SNPs and Genetic Association Studies

Carla Gallagher, PhD

Bioinformatics Course

April 6th, 2011
Genetic Association

- Searches for a population association between a disease and a particular allele of a genetic marker (frequency difference).
- Use case and control populations.
- Or an association between an allele and a quantitative trait (i.e. Search for an association between an allele of a SNP and carcinogen metabolite levels)
- Can use any type of polymorphism (marker), but most frequently use SNPs (single nucleotide polymorphisms)
Number of SNPs

• There are more than 10,000,000 SNPs in the human genome (available at NCBI’s SNP database – dbSNP)
• Even in one gene there are many SNPs to choose from (ie. UGT1A8 almost 1000 SNPs)
• Of course, genotyping all of the SNPs would give us the most information, but this is not usually reasonable do to cost and time (and it is not necessary)
How do we choose the SNPs to genotype in a gene of interest

| SNPs that are known to change the function of a gene | Excellent choice, but usually not available |
| SNPs that are in exons, UTR, promoter, or splice junctions | Good choice, but may require substantial sequencing first, and this could yield a large amount of SNPs as well. Also transcription factor binding sites are often not known and can be far (kb) away from ATG, sometimes even in intron 1. There is a chance you won’t detect association with the true functional variant |
| SNPs that tag the common variation in the region | Excellent choice to reduce number of SNPs without reducing the information (We will discuss this today) |

How to identify all reported* SNPs in a gene:
or UCSC genome browser: [http://genome.ucsc.edu/](http://genome.ucsc.edu/)

*these SNPs may not have been confirmed
Linkage Disequilibrium (LD)

• The non-random association of alleles at adjacent loci.
• 2 markers are in LD when an allele at one locus is found together on the same chromosome with an allele at a second locus more often than if they were segregating independently.
• So genotyping 1 marker SNP will give you information on the genotypes of other polymorphisms that are in LD with that marker SNP.
• Measured by D’ or $r^2$ (values range from 0 -1 where 1 = complete LD)
Tagger

- Chooses tagSNPs that represent all other SNPs in the region (identified by high LD values)
- By genotyping this group of SNPs you get information on all the SNPs that exhibit high LD with the genotyped SNPs
- Available using the program Haploview:
How to determine the htSNPs?

HapMap database & Haploview software
Haplotype Map of the Human Genome

www.hapmap.org

- Complete the genotyping of a dense set of SNPs across the human genome
- Define patterns of genetic variation across human genome (LD)
- Guide selection of SNPs efficiently to “tag” common variants across the genome
- Public release of all data (allele frequency, assays, genotypes)

**Phase I:** 1.3 M markers genotyped in 269 people
* ENCODE variation reference resource available

**Phase II:** +2.8 M markers genotyped in 270 people

~4,000,000 SNPs typed in total !!!
HapMap Samples
270 samples were genotyped across the genome

Phase 1 and 2
- 90 Yoruba individuals (30 parent-parent-offspring trios) from Ibadan, Nigeria (YRI)
- 90 individuals (30 trios) of European descent from Utah (CEU)
- 45 Han Chinese individuals from Beijing (CHB)
- 45 Japanese individuals from Tokyo (JPT)
HapMap Samples
Phase 3

- Population descriptors:
  ASW (A): African ancestry in Southwest USA
  CEU (C): Utah residents with Northern and Western European ancestry from the CEPH collection
  CHB (H): Han Chinese in Beijing, China
  CHD (D): Chinese in Metropolitan Denver, Colorado
  GIH (G): Gujarati Indians in Houston, Texas
  JPT (J): Japanese in Tokyo, Japan
  LWK (L): Luhya in Webuye, Kenya
  MEX (M): Mexican ancestry in Los Angeles, California
  MKK (K): Maasai in Kinyawa, Kenya
  TSI (T): Toscans in Italy
  YRI (Y): Yoruba in Ibadan, Nigeria
Using data from the HapMap to design a genetic SNP & Haplotype Association Study

Example:
Are SNPs in the ESR1 gene associated with cancer risk?
Finding HapMap SNPs in a Region of Interest

• Find the region of the genome containing the ESR1 gene \(\text{estrogen receptor alpha protein}\)
• Identify the characterized SNPs in the region.
• Download the region in Haploview format.
• View the patterns of LD in the region.
• Pick tag SNPs for genotyping in the association study.
1: HapMap Browser

1a. Go to www.hapmap.org

1b. Choose project data. When downloading data for use in haploview software, use Phase I and Phase II data only (Haploview isn’t updated to handle Phase III data yet).
2: Search for your gene of interest (ie. ESR1)

2. Type search term - “ESR1”
3: Examine Region

Chromosome-wide summary data is shown in overview.

