Genetic Variation and Transcriptional Regulation

BIOM262 - **Quantitative Methods in Genetics and Genomics**

Graham McVicker – Salk Institute

Overview

- Today:
  - Introduction to human genetics, quantitative traits, and genomewide association studies
- Thursday:
  - Mapping molecular quantitative trait loci

Which genetic variants affect human traits?

?

Alzheimer's Disease

Ankylosing Spondylitis

Leukemia

Multiple Sclerosis

Psoriasis

Breast Cancer

Schizophrenia

Rheumatoid Arthritis

Coronary Heart Disease

Celiac Disease

Autism

Crohn's Disease

Parkinson's Disease

Type I Diabetes

Systemic Lupus Erythematosus

ATCCTTTCCTAGAGAAAATAGAAAATAGAACTTTGAGGTTGAATCTCTTTTAATGTAATG
TTTTTCTCGAATCCAAGTGTTTTTACACTATACAATAGGAGTAGAAATTTGTCACCACTC
TGTGGCCAAAYTCACTTTTTCTTTCTTTTTTTATTTTYACATTAAAAAAAAATTTTACTT
TAAGTTCCAGGATACATGTGCAGGATGTGCAGGTTTGTTACATAGGTAAATGTTTTATTT
TAAATTTAATTTAACACTTTTTATYTTTAAGTCATACAACTCTCATAGCCAGTAGTTAAT
ATTACCTTGMAAGTTTGGTATGGTTGATGAATTGCATCCTGTTAATAATTGCTACAGATT
TTTGAATAATTGCAGACCAGTTTGATGRTCCTGGGTTGGCATAAGTACATGAAGATTTAC
TTTTTCCTGTGAGCTTTCTTGGGATGAAGAAATTTAGTGTTTTTTTTAATTTTTAAGAA
ATWTTTATTATTTTTTACATGATTTATTTCCCACTGAAAAATAAATCCCACCGGGYATAA
AGTGTATTTTTTTAAGTCACAGAGTAACCCAACTTGAAGCTAGTTTTTCAGACTTAGGCA
GTTCATGCTGTAAGCCCGAGATCTCATGGTCACCCTTGCAAGAGAAATATCTAATTGAAA
AAAAATATGAAGAGTATTAATTTTGWTAGTGCTAAAATGACATAAAGGGATCTCACTGGG
CTTGAGATATTAAGTATTAAAATTGTTARAGGTTTAAATTGTTAGTAACTTGTTATTGCA
TAGAAAATGTGCCAAATGTCAGTAAATAAAAAAACTTTTTTTAAAATAAAAATTTACAGA
ARAATTATGACRATACTACAAAGAGGTTCTGTACAACCCCCTCCCAGTTTCTCTTACTAT
TAACATCTTAAATTAGTATGTTACATTTGTCACAATTAGTGAACCAATATTGATACATTA
GTACKAACTAAAGTCAGTGTTCCTTTTACTGGAGAATGGTGTTAGAAACTAAGGTCTGGG
CACTGTGGTATGGTGGTTGCTATTGAGATGTYGTTATTTTTAGGTTCTTTCTCAGCTGAC
AGAGCAAAGAAATATATGTGTGTATATTAACCTATGTGTACACATACATCTATGATTATT
TCRATATGTAACATSTGTATCTTTATTAAGCTAAATATGAGTTCATATGGTGTCTTCAAT
TCTAATCAATTACTGTATAGATTATTCTAGCCTCTTCCTCTTGCTTATCTGTAACTTCCT
ATTTCAAACCGTGAAAAATCTGTCTTCCACCACCTACTATCTGCTTACCTAATTTCTYAT
TTCCAGTTTAKGTATACAGTGGCTTCAGAATTATTACATATARCCCTGTGGGATACAACT
TTGTCAACTAGAGTGGTGCTTATGTAAGTTCTTCTATCTTTAGTTTTACTGACTCTACTC

Topics

- Genome-wide association studies
- Molecular quantitative trait loci (QTLs)
- Gene expression QTLs
- Chromatin QTLs
- DNA methylation QTLs
- Intersection of molecular QTLs and GWAS

Testing for genetic association

TTGTTAGAGGTTTAAATT
TTGTTAGAGGTTTAAATT

G G

TTGTTAGAGGTTTAAATT
TTGTTAAAGGTTTAAATT

G A

TTGTTAAAGGTTTAAATT
TTGTTAAAGGTTTAAATT

A A

Case-Control Association Study

Case-Control Association Study

	# A alleles	# B alleles
Cases	750	250
Controls	800	200

χ^2 p-value = 0.009

Quantitative Trait Association Study

Discussion

- What are examples of human traits?
  - Quantitative Traits?
  - Binary (Case/Control) Traits?

