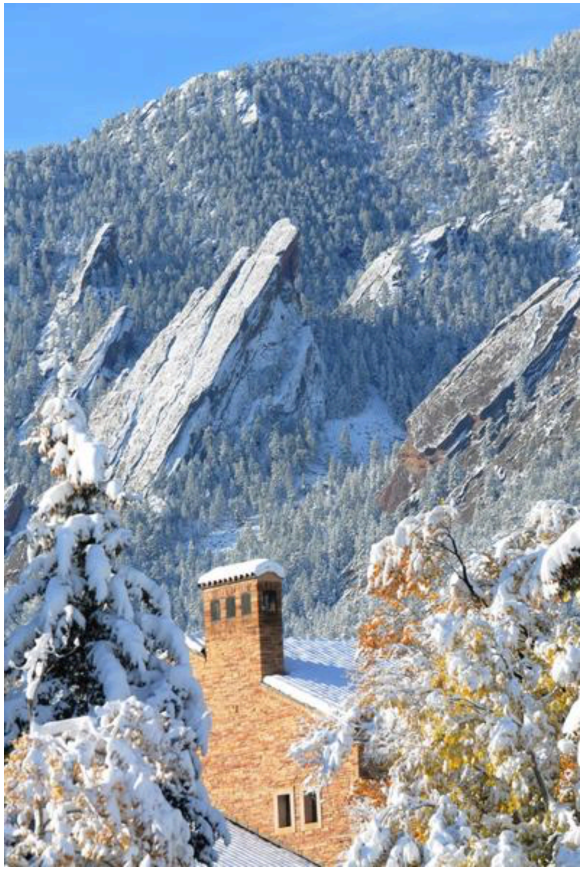
University of Colorado
Boulder

Lecture 1: Introduction

Matt McQueen | Associate Professor

Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado



Personal Introduction

- Biostatistics (post-doctoral fellowship)
  - Harvard School of Public Health
- Epidemiology (doctoral)
  - Harvard School of Public Health
- Public Health (master’s)
  - University of Washington School of Public Health
- Neuroscience (bachelor’s)
  - University of Colorado Boulder

What I do

- “Big Data Epidemiology”
- Risk Factors
  - Molecular and genetic factors
  - Traditional epidemiological factors
- Health and behavioral outcomes
  - Behavioral and Psychiatric outcomes
  - Cardio-metabolic outcomes (obesity)

Course Overview – Day 1

- Lecture 1: Introduction
  - Tutorial 1: Getting Started

- Lecture 2: Quality Control Procedures
  - Tutorial 2: Data Cleaning and Genetic Ancestry

Course Overview – Day 2

- Lecture 3: Genome-Wide Association Approaches
  - Tutorial 3: Genome-Wide Association Analysis

- Lecture 4: Aggregation of GWAS Data
  - Tutorial 4: Heritability and Polygenic Scores

Course Overview – Day 3

- Lecture 5: Family-Based Approaches
  - Tutorial 5: Family-Based Association Testing

- Lecture 6: Meta-Analysis
  - Tutorial 6: Conducting a meta-analysis

Course Objectives

- Have a working knowledge of the steps necessary to carry out a genome-wide analysis
- Appreciate the inherent limitations to genome-wide analysis
- Gain hands-on experience working with genome-wide data

Background

Measurable Genetic Variation

- DNA
  - Frequencies of alleles at base-pair locations
- Gene Expression
  - Amount of gene product being expressed
- Epigenetics
  - Methylation patterns, etc.

DNA Variation

- DNA
  - Adenine (A)
  - Guanine (G)
  - Cytosine (C)
  - Thymine (T)
- DNA double helix
  - A pairs with T and G pairs with C
- Codons
  - Triplets of bases
  - 64 possible codons
    - 20 amino acids

Mutations:

A Source of DNA Variation

- Point
  - Substitute one base for another
- Deletions
  - Base removed entirely
- Insertions
  - Base inserted
- Duplications
  - Base and/or sequence duplicated

