Linking Genetic Variation to Important Phenotypes:

SNPs, CNVs, GWAS, and eQTLs

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Colin Dewey

cdewey@biostat.wisc.edu

Spring 2015

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 genetics variations and complex phenotypes/diseases?

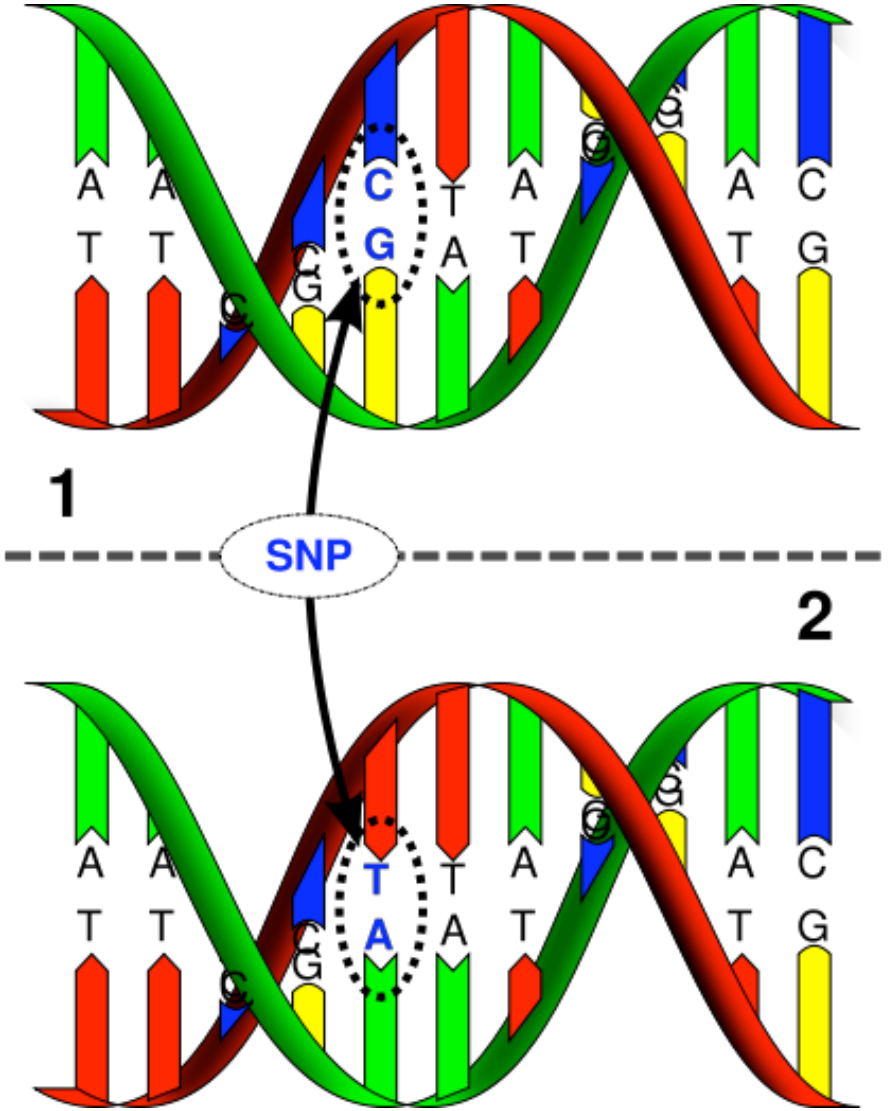
1. Understanding Human Genetic Variation

- 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 1000 individuals from diverse populations

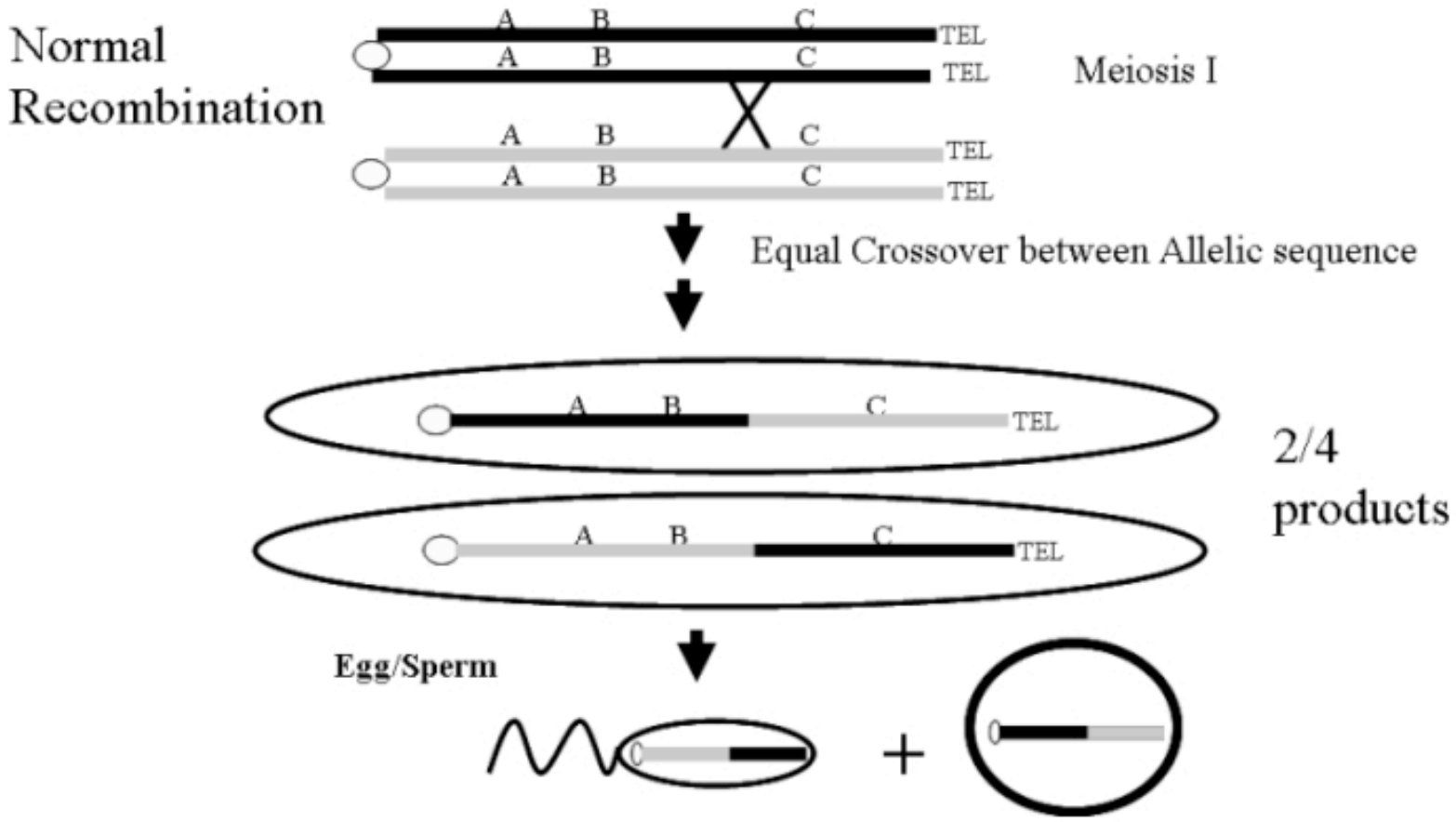
Classes of variants

- Single Nucleotide Polymorphisms (SNPs)
- Indels (insertions/deletions)
- Copy number variants (CNVs)
- Other structural variants
  - Inversions
  - Transpositions

