Week 8:
Related Questions in the Domain

1. groups/projects
2. vcftools

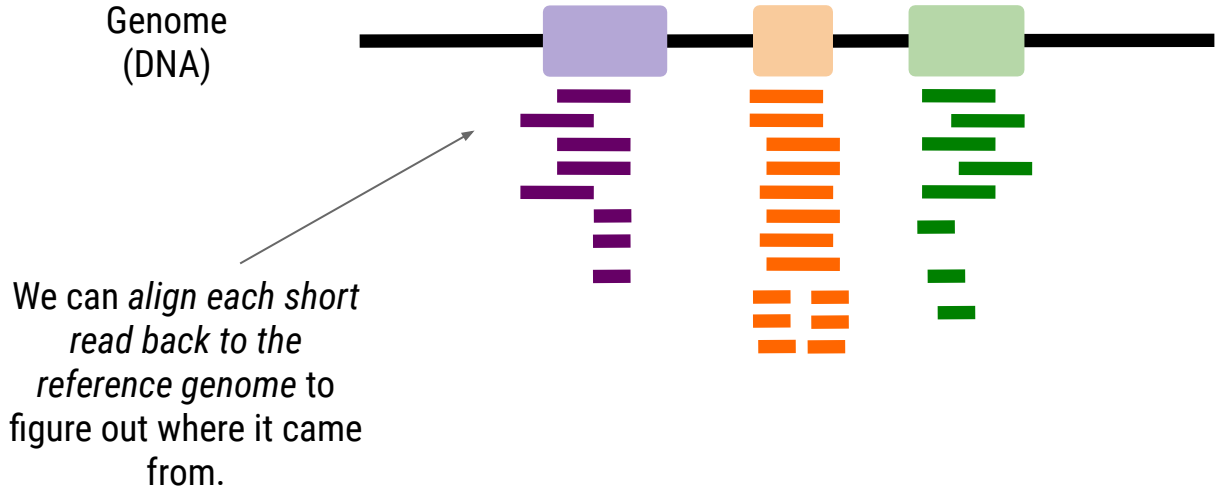
Data Types

- Genotypes (ATCG)
- Sequencing/Microarrays:
  - WGS (Whole Genome Sequencing) - sequence entire genome
  - Exome - sequence genes
  - RNA-Seq - expression level (continuous); expression microarrays
  - ATAC-Seq (chromatin accessibility), ChIP-Seq (DNA-protein interactions)
  - WGBS (whole-genome bisulfite sequencing), methyl-Seq, RRBS
  - CRISPR + sequencing

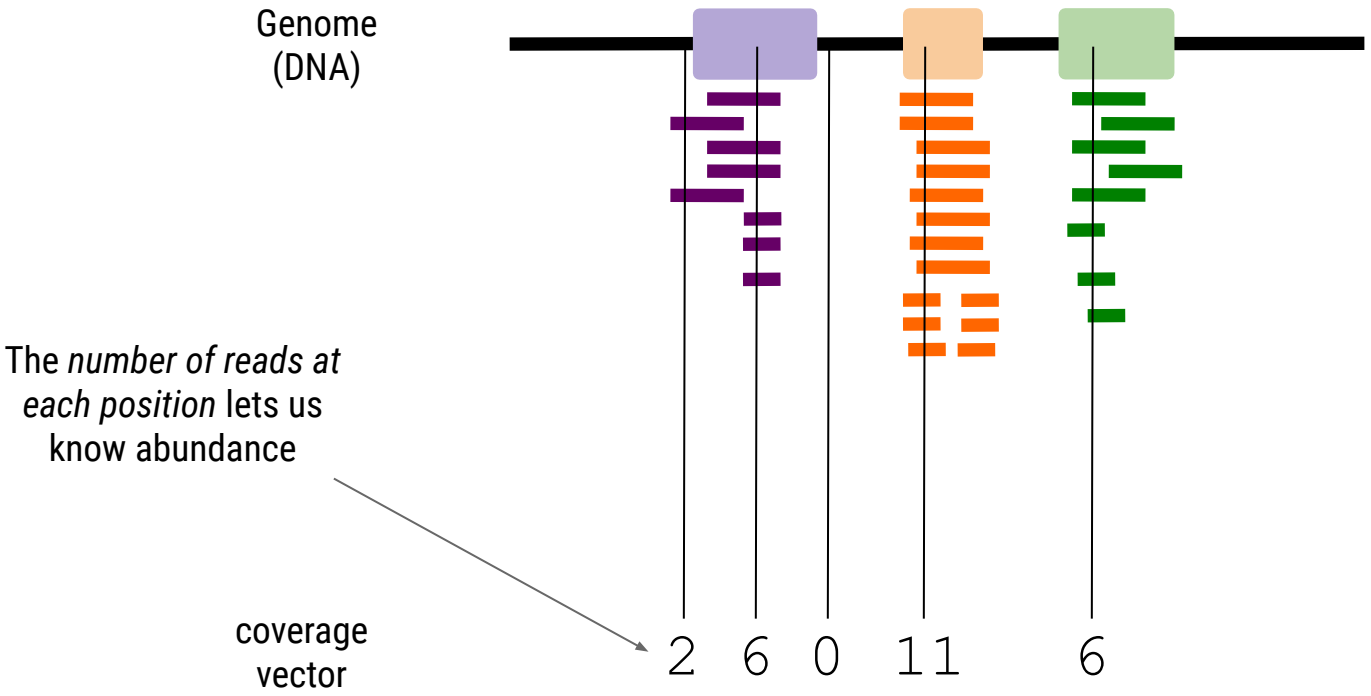
Next Generation Sequencing

A short read tells you the sequence of the RNA in that read

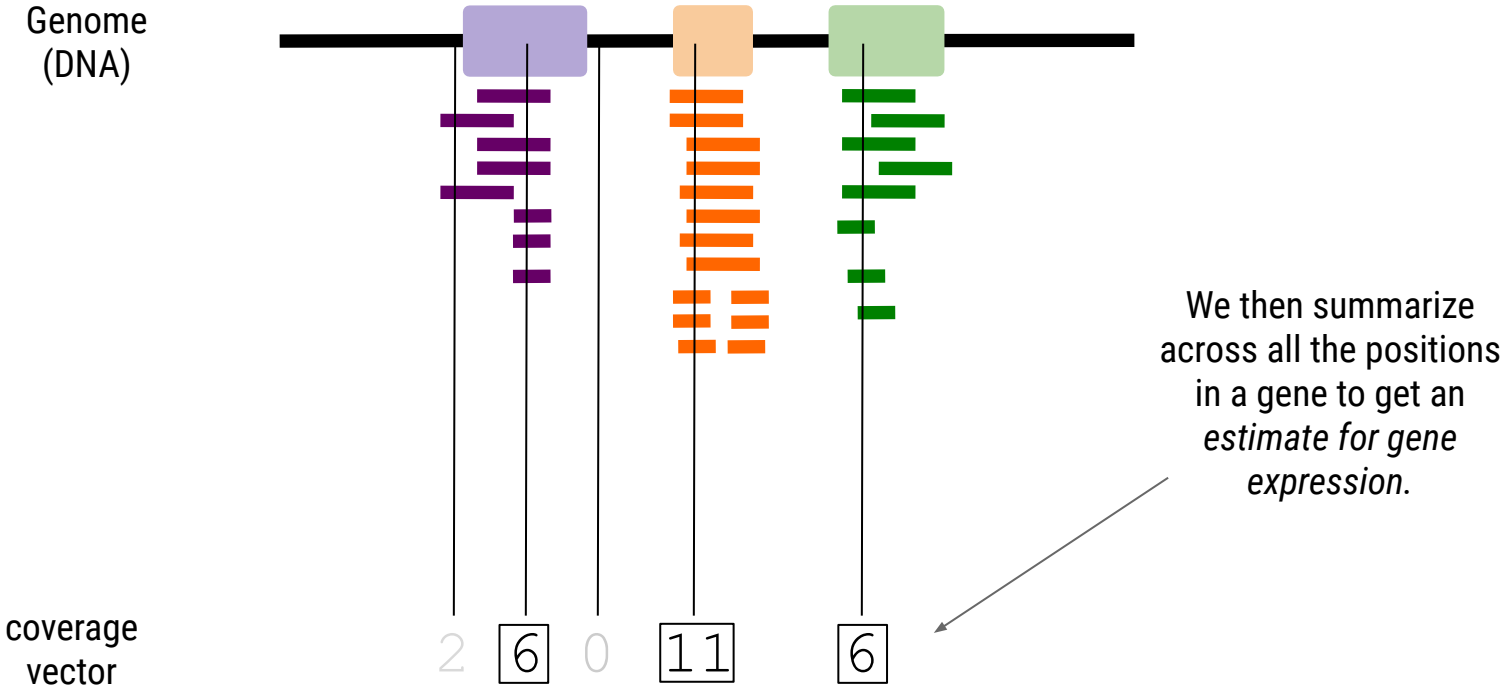
We first need to align these reads back to the genome



We first need to align these reads back to the genome



We first need to align these reads back to the genome



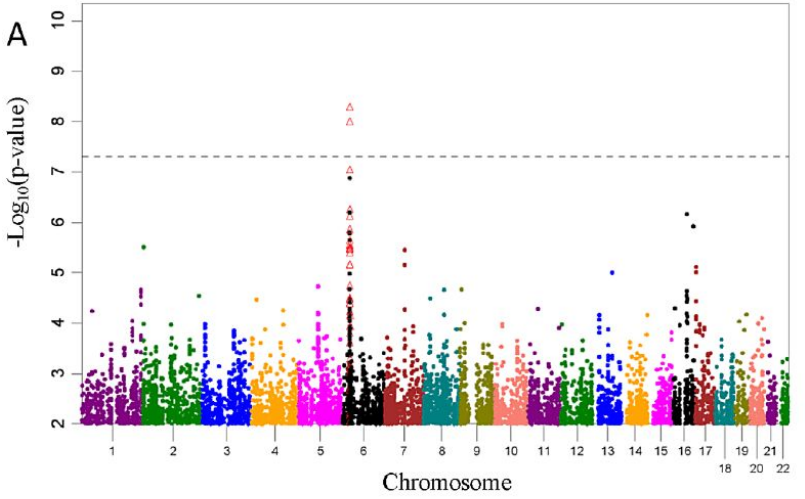
expression ≅ # RNA-Seq reads

Analysis Techniques

- **GWAS** : genome-wide association studies
  - Data: genotypes from SNP microarray/sequencing
  - Detect places of the genome associated with trait/disease
  - linear/logistic regression across the genome; multi-testing correction
  - Option: meta-analysis - analyzing summary statistics across GWAS

GWAS

- Check each SNPs association with trait
- Associated regions are “peaks” in Manhattan Plot
- LD explains why associated SNPs are not “smoking gun”



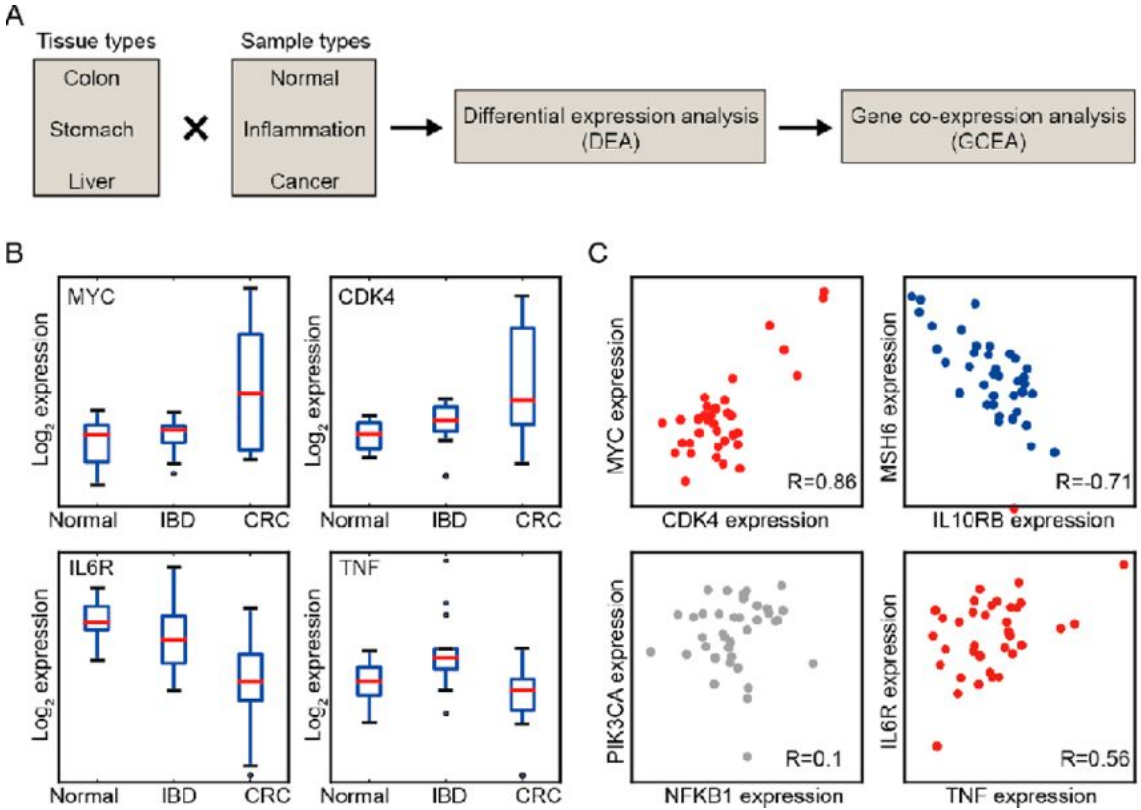
https://www.researchgate.net/figure/NPC-associations-of-GWAS-and-Taqman-validation-A-Manhattan-plot-of-GWAS-P-value_fig1_233840572

Analysis Techniques

- **GWAS** : genome-wide association studies
  - Data: genotypes from SNP microarray/sequencing
  - Detect places of the genome associated with trait/disease
  - linear/logistic regression across the genome; multi-testing correction
  - Option: meta-analysis - analyzing summary statistics across GWAS

- **DGEA** : Differential Gene Expression Analysis
  - Data: expression levels from expression microarray/RNA-Seq
  - Detect genes that have different levels between case and control
  - Related: gene set enrichment analysis, WGCNA (whole genome coexpression network analysis), GO (Gene ontology) analysis
  - Option: look at variance rather than the mean between groups; compare/contrast with difference in means

DGEA (Differential Gene Expression Analysis)



- Check each gene's expression between cases and controls
- Report genes differentially expressed
- Optional: co-expression analysis

https://www.researchgate.net/figure/Differential-gene-expression-analysis-and-coexpression-network-analysis-A-Flow-diagram_fig2_331102144

Available Datasets

Dataset	URL	Genotype	Expression	Disease
1000 Genomes + HGDP (more populations)	https://www.internationalgenome.org/	Y		
UK Biobank	https://biobankengine.stanford.edu/ summary stats	Upon request	Upon request	
UK10K	https://www.uk10k.org/ (not much phenotype info)	Y		
SRA (Sequence Read Archive)	https://www.ncbi.nlm.nih.gov/sra	Y	Y	Y
Recount	https://jhubiostatistics.shinyapps.io/recount/		Y	Y
TCGA (The Cancer Genome Atlas)	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga	Y	Y	Y
GTEx (Genotype-Tissue Expression Project)	https://gtexportal.org/home/	Y	Y	
GWAS Catalog	https://www.ebi.ac.uk/gwas/	Summary statistics		Y
gnomAD	https://gnomad.broadinstitute.org/	catalog of SNVs/INDELs		

Possible Questions

1. Which genotypes/markers are most informative for each population
    a. Could use 1000G data
    b. Build predictive model for genetic ancestry
2. What genes are differentially expressed in disease/trait X
    a. This could look at difference in mean expression or variance between groups
    b. RNA-Seq data available from recount2 in processed format
3. Are there parts of the genome that are associated more frequently across diseases? What regions are these?
4. Are results consistent across studies of the same disease/trait?
5. Comparison of analytical methods - where does one under/over-perform relative to other approaches?