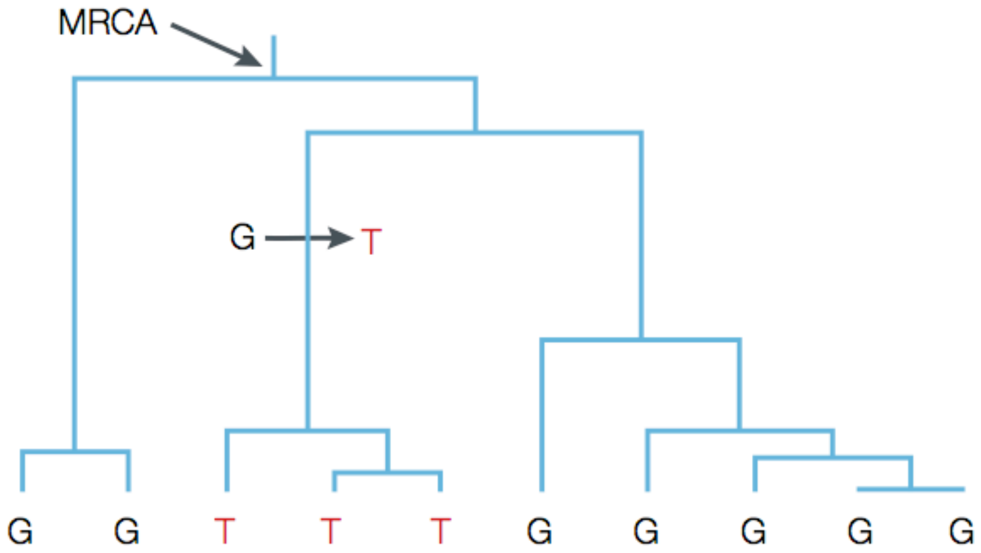
Point Mutations:

A Source of Single Nucleotide Polymorphisms

- Base pair substitution
  - (e.g. replication error)
- Synonymous
  - *No change* in amino acid
- Nonsynonymous
  - Amino acid change
- Transitions
  - Between A and G or C and T
  - Change within same base (purines and pyramidines)
- Transversions
  - All others
  - Change between bases

Population Genetics

- Infinite Sites Model
  - Each mutation creates a unique polymorphic site
  - Mutation rate ~ 10^{-6}

