University of Michigan
Medical School

Introduction to eQTLs and Overview of Group Projects

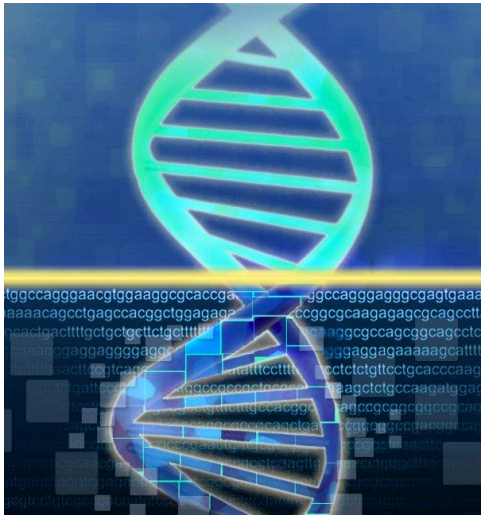
Ryan E. Mills, Ph.D.

Department of Computational Medicine & Bioinformatics
Department of Human Genetics
University of Michigan Medical School
Ann Arbor, MI, USA

University of Michigan
Medical School

Genotypes and Phenotypes

An individual's genotype is their heritable genetic information, as encoded in their respective genome



A phenotype is an observable property resulting from that genotype as well as any environmental effects

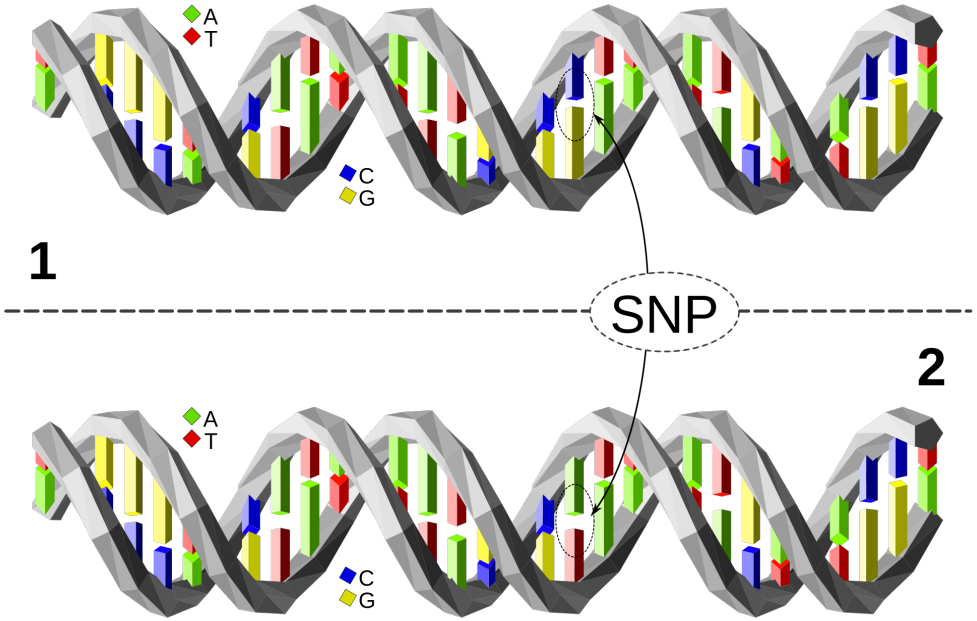
Jonathan Bailey, Gregg Langham, Courtesy: National Human Genome Research Institute.

University of Michigan
Medical School

Individual Human Genotypes

When we speak of 'genotypes', we are typically referring to the millions of positions within the human genome that have been observed to be different between some individuals.

We refer to these differences as *polymorphisms.* A **S**ingle **N**ucleotide **P**olymorphism is denoted as a *SNP* (*<snip>*)