Default tracks show HapMap genotyped SNPs, named genes from Entrez, and alternative mRNA splicing patterns.
3: Examine Region (cont)

Use the Scroll/Zoom buttons and menu to change position & magnification.

As you zoom in, the display changes to indicate more detail.

Population descriptions:
- YRI: Yoruba in Ibadan, Nigeria
- JPT: Japanese in Tokyo, Japan
- CHB: Han Chinese in Beijing, China
- CEU: CEPH (Utah residents with ancestry from northern and western Europe)

Overview

Details

Genotyped SNPs
Change tracks to your preference

Click checkmarks to add tracks, then click Update Image
Look at SNPs in Exons
9: Generate Reports

9. Select the desired "Download" option and press "Go" or "Configure". Configure will let you choose your population.
9: Save data as a .txt file
(I usually do this in excel)

The Genotype download format can be saved as a .txt file and loaded into Haploview.

10. Delete the 2 comment lines (begin with #). Although those are comments, haploview doesn’t view them that way and they interfere with analysis (you will get an error if you leave them in the file).

Your first line should start with rs#
Open hapmap data (.txt) in Haploview
File, Open new data, Hapmap format, browse to find file, ok
Check markers tab
Info on allele frequency, Hardy-Weinberg, etc (can do this with your own data too)