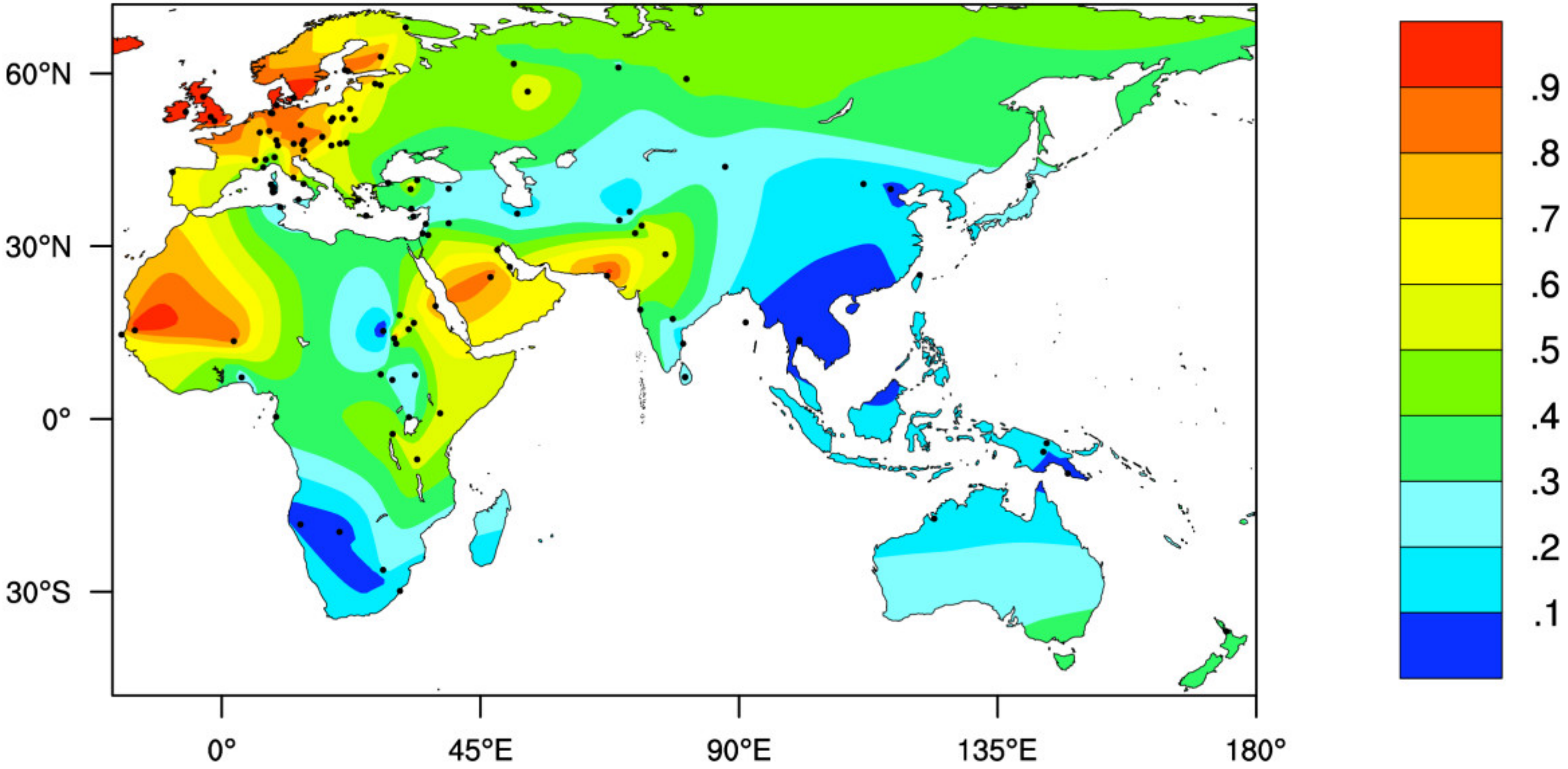


Life After Mutation

- Mutation is neutral
  - Random Genetic Drift
    - Eventually, the allele will “drift” out
- Mutation is harmful
  - Selective Pressure
    - Allele may quickly disappear
- Mutation is beneficial
  - Selective Pressure
    - Allele frequency may increase rapidly

Beneficial Mutation

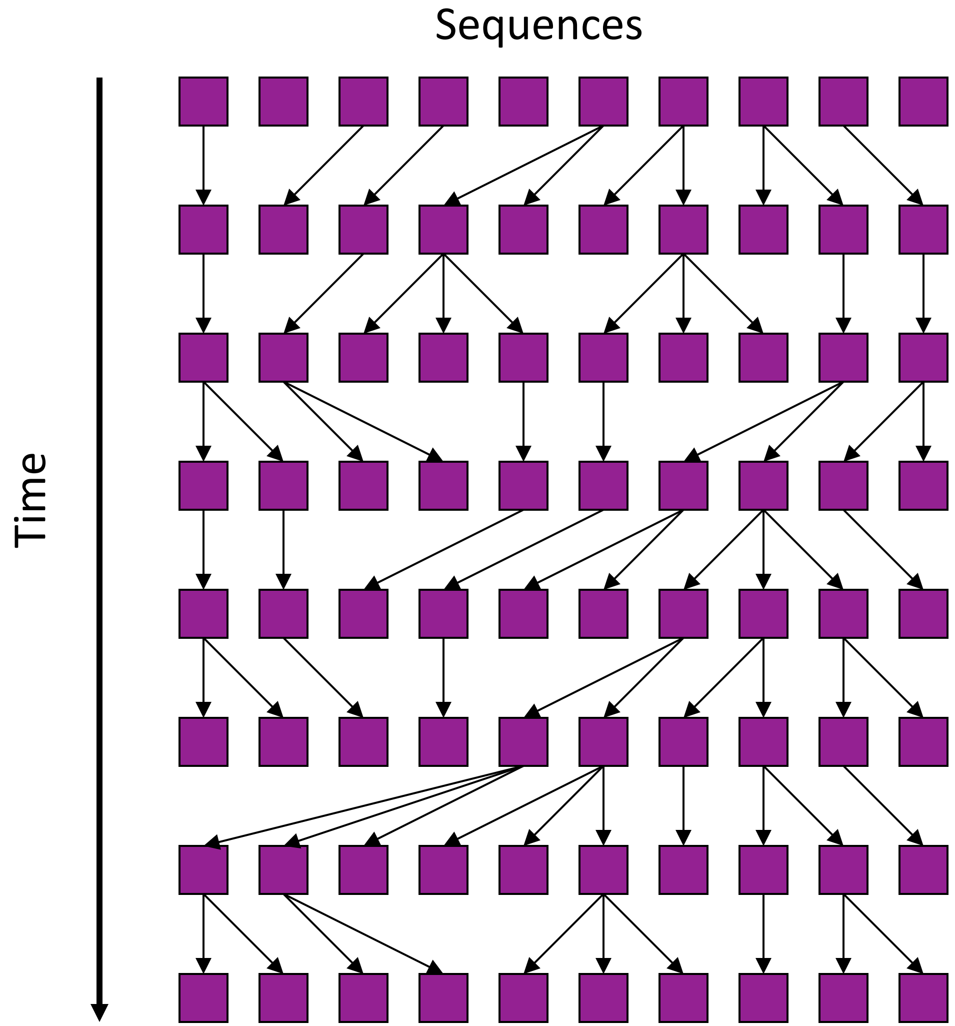
- Lactase Persistence (~7500 years)