In class exercise

- GWAS study results from Rheumatoid Arthritis (Okada et al. 2014)
- Meta-analysis of European and Asian individuals
- 29,880 RA cases and 73,758 controls
- ~10 million SNPs tested
- **Datafile:** RA_GWAS.txt
  (reduced to random set of 100k SNPs)

Exercise

- Read RA_GWAS.txt table into R
- Make a histogram of the p-values

P-value distribution

Exercise

- Make a quantile-quantile plot of -log10 expected vs. observed p-values
- Expected p-values are uniformly distributed

P-value QQPlot

truncated -log10 P-value QQPlot

example of deflated p-values

Discussion

- What can cause ”deflated” p-values?
- Inflated p-values?
- What other information can QQPlot of –log10 P-values provide?

Exercise

- Make a “manhattan” plot with SNP genome position on X-axis, -log10 p-values on Y-axis
- Try to make neighboring chromosomes different colors

What region is this signal from?

Exercise

- What SNP is giving the lowest p-value?
- What gene(s) are in this region (look up in genome browser)

-log10(P-value)
0
50
100
150
200
250
32000000
32200000
32400000
32600000
32800000
33000000
chr6 position

- Why do so many SNPs in this region have low P-values?

MHC Class 2

chr6 (p21.33-p21.32)
p22.3
q12
q13
15
6q21
q27
Scale
500 kb
hg38
chr6:
32,500,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
C4A
C4A
C4A-AS1
AL645922.1
C4B
C4B
C4B-AS1
CYP21A2
CYP21A2
RNA5SP206
TNXB
TNXB
ATF6B
ATF6B
FKBPL
PRRT1
PRRT1
PPT2
PPT2
PPT2
EGFL8
PPT2-EGFL8
EGFL8
RNF5
AGER
PBX2
AGER
C6orf10
RNU6-603P
HCG23
BTNL2
BTNL2
BTNL2
HLA-DRA
HLA-DRA
HLA-DRB5
RNU1-61P
HLA-DRB1
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DQA1
HLA-DQA1
HLA-DQB1
HLA-DQB1
HLA-DQB1
HLA-DQB1
HLA-DQB1-AS1
XXbac-BPG254F23.7
HLA-DQA2
HLA-DQB2
MIR3135B
HLA-DQB2
HLA-DQB2
HLA-DOB
TAP2
TAP2
TAP2
PSMB8
PSMB8
PSMB8
PSMB8-AS1
PSMB8-AS1
PSMB9
TAP1
PSMB9
XXbac-BPG246D15.9
HLA-DMB
HLA-DMA
HLA-DMA
HLA-DMA
BRD2
AL645941.1
BRD2
BRD2-IT1
BRD2
BRD2
BRD2
XXbac-BPG181M17.5
XXbac-BPG181M17.6

Exercise

- What about outside of the MHC region?
- remake manhatten plot, but threshold low p-values to 1e-20
- Draw a line indicating genome-wide significance at p=5e-8
- Roughly how many significant hits are there?

-log10(P-value)
0
5
10
15
20
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
2.5e+09
3.0e+09
genome position

Discussion

- Why use 5e-8 as a threshold for genome-wide significance?
- Why not simply correct for the number of SNPs tested?

Wellcome Trust Case Control Study

- 14,000 cases for 7 diseases
- 3,000 shared controls

WTCCC 2007

Bipolar Disorder
Coronary Artery Disease
Crohn's Disease
Hypertension
Rheumatoid Arthritis
Type 1 Diabetes
Type 2 Diabetes
Observed test statistic
Expected chi-squared value
0
5
10
15
20
25
30

WTCCC 2007

Manhattan Plot of Association Signals

WTCCC 2007

Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
$-\log_{10}(P)$
Chromosome

WTCCC 2007

QTLs for human height

Lango-Allen et al. 2010

Crohn's Disease Hit Region, Chromosome 3

WTCCC 2007

Discussion

- What are advantages of GWAS over candidate gene studies?
- What are some disadvantages / limitations of GWAS?

Difficulty in interpreting GWAS hits

- Most associated variants (hits) are far from known genes
- Which gene(s) do they effect?
- What is their function / mechanism?
- What cell types are the variants active in?
- Which variants are causal?