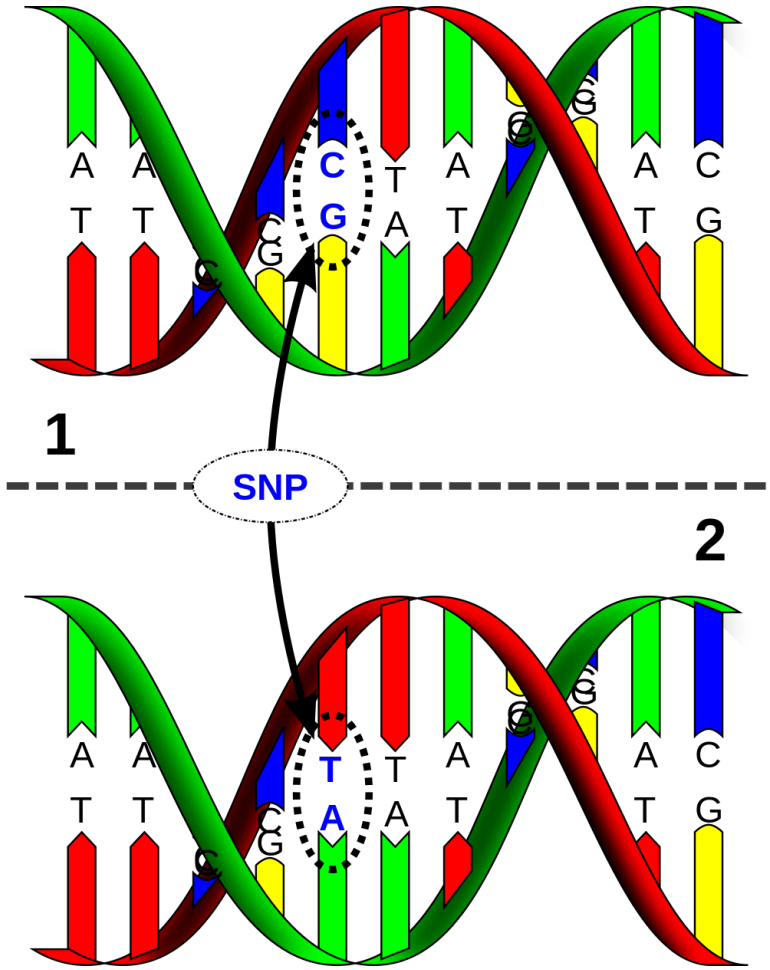


Describing Genotypes

An individual position in the genome, termed a *nucleotide* or *base*, can be one of 4 *alleles* (A,C,T,G).

Diploid organisms such as humans have both maternal and paternal alleles.

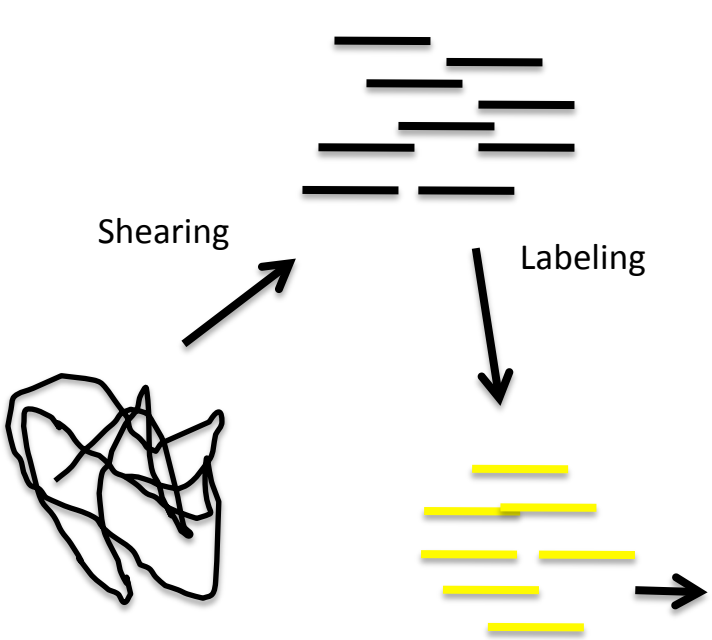
Genotypes at these positions are thus described as combinations of these alleles, e.g. C/C, C/G, and G/G.