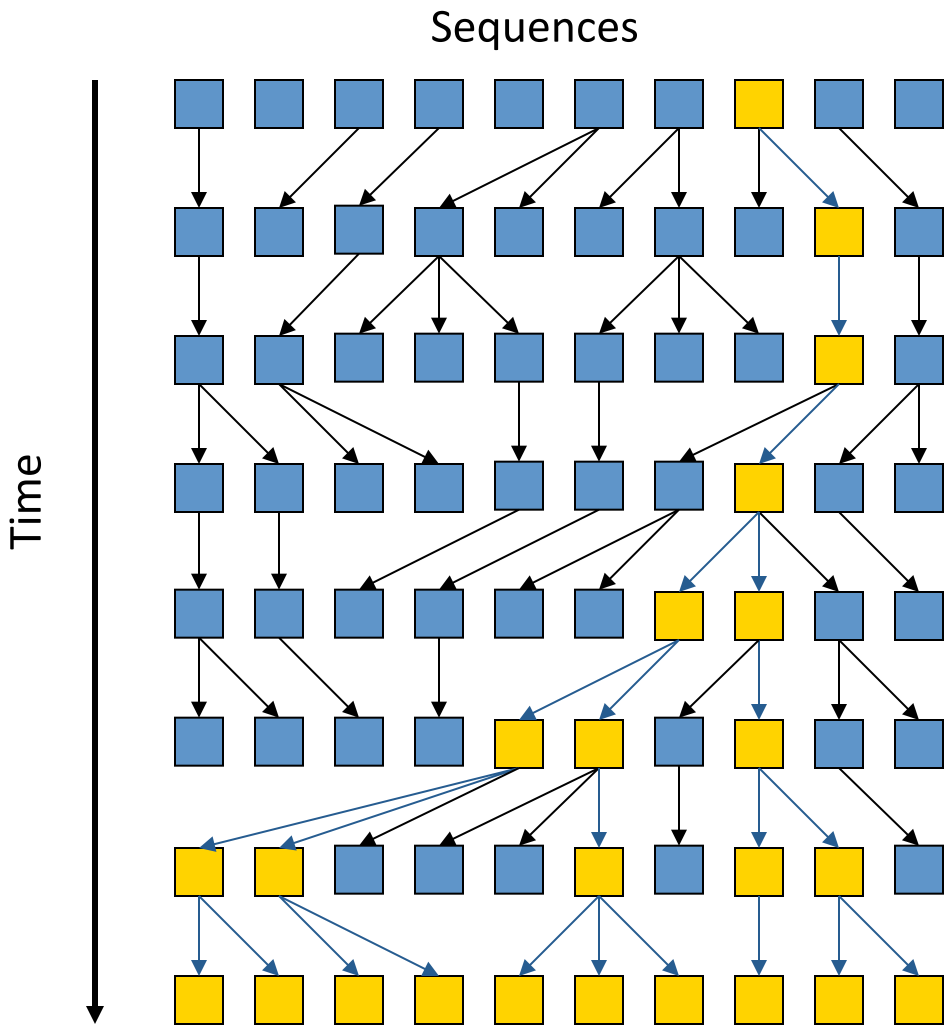
Beneficial Mutation

- Thriving at Altitude (~3000 years)