Molecular traits as an intermediate phenotype

Why use molecular traits?

- Close to underlying genetics
  - affected by small number of polymorphisms
  - require smaller sample sizes
- Can measure 1000s of traits in single experiment (e.g. RNA-seq)
- Reveal molecular basis of organismal traits
  - implicate specific cell types

Discussion

- What else could be used as molecular or cellular traits?
- How would you measure these?

The first gene expression QTL mapping study

Cross two diverged yeast strains “RM” and “BY”

Measure Gene Expression with microarrays

Genotype “segregants” to identify parental genome segments

Brem et al. 2002

Quantitative Trait Locus Mapping with recombinant inbred lines

Lango-Allen et al. 2010

A yeast eQTL

Brem et al. 2002

eQTL mapping in humans

Pedigree-based mapping of eQTLs in human B cell lines

Morley et al. 2004

eQTL mapping with RNA-seq

Pickrell et al. 2010

Splicing QTLs (sQTLs)

Pickrell et al. 2010

eQTLs are enriched near transcription start sites

Battle et al. 2014

eQTLs are enriched in regions with open/active chromatin

Gaffney et al. 2012

Distance from TSS and DNase sensitivity are predictive of eQTLs

Gaffney et al. 2012

Break for exercises

Chromatin as a molecular trait

Ecker et al. 2012

Discussion

- What aspects of chromatin could be treated as a molecular trait?
- Do these traits have discrete or continuous values?

Nucleosomes are depleted in regulatory regions

Raj & McVicker 2013

Measuring chromatin accessibility with DNase-seq

digest chromatin with Deoxyribonuclease 1 (DNase1)

Measuring chromatin accessibility with DNase-seq

Measuring chromatin accessibility with DNase-seq

sequence
DNA fragments

TTGTTAAGGTTTATGTTCGTTTAATGGTG
TTTCTTACGACTGTACGATCAAAACGGGG
AGGCGGCAAGCAGGTGCAGCGTTTTTATA
GGGCTACAACACGTTGGTGCACCCAACAC

Measuring chromatin accessibility with DNase-seq

TTGTTAAGGTTTATGTTCGTTTAATGGTG

TTTCTTACGACTGTACGATCAAAACGGGG

AGGCGGCAAGCAGGTGCAGCGTTTTTATA

GGGCTACAACACGTTGGTGCACCCAACAC

map to genome

GTTCGTTTAATGGTGGCCGGAGGG
GTTTATGTTCGTTTAATGGTGCCG
GTTTATGTTCGTTTAATGGTGCCG
AAGGTTTATGTTCGTTTAATGGTG
TTGTTAAGGTTTATGTTCGTTTAA

...ACCATTGTTAAGGTTTATGTTCGTTTAATGGTGCCGGGAGGGATTTA...

Measuring chromatin accessibility with DNase-seq

TTGTTAAGGTTTATGTTCGTTTAATGGTG

TTTCTTACGACTGTACGATCAAAACGGGG

AGGCGGCAAGCAGGTGCAGCGTTTTTATA

GGGCTACAACACGTTGGTGCACCCAACAC

DNase hypersensitive sites

DNase sensitivity as a quantitative trait

distance from center of DNase peak

Degner et al. 2012

DNase sensitivity as a quantitative trait

NF-κB motif

Degner et al. 2012

Determining nucleosome positions with MNase-seq

digest chromatin with micrococcal nuclease (MNase)

Determining nucleosome positions with MNase-seq

isolate nucleosome-sized fragments and sequence ends

Determining nucleosome positions with MNase-seq

peaks are consistently-positioned nucleosomes

dsQTLs are associated with nucleosome positions

Gaffney*, McVicker* et al. 2012

dsQTLs are associated with nucleosome positions

Gaffney*, McVicker* et al. 2012

Histone modifications

H3K4me1: active/open chromatin outside of promoters

H3K4me3: active promoters

H3K27ac: active promoters & enhancers

H3K27me3: silenced genes

Measuring histone marks with ChIP-seq

Measuring histone marks with ChIP-seq

Chromatin Immunoprecipitation (ChIP)