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Position</th>
<th>ObsHet</th>
<th>PredHet</th>
<th>HWval</th>
<th>%Geno</th>
<th>FamTrio</th>
<th>MendErr</th>
<th>MAF</th>
<th>Alleles</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2077647</td>
<td>15</td>
<td>7033790</td>
<td>0.397</td>
<td>0.406</td>
<td>0.219</td>
<td>95.7</td>
<td>26</td>
<td>0.422</td>
<td>A-G</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>rs716432</td>
<td>15</td>
<td>7171001</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>C-C</td>
<td>✓</td>
</tr>
<tr>
<td>3</td>
<td>rs323010</td>
<td>15</td>
<td>7216111</td>
<td>0.4</td>
<td>0.425</td>
<td>0.719</td>
<td>100.0</td>
<td>30</td>
<td>0.317</td>
<td>T-C</td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td>rs12660738</td>
<td>15</td>
<td>7239134</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.9</td>
<td>29</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>rs2241492</td>
<td>15</td>
<td>7390144</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>rs17001990</td>
<td>15</td>
<td>7397928</td>
<td>0.103</td>
<td>0.219</td>
<td>0.422</td>
<td>100.0</td>
<td>30</td>
<td>0.125</td>
<td>T-G</td>
<td>✓</td>
</tr>
<tr>
<td>7</td>
<td>rs575338</td>
<td>15</td>
<td>7388668</td>
<td>0.069</td>
<td>0.067</td>
<td>1.0</td>
<td>95.6</td>
<td>26</td>
<td>0.034</td>
<td>C-T</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td>rs1216113</td>
<td>15</td>
<td>7398001</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>rs3862744</td>
<td>15</td>
<td>7397007</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>C-C</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>rs1909700</td>
<td>15</td>
<td>7397957</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>rs10949222</td>
<td>15</td>
<td>7401040</td>
<td>0.15</td>
<td>0.193</td>
<td>0.248</td>
<td>98.9</td>
<td>29</td>
<td>0.108</td>
<td>C-T</td>
<td>✓</td>
</tr>
<tr>
<td>12</td>
<td>rs1053993</td>
<td>15</td>
<td>7422989</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>95.7</td>
<td>27</td>
<td>0.0</td>
<td>T-T</td>
<td>✓</td>
</tr>
<tr>
<td>13</td>
<td>rs3776610</td>
<td>15</td>
<td>7505044</td>
<td>0.017</td>
<td>0.017</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>0.0000</td>
<td>✓</td>
</tr>
<tr>
<td>14</td>
<td>rs975120</td>
<td>15</td>
<td>7557600</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.9</td>
<td>29</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>15</td>
<td>rs7067110</td>
<td>15</td>
<td>7631114</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>T-T</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>rs1940776</td>
<td>15</td>
<td>7714444</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.9</td>
<td>29</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>17</td>
<td>rs1940777</td>
<td>15</td>
<td>7717477</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>95.7</td>
<td>27</td>
<td>0.0</td>
<td>C-C</td>
<td>✓</td>
</tr>
<tr>
<td>18</td>
<td>rs1940778</td>
<td>15</td>
<td>7721902</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>19</td>
<td>rs1940779</td>
<td>15</td>
<td>7730002</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.7</td>
<td>29</td>
<td>0.0</td>
<td>C-C</td>
<td>✓</td>
</tr>
<tr>
<td>20</td>
<td>rs1940780</td>
<td>15</td>
<td>7745252</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.7</td>
<td>27</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>21</td>
<td>rs1940781</td>
<td>15</td>
<td>7837038</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>98.9</td>
<td>29</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>22</td>
<td>rs1940782</td>
<td>15</td>
<td>7850696</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>93.9</td>
<td>28</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>23</td>
<td>rs1952642</td>
<td>15</td>
<td>7909907</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>A-A</td>
<td>✓</td>
</tr>
<tr>
<td>24</td>
<td>rs1708178</td>
<td>15</td>
<td>7934815</td>
<td>0.19</td>
<td>0.25</td>
<td>0.172</td>
<td>94.1</td>
<td>26</td>
<td>0.147</td>
<td>A-G</td>
<td>✓</td>
</tr>
<tr>
<td>25</td>
<td>rs6093222</td>
<td>15</td>
<td>7968300</td>
<td>0.017</td>
<td>0.017</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>0.0000</td>
<td>✓</td>
</tr>
<tr>
<td>26</td>
<td>rs1168236</td>
<td>15</td>
<td>7979312</td>
<td>0.059</td>
<td>0.057</td>
<td>1.0</td>
<td>97.8</td>
<td>21</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>27</td>
<td>rs3907166</td>
<td>15</td>
<td>8005181</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>97.8</td>
<td>21</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
<tr>
<td>28</td>
<td>rs3555240</td>
<td>15</td>
<td>8101579</td>
<td>0.167</td>
<td>0.251</td>
<td>0.109</td>
<td>100.0</td>
<td>30</td>
<td>0.130</td>
<td>C-C</td>
<td>✓</td>
</tr>
<tr>
<td>29</td>
<td>rs9835299</td>
<td>15</td>
<td>8186618</td>
<td>0.345</td>
<td>0.385</td>
<td>0.696</td>
<td>97.8</td>
<td>28</td>
<td>0.259</td>
<td>A-C</td>
<td>✓</td>
</tr>
<tr>
<td>30</td>
<td>rs3944508</td>
<td>15</td>
<td>8187135</td>
<td>0.167</td>
<td>0.231</td>
<td>0.109</td>
<td>100.0</td>
<td>30</td>
<td>0.130</td>
<td>T-C</td>
<td>✓</td>
</tr>
<tr>
<td>31</td>
<td>rs9215517</td>
<td>15</td>
<td>8187902</td>
<td>0.05</td>
<td>0.049</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.025</td>
<td>A-G</td>
<td>✓</td>
</tr>
<tr>
<td>32</td>
<td>rs1849509</td>
<td>15</td>
<td>8202157</td>
<td>0.167</td>
<td>0.251</td>
<td>0.109</td>
<td>100.0</td>
<td>30</td>
<td>0.130</td>
<td>T-C</td>
<td>✓</td>
</tr>
<tr>
<td>33</td>
<td>rs3952398</td>
<td>15</td>
<td>8223947</td>
<td>0.017</td>
<td>0.017</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>0.0000</td>
<td>✓</td>
</tr>
<tr>
<td>34</td>
<td>rs9479121</td>
<td>15</td>
<td>8285600</td>
<td>0.017</td>
<td>0.017</td>
<td>1.0</td>
<td>100.0</td>
<td>30</td>
<td>0.0</td>
<td>0.0000</td>
<td>✓</td>
</tr>
<tr>
<td>35</td>
<td>rs9479122</td>
<td>15</td>
<td>8292633</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>93.9</td>
<td>29</td>
<td>0.0</td>
<td>G-G</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Mantel-Haenszel p-value cutoff:** 0.0010

**Min genotype %:** 75

**Max # mendel errors:** 1

**Minimum minor allele freq.** 0.0010

Select All | Deselect All | Reset Values | Rescore Markers
LD plot tab

D’ values are displayed in the squares (empty squares have a pairwise D’=1.00). Red squares show high pairwise LD, gradually coloring down to white squares of low pairwise LD. Blue squares indicate high LD, but low significance. The black triangles indicate the LD haplotype blocks. There are many ways to define blocks (see below).
Tags in blocks

Haploview can determine the htSNPs - indicated with the triangles.