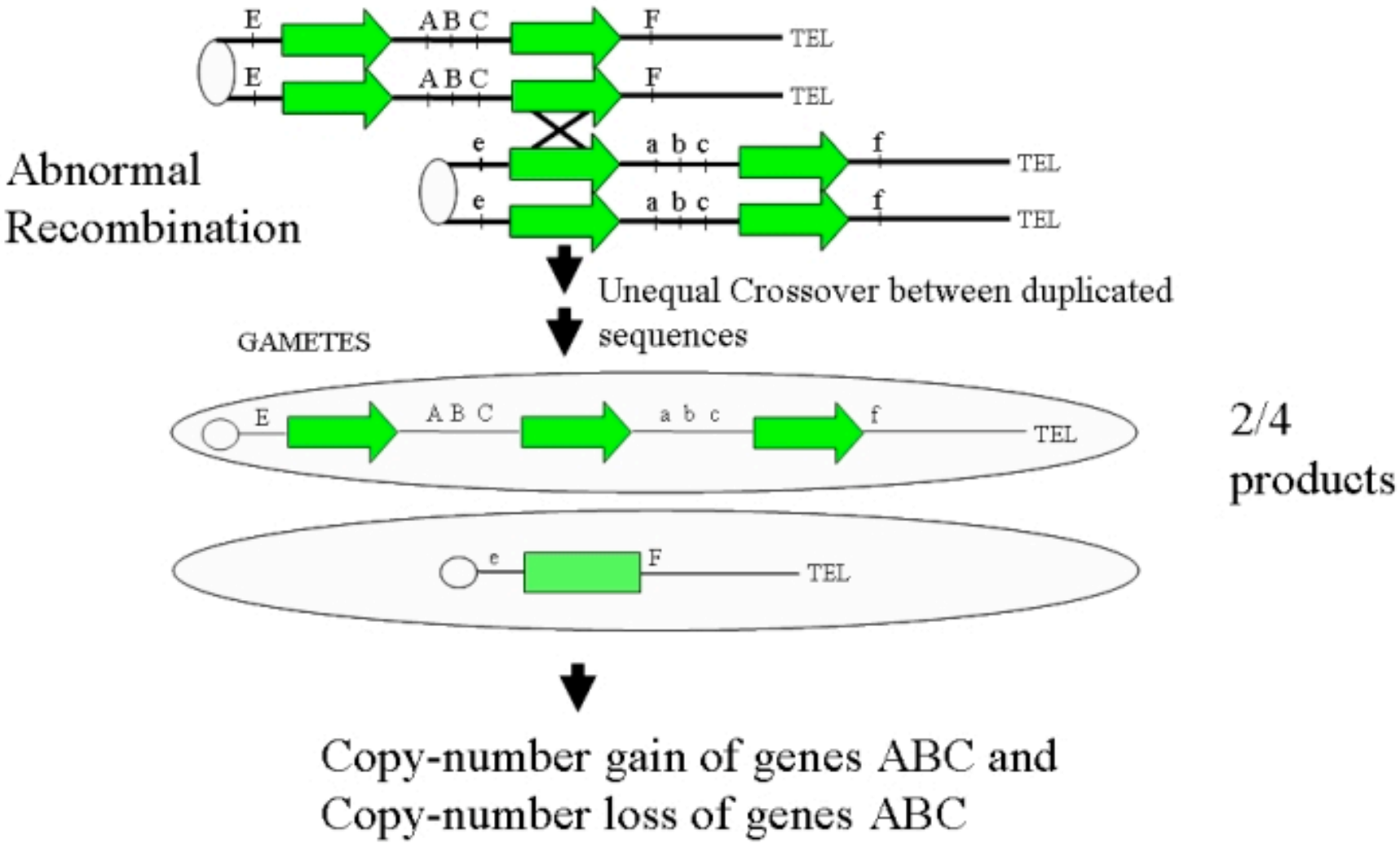
Genetic Variation:
Single Nucleotide Polymorphisms (SNPs)



Genetic Variation:
Copy Number Variants (CNVs)

