GWAS for Compound Heterozygous Traits: Phenotypic Distance and Integer Linear Programming
Dan Gusfield, Rasmus Nielsen

December 11, 2016
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

\[ X_g \text{ (SNPs): } 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \]

\[ H_{1,1} : \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \]
\[ H_{1,2} : \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \]

\[ H_{2,1} : \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \]
\[ H_{2,2} : \ 0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \]

**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 $CH(i)$:

$$CH(i) = \left[ \bigvee_{c} (X_g(c) \land H_{i,1}(c)) \right] \land \left[ \bigvee_{c} (X_g(c) \land H_{i,2}(c)) \right]$$  \hspace{1cm} (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.
We let $CH$ denote the vector of length $n$, containing the values $CH(1), \ldots, CH(n)$.

**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.

$CH + \text{noise} \rightarrow T$.

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 discuss 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.
True, but unknown SNPs

Low statistical power.
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

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

\[
\begin{align*}
X_g \text{ (SNPs)}: & \quad 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \\
X_g \text{ (SNPs)}: & \quad \begin{array}{cccccc}
H_{1,1}: & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\
H_{1,2}: & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\
H_{2,1}: & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
H_{2,2}: & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1\end{array}
\end{align*}
\]
The Phenotypic Distance Problem

Definition: Given only $H$ and $T$, the Phenotypic Distance Problem is the problem of determining a vector $\tilde{X}_g$, which induces a vector $\tilde{CH}$ to Minimize the Hamming Distance between $\tilde{CH}$ 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 $\tilde{X}_g$ 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).
In ILP, we translate the problem of computing \( PD(H,T) \) into a set of \textit{linear inequalities} on a set of \textit{binary variables}, and a linear \textit{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)
The ILP formulation is almost an immediate restatement of the problem requirements, with only a couple of subtleties.

Recall that an $H_i$ pair with $T(i) = 1$ is called a case, even though, due to false-positives, individual $i$ 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:

$$\tilde{C}H(i) \leq \sum_{c: \ H_i,1(c)=1} \tilde{X}(c)$$

$$\tilde{C}H(i) \leq \sum_{c: \ H_i,2(c)=1} \tilde{X}(c)$$

The first inequality ensures that for any case, $\tilde{C}H(i)$ can be set to 1 only if some $\tilde{X}(c)$ 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 $H_i$, $\tilde{C}H(i)$ will be set to 1 only if the values of $\tilde{X}$ and $H_i$ satisfy equation 1.
Errors from Cases

It follows that in an ILP solution,

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

is the number of cases (i.e., \(T(i) = 1\)), where \(\widehat{CH}(i)\) 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:

\[
\sum_{c: H_{i,1}(c)=1} \tilde{X}(c) \leq f_i \times Z_{i,1}
\]

\[
\sum_{c: H_{i,2}(c)=1} \tilde{X}(c) \leq s_i \times Z_{i,2}
\]

\[
Z_{i,1} + Z_{i,2} - \overline{CH}(i) \leq 1
\]

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 $\tilde{X}(c)$ 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 $\tilde{X}(c)$ is set to 1 and $H_{i,2}(c) = 1$. The third inequality ensures that $\overline{CH}(i)$ 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, $\bar{C}H(i)$ will be set to 1 only if those inequalities are satisfied, is not needed because the objective function has the term $+ \sum_{H_i \text{ a control}} \bar{C}H(i)$, and since the objective is a minimization, $\bar{C}H(i)$ 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, $\sum_{H_i \text{ a control}} \bar{C}H(i)$ is the number of $H_i$ pairs where $T(i) = 0$, but $\bar{C}H(i)$ is set to 1.
The Objective Function

It follows that in an optimal ILP solution, the Hamming Distance between $\tilde{CH}$ and $T$ is

$$[(\text{The number of cases}) - \sum_{H_i \text{ case}} \tilde{CH}(i)] + \sum_{H_i \text{ control}} \tilde{CH}(i).$$

So, the ILP formulation optimizes the objective function:

Minimize$$[(\# \text{of Cases}) - \sum_{H_i \text{ case}} \tilde{CH}(i)] + \sum_{H_i \text{ control}} \tilde{CH}(i),$$