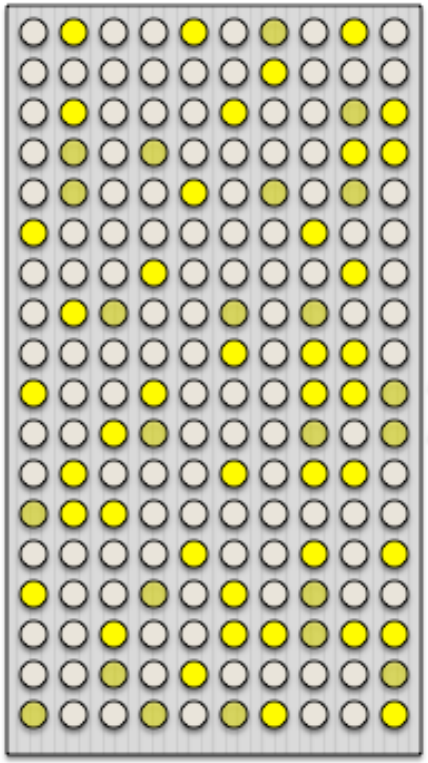


Ascertaining Genotypes

University of Michigan Medical School

SNP Microarrays





TAACGATGAATC**T**TAGGCATCGCGC

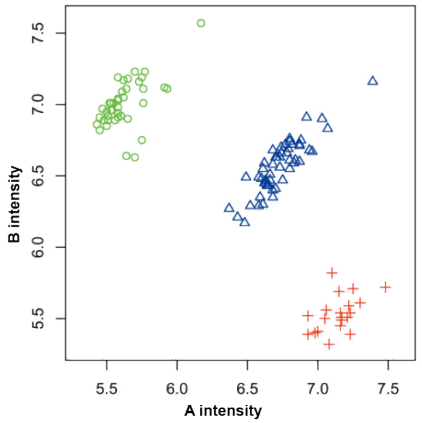
TAACGATGAATC**G**TAGGCATCGCGC

genotype: T/T

GGCTTAAGTACC**C**TATGGATTACGG

GGCTTAAGTACC**T**TATGGATTACGG

genotype: C/T

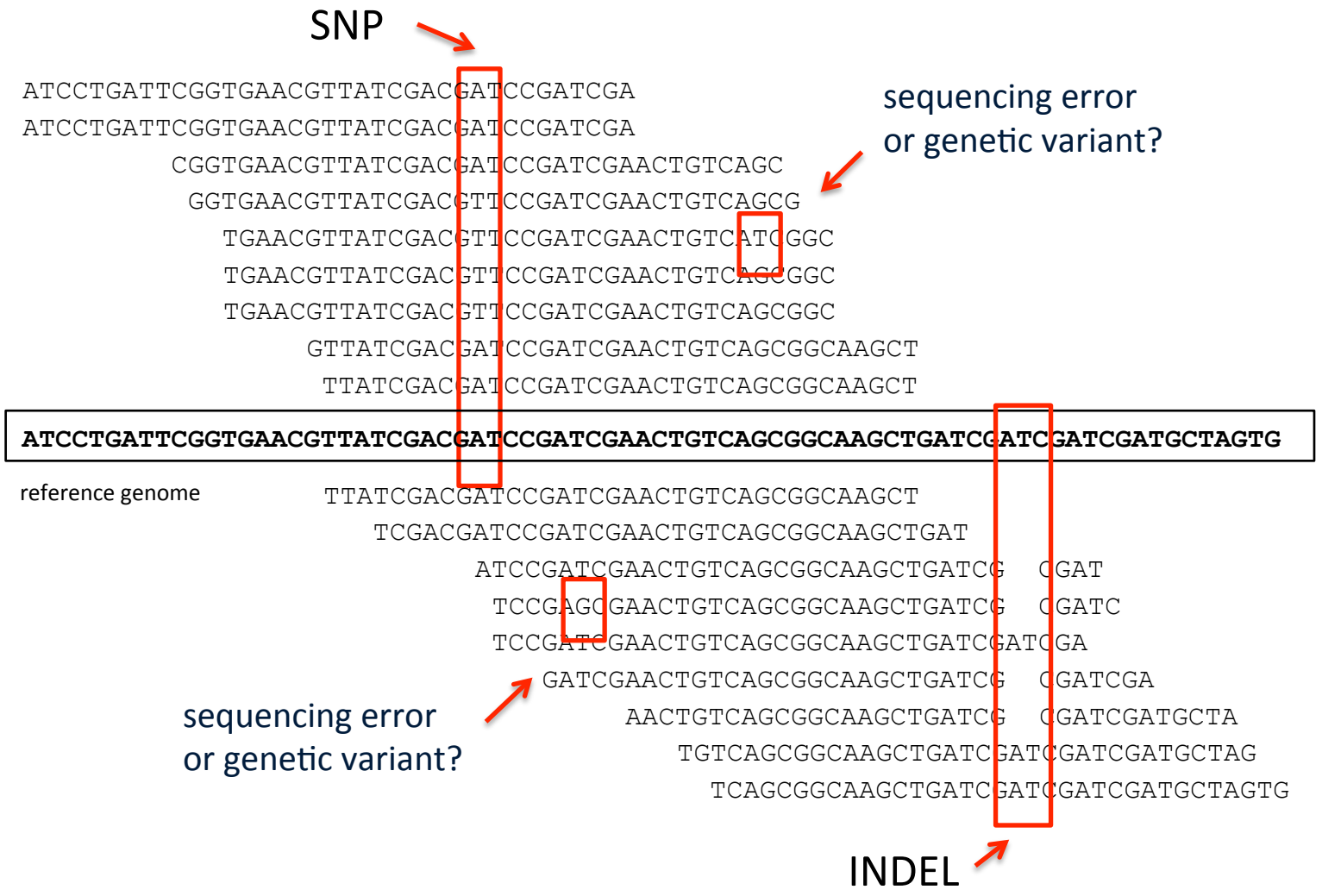


Maggie Bartlett, Courtesy: National Human Genome Research Institute.

University of Michigan
Medical School

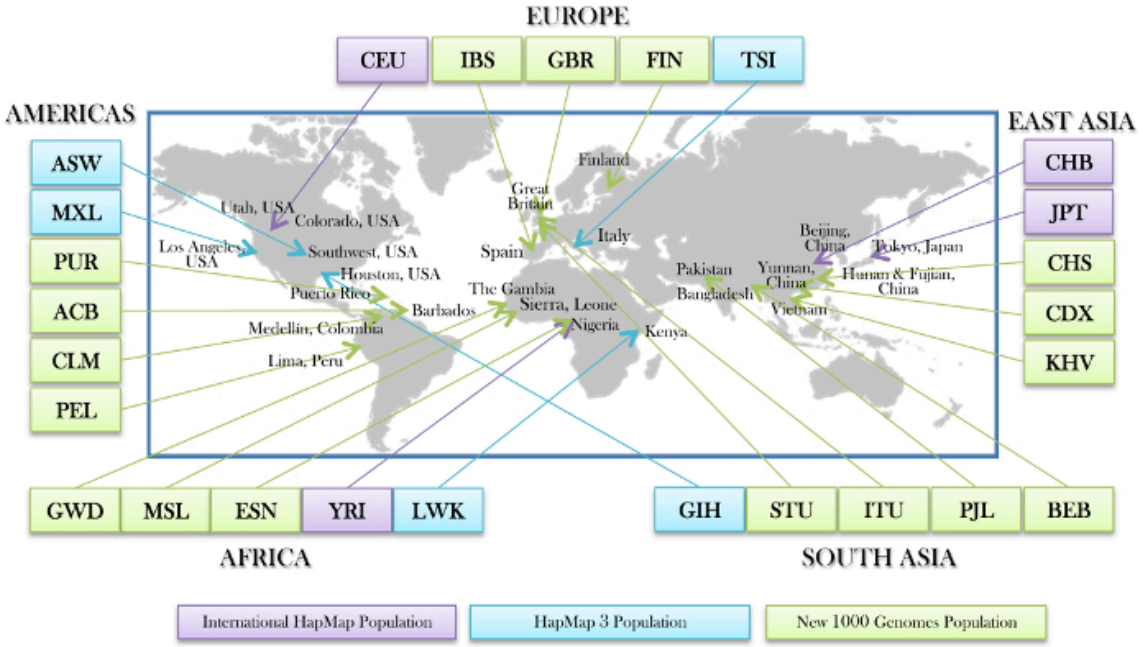
Ascertaining Genotypes

Whole Genome/Exome Sequencing



University of Michigan
Medical School

1000 Genomes Project



Generated whole genome sequences for 2504 individuals across 26 global populations, finding over 88 million points of genetic variation.