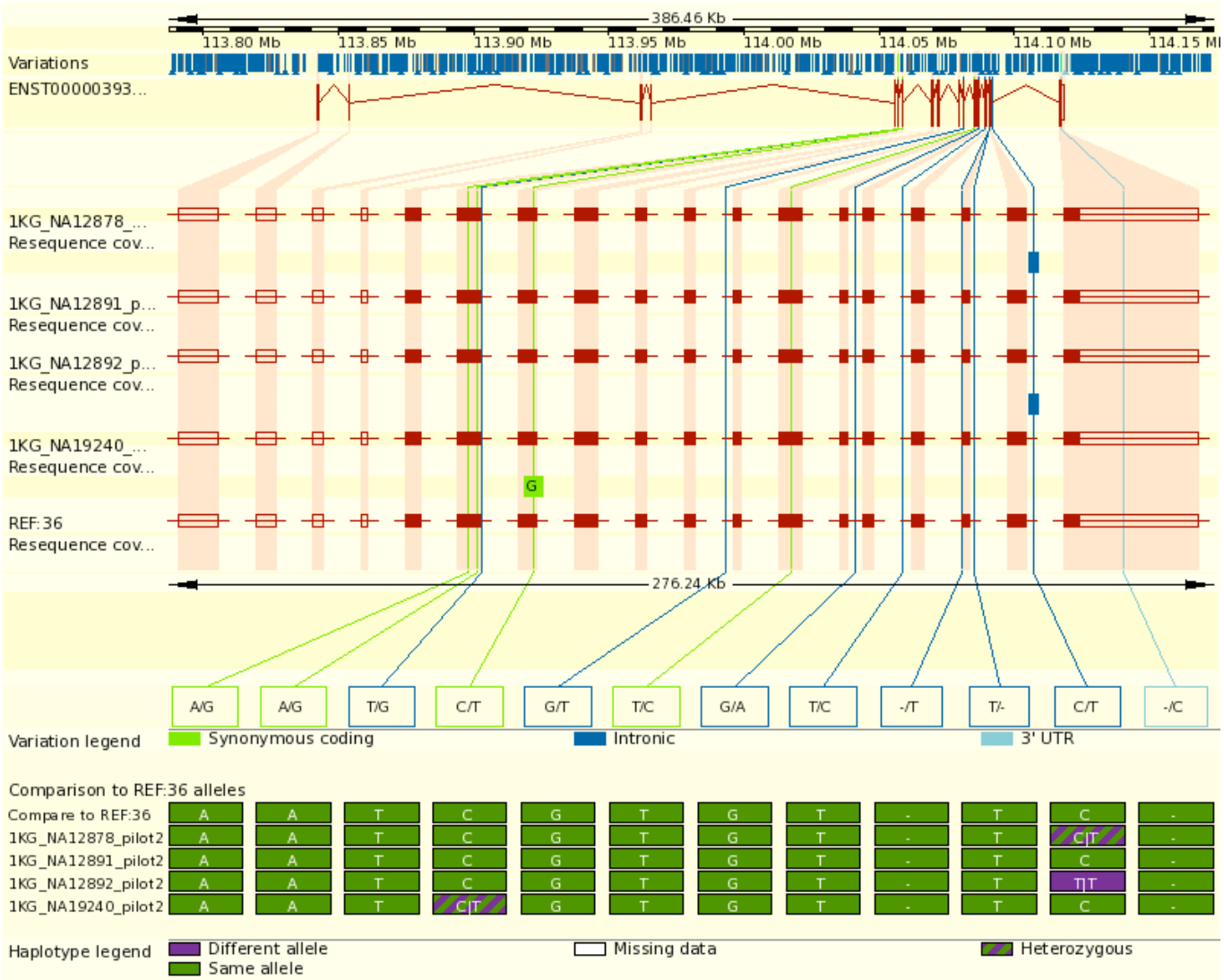


Genetic Variation:
Copy Number Variants (CNVs)



1000 Genomes Project

project goal: produce a catalog of human variation down to variants that occur at >= 1% frequency over the genome

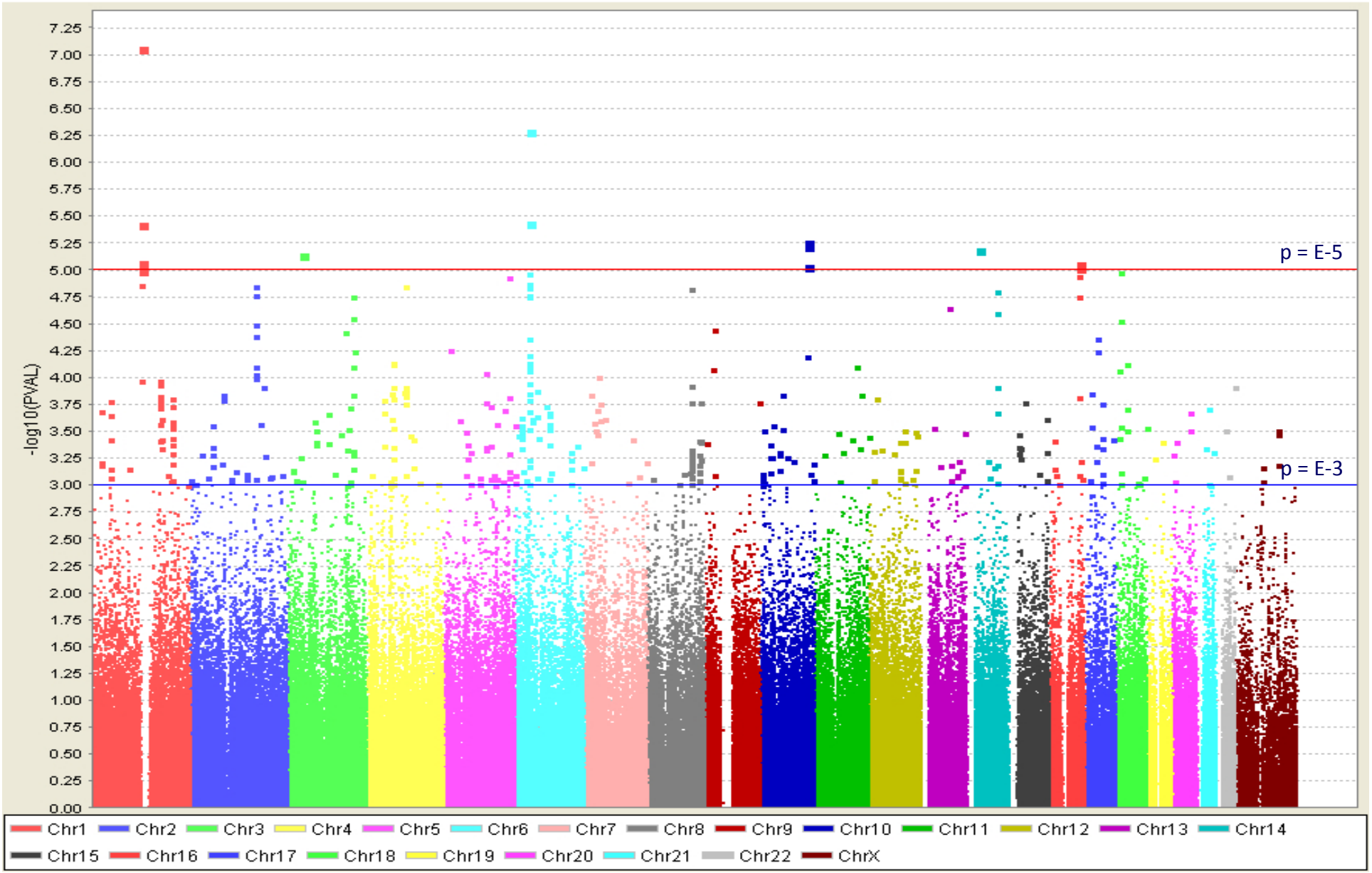


2. Understanding Associations Between Genetic Variation and Disease

genome wide association study (GWAS)