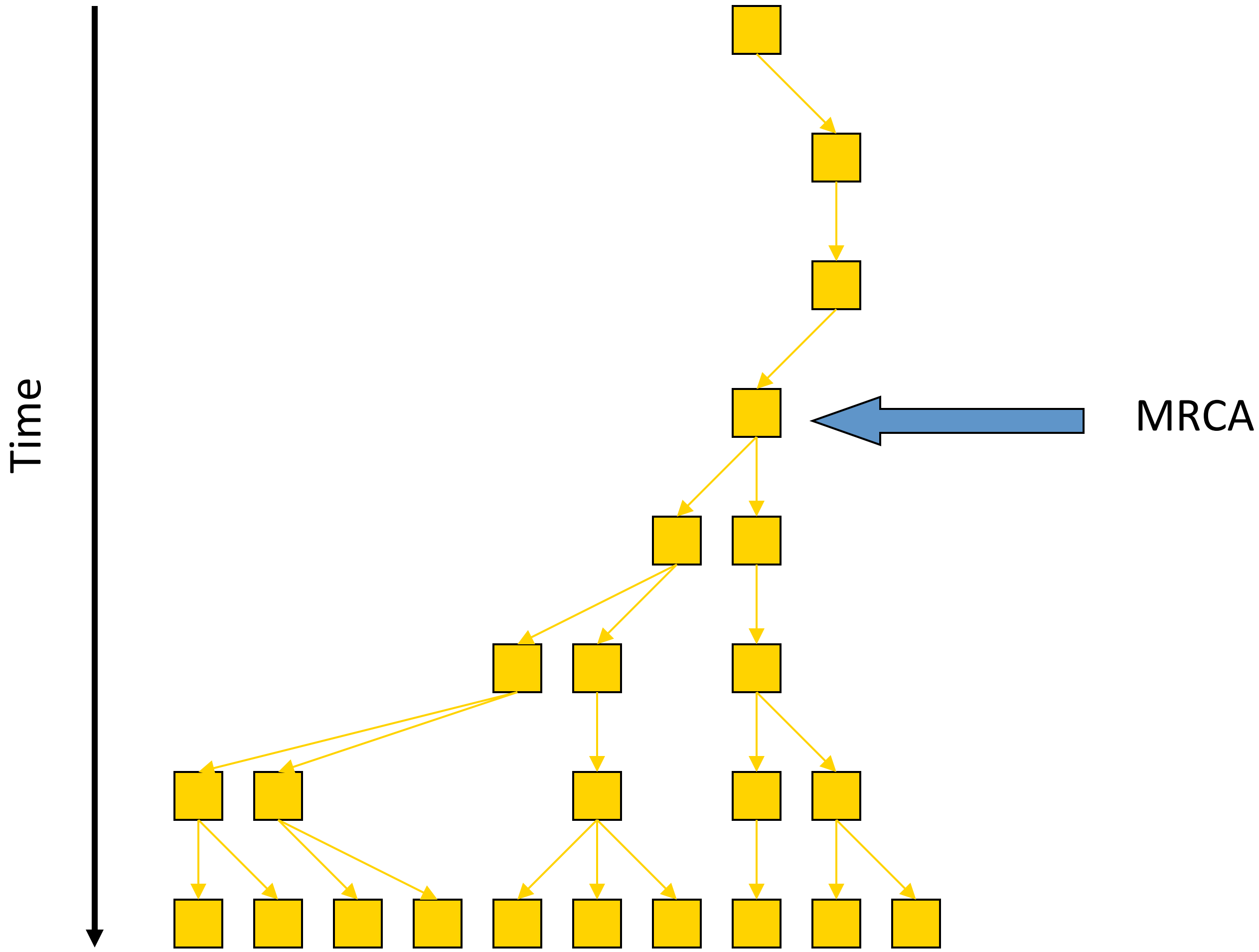
Common Ancestry



Common Ancestry



Common Ancestry



Ancestry and Genetic Variation

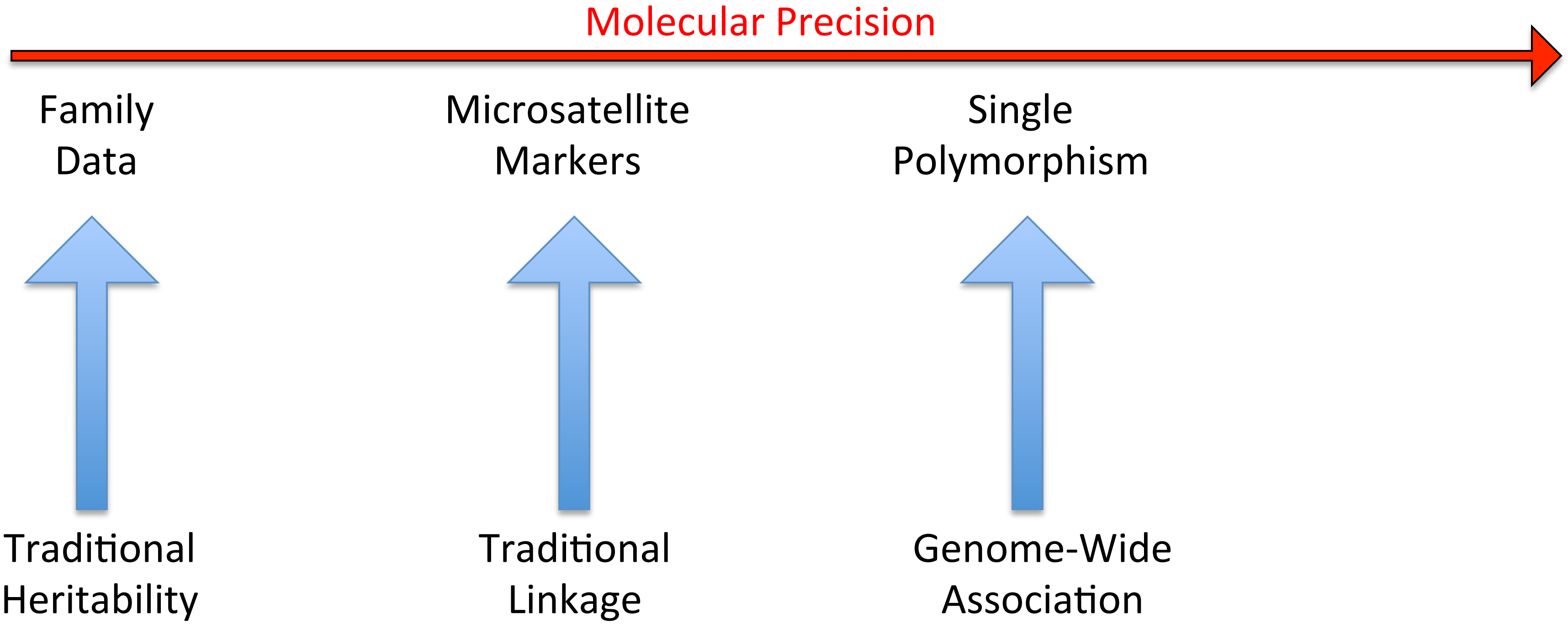
- All DNA sequences are derived from others
  - Every sample has a genealogy
- Eventually, all lineages coalesce
  - Most Recent Common Ancestor (MRCA)
- Mutations may become polymorphisms

Measuring the Genome

Rapid pace of technology

- Advantages
  - Unprecedented look into biological systems
  - Efficiency is up
  - Cost is down
- Disadvantages
  - Technology is driving the bus
  - We do things because we can
  - We often have no idea what we’re looking for