A typical genome differs from a consensus human genome at 4.09 - 5.02 million sites, affecting ~20 million bases of sequence

University of Michigan
Medical School

Storing Genotypes

VCF files are used to store genotypes for specific genomic positions across multiple individuals

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
21	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

The consensus (reference) allele is indicated in the REF column, while the alternative allele (s) are indicated in the ALT column.

Alleles are then labeled as 0 (REF) or 1 (ALT) for each site and encoded as genotypes 0|0, 0|1, and 1|1 for different allelic combinations observed in specific samples

University of Michigan
Medical School

VCF File Access

VCF files can be very large and are typically dividing into individual files for each chromosome. Even then, depending on the project each file can contain millions of rows with genotypes for thousands of sample.

VCF files are therefore typically *compressed* and *indexed* to reduce their size while still allowing random access. However, a specialized compression algorithm needs to be used in order to do this.

bgzip: modified version of gzip which compresses VCF file

```
bgzip GEUVADIS.chr1.PH1PH2_465.IMPFRQFILT_BIALLELIC_PH.annotv2.genotypes.vcf
```

tabix: indexes files compressed with bgzip

```
tabix -p vcf GEUVADIS.chr1.PH1PH2_465.IMPFRQFILT_BIALLELIC_PH.annotv2.genotypes.vcf.gz
```

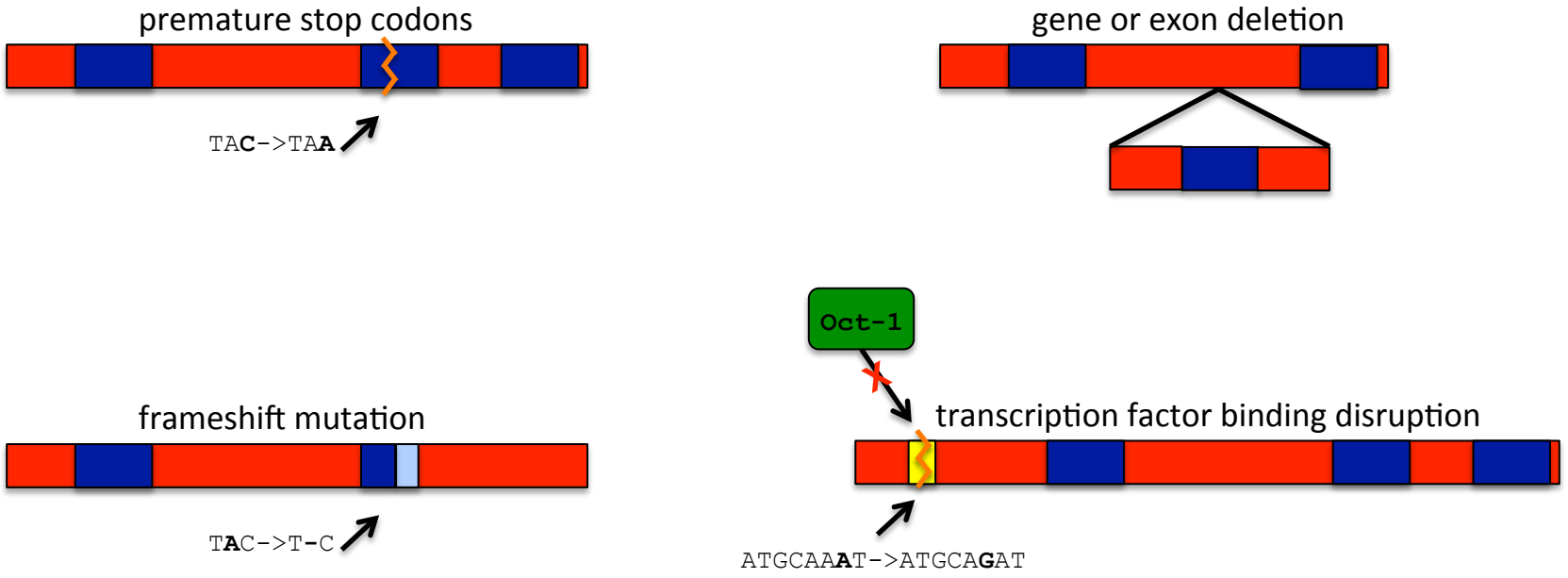
subsets files with coordinate ranges

```
tabix –h GEUVADIS.chr1.PH1PH2_465.IMPFRQFILT_BIALLELIC_PH.annotv2.genotypes.vcf.gz chr1:5363-5463
```