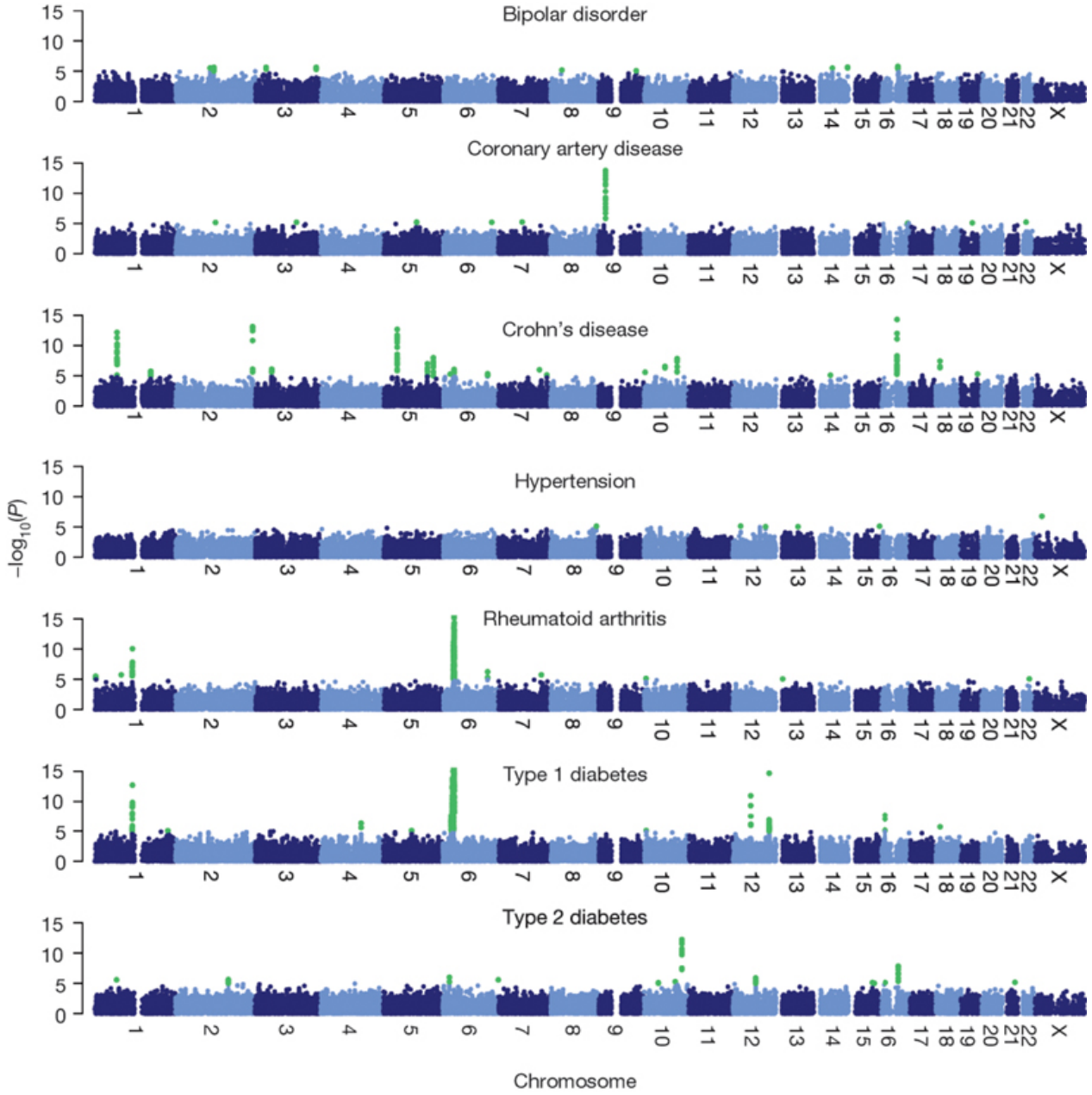
- gather some population of individuals
- genotype each individual at polymorphic markers (usually SNPs)
- test association between *state* at marker and some variable of interest (say disease)
- adjust for multiple comparisons

Type 2 Diabetes Results: 386,731 markers



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



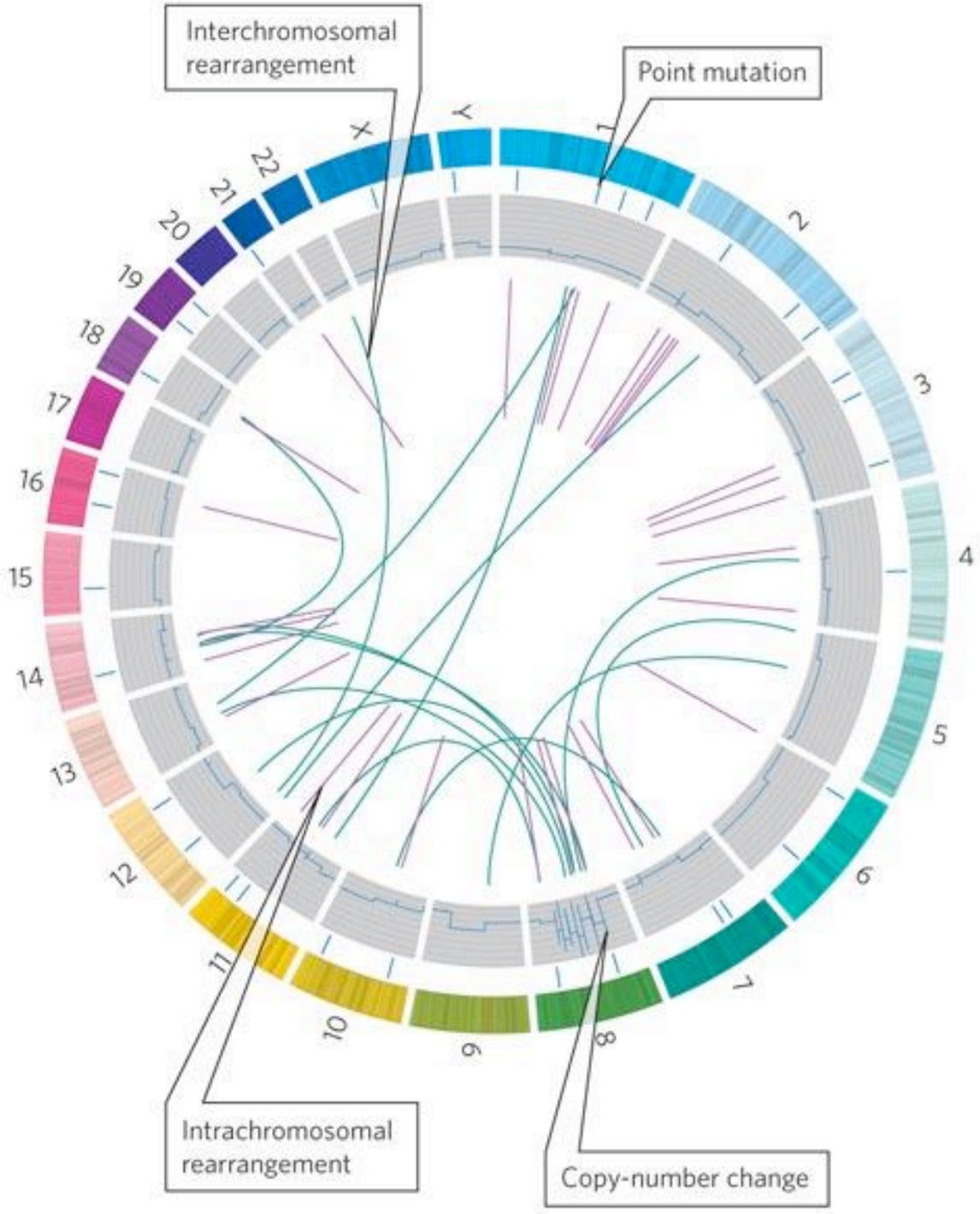
2. Understanding Associations Between Genetic Variation and Disease

International Cancer Genome Consortium

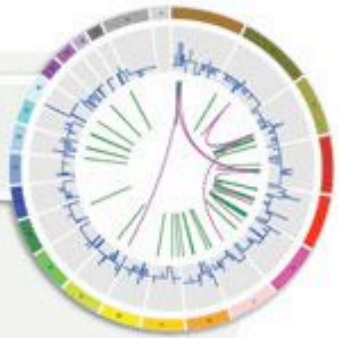
- includes the NIH's *Cancer Genome Atlas*
- sequencing DNA from 500 tumor samples for each of 50 different cancers

