Quantitative Trait Loci (QTLs)
6.874 Lecture 19

David Gifford
MIT, Cambridge, MA
Broad heritability of a trait

\[ H^2 = \frac{\sigma_g^2}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]
Additive polygenic traits
Dissecting the architecture of a quantitative trait locus in yeast

Lars M. Steinmetz†, Himanshu Sinha‡, Dan R. Richards*, Jamie I. Spiegelman†, Peter J. Oefner†§, John H. McCusker‡ & Ronald W. Davis*†§

NATURE | VOL 416 | 21 MARCH 2002 | www.nature.com
What causes a hybrid to have increased fitness?

**Figure 1** Analysis of the high-temperature-growth phenotype (Htg). **a**, Qualitative differences measured by colony size; **b**, quantitative differences in growth measured by competition assay after 48 h at 30 and 41 °C: (i) S1029, an Htg− S288c strain, (ii) YAG040, an Htg+ YJM145 strain, and (iii) XHS123, a YJM145/S288c hybrid. Bars indicate s.e.m. (n = 6).
Figure 2 Detection by genome-wide mapping and detailed analysis by fine-structure mapping of the chromosome XIV Htg QTL. **a**, Calculated probability assuming random segregation (y-axis) for 3,444 markers in a whole-genome allelic variation scan of 19 Htg+ segregants, shown here for chromosome XIV. A region with low probability was identified between bp 434,194 and 485,856. **b**, Relative-risk plot representing the fine-structure map of the low-probability region. From 104 Htg+ segregants, the recombination breakpoints defining an interval with high relative risk were placed to within 32 kb (bp 445,003–477,059 of chromosome XIV).
Chromosome XIV QTL Sequence Data
6 Htg+ and 7 Htg-
Locates candidate genes with non-synomous changes
Reciprocal-hemizygosity analysis
Tested 15 genes; found MKT1, END3, ROH2

All data at 41° C (High temperature)
Regression trees can incorporate genotypic information (Lee et. al)

- **Design**
  - 112 individuals from a BY and RM strain cross
  - 304 putative expression regulators selected

- **For each individual**
  - 581 genetic markers with low-correlation (from 2957)
  - expression profiles for 3152 genes

- **Result**
  - 165 programs
  - 50% of variance of gene expression explained for 828 genes
Example eQTL regression tree
Percentage of gene variance explained
(average of 10 random splits for the detection and estimation sets)
Example chromosomal module
Model based on 23 chromatin markers
(markers linked to chromatin regulators)
QTLs that describe transcription factor binding

• Design
  – S288c (S96) and YJM789 (HS959) cross
  – 43 Mat alpha segregants
• Profile parents and segregants
  – Ste12 ChIP Seq
  – Expression profiles
  – 2592 non-redundant genotypic markers (from 53,460)

Genetic analysis of variation in transcription factor binding in yeast

Wei Zheng¹, Hongyu Zhao², Eugenio Mancera⁴, Lars M. Steinmetz⁴ & Michael Snyder¹,³

doi:10.1038/nature08934
QTLs that describe transcription factor binding

- **Design**
  - S288c and YJM789 cross
  - 43 Mat alpha haploid segregants
- **Profile parents and segregants**
  - Ste12 ChIP Seq
  - Expression profiles
  - 2592 non-redundant genotypic markers (from 53,460)

Genetic analysis of variation in transcription factor binding in yeast

Wei Zheng\textsuperscript{1,4}, Hongyu Zhao\textsuperscript{2,4}, Eugenio Mancera\textsuperscript{4}, Lars M. Steinmetz\textsuperscript{4} & Michael Snyder\textsuperscript{1,5}

[doi:10.1038/nature08934]
Ste12 binding overlap

- Cell wall biosynthesis ($P = 1 \times 10^{-3}$)
- Pyridine nucleotide metabolism ($P = 0.019$); alcohol metabolism ($P = 0.048$)
- Conjugation ($P = 1.4 \times 10^{-6}$)
Mendelian and transgressive Ste12 binding
Ste12 variable binding regions across strains

- Defined by top 2.5% of binding variance across strains
- Grouped into 930 quantitative genomic “traits”
- Each trait is the regional sum of normalized difference scores of Ste12 binding
- 78% of the traits are Mendelian
Single marker regression of 930 binding traits with 2,592 genetic markers

• 195 trait / marker associations (FDR 0.01)
• 166 traits / cis-marker(s)
• 35 traits / trans-marker(s)
• 6 traits / cis and trans markers
• 85% associated traits have a cis associated marker
• 102 / 166 cis regions contain a Ste12 binding site
• 31 / 102 had mutations in one or more Ste12 sites
• 72 / 102 had mutations in one or more known TF motifs
• 2 of 177 Ste12 motifs in non-variable regions had SNPs
Marker / binding variation association

![Graph showing genomic position of markers against genomic position of binding variation]