Both are part of the HTSlib software package: http://www.htslib.org/

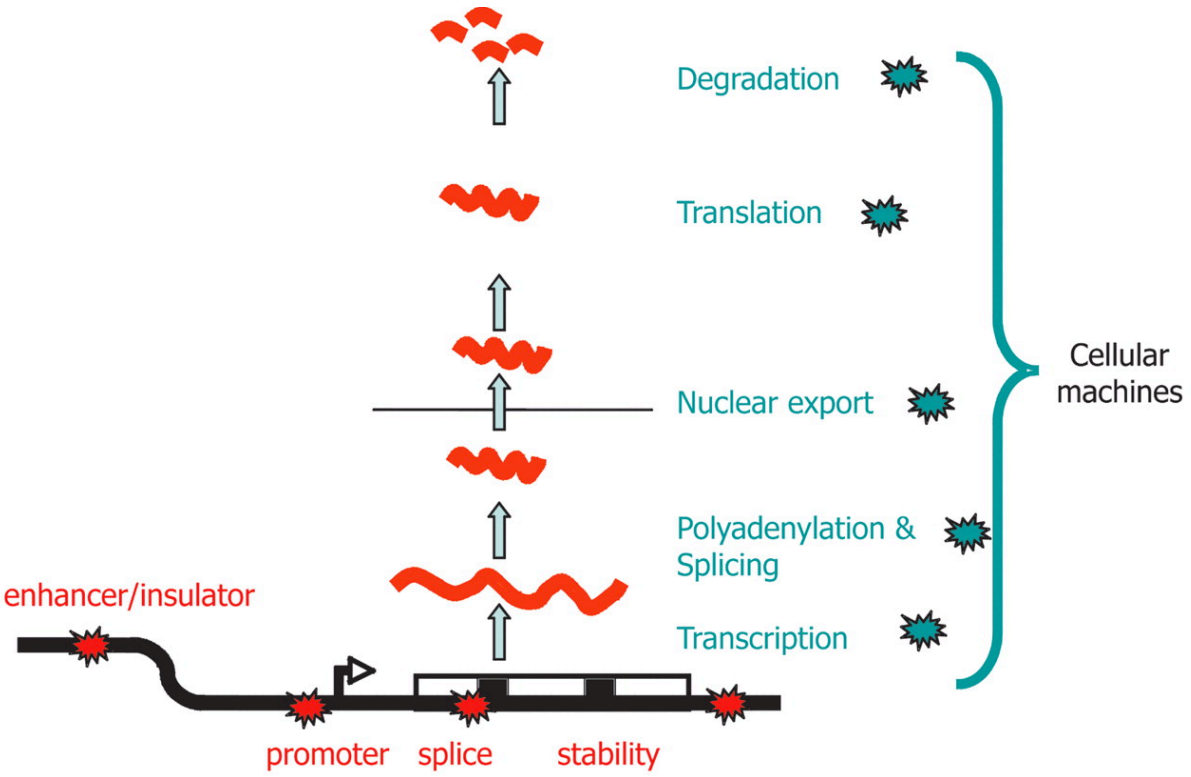
Impact of Genetic Variation

There are numerous ways genetic variation can exhibit functional effects



University of Michigan
Medical School

Effects on Gene Expression



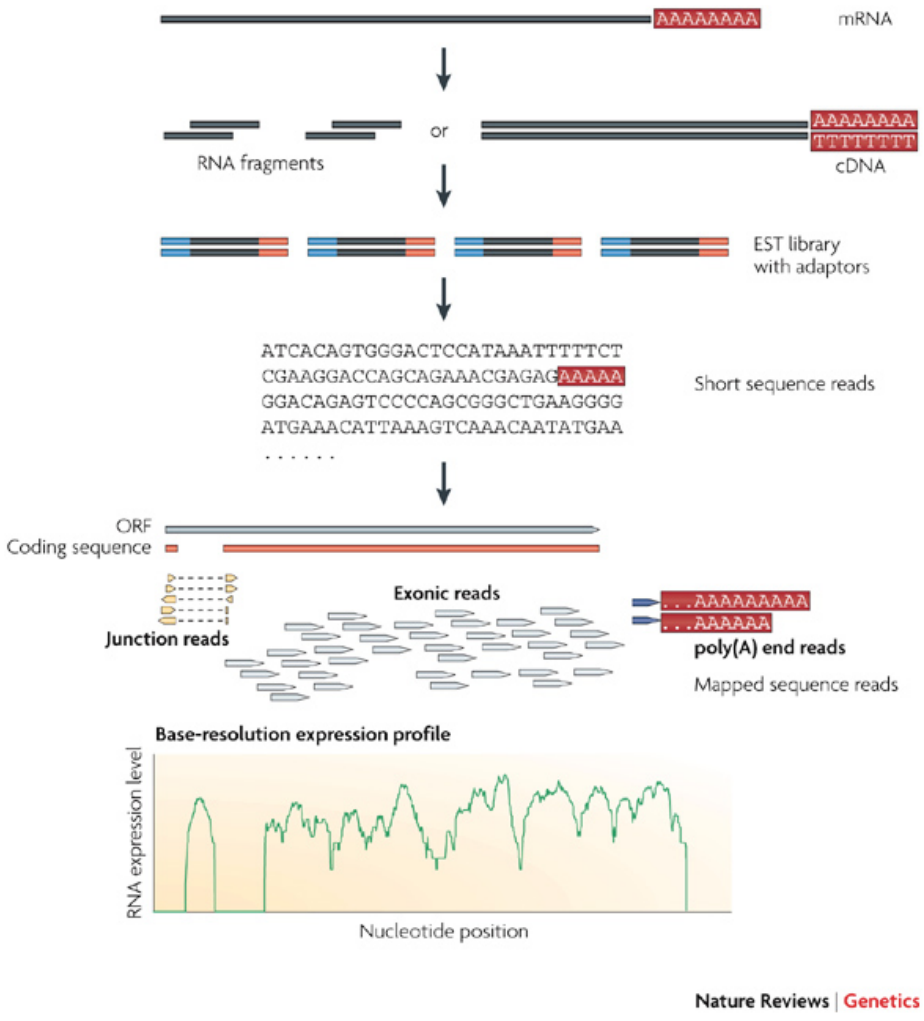
Genetic variants influencing gene expression may reside within:

- Regulatory sequences
- Promoters
- Enhancers
- Splice sites
- Secondary Structure Motifs

Williams et al, 2007, Genome Research 17:1707-1716

Measuring Gene Expression

University of Michigan
Medical School



RNA-Seq is a common, modern procedure for measuring gene expression. The number of sequence reads that align to a gene is called a *tag or read count*

These counts are typically normalized across samples in terms of reads/ fragments per kilobase million reads (R/FPKM):

$$FPKM = \frac{(read\,count \;\; x \;\; 1x10^{9})}{(gene\;length \;\; x \;\; total\,mapped\,reads)}$$

Alternative normalization can use transcripts per million (TPM):

$$TPM = \frac{(read\,count \;\; x \;\; read\,length \;\; x \;\; 1x10^{6})}{(gene\;length \;\; x \;\; total\,transcript\,count)}$$

Wang et al, 2009, Nat Rev Genet, 10:57-63

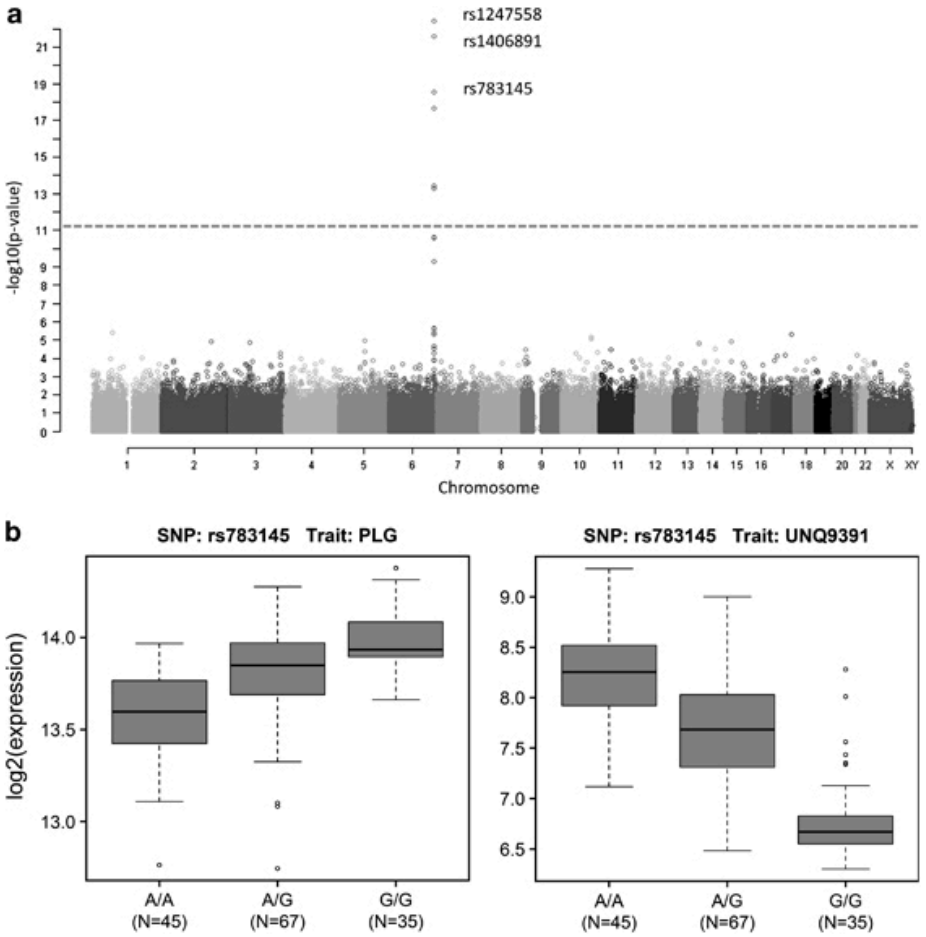
Linking Genotypes with Gene Expression

University of Michigan Medical School

An expression quantitative trait locus (eQTL) is a position in the genome that explains a fraction of the variance associated with gene expression.

Standard eQTL analysis involves the direct pairwise testing between genetic markers (e.g. SNPs) and gene expression levels across many (>100) individuals.

This can be done proximal (*cis*) or distal (*trans*) to each gene.



Association of SNPs with PLG (cis) and UNQ9391 (trans) expression in the context of drug absorption, distribution, metabolism and excretion in the liver.

Schröder et al, 2011, Pharmacogenomics, 13:12-20

University of Michigan
Medical School

Testing for eQTL Associations

Typical approaches make use of Generalized Linear Models (GLMs).