- goal is to distinguish *drivers* (mutations that cause and accelerate cancers) from *passengers* (mutations that are byproducts of cancer's growth)

A Circos Plot

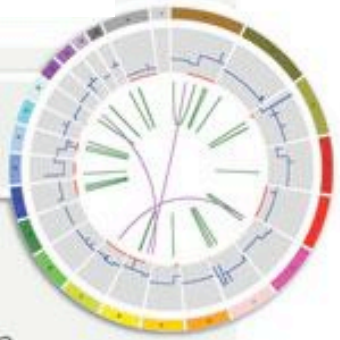


Some Cancer Genomes



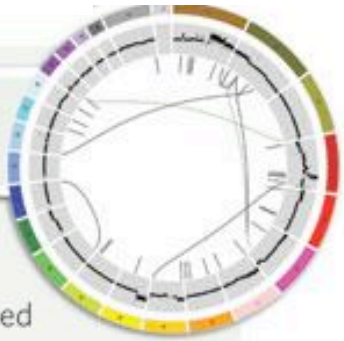
LUNG CANCER
Cancer: small-cell lung carcinoma

- 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

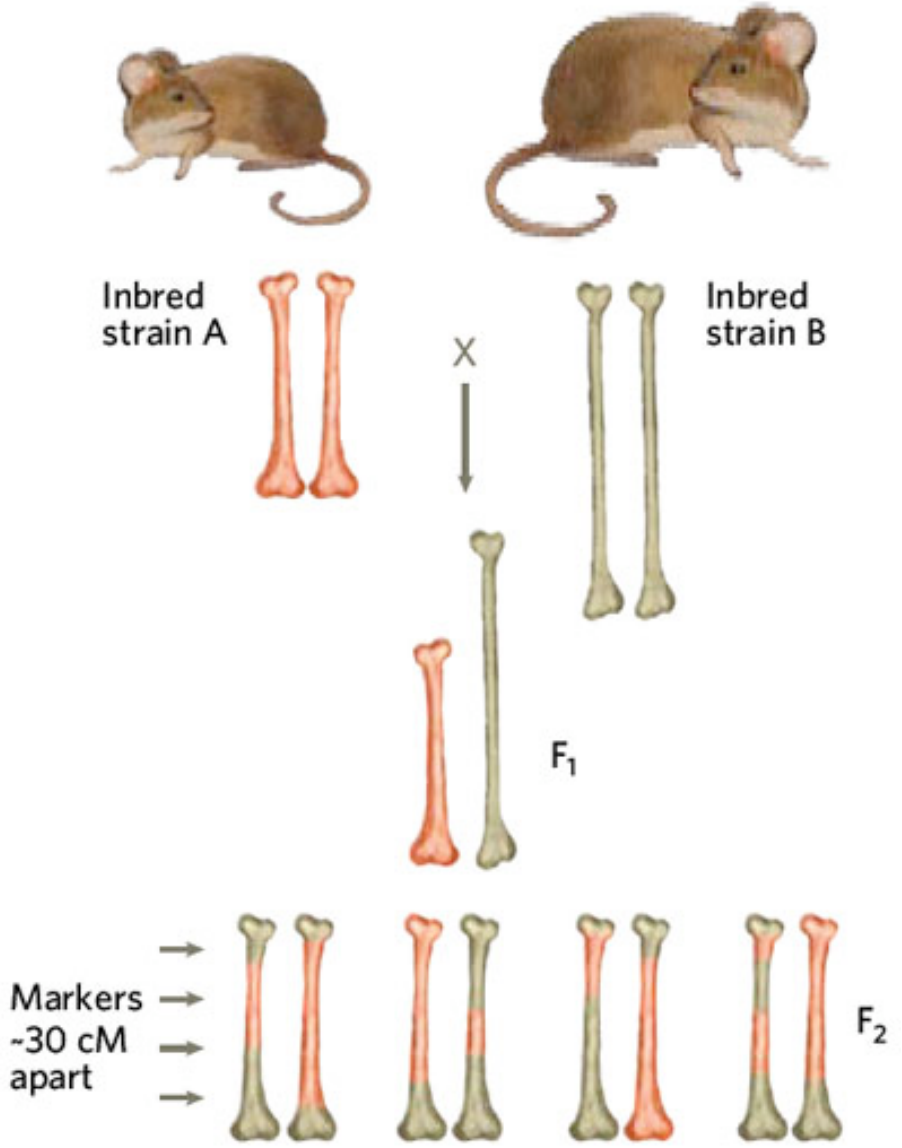
- Sequenced: full genome
- Source: primary tumour, brain metastasis, and tumours transplanted into mice
- Point mutations: 27,173 in primary, 51,710 in metastasis and 109,078 in transplant
- Point mutations in gene regions: 200 in primary, 225 in metastasis, 328 in transplant
- Genomic rearrangements: 34
- Copy-number changes: 155 in primary, 101 in metastasis, 97 in transplant

3. Understanding Associations Between Genetic Variation and Complex Phenotypes

quantitative trait loci (QTL) mapping

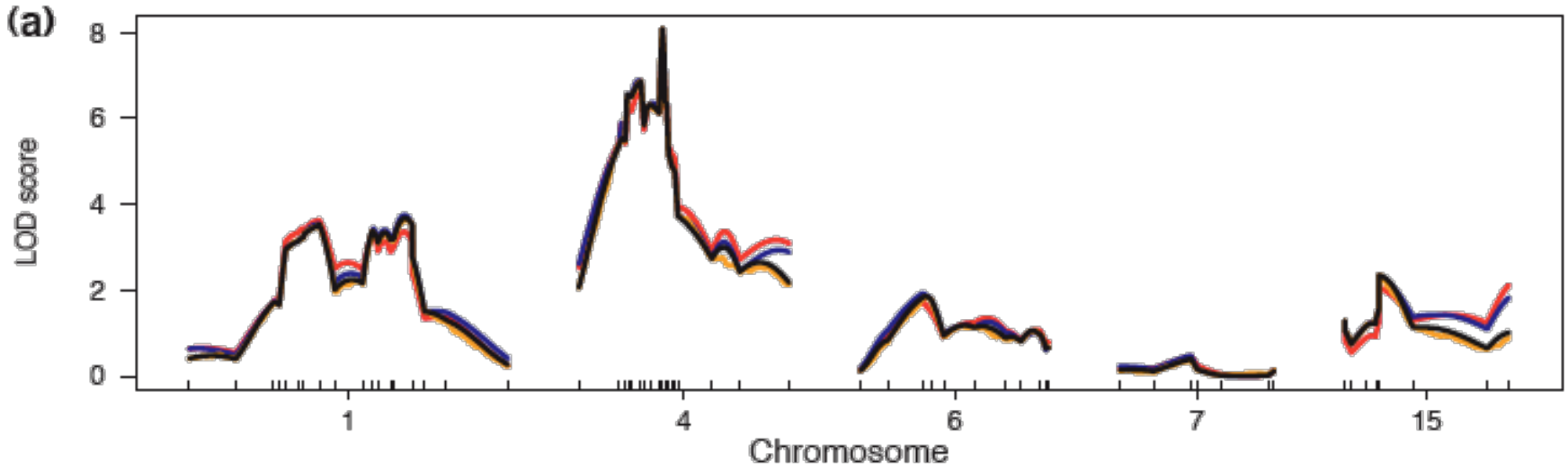
- gather some population of individuals
- genotype each individual at polymorphic markers
- map quantitative trait(s) of interest to chromosomal locations that seem to explain variation in trait