Eg. Block 5 – By genotyping only 4 of the 15 SNPs you can distinguish each of the 5 common haplotypes.

ACAA
GCAA
ATGG

ATGA
ACGA
Pairwise tags
Tagger tab, configuration tab, run tagger, export current tab as text
Exported tagger output

tagSNPs significantly reduce the number of SNPs to genotype (i.e. Getting information from 382 SNPs by genotyping 109 SNPs)

Actually getting info from many more SNPs (even the ones that aren’t genotyped here)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># captured 382 of 382 alleles at $r^2 \geq 0.8$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td># captured 100 percent of alleles with mean $r^2$ of 0.976</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td># using 109 Tag SNPs in 109 tests.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allele</td>
<td>Best Test</td>
<td>$r^2$ w/test</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>rs2077647</td>
<td>rs38232433</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>rs532010</td>
<td>rs3853249</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>rs1708169</td>
<td>rs7761133</td>
<td>0.849</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>rs576330</td>
<td>rs576330</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>rs1046492</td>
<td>rs9340788</td>
<td>0.831</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>rs3778610</td>
<td>rs7759411</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>rs1708171</td>
<td>rs7761133</td>
<td>0.851</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>rs6930222</td>
<td>rs7759441</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>rs1196823</td>
<td>rs1196823</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>rs3853248</td>
<td>rs7761133</td>
<td>0.924</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>rs3853249</td>
<td>rs3853249</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>rs3844508</td>
<td>rs7761133</td>
<td>0.924</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>rs9371557</td>
<td>rs9371557</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>rs3844509</td>
<td>rs7761133</td>
<td>0.924</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>rs3892535</td>
<td>rs7759411</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>rs9479121</td>
<td>rs9479121</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>rs6917746</td>
<td>rs7761133</td>
<td>0.924</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>rs9340788</td>
<td>rs9340788</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>rs9340789</td>
<td>rs9340789</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>rs1175309</td>
<td>rs1175309</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Genome-wide association studies (GWAS)

- Make use of linkage disequilibrium and tagSNPs across the whole genome
- Good for studies where there aren’t obvious candidate genes so that every gene (and all intergenic regions where there might be an undiscovered gene) are tested for association with disease
Positive association to a SNP or haplotype requires detailed interpretation

• When you find association you are most likely not finding the functional SNP!!! You are finding a marker associated with disease, so the functional SNP is nearby (within region of LD). Now that you know this region is involved in your disease (or trait) of interest, you can try to figure out why.

  – How many other SNPs are in LD with this SNP?
  – What genes are in LD with this SNP?
  – What coding variants and putative functional variants are in LD with this SNP?
  – Maybe sequencing the region of LD will be required to discover the functional variant.
End of class

• Additional material for those interested in genetics research follows
• I’d be happy to meet with individuals to discuss further
Validation of HapMap Data

Use of data from the ENCODE project (representing most variations in the genome) to determine the efficiency & power of HapMap
ENCODE-HapMap variation project

A much more complete variation resource by which the genome-wide map can evaluated

• Ten “typical” 500kb regions
• 48 samples sequenced for SNP discovery
• All discovered SNPs (and any others in dbSNP) typed in all 270 HapMap samples
• Current data set – 1 SNP every 279 bp

Sequenced to discover all common variants, then looked at HapMap data to see if it was a good representation of all of the variants

* One of the ten regions sequenced, includes the UGT1A gene cluster
Coverage of HapMap (estimated from ENCODE data)

Panel     %r^2 > 0.8
YRI        81
CEU        94
CHB+JPT    94

Percentage of deeply ascertained common variants highly correlated with a HapMap SNP

From Table 6 – “A Haplotype Map of the Human Genome”, Nature
Tagging from HapMap

- Since HapMap describes the majority of common variation in the genome, choosing non-redundant sets of SNPs from HapMap offers considerable efficiency without power loss in association studies
Efficiency and power