Measuring the Genome



Molecular Precision

Family
Data

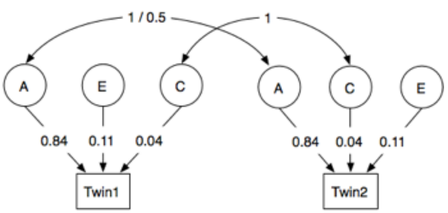
Microsatellite
Markers

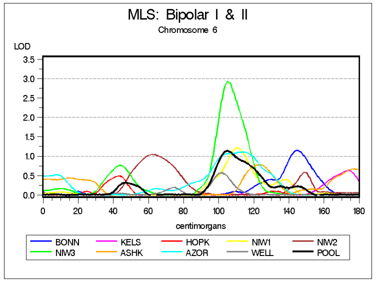
Single
Polymorphism

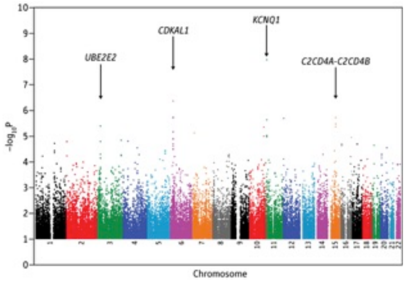
Traditional
Heritability

Traditional
Linkage

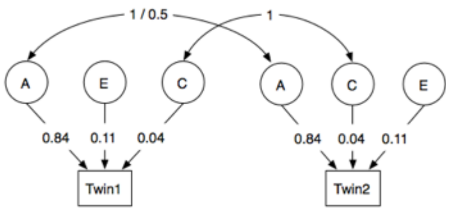
Genome-Wide
Association





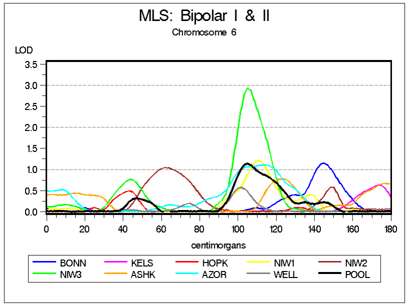


Traditional Heritability



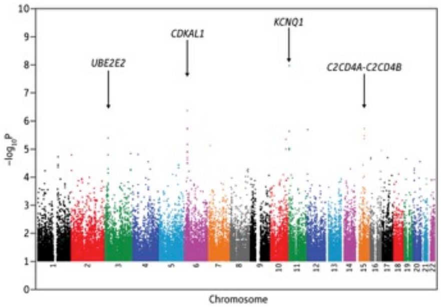
- Objective
  - “Does it run in families?”
- Design
  - Family, twin and adoption studies
- Molecular data
  - None
- Desired outcome
  - Gives us a clue that genetics might be important

Traditional Linkage



- Objective
  - – Find **genomic loci** linked to disease
- Design
  - – Family-based
- Molecular information
  - – 300 - 600 markers (often short tandem repeats)
- Desired outcome
  - – Find genetic variation linked to disease

Genome-Wide Association



- Objective
  - – Find **common alleles** associated with disease
- Design
  - – Cohort, case-control, family-based
- Molecular information
  - – 500,000 – 2.5M single nucleotide polymorphisms
- Desired outcome
  - – Find genetic variation associated with disease

Why Genome-Wide Association?

- Molecular precision
  - Measure the genome on a more refined scale
- More powerful to detect common alleles
  - Common disease, common variant
- A result of the human genome project effort

Approaches to Genetic Research

- Genome-Wide
  - Linkage Analysis
  - **Genome-Wide Association Analysis**
  - Whole Genome Sequence Analysis
- Targeted
  - Candidate Gene(s) Association Analysis

Gene Hunters

- By definition, "integrative"
  - Combines epidemiological, sociological, statistical, clinical, genetic and molecular approaches

- The Goal…
  - FIND GENES INVOLVED WITH DISEASE

Why Hunt for Genes?

Why Hunt for Genes?

- Disease etiology
- Refined diagnosis and/or prognosis
- Drug development
- Disease prediction

Gene-Mapping

- Monogenic ‘Mendelian’ Diseases
  - Rare disease
  - Rare variants
    - Highly penetrant
- Complex Disease
  - Rare/Common disease
  - Rare/Common variants
    - Variable penetrance

Gene-Mapping

- Monogenic ‘Mendelian’ Diseases
  - Rare disease
  - Rare variants
    - Highly penetrant

Linkage!