QTL Mapping Example



QTL Mapping Example

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



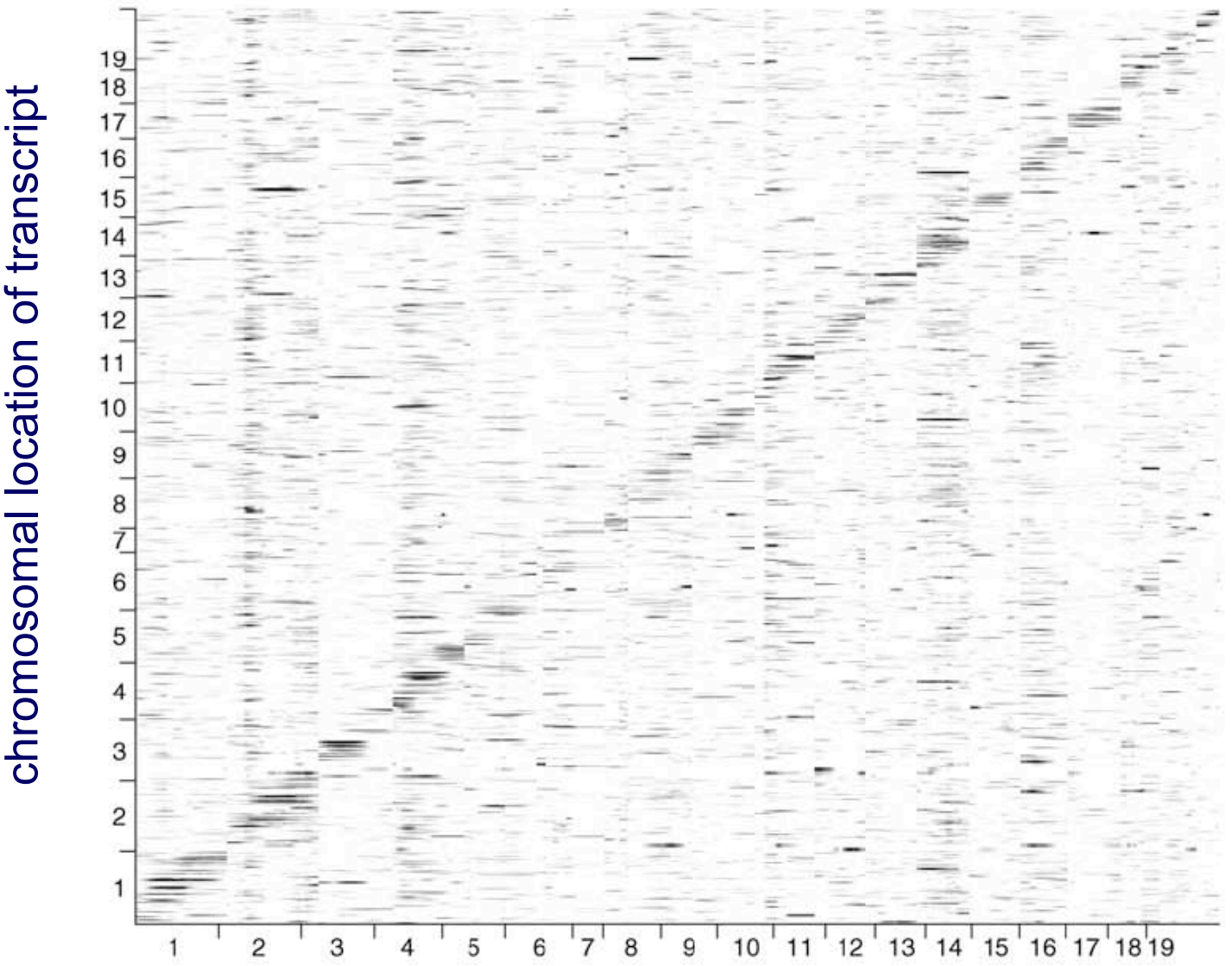
$$\mathrm{LOD}(q) = \log_{10} \frac{P(q \mid \text{QTL at } m)}{P(q \mid \text{no QTL at } m)}$$

QTL Mapping

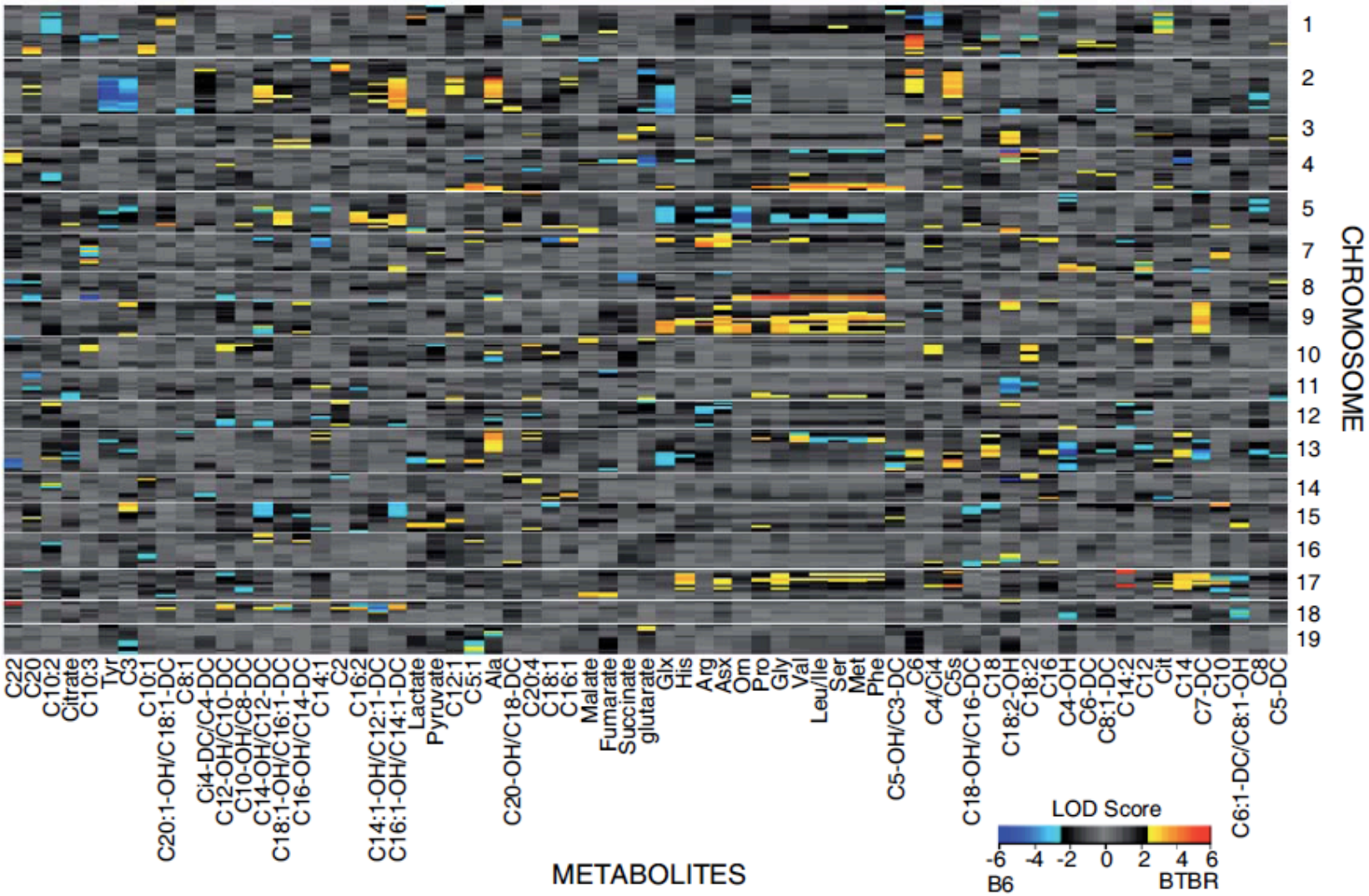
- QTL mapping can be done for large-scale quantitative data sets
  - expression QTL: traits are expression levels of various genes
  - metabolic QTL: traits are metabolite levels

