CH}(i)$ will be set to 1 *if both* $Z_{i,1}$ and $Z_{i,2}$ are set to 1.

The Converse

The converse, that for H_i a control, $\widetilde{CH}(i)$ will be set to 1 *only if* those inequalities are satisfied, is not needed because the objective function has the term $+\sum_{H_i \text{a control}} \widetilde{CH}(i)$, and since the objective is a *minimization*, $\widetilde{CH}(i)$ *will* be set to 0 for any control H_i, *unless* doing so violates one of the three inequalities above.

The result is that in an optimal ILP solution, $\sum_{H_i \text{ a control}} \widetilde{CH}(i)$ is the number of H_i pairs where $T(i) = 0$, but $\widetilde{CH}(i)$ is set to 1.

The Objective Function

It follows that in an optimal ILP solution, the Hamming Distance between $\widetilde{CH}$ and T is
$[(\text{The number of cases}) - \sum_{H_i \text{a case}} \widetilde{CH}(i)] + \sum_{H_i \text{a control}} \widetilde{CH}(i)$.

So, the ILP formulation optimizes the objective function:

$$\text{Minimize}[(\#ofCases) - \sum_{H_i \text{a case}} \widetilde{CH}(i)] + \sum_{H_i \text{a control}} \widetilde{CH}(i),$$

and hence the optimal solution has value exactly PD(H,T).

The formulation has at most $3n + m$ variables and at most $3n$ inequalities, and so has modest size.

min Z

subject to:

$+X5 + X109 + X131 + X167 + X215 + X227 - 6Z0,1 \leq 0$

$+X129 - 1Z0,2 \leq 0$

$Z0,1 + Z0,2 - W0 \leq 1$

$+X118 + X139 - 2Z1,1 \leq 0$

$-0Z1,2 \leq 0$

$Z1,1 + Z1,2 - W1 \leq 1$

$+X52 + X77 + X210 - 3Z2,1 \leq 0$

$+X196 - 1Z2,2 \leq 0$

$Z2,1 + Z2,2 - W2 \leq 1$

$-0Z3,1 \leq 0$

$+X7 + X127 + X130 + X167 + X215 + X227 - 6Z3,2 \leq 0$

$Z3,1 + Z3,2 - W3 \leq 1$

$-X13 - X63 - X80 - X114 - X192 - X217 + W4 \leq 0$

$-X14 - X36 - X44 - X45 - X59 - X74 - X102 - X115 + W5 \leq 0$

$\vdots$ about 10,000 more inequalities

The GWAS Context and Simulations

Let T_g denote the phenotype vector at a *causal* gene. In GWAS simulations, we generate data for many genes, but use T_g as the phenotype vector for all genes.

Identifying Causal Genes with GWAS

The computations distinguish *causal* genes from *non-causal* genes.

1. The phenotypic distance computed at a *causal* gene is consistently and significantly less than the phenotypic distance computed at every *non-causal* gene. Meaning: Causal genes fit the CH-model better than non-causal ones.

2. The *time* needed at a causal gene is generally less than at a non-causal gene.

3. Permutation tests (of T_g) behave very differently at *causal* and *non-causal* genes. Large increase in $PD(H, T_g)$ when T_g is permuted at a causal gene g, but little change when permuted at a non-causal gene.

sites	HPD	PD	PD/SNP	cases controls	secs	SNP-dist
233	890	879	3.77	2000, 2000	1.50	64
255	1374	1340	5.25	2000, 2000	4.11	85
239	1324	1301	5.44	2000, 2000	17.33	71
238	1348	1319	5.54	2000, 2000	4.77	77
241	1333	1313	5.44	2000, 2000	47.86	88
242	1337	1305	5.39	2000, 2000	9.67	79
269	1344	1300	4.83	2000, 2000	33.26	88
237	1381	1345	5.67	2000, 2000	13.39	82
236	1378	1345	5.69	2000, 2000	7.58	74
236	1320	1281	5.42	2000, 2000	5.27	83
237	1353	1320	5.56	2000, 2000	1.84	79
247	1297	1280	5.18	2000, 2000	9.61	61

Table : The first few datasets in a GWAS simulation. The first one is the causal gene, and the others are non-causal.

c/p	hp	sites	HPD	PD/ub	case con	secs	SNP-dist
c	400	153	91	86	200, 200	0.02	46
p				159	200, 200	0.10	71
c	400	155	96	85	200, 200	0.02	63
p				165	200, 200	0.19	67
c	400	162	80	75	200, 200	0.03	59
p				156	200, 200	0.19	63
c	400	145	100	90	200, 200	0.04	44
p				162	200, 200	0.30	59
c	400	178	91	75	200, 200	0.02	70
p				149	200, 200	0.11	71
c	400	172	99	85	200, 200	0.04	59
p				142	200, 200	0.32	78
AVG c	400	157.6	91.6	83.2	200, 200	0.025	54.7
AVG p				153.8	200, 200	0.172	68.8

Table : Simulations at causal loci, each followed by a test of the permuted T vector.

Thank You!