Linking Genetic Variation to Important Phenotypes

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2017
Anthony Gitter
gitter@biostat.wisc.edu

These slides, excluding third-party material, are licensed under CC BY-NC 4.0 by Mark Craven, Colin Dewey, and Anthony Gitter
Outline

• How does the genome vary between individuals?
• How do we identify associations between genetic variations and simple phenotypes/diseases?
• How do we identify associations between genetic variations and complex phenotypes/diseases?
Understanding Human Genetic Variation

- The “human genome” was determined by sequencing DNA from a small number of individuals (2001)
- The HapMap project (initiated in 2002) looked at polymorphisms in 270 individuals (Affymetrix GeneChip)
- The 1000 Genomes project (initiated in 2008) sequenced the genomes of 2500 individuals from diverse populations
- 23andMe genotyped its 1 millionth customer in 2015
- Genomics England plans to sequence 100k whole genomes by 2017 and link with medical records
Classes of Variants

- Single Nucleotide Polymorphisms (SNPs)
- Indels (insertions/deletions)
- Structural variants

Formal definitions: https://www.snpedia.com/index.php/Glossary
Single Nucleotide Polymorphisms (SNPs)

One nucleotide changes

Variation occurs with some minimal frequency in a population

Pronounced “snip”
Insertions and Deletions

Black box: DNA template strand
White box: newly replicated DNA

Insertion: slippage inserts extra nucleotides

Deletion: slippage excludes template nucleotides

Structural Variants

- **Copy number variants (CNVs)**
  - Gain or loss or large genomic regions, even entire chromosomes
- **Inversions**
  - DNA subsequence is reversed
- **Translocations**
  - DNA subsequence is moved to a different chromosome
Genetic Recombination

Normal Recombination

Meiosis I

Equal Crossover between Allelic sequence

2/4 products

Egg/Sperm
Recombination Errors Lead to Copy Number Variants (CNVs)

Abnormal Recombination

Unequal Crossover between duplicated sequences

Copy-number gain of genes ABC and Copy-number loss of genes ABC

2/4 products
1000 Genomes Project

Project goal: produce a catalog of human variation down to variants that occur at \( \geq 1\% \) frequency over the genome.
Understanding Associations Between Genetic Variation and Disease

*Genome-wide association study* (GWAS)

- Gather some population of individuals
- Genotype each individual at polymorphic markers (usually SNPs)
- Test association between state at marker and some variable of interest (say disease)
- Adjust for multiple comparisons

- Phenotypes: observable traits
Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.
Wellcome Trust GWAS
Morning Person GWAS

Hu et al. *Nature Communications* 2016
Understanding Associations Between Genetic Variation and Disease

International Cancer Genome Consortium

- Includes NIH’s *The Cancer Genome Atlas*
- Sequencing DNA from 500 tumor samples for each of 50 different cancers

- Goal is to distinguish *drivers* (mutations that cause and accelerate cancers) from *passengers* (mutations that are byproducts of cancer’s growth)
A Circos Plot
Some Cancer Genomes

**LUNG CANCER**
Cancer: small-cell lung carcinoma
- Sequenced: full genome
- Source: NCI-H209 cell line
- Point mutations: 22,910
- Point mutations in gene regions: 134
- Genomic rearrangements: 58
- Copy-number changes: 334

**SKIN CANCER**
Cancer: metastatic melanoma
- Sequenced: full genome
- Source: COLO-829 cell line
- Point mutations: 33,345
- Point mutations in gene regions: 292
- Genomic rearrangements: 51
- Copy-number changes: 41

**BREAST CANCER**
Cancer: basal-like breast cancer
- Sequenced: full genome
- Source: primary tumour, brain metastasis, and tumours transplanted into mice
- Point mutations: 27,173 in primary, 51,710 in metastasis and 109,078 in transplant
- Point mutations in gene regions: 200 in primary, 225 in metastasis, 328 in transplant
- Genomic rearrangements: 34
- Copy-number changes: 155 in primary, 101 in metastasis, 97 in transplant
Understanding Associations Between Genetic Variation and Complex Phenotypes

Quantitative trait loci (QTL) mapping
- Gather some population of individuals
- Genotype each individual at polymorphic markers
- Map quantitative trait(s) of interest to chromosomal locations that seem to explain variation in trait
QTL Mapping Example

Inbred strain A

X

Inbred strain B

F_1

Markers ~30 cM apart

F_2
QTL Mapping Example

QTL mapping of mouse blood pressure, heart rate
[Sugiyama et al., Broman et al.]

\[ \text{LOD}(q) = \log_{10} \frac{P(q \mid \text{QTL at } m)}{P(q \mid \text{no QTL at } m)} \]

Logarithm of Odds

quantitative trait

position in the genome
QTL Example: Genotype-Tissue Expression Project (GTEx)

• Expression QTL (eQTL): traits are expression levels of various genes

• Map genotype to gene expression in different human tissues
QTL Example: GTEx

https://www.genome.gov/27543767/
GWAS Versus QTL

• Both associate genotype with phenotype

• GWAS pertains to discrete phenotypes
  – For example, disease status is binary

• QTL pertains to quantitative (continuous) phenotypes
  – Height
  – Gene expression
  – Splicing events
  – Metabolite abundance
Determining Association is Not Enough

A simple case: CFTR
Many Measured SNPs Not in Coding Regions

• Genes encoding CD40 and CD40L with relative positions of the SNPs studied

Chadha et al. *Eur J Hum Genet* 2005
Computational Problems

• Assembly and alignment of thousands of genomes

• Data structures to capture extensive variation

• Identifying functional roles of markers of interest (which genes/pathways does a mutation affect and how?)

• Identifying interactions in multi-allelic diseases (which combinations of mutations lead to a disease state?)

• Identifying genetic/environmental interactions that lead to disease

• Inferring network models that exploit all sources of evidence: genotype, expression, metabolic