- case study: uncovering the genetic/metabolic basis of diabetes (Attie lab at UW)

Expression QTL (eQTL) Mapping

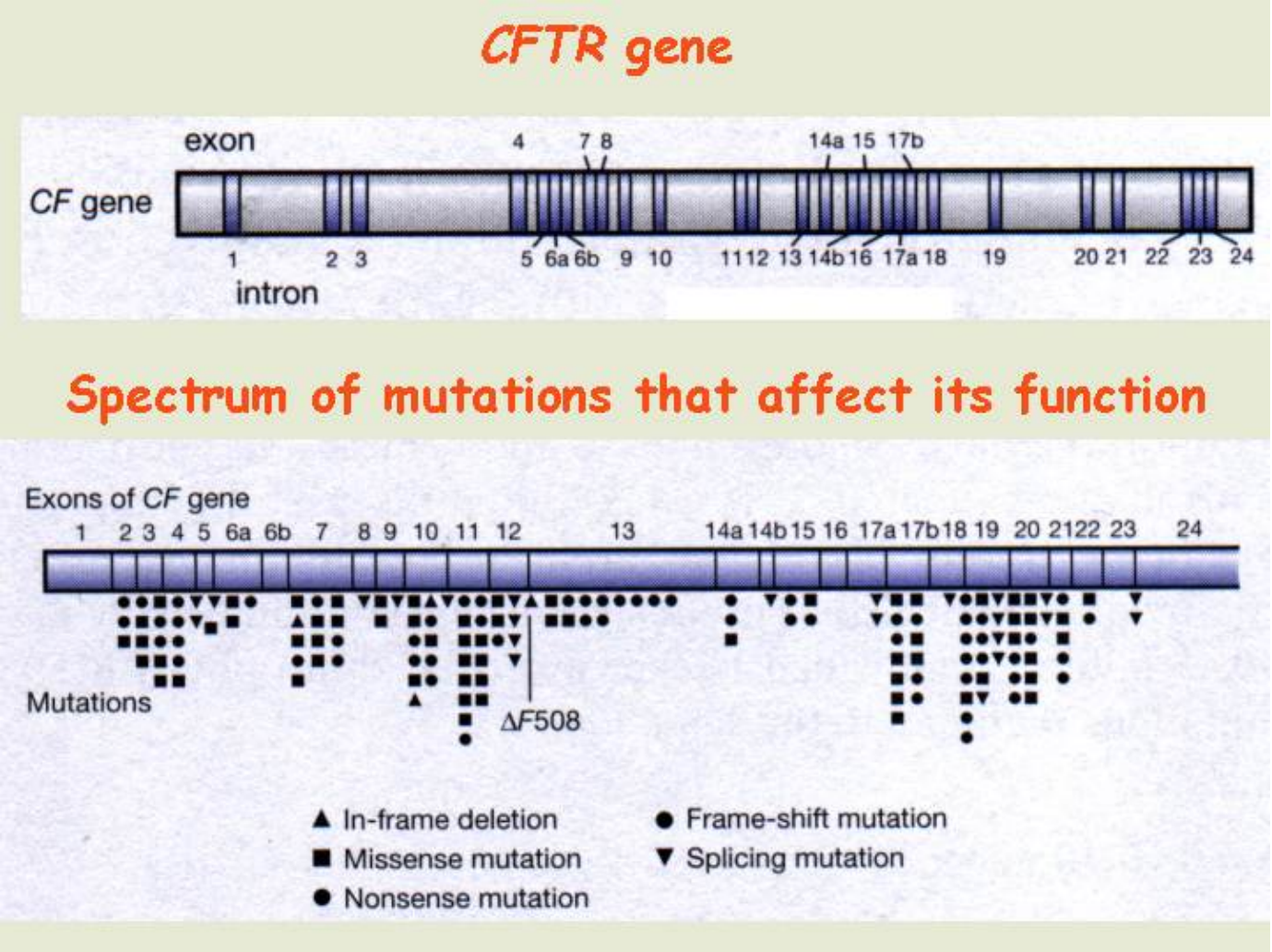


Metabolite QTL (mQTL) Mapping



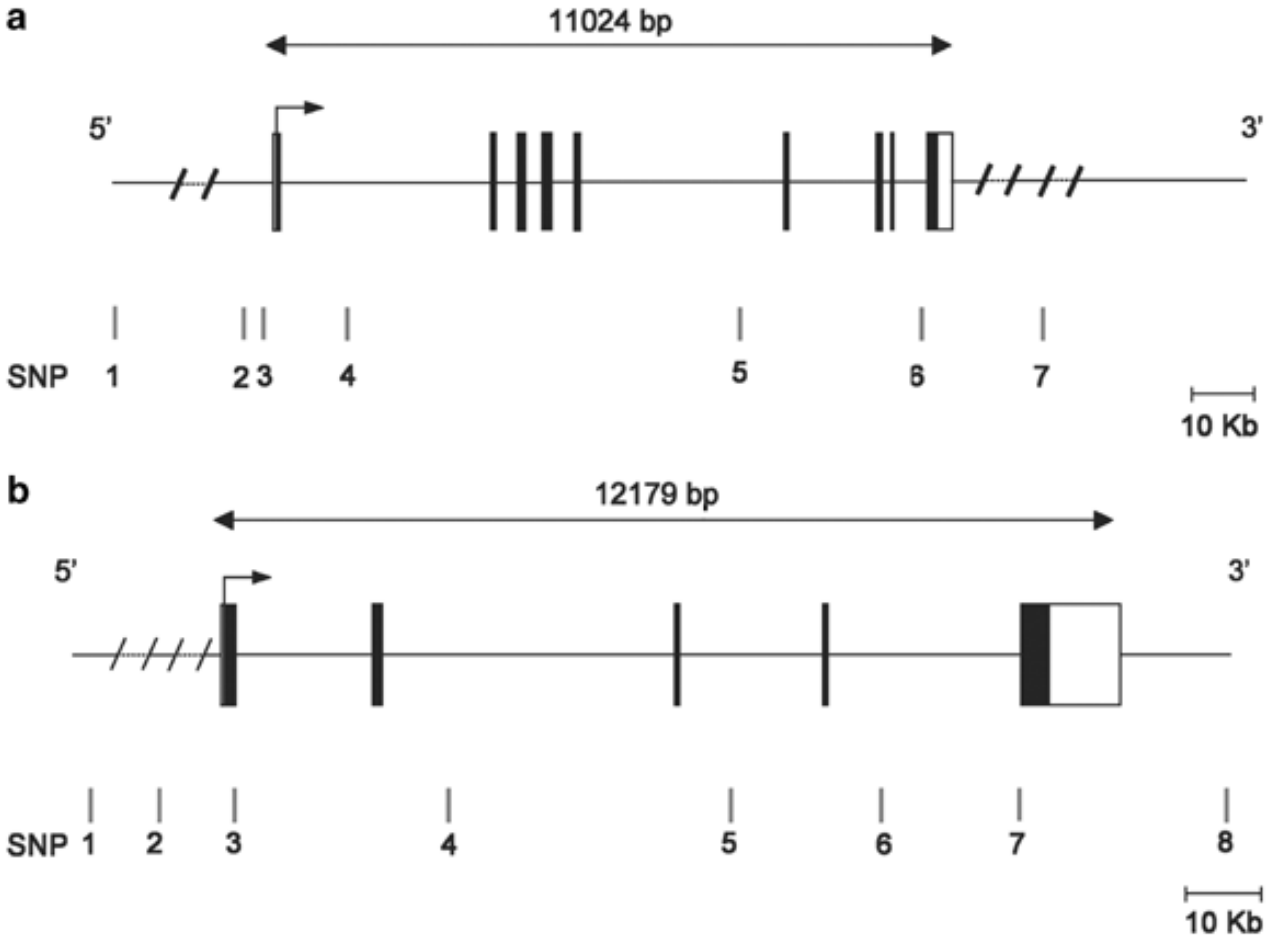
Is determining association the end of the story?

a simple case: CFTR