For simplicity, we will consider simple linear models of the form:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

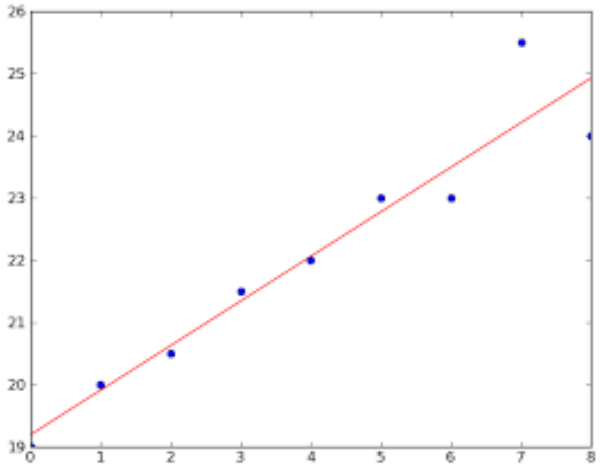
where Y_i is the normalized gene expression (FPKM) for individual i and X_i indicates the genotype (0,1,2) at a given SNP. ε_i can be set as a randomly distributed random variable

University of Michigan
Medical School

Testing for Associations (Python)

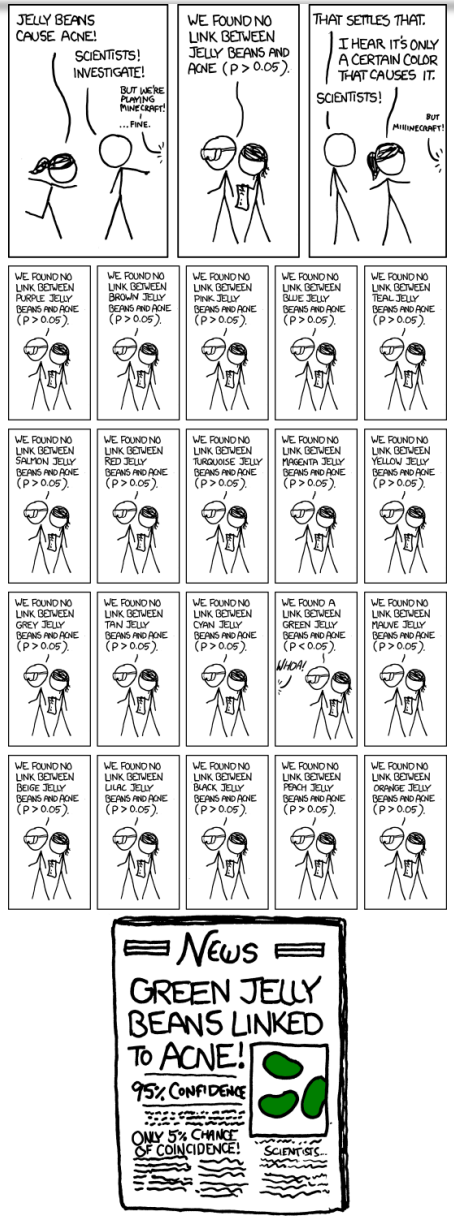
We can use functions available in *numpy* and *scipy* to construct our linear model

```
from numpy import arange,array,ones#,random,linalg
from pylab import plot,show
from scipy import stats

xi = arange(0,9)
# linearly generated sequence
y = [19, 20, 20.5, 21.5, 22, 23, 23, 25.5, 24]

slope, intercept, r_value, p_value, std_err = stats.linregress(xi,y)

print 'r value', r_value
print  'p_value', p_value
print 'standard deviation', std_err

line = slope*xi+intercept
plot(xi,line,'r-',xi,y,'o')
show()
```



University of Michigan
Medical School

Multiple Test Correction



Different methods available for correcting p-values for multiple tests.

Bonferroni corrections weight the p-values by the number of tests done

$$p < \frac{a}{m}$$

with *a* as the significance level (0.05) and *m* is the number of tests conducted. This controls the Familywise error rate

Other correction methods are also useful, for example *Benjamani-Hochberg* which can be used to control the false discovery rate

https://xkcd.com/882/

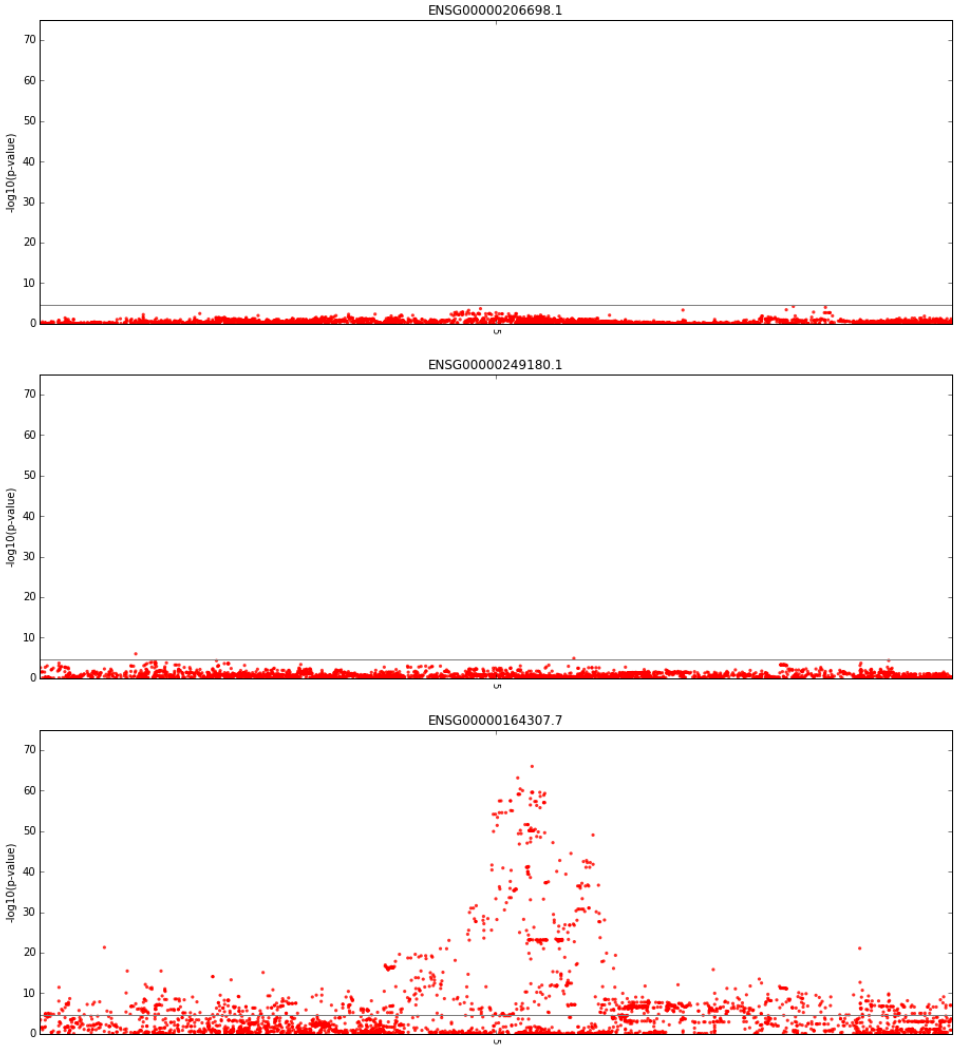
University of Michigan
Medical School

Visualizing Results

Association results for each SNP can be visualized by plotting their corrected p-values at each SNP position across a chromosome.

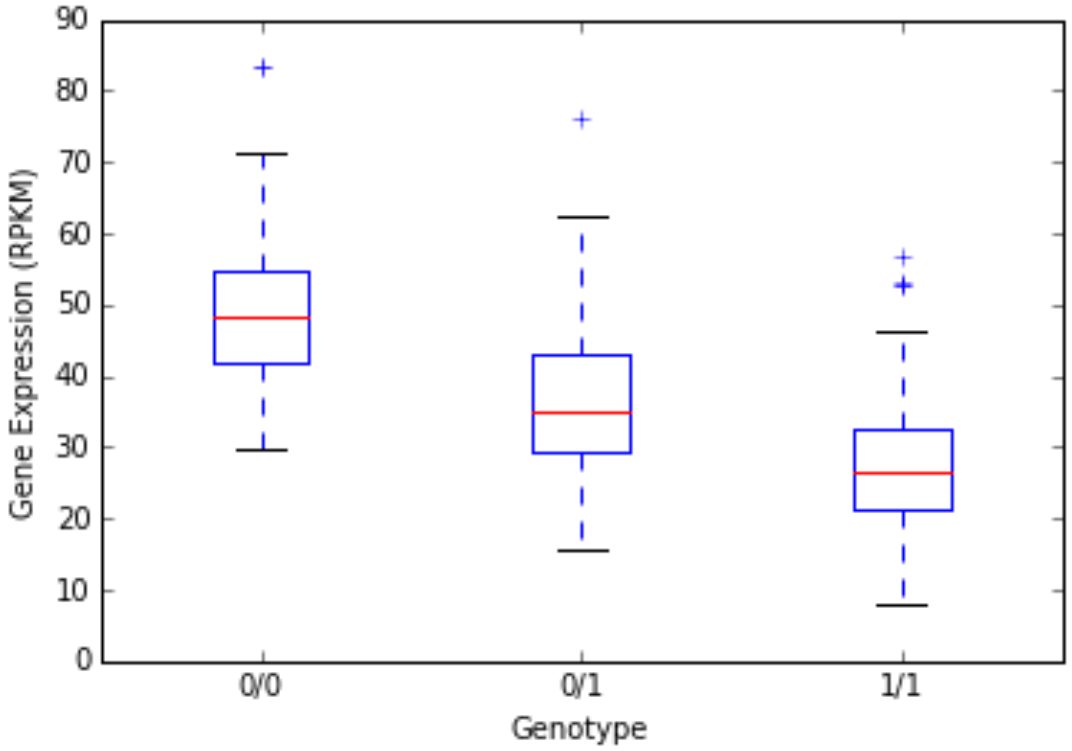
These types of plots are called ‘Manhattan’ plots due to their (somewhat) resemblance to a city skyline.

For ease of viewing, p-values are typically converted by multiplying them by $-\log_{10}$(p-value)



University of Michigan
Medical School

Visualizing Results



Highly significant associations can be further examined using categorical boxplots of the genotypes and their associated gene expression levels.

University of Michigan
Medical School

Geuvadis Data Set

Genotype data for 465 individuals:

Remote: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/genotypes/

Local: /scratch/biobootcamp_fluxod/remills/bioboot/geuvadis/genotypes

Contains: Indexed VCF files for each chromosome across all individuals

Expression data for 465 individuals:

Remote: ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/

Local: /scratch/biobootcamp_fluxod/remills/bioboot/geuvadis/analysis_results

Contains: FPKM values for each gene across all individuals

Format: Gene Name, Gene Symbol, Chr, Position, Sample1, Sample2…, SampleN

University of Michigan
Medical School

Regions

- chr5:95984676-96185176
- chr11:47426802-47627302
- chr1:205620478-205820978
- chr21:38432812-38633312
- chr20:5883504-6084004
- chr5:156807522-157008022