Measuring histone marks with ChIP-seq

High-throughput
DNA sequencing

TTGTTAAGGTTTATGTTCGTTTAATGGTG

TTTCTTACGACTGTACGATCAAAACGGGG

AGGCGGCAAGCAGGTGCAGCGTTTTTATA

GGGCTACAACACGTTGGTGCACCCAACAC

Measuring histone marks with ChIP-seq

TTGTTAAGGTTTATGTTCGTTTAATGGTG

TTTCTTACGACTGTACGATCAAAACGGGG

AGGCGGCAAGCAGGTGCAGCGTTTTTATA

GGGCTACAACACGTTGGTGCACCCAACAC

Map reads to genome

GTTCGTTTAATGGTGGCCGGAGGG
GTTTATGTTCGTTTAATGGTGCCG
GTTTATGTTCGTTTAATGGTGCCG
AAGGTTTATGTTCGTTTAATGGTG
TTGTTAAGGTTTATGTTCGTTTAA

...ACCATTGTTAAGGTTTATGTTCGTTTAATGGTGCCGGGAGGGATTTA...

Measuring histone marks with ChIP-seq

TTGTTAAGGTTTATGTTCGTTTAATGGTG

TTTCTTACGACTGTACGATCAAAACGGGG

AGGCGGCAAGCAGGTGCAGCGTTTTTATA

GGGCTACAACACGTTGGTGCACCCAACAC

Identify ChIP-seq peaks

ChIP-seq data from 10 individuals

H3K4me1

H3K4me3

H3K27ac

H3K27me3

Are eQTLs also associated with histone modifications?

Are eQTLs also associated with histone modifications?

McVicker*, van de Geijn* et al. 2013

Are dsQTLs also associated with histone modifications?

Are dsQTLs also associated with histone modifications?

McVicker*, van de Geijn* et al. 2013

Can we identify Histone Mark QTLs?

Combined Haplotype Test

Read Depth Association Test

Model read counts with Poisson distribution:

λ	=	2α	$\alpha+\beta$	2β

Allelic Imbalance Test

Model reference proportion with binomial distribution:

$$p = \frac{\alpha}{\alpha+\beta}$$

Haplotype Imbalance Test

- Phase SNPs
- Test allelic imbalance across entire haplotype

Combined Haplotype Test

Read Depth Association Test

$$\lambda = \begin{cases} 2\alpha & \text{if homozygous AA} \\ \alpha+\beta & \text{if heterozygous AB} \\ 2\beta & \text{if homozygous BB} \end{cases}$$

+

Haplotype Imbalance Test

$$p = \frac{\alpha}{\alpha+\beta}$$

null hypothesis, $\mathbf{H_0}$: $\alpha = \beta$

alternative hypothesis, $\mathbf{H_1}$: $\alpha \neq \beta$

Standard mapping of H3K27ac QTLs with 10 Individuals

Combined Haplotype Test Mapping of H3K27ac QTLs

van de Geijn*, McVicker*, et al. 2015

Example of a histone mark QTL

McVicker*, van de Geijn* et al. 2013

Example of a histone mark QTL

McVicker*, van de Geijn* et al. 2013

Example of a histone mark QTL

McVicker*, van de Geijn* et al. 2013

Example of a histone mark QTL

McVicker*, van de Geijn* et al. 2013

DNA methylation QTLs

Banovich et al. 2014

DNA Methylation QTLs are often also DNase sensitivity QTLs

Banovich et al. 2014

Histone mark QTLs are often DNA methylation QTLs

Banovich et al. 2014

Discussion

- How can molecular traits be connected to organismal traits?
- What can we learn from molecular traits?

Intersecting eQTLs with GWAS

Raj et al. 2014

GWAS enrichment in cell-type specific annotations

Finucane et al. 2015

Finucane et al. 2015

Cytokine induced cell states
0
16
120 h
Naive
Memory
CD4+ T cells
0
6
24 h
Monocytes
Monocyte-derived macrophage
Chromatin activity (ATAC-seq and H3K27ac ChIPmentation-seq)
Sequence reads
Genomic position
CD4 T cells
Memory
Naive
16 h
120
16 h
120 h
Resting
TH0
TH1
TH2
TH17
iTreg
IL-10
IL-21
IL-27
IFNβ
TNFα
Macrophages
6 h
24 h
Resting
IFNγ
IL-4
IL-23
IL-26
TNFα

Soskic et al. 2019

Soskic et al. 2019

Summary

- Many molecular traits can be associated with genetic variants
- Molecular QTLs can reveal mechanism underlying organismal traits
- Smaller samples are needed to map molecular QTLs that organismal traits
- Genetic associations are challenging to interpret