- chrXV:957535..966064
- chrXII:643162-759987
- chrX:24374..32847
- chrV:361239..377186, FLO8
- chrII:552028..563312, AMN1
- chrII:215949..216634
Explanation of variable binding traits
Number of markers and location of markers

(b) Number of markers vs. No. of associated variable binding traits

(c) Explained proportion of variance

- Black: Cis < 10 kb
- Red: Cis > 10 kb
- Green: Trans
Trans-markers linked to binding variation traits

- 194 trans-markers – too many to test
- Clustered 930 quantitative traits into 121 clusters
- Tested association between clusters and markers (FDR 0.01)
- Seven trans-markers were in common
- Tested 12 candidate genes in these regions (two genes, AMN1 and FLO8 were in the same region)
Ste12 ChIP-Seq for all strains

Calculate NormDiff for each strain

Select and quantify highly variable binding regions

QTL scan for individual binding traits

Identify cis-QTLs

Motif analysis for cis-QTLs

Identify trans-QTLs

Identify common trans-QTLs from both methods

Measure Gene Expression for all strains

Identify functional Ste12 binding variation

Hierarchical clustering of binding traits

QTL scan for cluster-level traits
Validation of two causative quantitative trait genes

(a) S96 WT
   - HS959 WT
   - S96 amn1Δ
   - HS959 amn1Δ
   - S96 ics2Δ
   - HS959 ics2Δ
   - S96 input
   - HS959 input

(b) S96 WT
   - HS959 WT
   - S96 flo8Δ
   - HS959 flo8Δ
   - S96 gle2Δ
   - HS959 gle2Δ
   - S96 input
   - HS959 input

YLR2815c   YLR286C   YLR2816c
708,000    710,000

YAR0506w
30    202,000    204,000
Ste12 binding variation correlates with expression
Ste12 binding variation correlates with expression
Allele Frequency

High-frequency polymorphisms
Eg: many now known
HapMap
First generation arrays

Lower-frequency polymorphisms
Eg: CFTR delta S08
PCSK9 C679X
1000 Genomes Project
New arrays, imputation

Rare Mutations
Eg: most mendelian
MC4R, ABCA1
1q21.1 in SCZ
Direct sequencing
Array-based detection (CNV)

Rarer Alleles, Stronger Effects

50%
5%
0.5%
0.05%
Animation: Itsik Pe’er, Columbia
Disease cases 😞  Healthy controls 😊
Association between genotype and phenotype
Age-related macular degeneration

Cohort – 2172 unrelated European descent individuals at least 60 years old

2004: Little known about cause of AMD

934 controls

1238 cases
SNP rs1061170
1238 individuals with AMD and 934 controls
2172 individuals / 4333 alleles

<table>
<thead>
<tr>
<th>Allele</th>
<th>Cases (with AMD)</th>
<th>Controls (without AMD)</th>
<th>Total Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1522 (a)</td>
<td>670 (b)</td>
<td>2192</td>
</tr>
<tr>
<td>T</td>
<td>954 (c)</td>
<td>1198 (d)</td>
<td>2152</td>
</tr>
<tr>
<td>Total</td>
<td>2476</td>
<td>1868</td>
<td>4344</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \]

\[ \chi^2 = 279 \quad \text{Df} = (2 \text{ rows}-1)(2 \text{ columns}-1) = 1 \]

P-value = 1.2 x 10^{-62}
Does the affected or control group exhibit Population Stratification?

- Population stratification is when subpopulations exhibit allelic variation because of ancestry
- Can cause false positives in an association study if there are SNP differences in the case and control population structures
- Control for this artifact by testing control SNPs for general elevation in $\chi^2$ distribution between cases and controls
Age-related macular degeneration

2004: Little known about cause of AMD

2006: Three genes (5 common variants)
Together explain >50% of risk

Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases, $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with $P$ values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P\text{ value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.
Linkage Disequilibrium (LD) between two loci L1 and L1 in gametes

At locus L1
- $p_A$ probability L1 is A
- $q_a$ probability L1 is a

At locus L2
- $p_B$ probability L2 is B
- $q_b$ probability L2 is b

<table>
<thead>
<tr>
<th>L2 B</th>
<th>L2 b</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 A</td>
<td>$P_{AB} = p_A p_B + D$</td>
</tr>
<tr>
<td>L1 a</td>
<td>$P_{aB} = q_a p_B - D$</td>
</tr>
</tbody>
</table>

$D = \text{Measure of linkage disequilibrium}$

$= 0 \text{ when L1 and L2 are in equilibrium}$

$$r^2 = \frac{D^2}{(p_A q_a p_B q_b)}$$

$r$ is $[0,1]$ and is the correlation coefficient between allelic states in L1 and L2
The length of haplotype blocks vs time
\( r^2 \) from human chromosome 22
LD organizes the genome into haplotype blocks

Human genome 5q31 region (associated with Inflammatory Bowel Disease)
Visualizing Haplotype Blocks
http://benfry.com/isometricblocks/
Proxy SNPs

$r^2 = 1$

$r^2 = 0.75$