~300,000 tag SNPs needed to cover common variation in whole genome in CEU

Can incorporating tests of haplotypes of SNPs on the genome-wide product improve coverage?
Haplotypes increase coverage for tests of genetic association

Data represents genome-wide association tests

Example: 500K data generated by Affymetrix and recently submitted to HapMap DCC
SNP genotyping methods available through core facilities at PSCOM

- **TaqMan**
  (single SNP) – real-time PCR
- **SNaPshot**
  (2-10 SNPs multiplex) – capillary electrophoresis
- **SNPlex**
  (12-48 SNPs multiplex) – capillary electrophoresis
- **Illumina BeadStation500**
  Can perform GoldenGate or Infinium Assays (price becomes better than BeadXpress for >384 SNPs)
- **Illumina BeadXpress**
  Can genotype 96 or 384 SNPs with GoldenGate, or 1-96 SNPs with allele specific primer extension

**Illumina Assays**
GoldenGate Assay: 96-1536 SNP multiplex
Infinium Assay: 7600-60800 SNP multiplex (used for genome-wide association studies)
What is the phase?
For two SNPS A and B, each having 2 alleles (Aa and Bb):

SNP1  SNP2

Genotypes:  AA   BB

In this case the haplotypes are known
What is the phase?
For two SNPs A and B, each having 2 alleles (Aa and Bb):

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>BB</td>
</tr>
</tbody>
</table>

In this case the haplotypes are known

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa</td>
<td>Bb</td>
</tr>
</tbody>
</table>

In this case the haplotypes are unknown

OR

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>B</td>
<td>b</td>
</tr>
</tbody>
</table>

OR

<table>
<thead>
<tr>
<th>SNP1</th>
<th>SNP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>B</td>
</tr>
</tbody>
</table>
Ways to estimate unknown phase

- The **EM algorithm** estimates how likely a haplotype is based on allele frequencies and previous haplotype frequencies – used in the program haplo.stats.

- **Bayesian** methods use a coalescent model (a genetic tree of haplotypes) to estimate how likely a haplotype is – used in the program PHASE.
Association Analysis

• Those programs (haplo.stats and PHASE) weight the possible haplotypes for a person and use that information in association analysis.
• We are looking for a SNP or haplotype that has a frequency difference in people with low compared to high irinotecan toxicity
• For example people with A/A genotype or AGTT haplotype have a lower mean irinotecan toxicity than people with the G/G genotype or any other haplotype.
When the tagSNP approach will work

Common Disease – Common Variant Hypothesis

- The common disease – common variant hypothesis states that diseases that are common in the population (i.e., diabetes, cancer, asthma, heart disease) will be caused by variants common in the population.

- So since we are choosing SNPs that represent the **COMMONT** haplotypes we should be able to detect association with these common variants that affect the trait.
When the tagSNP approach is less effective

• If the trait of interest is caused by rare variants, htSNPs may not identify the association.
• When choosing htSNPs we pick SNPs that tag the **COMMON** haplotypes, if the causal variant is only present on a rare haplotype it is unlikely that these SNPs will detect the association.
• It might be a mixture of rare and common variants that are responsible for these traits, but it will be good to identify the common variants first.
Evidence Supporting the Common Disease - Common Variant Hypothesis

Common Gene Variation in Complex Disease

- Case-control studies, comparing the frequencies of common gene variants can identify susceptibility and protective alleles
- Some have multiple identified genes (*)

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Gene</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptic ulcer</td>
<td>ABO</td>
<td>B</td>
</tr>
<tr>
<td>IDDM*</td>
<td>HLA</td>
<td>DR3,4</td>
</tr>
<tr>
<td>Alzheimer dementia</td>
<td>APOE</td>
<td>E4</td>
</tr>
<tr>
<td>Deep venous thrombosis</td>
<td>F5</td>
<td>Leiden</td>
</tr>
<tr>
<td>Falciparum malaria*</td>
<td>HBBE</td>
<td>$\beta^s$</td>
</tr>
<tr>
<td>AIDS*</td>
<td>CCR5</td>
<td>$\Delta32$</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>APC</td>
<td>3920A</td>
</tr>
<tr>
<td>NIDDM</td>
<td>PPAR$_\gamma$</td>
<td>12A</td>
</tr>
</tbody>
</table>
Questions?

• My contact information:
  Carla Gallagher
  cjc17@psu.edu
  x2973

• Further Information:

