CS273B: Deep Learning in Genomics and Biomedicine.

Recitation 1

30/9/2016

Topics

- Genetic variation
- Population structure
- Linkage disequilibrium
- Natural disease variants
- Genome Wide Association Studies
- Gene regulation
- Exome sequencing
- Data formats and sources

Genetic Variation

Genetic Variation

- Human Genome: 2 x 3 billion bp, 20K genes
- Diploid organism
- Consider genotype at a single locus:
  - Heterozygous: contains two different alleles
  - Homozygous: contains the same allele

Measuring Diversity

- pi: Average pairwise heterozygosity (per bp)
- Human average pi ~ 0.08%: a typical individual is heterozygous in 0.8 sites per kilobase
- What determines the amount of variation:
    - Mutation rate
    - Population size
    - Natural selection
- Genetic Drift: random changes in allele frequencies of neutral variants

Why Genetic Variation?

- Variation of the DNA sequences in our genome:
  - Understand biological processes and mutations
  - Plays a central role in human disease
  - Study human history
  - Natural selection and adaptation

Variation Model

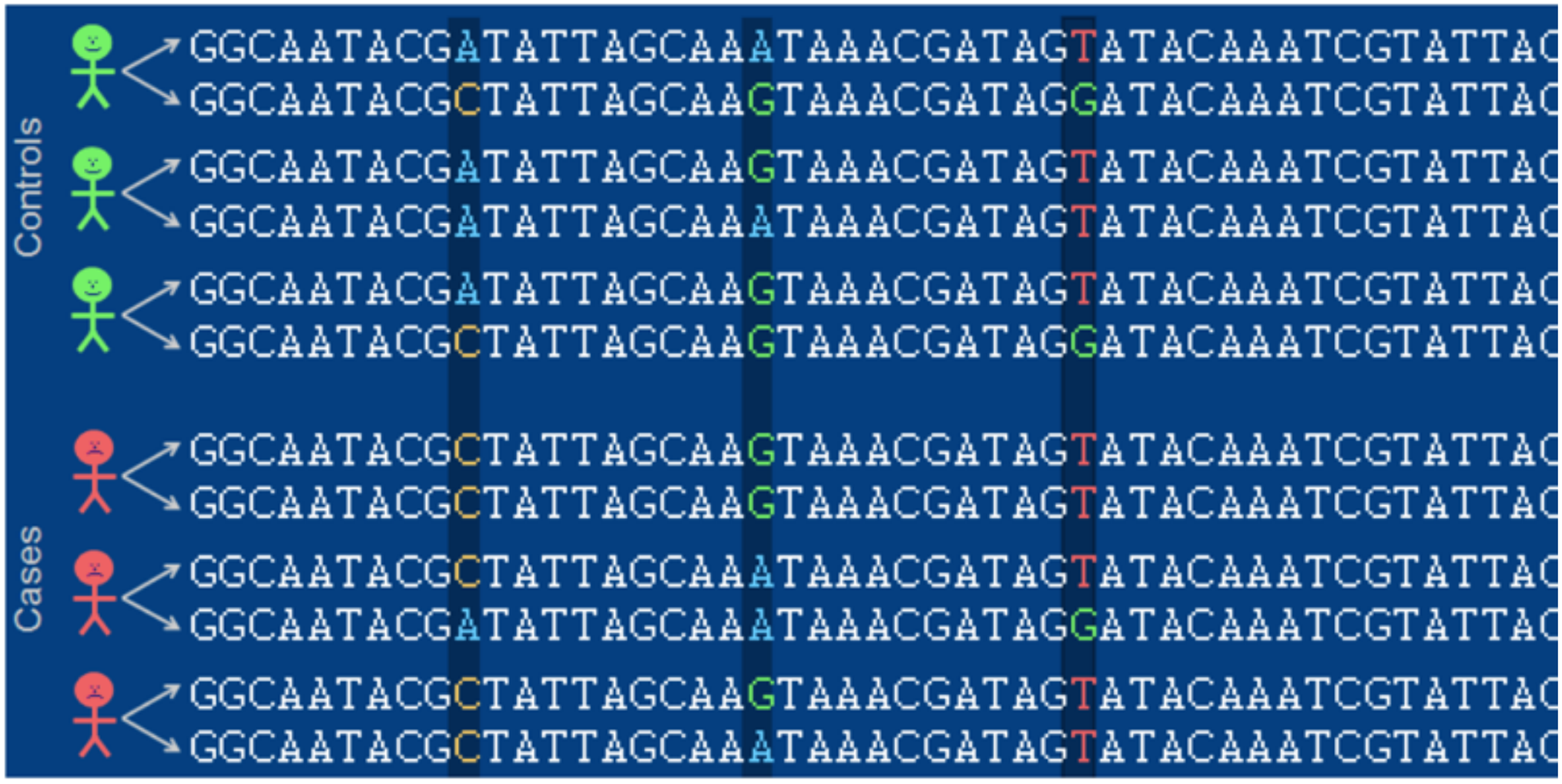
- Hardy-Weinberg
  - Site with 2 alleles: A and B
    - A at frequency p
    - B at frequency q=1-p
- Three genotypes:
  - AA with frequency p^2
  - AB with frequency 2pq
  - BB with frequency q^2

Genetic Variation

- Mutations: alteration of the nucleotide sequence
    - Small scale:
        - Single Nucleotide Polymorphisms (SNPs):
            - 3x10^6 common SNPs (>5% frequency in human population)
            - Rare SNPs
        - Insertion/Deletion of a few nucleotides
    - Large scale:
        - Copy Number Variations (CNVs)
        - Insertions, Inversions and Translations

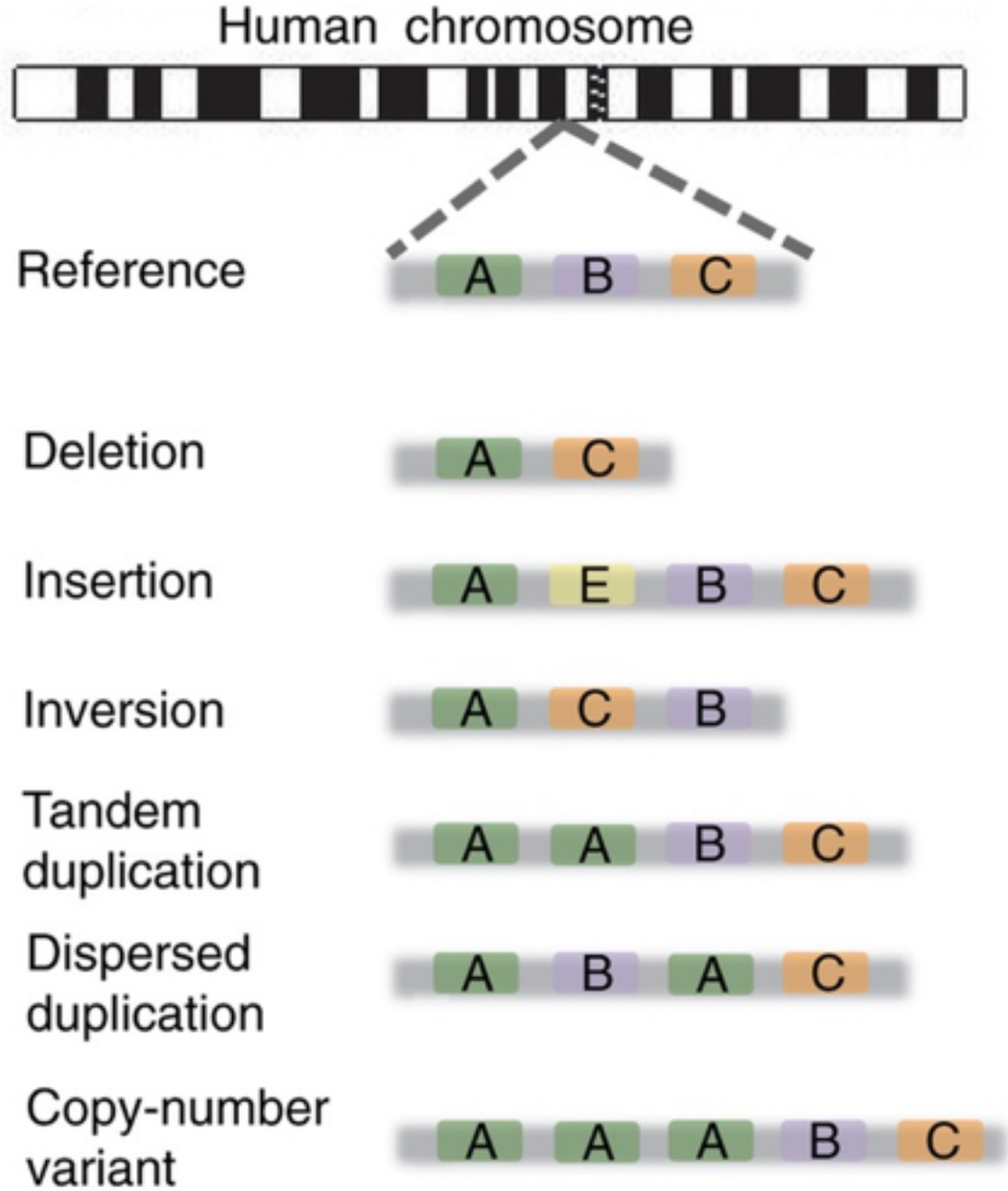
Genetic Variation

- Sample common SNPs

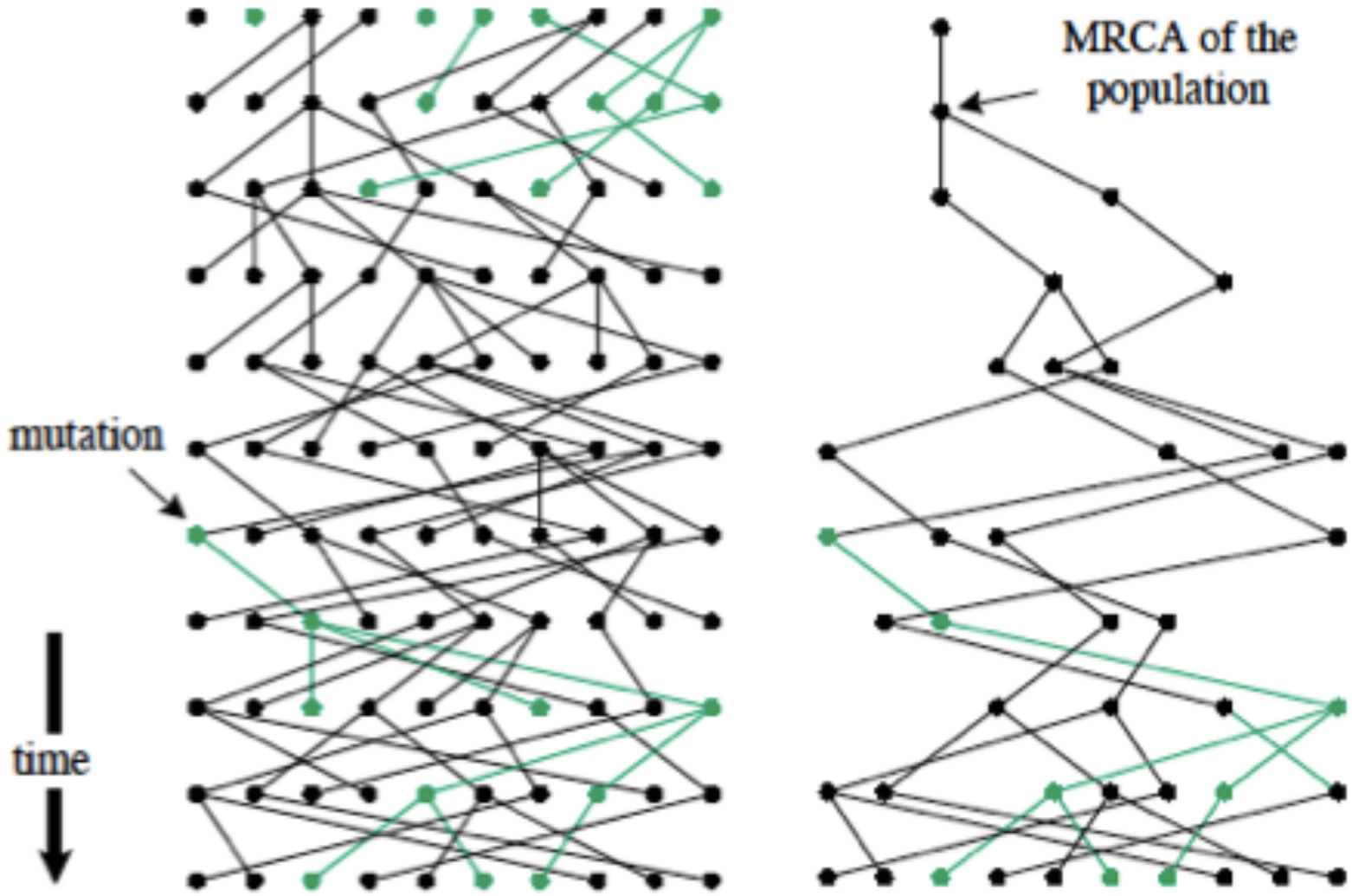


Genetic Variation

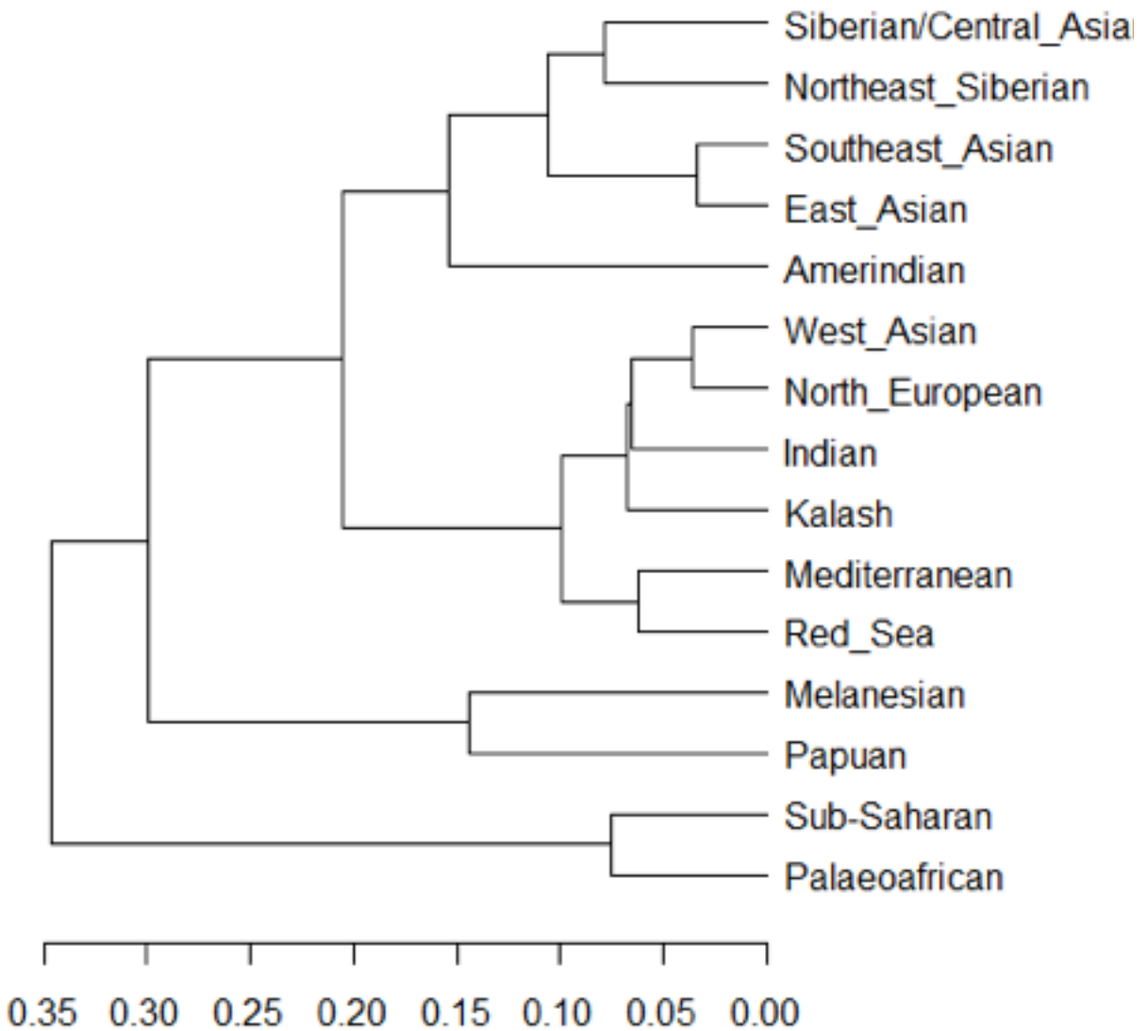
- Structural Variations:
  - Insertion
  - Inversion
  - Translocation
- CNVs:
  - Duplication
  - Deletion



Population Structure



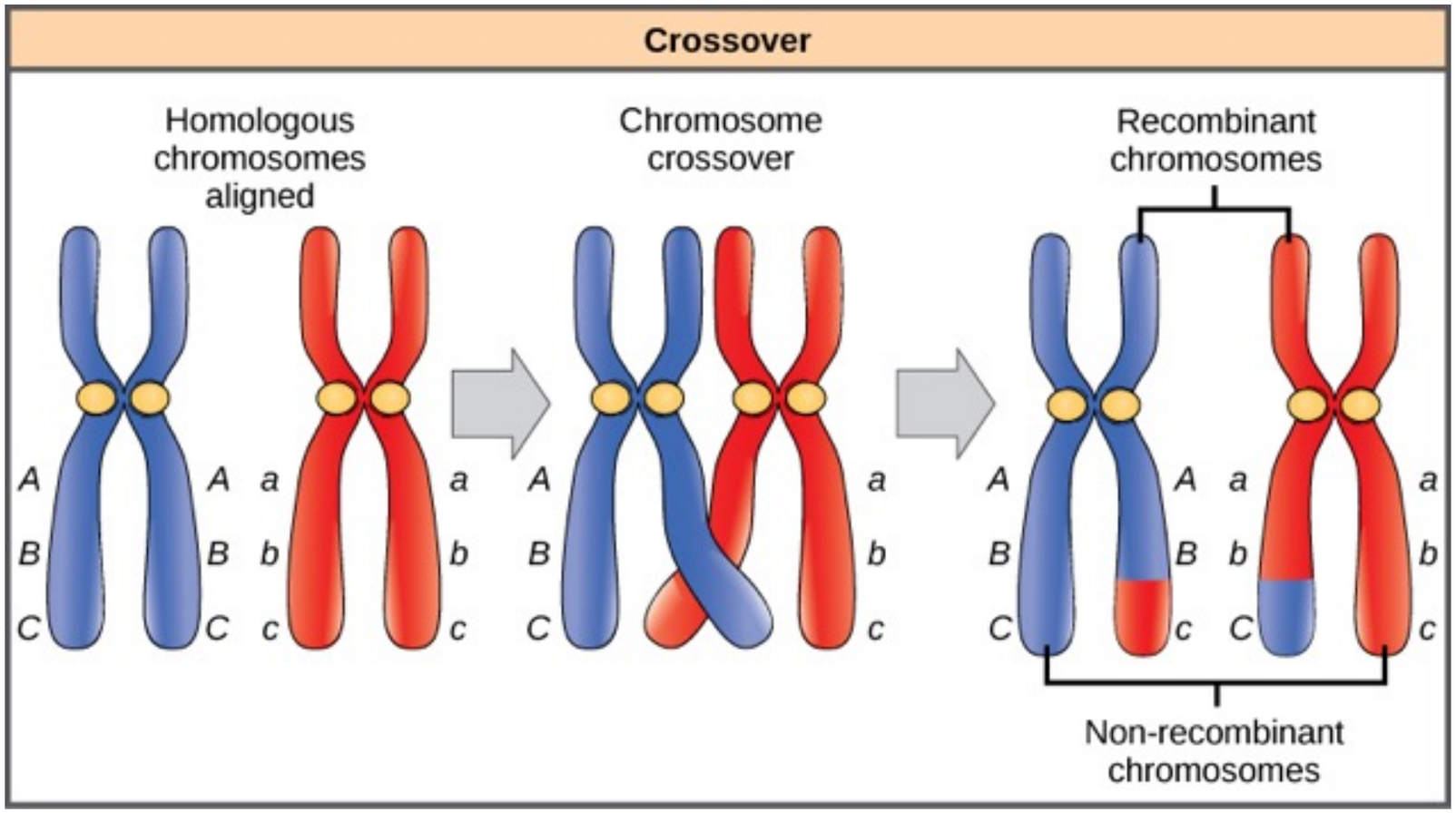
Population Structure



Gene Recombination

- Production of offspring with combinations of traits that differ than those found in the parents

- Genes that typically stay together during recombination are said to be linked



Linkage Disequilibrium

- Alleles are in LE if the frequency of a genotype is equal to the product of the frequencies of the individual alleles

- Measure deviation from LE by comparing the observed and expected genotype frequencies

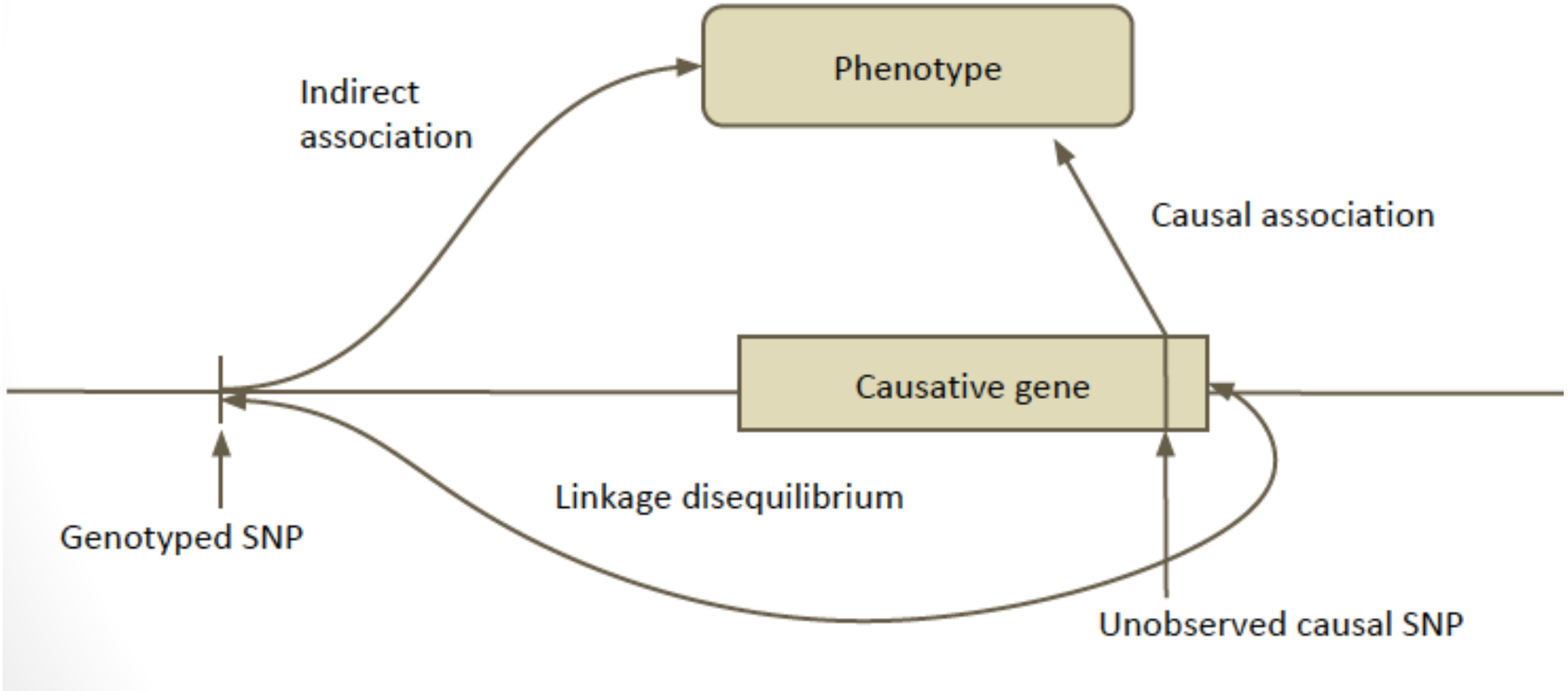
- Haplotype: set of SNPs that always happen together

- In a short chromosome segment there are only a few distinct haplotypes

Linkage Disequilibrium

- Chromosomes are like mosaics
- Extent of conservation varies according to:
    - Natural selection
    - Mutation rate
    - Recombination rate
    - Population size
- Carefully selected SNPs can determine the status of other SNPs

Linkage Disequilibrium



- Neighboring markers tend to be inherited together
- Genotypes are redundant because LD causes correlations between the markers

Linkage Disequilibrium

- Basic descriptors:
  - Haplotype frequency of each type of chromosome
  - Common summary measures
    - D
    - D’
    - r^2

LD Measures

- Disequilibrium coefficient: D_AB=p_AB-p_Ap_B
- D_AB is hard to interpret
  - arbitrary sign
  - range depends on allele frequencies

		LocusB		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1

LD Measures

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

- D’_AB in [-1,1]
- D’ = +1/-1 means there is no recombination evidence
- High D’ means markers are good surrogates for each other
- Disadvantages: Inflated if sample is small or one allele is rare

LD Measures

$$r^2 = \frac{(D_{AB})^2}{p_A(1-p_A)p_B(1-p_B)}$$

- Ranges in [0,1]: 0 if in equilibrium, 1 if identical information
- Measures loss in efficiency when marker A is replaced by marker B in an association study

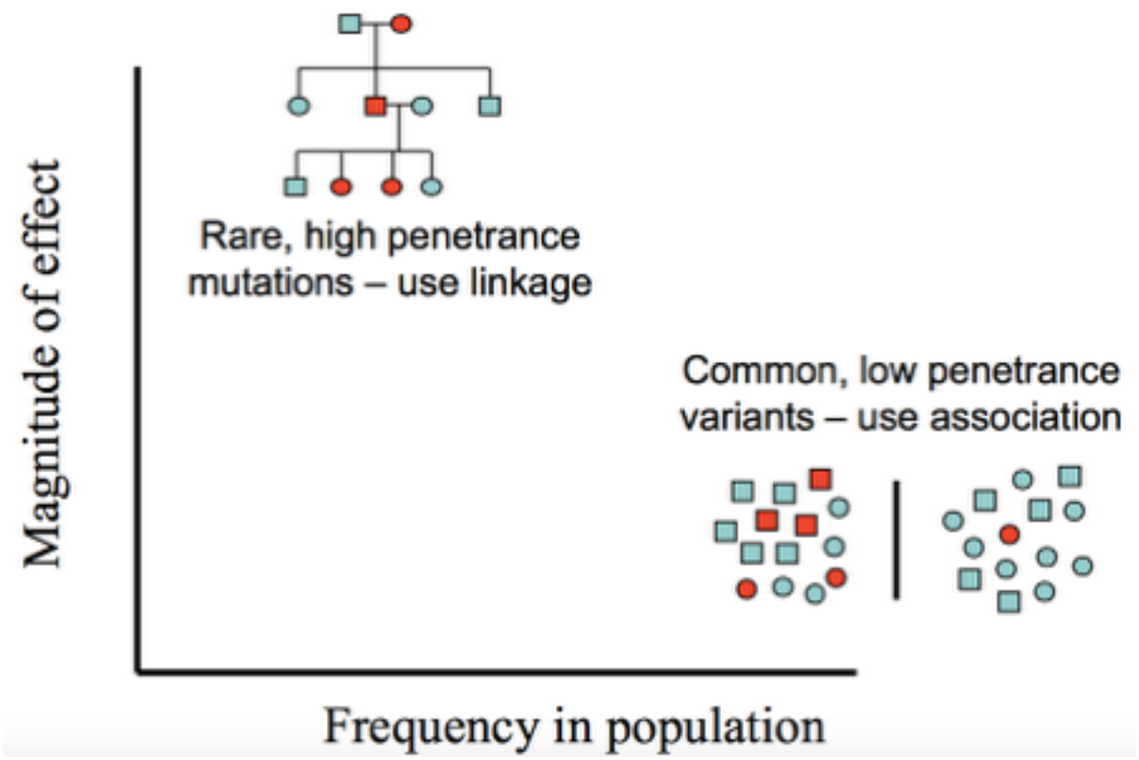
Natural Disease Variation

- Interest in finding genetic factors underlying disease
    - personalized treatment
    - identify druggable targets
    - insight into biological pathways of disease
- Two main classes of diseases:
    - Mendelian: mutations in a single disease gene produce phenotype (eg cystic fibrosis)
    - Complex: multifactorial, many genes and environmental factors (eg diabetes)

Two types of studies

- Linkage: find markers that are transmitted with disease in families (powerful for Mendelian cases)

- Association: identify markers with frequency differences between cases and controls (more common)



Genome-wide association studies

- Examine set of genetic variants in many individuals to associate them with a trait

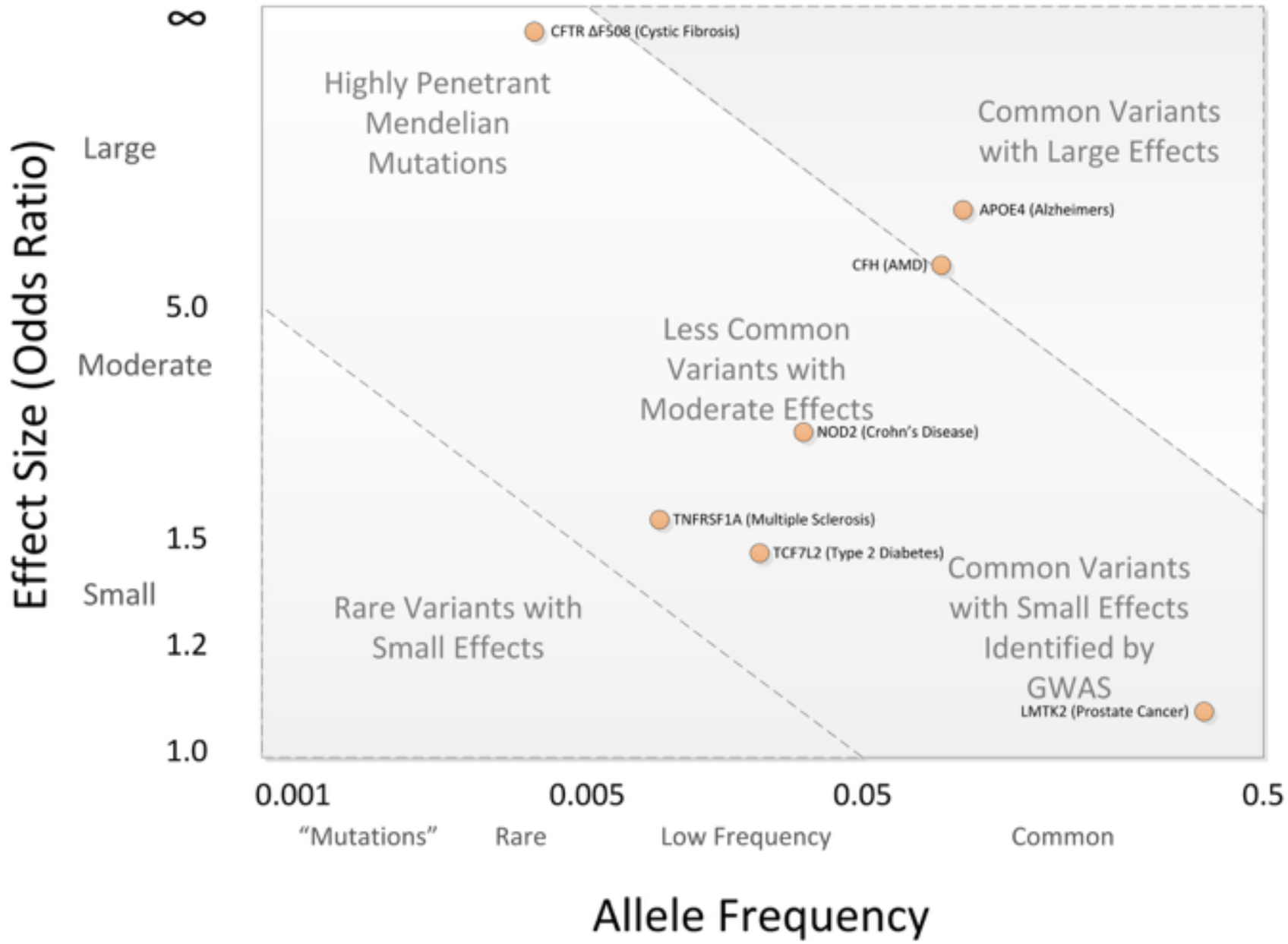
- Identify large amounts of associations efficiently to understand genetics of diseases and traits

- Focused on associations between SNPs and diseases

- If one type of the variant is more frequent in people with the disease it is associated with the disease

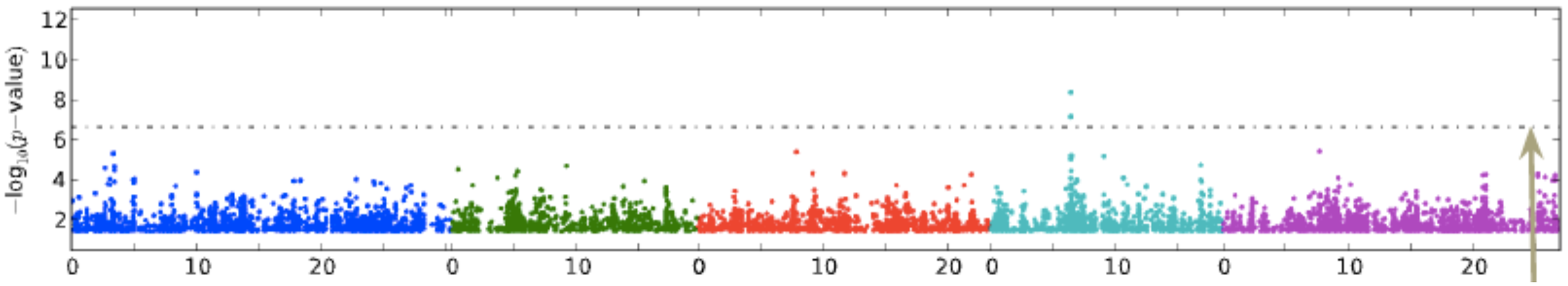
- Use summary association statistics in conjunction with linkage disequilibrium

Disease Allele Effects



GWAS Example

- Manhattan plot:
  - x-axis: SNP locations
  - y-axis: -log(p-value)
- Multiple testing problem:
  - At 5% significance threshold, will expect 5% of markers that have true effect of 0 to be significant
  - Bonferroni correction

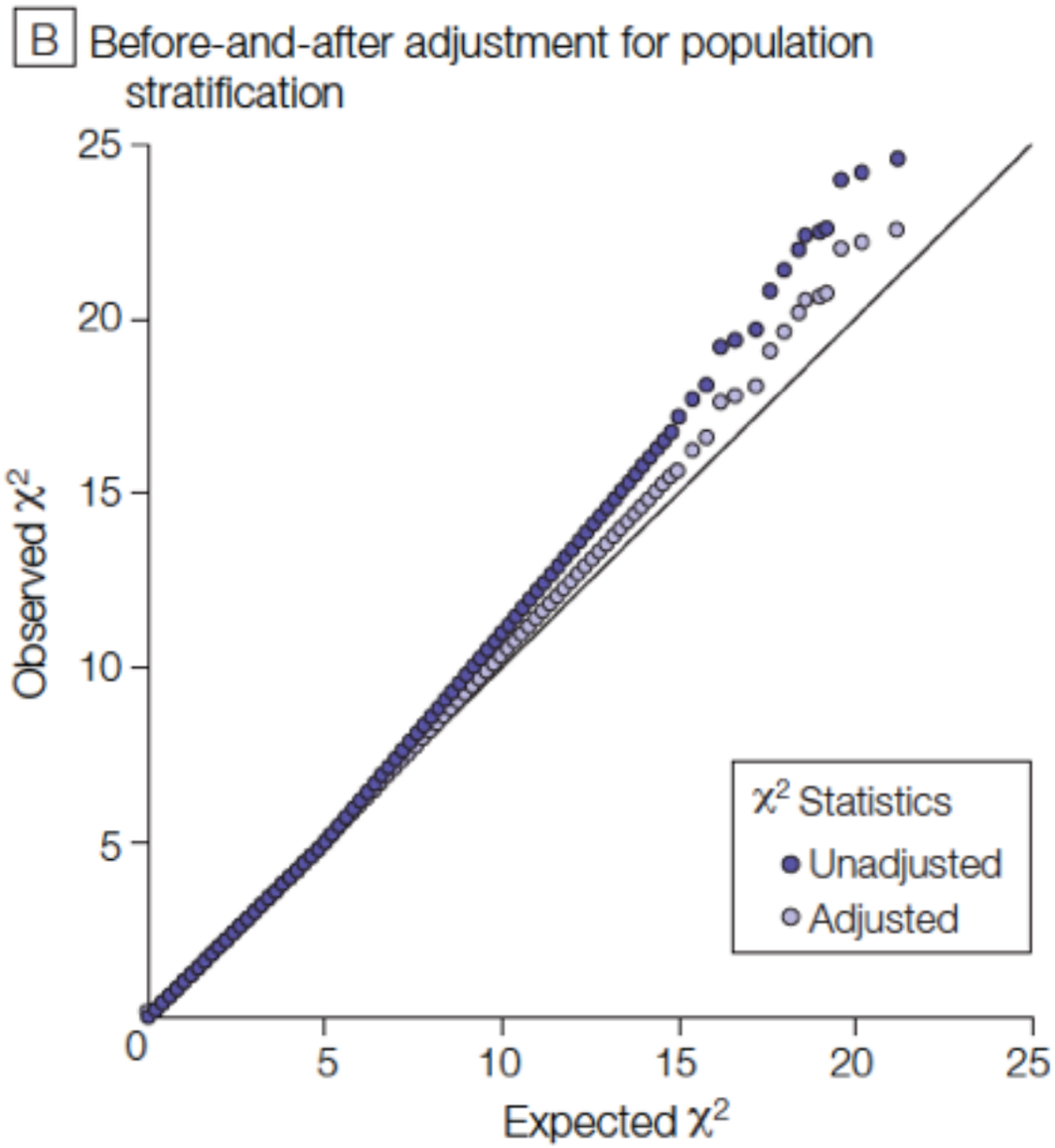


GWAS Confounders

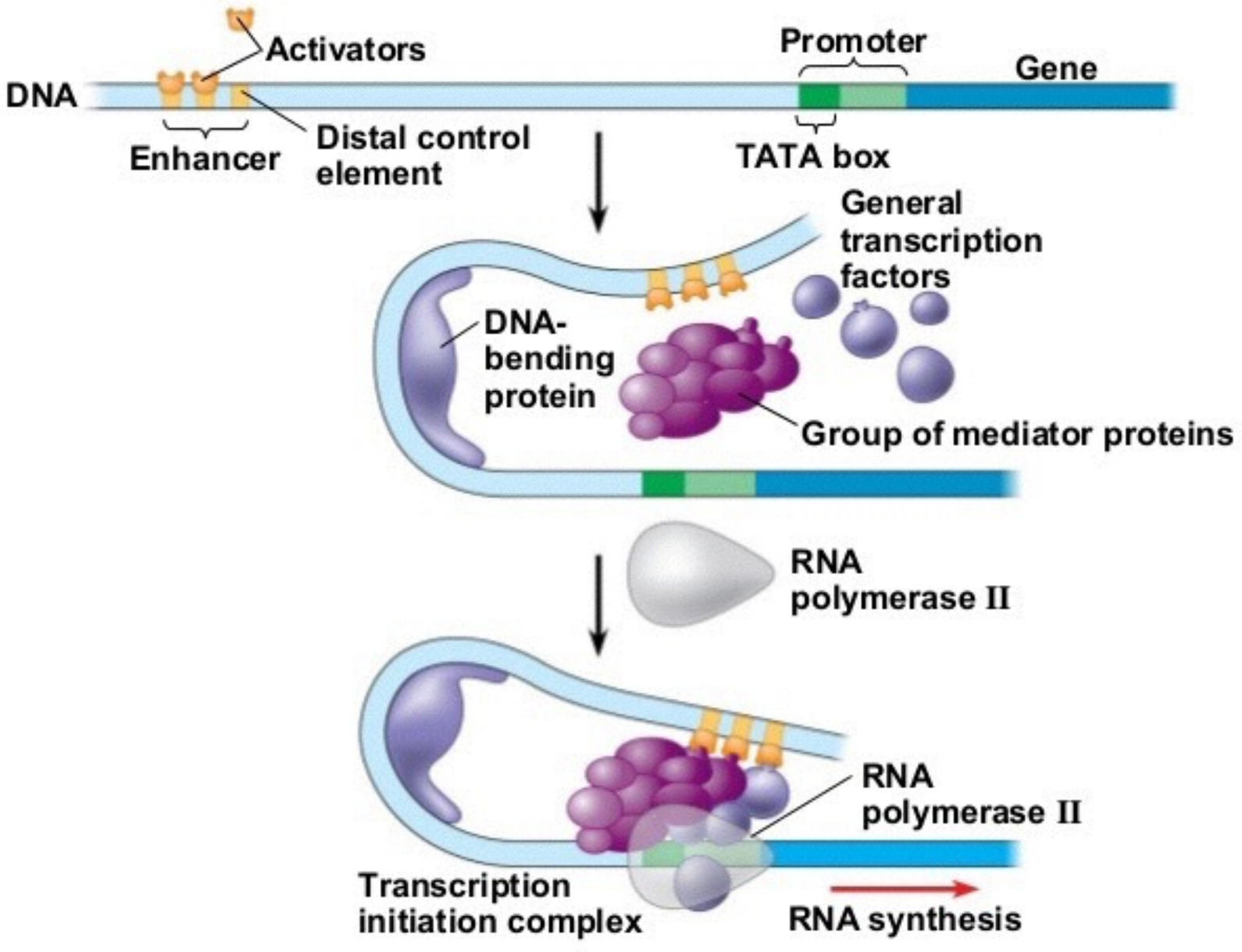
- Careful about variables that could be different between the cases and controls other than the disease itself:

- Population structure is the most common confounder. Example: marker for skin color might be associated with malaria resistance

- Use QQ plot to show confounders aren't at work



Gene Regulation



Exome Sequencing

- Exon: segment of DNA containing information coding for a protein
- Exome consists of the exons of all our genes
- WES is technique for sequencing all expressed genes: capture exon after amplification
  - micro-array based capture using cDNA library
  - hybridization based using cDNA that only bind to exons
  - molecular inversion probe
- Goal: avoid high cost of whole genome sequencing

Key Challenges

- Noisy labels: often better frame classification problems as positive/ambiguous/negative

- Feature extraction is not straightforward:sequences can have many patterns in addition to what we are modeling

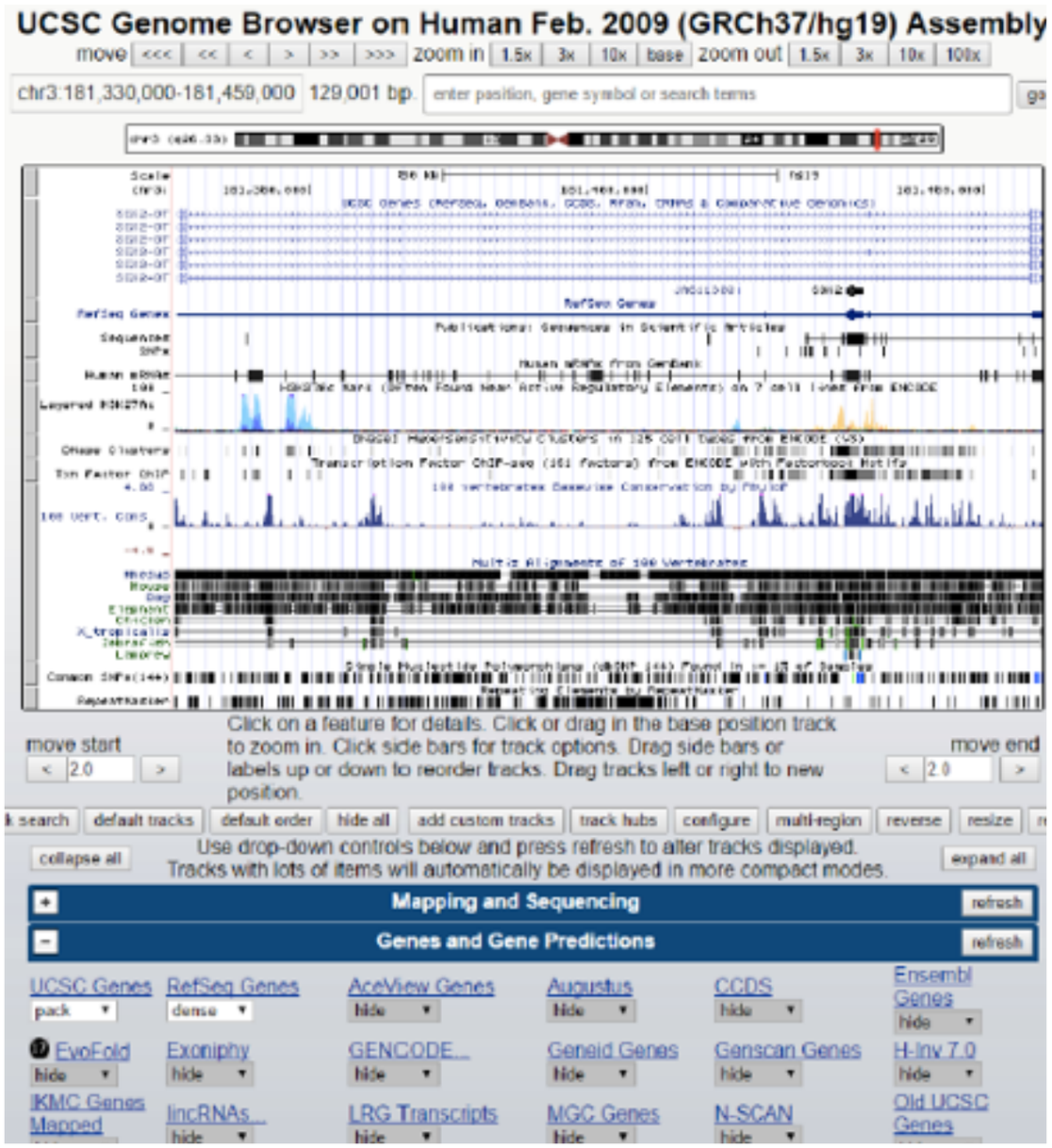
- Comparisons between different sequencing depths datasets

- p-values are not calibrated: get different peaks on similar experiments if just use simple thresholds

Genome Browsers

USCS Genome Browser

- Old browser: clunky and a bit ugly
- High utility: largest unified collection of built in data tracks



WashU Epigenome Browser

- Next-gen browser: dynamic, pretty, more responsive and can handle large-scale data.
- Visualizing multiple loci simultaneously
- Visualizing long-range genome interaction data
- Widgets for built-in data analysis (scatter plots, correlations aggregate plots)

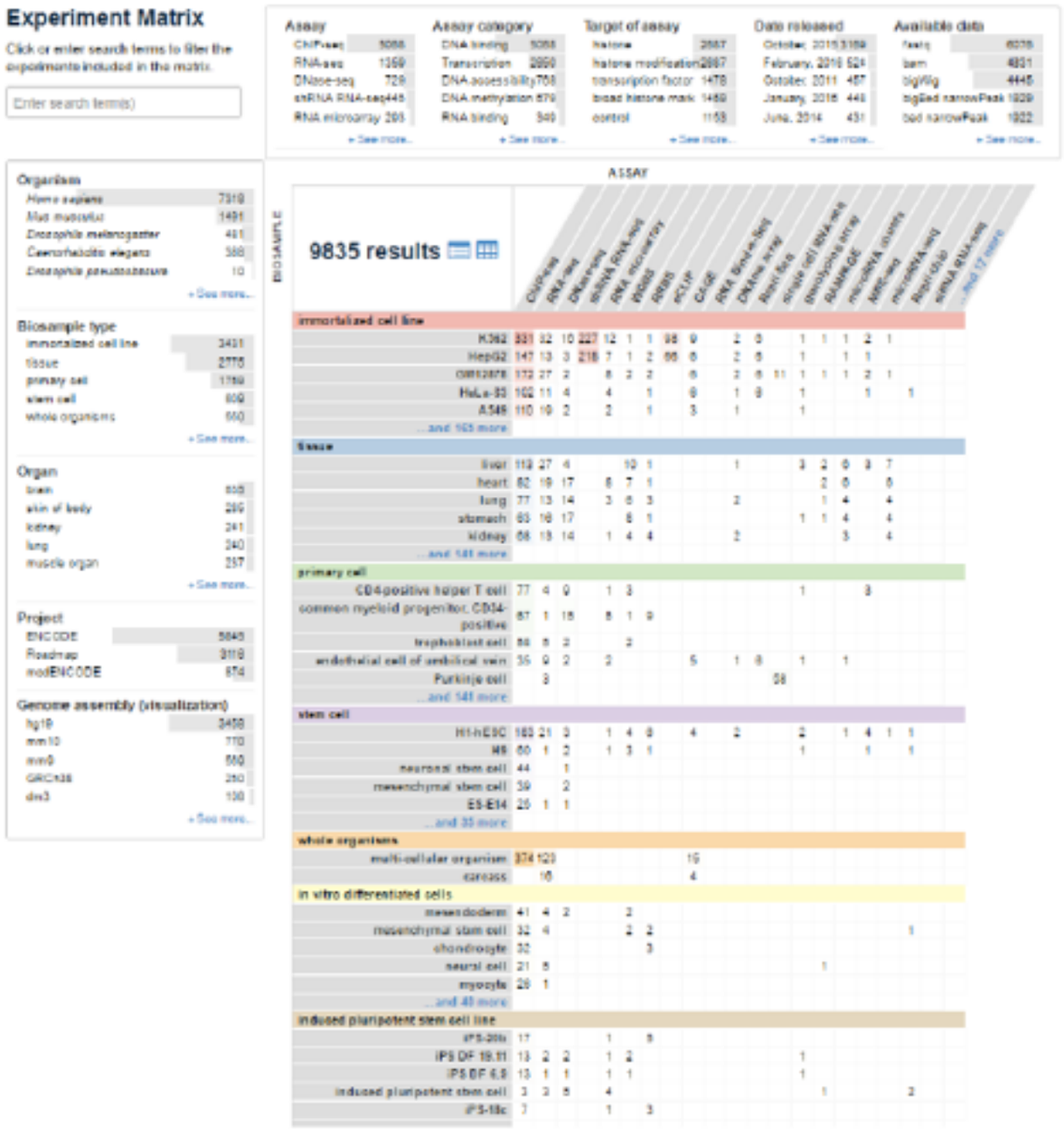
Data Sources

- ENCODE
- Roadmap epigenomics project
- Cancer Genome Atlas
- 1000 Genomes project
- Data Deluge