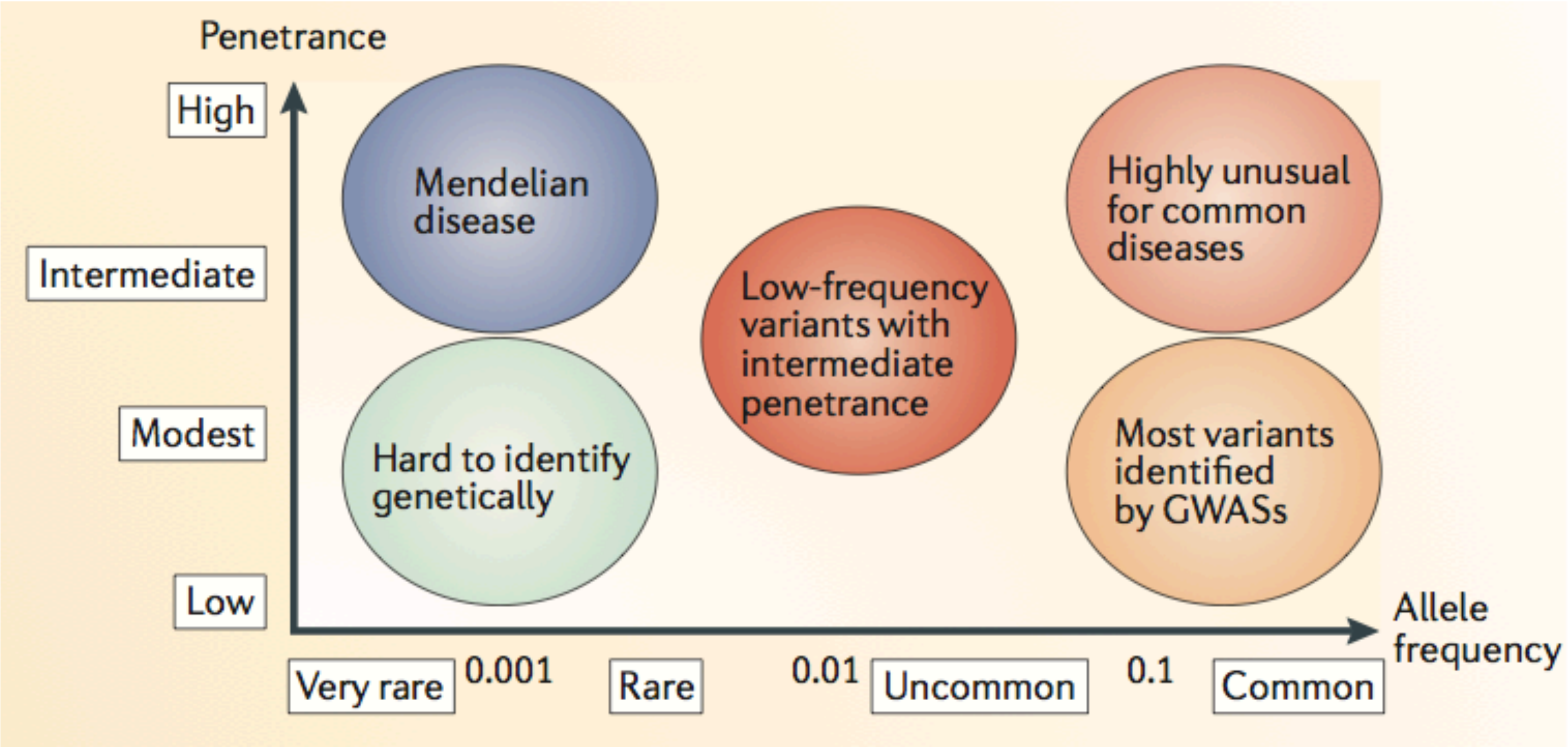
- Complex Disease
  - Rare/Common disease
  - Rare/Common variants
    - Variable penetrance

Gene-Mapping

- Monogenic 'Mendelian' Diseases
  - Rare disease
  - Rare variants
    - Highly penetrant
- Complex Disease
  - Rare/Common disease
  - Rare/Common variants
    - Variable penetrance

Association

Disease and DNA Variation


Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWASs
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Penetrance: P(D | G)

2012 Nature Reviews | Genetics

“Where in the genome…?”

- 1980s - 2005
    - Linkage (LOD Scores, etc.)
- 2006 -
    - Association

Why Now?

The “-omics” Age

c. 2000

- –Pre-genomic era
- –100’s of Markers
  - STRs
- –Genome-wide linkage

c. 2016

- –Post-genomic era
- –1M+ markers
  - SNPs
- –Genome-wide association

Genetic Information

- Human Genome Project
  - One human genome (3B base pairs)
- HapMap Project
  - 100s of human genomes (millions of base pairs)
- ENCODE Project
  - 100s of human genomes (functional data)
- 1000 Genomes Project
  - 1000s of human genomes (12M base pairs)
- Large-Scale Whole Genome Sequencing
  - 1000s of human genomes (3B base pairs)

Next up…

- Tutorial 1
  - Getting started