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:
$+X_5 + X_{109} + X_{131} + X_{167} + X_{215} + X_{227} - 6Z_0, 1 \leq 0$
$+X_{129} - 1Z_0, 2 \leq 0$
$Z_0, 1 + Z_0, 2 - W_0 \leq 1$
$+X_{118} + X_{139} - 2Z_1, 1 \leq 0$
$-0Z_1, 2 \leq 0$
$Z_1, 1 + Z_1, 2 - W_1 \leq 1$
$+X_{52} + X_{77} + X_{210} - 3Z_2, 1 \leq 0$
$+X_{196} - 1Z_2, 2 \leq 0$
$Z_2, 1 + Z_2, 2 - W_2 \leq 1$
$-0Z_3, 1 \leq 0$
$+X_{7} + X_{127} + X_{130} + X_{167} + X_{215} + X_{227} - 6Z_3, 2 \leq 0$
$Z_3, 1 + Z_3, 2 - W_3 \leq 1$
$-X_{13} - X_{63} - X_{80} - X_{114} - X_{192} - X_{217} + W_4 \leq 0$
$-X_{14} - X_{36} - X_{44} - X_{45} - X_{59} - X_{74} - X_{102} - X_{115} + W_5 \leq 0$
\vdots about 10,000 more inequalities
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.
<table>
<thead>
<tr>
<th>sites</th>
<th>HPD</th>
<th>PD</th>
<th>PD/SNP</th>
<th>cases controls</th>
<th>secs</th>
<th>SNP-dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>233</td>
<td>890</td>
<td>879</td>
<td>3.77</td>
<td>2000, 2000</td>
<td>1.50</td>
<td>64</td>
</tr>
<tr>
<td>255</td>
<td>1374</td>
<td>1340</td>
<td>5.25</td>
<td>2000, 2000</td>
<td>4.11</td>
<td>85</td>
</tr>
<tr>
<td>239</td>
<td>1324</td>
<td>1301</td>
<td>5.44</td>
<td>2000, 2000</td>
<td>17.33</td>
<td>71</td>
</tr>
<tr>
<td>238</td>
<td>1348</td>
<td>1319</td>
<td>5.54</td>
<td>2000, 2000</td>
<td>4.77</td>
<td>77</td>
</tr>
<tr>
<td>241</td>
<td>1333</td>
<td>1313</td>
<td>5.44</td>
<td>2000, 2000</td>
<td>47.86</td>
<td>88</td>
</tr>
<tr>
<td>269</td>
<td>1344</td>
<td>1300</td>
<td>4.83</td>
<td>2000, 2000</td>
<td>33.26</td>
<td>88</td>
</tr>
<tr>
<td>237</td>
<td>1381</td>
<td>1345</td>
<td>5.67</td>
<td>2000, 2000</td>
<td>13.39</td>
<td>82</td>
</tr>
<tr>
<td>236</td>
<td>1378</td>
<td>1345</td>
<td>5.69</td>
<td>2000, 2000</td>
<td>7.58</td>
<td>74</td>
</tr>
<tr>
<td>236</td>
<td>1320</td>
<td>1281</td>
<td>5.42</td>
<td>2000, 2000</td>
<td>5.27</td>
<td>83</td>
</tr>
<tr>
<td>237</td>
<td>1353</td>
<td>1320</td>
<td>5.56</td>
<td>2000, 2000</td>
<td>1.84</td>
<td>79</td>
</tr>
<tr>
<td>247</td>
<td>1297</td>
<td>1280</td>
<td>5.18</td>
<td>2000, 2000</td>
<td>9.61</td>
<td>61</td>
</tr>
</tbody>
</table>

**Table:** The first few datasets in a GWAS simulation. The first one is the causal gene, and the others are non-causal.
<table>
<thead>
<tr>
<th>c/p</th>
<th>hp</th>
<th>sites</th>
<th>HPD</th>
<th>PD/ub</th>
<th>case con</th>
<th>secs</th>
<th>SNP-dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>400</td>
<td>153</td>
<td>91</td>
<td>86</td>
<td>200, 200</td>
<td>0.02</td>
<td>46</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>159</td>
<td></td>
<td>200, 200</td>
<td>0.10</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>400</td>
<td>155</td>
<td>96</td>
<td>85</td>
<td>200, 200</td>
<td>0.02</td>
<td>63</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>165</td>
<td></td>
<td>200, 200</td>
<td>0.19</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>400</td>
<td>162</td>
<td>80</td>
<td>75</td>
<td>200, 200</td>
<td>0.03</td>
<td>59</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>156</td>
<td></td>
<td>200, 200</td>
<td>0.19</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>400</td>
<td>145</td>
<td>100</td>
<td>90</td>
<td>200, 200</td>
<td>0.04</td>
<td>44</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>162</td>
<td></td>
<td>200, 200</td>
<td>0.30</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>400</td>
<td>178</td>
<td>91</td>
<td>75</td>
<td>200, 200</td>
<td>0.02</td>
<td>70</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>149</td>
<td></td>
<td>200, 200</td>
<td>0.11</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>400</td>
<td>172</td>
<td>99</td>
<td>85</td>
<td>200, 200</td>
<td>0.04</td>
<td>59</td>
</tr>
<tr>
<td>p</td>
<td></td>
<td>142</td>
<td></td>
<td>200, 200</td>
<td>0.32</td>
<td>78</td>
<td></td>
</tr>
</tbody>
</table>

**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!