Encyclopedia of DNA elements (ENCODE)

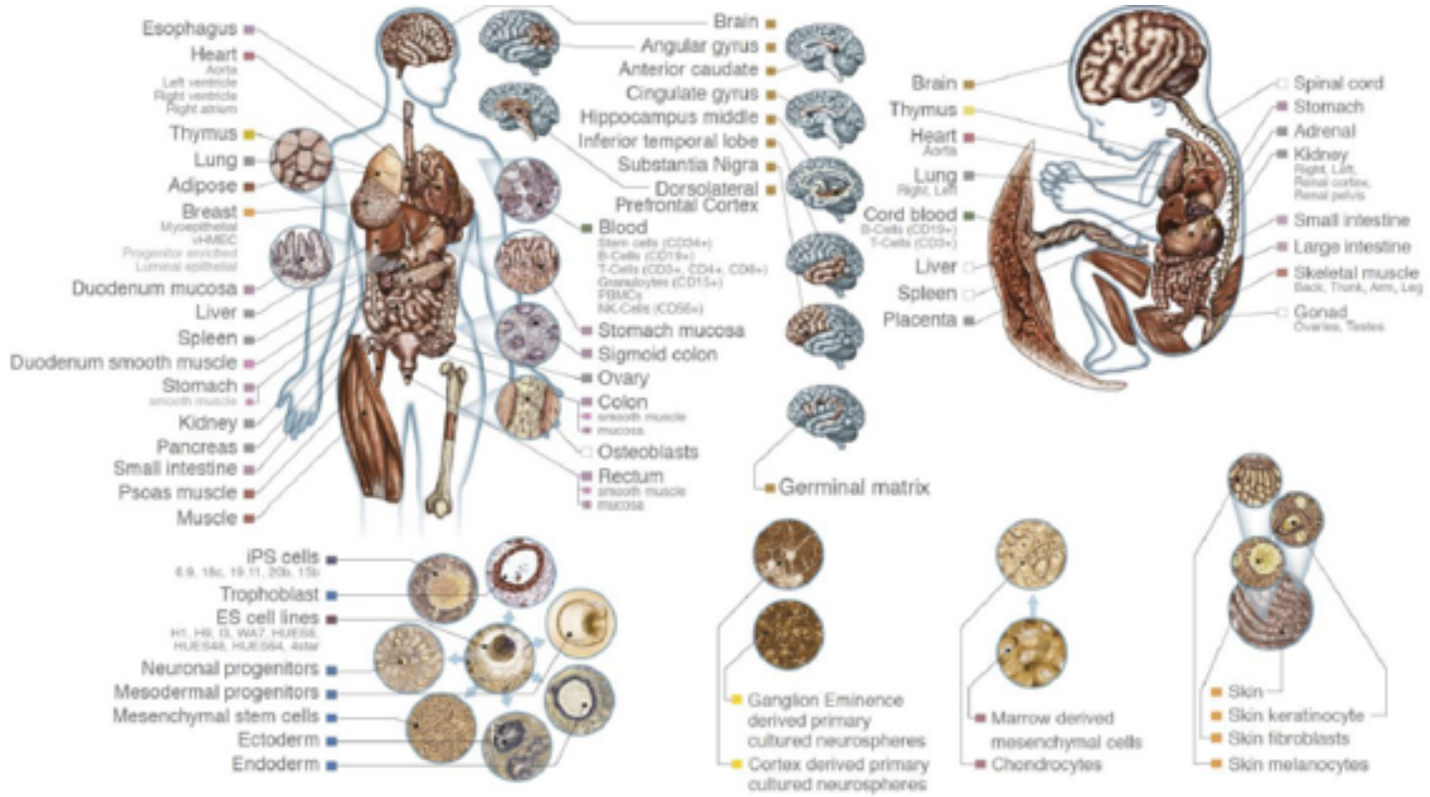
- https://www.encodeproject.org

- https://www.nature.com/encode



Roadmap Epigenomics project

- >150 Primary cells / tissues
  - 6 histone marks
  - open chromatin
  - DNA methylation
  - Gene Expression



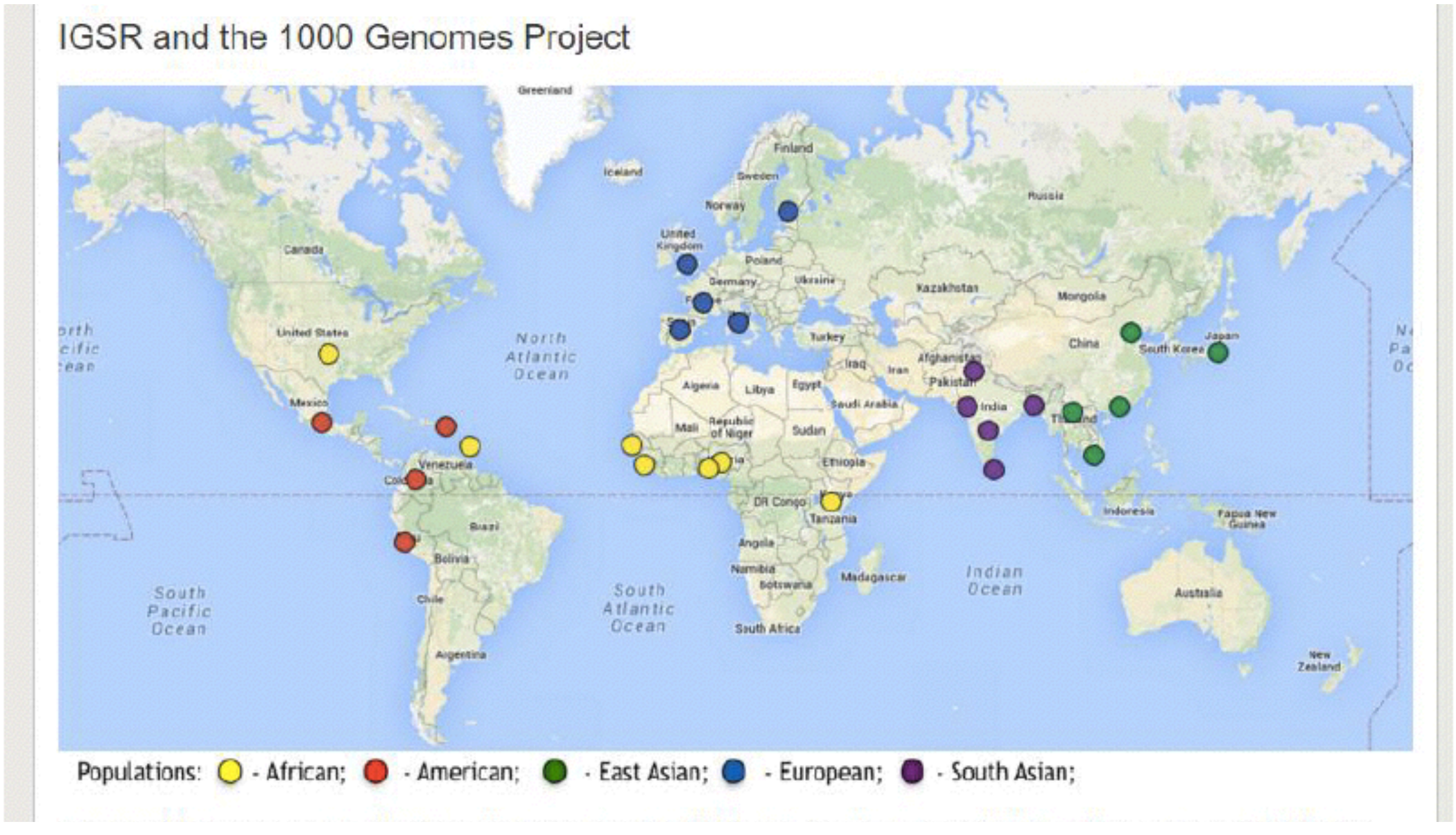
The Cancer Genome Atlas (TCGA)

- Provides platform for researchers to search, download and analyze datasets generated by TCGA

- Contains clinical information, genomic data characterization and high level analysis of tumor genomes