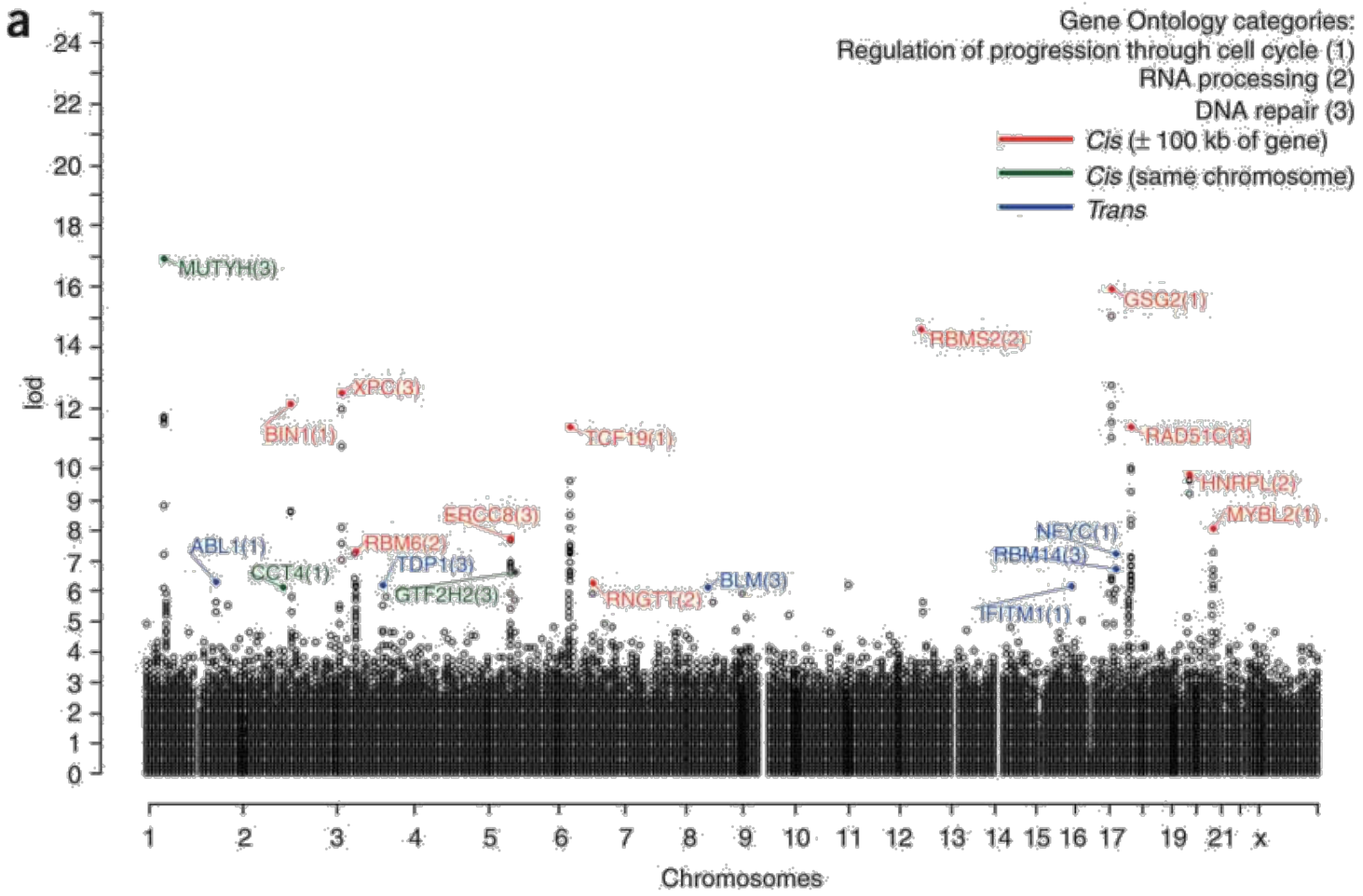


Many measured SNPs not in coding regions

- genes encoding CD40 and CD40L with relative positions of the SNPs studied



Expression QTL (eQTL) Mapping



Computational Problems

- assembly and alignment of thousands of genomes
- 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