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	03/23/16
Adrenocortical carcinoma [ACC]	80	80	03/14/16
Bladder Urothelial Carcinoma [BLCA]	412	412	03/14/16
Brain Lower Grade Glioma [LGG]	516	516	03/14/16
Breast invasive carcinoma [BRCA]	1100	1097	03/14/16
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	03/14/16
Cholangiocarcinoma [CHOL]	36	36	03/14/16
Colon adenocarcinoma [COAD]	461	461	03/16/16
Esophageal carcinoma [ESCA]	185	185	03/14/16
FFPE Pilot Phase II [FPPP]	38	38	01/25/16
Glioblastoma multiforme [GBM]	529	528	03/16/16

1000 Genomes Project



- http://www.1000genomes.org

Data Deluge Summary

Genomic sequence variation

1000 Genomes Project
http://www.1000genomes.org/
Data collection and a catalog of human variation

dbSNP
http://www.ncbi.nlm.nih.gov/projects/SNP/
A catalog of SNPs and short indels

dbVar and Database of Genomic Variants
http://www.ncbi.nlm.nih.gov/dbvar/
http://dgv.tcag.ca/dgv/app/home?ref=GRCh37/hg19
http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&g=dgvPlus (browser track)
A catalog of structural variants

Online Mendelian Inheritance in Man
http://www.omim.org/about
OMIM is a comprehensive, authoritative compendium of human genes and genetic p

The Exome Aggregation Consortium (ExAC)
http://exac.broadinstitute.org/
ExAC is a coalition of investigators seeking to aggregate and harmonize exome seq
useful reference set of allele frequencies for severe disease studies. All of the raw da

Molecular function

Encyclopedia Of DNA Elements (ENCODE) Project
http://genome.ucsc.edu/ENCODE/
Links to ENCODE2 uniformly processed histone mark data: https://sites.google.com/site/anshulkundaje/projects/epigenomeroadmap
Links to other ENCODE2 uniformly processed data: http://genome.ucsc.edu/ENCODE/downloads.html
Data collection, integrative analysis, and a comprehensive catalog of
all sequence-based functional elements

Roadmap Epigenomics Project (NIH Common Fund)
http://compbio.mit.edu/roadmap (Uniformly processed data)
http://www.roadmapepigenomics.org/
https://commonfund.nih.gov/epigenomics/
Data collection, integrative analysis and a resource of human epigenomic data

International Human Epigenome Consortium (IHEC)
http://www.ihec-epigenomes.org/
Data collection and reference maps of human epigenomes for key
cellular states relevant to health and diseases

BLUEPRINT Epigenome
http://www.blueprint-epigenome.eu/
http://epigeneticsandchromatin.biomedcentral.com/articles/1756-8935-5-13
Data collection on the epigenome of blood cells

Human BodyMap
Viewable with Ensembl (http://www.ensembl.org/index.html) or the
Integrated Genomics Viewer (http://www.broadinstitute.org/igv)
Gene expression database from Illumina, from RNA-seq data

Cancer Cell Line Encyclopedia (CCLE)
http://www.broadinstitute.org/ccle/home
Array based expression data, CNV, mutations, perturbations over huge collection of cell lines

FANTOM5 Project
http://fantom.gsc.riken.jp/
http://fantom.gsc.riken.jp/5/data/Data_source
Large collection of CAGE based expression data across multiple species (time-series and perturbations)

Array Express
http://www.ebi.ac.uk/arrayexpress/
Database of gene expression experiments

Gene Expression Atlas
http://www.ebi.ac.uk/gxa/
Database supporting queries of condition-specific gene expression on
a curated subset of the Array Express Archive.

GNF Gene Expression Atlas
Viewable at BioGPS (http://biogps.org/#goto=welcome)
GNF (Genomics Institute of the Novartis Research Foundation) human and
mouse gene expression array data.

The Human Protein Atlas
http://www.proteinatlas.org/
Protein expression profiles based on immunohistochemistry for a large
number of human tissues, cancers and cell lines; subcellular
localization; transcript expression levels

UniProt
http://www.uniprot.org/
A comprehensive, freely accessible database of protein sequence and
functional information

InterPro
http://www.ebi.ac.uk/interpro/
An integrated database of protein classification, functional domains,
and annotation (including GO terms).

Protein Capture Reagents Initiative
http://commonfund.nih.gov/proteincapture/
Resource generation: renewable, monoclonal antibodies and other
reagents that target the full range of proteins

Knockout Mouse Program (KOMP)
http://www.nih.gov/science/models/mouse/knockout/index.html
Resource generation: create knockout strains for all mouse genes.
Trans-NIH project

The Connectivity Map (CMAP)
http://www.broadinstitute.org/cmap/
The Connectivity Map (also known as cmap) is a collection of genome-wide transcriptional expression data from cultured human cells treated with t

Library of Integrated Network-based Cellular Signatures (LINCS)
http://commonfund.nih.gov/LINCS/
Data collection and analysis of molecular signatures that describe how
different types of cells respond to a variety of perturbing agents

Genomic of drug sensitivity in cancer
http://www.cancerrxgene.org/
Mutation, CNV, Affy expression and drug sensitivity in ~300 cancer cell-lines
Papers: http://cancerdiscovery.aacrjournals.org/content/4/1/CD-13-0655.long , http://www.nature.com/nature/journal/v483/n7391/full/nature11005.html

Phenotypes and disease

Human Ageing Genomic Resources
http://genomics.senescence.info/

The Cancer Genome Atlas (TCGA)
http://cancergenome.nih.gov/
Data collection and a data repository including cancer genome sequence data

International Cancer Genome Consortium (ICGC)
http://www.icgc.org/
Data collection and a data repository for a comprehensive description
of genomic, transcriptomic and epigenomic changes of cancer

Genotype-Tissue Expression (GTEx) Project
https://commonfund.nih.gov/GTEx/
Data collection, data repository, and sample bank for human gene
expression and regulation in multiple tissues, compared to genetic
variation

Knockout Mouse Phenotyping Program (KOMP2)
https://commonfund.nih.gov/KOMP2/
Data collection for standardized phenotyping of a genome-wide
collection of mouse knockouts

Database of Genotypes and Phenotypes (dbGaP)
http://www.ncbi.nlm.nih.gov/gap
Data repository for results from studies investigating the interaction
of genotype and phenotype

NHGRI Catalog of Published GWAS
http://www.genome.gov/gwastudies/
Public catalog of published Genome-Wide Association Studies

Clinical Genomic Database
http://research.nhgri.nih.gov/CGD/
A manually curated database of conditions with known genetic causes, focusing on medically significant genetic data with available in

NHGRI's Breast Cancer Information core
http://research.nhgri.nih.gov/bic/
Breast Cancer Mutation database

ClinVar
http://www.ncbi.nlm.nih.gov/clinvar/
ClinVar is designed to provide a freely accessible, public archive of reports of the relationships among human variations and phenoty
presents the data for interactive users as well as those wishing to use ClinVar in daily workflows and other local applications. ClinVar w

Human Gene Mutation Database (HGMD)
http://www.hgmd.cf.ac.uk/ac/
The Human Gene Mutation Database (HGMD) represents an attempt to collate known (published) gene lesions responsible for hur

NHLBI Exome Sequencing Project (ESP) Exome Variant Server
http://evs.gs.washington.edu/EVS/
The goal of the NHLBI GO Exome Sequencing Project (ESP) is to discover novel genes and mechanisms contributing to heart, lung a
disorders.

Genetics Home Reference
http://ghr.nlm.nih.gov/
Genetics Home Reference is the National Library of Medicine's web site for consumer information about genetic conditions and the ge

GeneReviews
http://www.ncbi.nlm.nih.gov/books/NBK1116/
GeneReviews are expert-authored, peer-reviewed disease descriptions presented in a standardized format and focused on clinically r

Global Alzheimer's Association Interactive Network (GAAIN)
http://www.gaain.org/
The Global Alzheimer's Association Interactive Network (GAAIN) is a collaborative project that will provide researchers around the gl
and other neurodegenerative diseases.
In 2013, obtained WGS data for the largest cohort of 800 Alzheimer's patients

The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium
http://web.chargeconsortium.com/
The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium was formed to facilitate genome-wide a

The NIMH Center for Collaborative Genomic Studies on Mental Disorders
(Include Psychiatric Disease Consortium http://pgc.unc.edu/)
http://www.nimhgenetics.org/
The NIMH Center, now known as NIMH Repository and Genomics Resource (NIMH-RGR) plays a key role in facilitating psychiatric gen

References

- Kundaje and Pritchard lecture notes, GENE245
- Linkage Disequilibrium Part 1, University of Washington
- Statistical Challenges in genome-wide associations, Lecture14 University of Oslo
- How to interpret a Genome-wide association study, by Thomas Pearson and Teri Manolio
- Hands-on tutorial to Genome-wide Association studies, by Umit Seren