

Human variation and allele frequency spectrum



In human genetics, experiments are all *post hoc*

No controlled crosses

No defined genetic backgrounds

Large genome (haploid three gigabase pairs)

Good phenotyping!

Lots of \$\$\$

How do we identify genes in humans?

Draft human genome announced in June 2000

ws
Print"

The New York Times

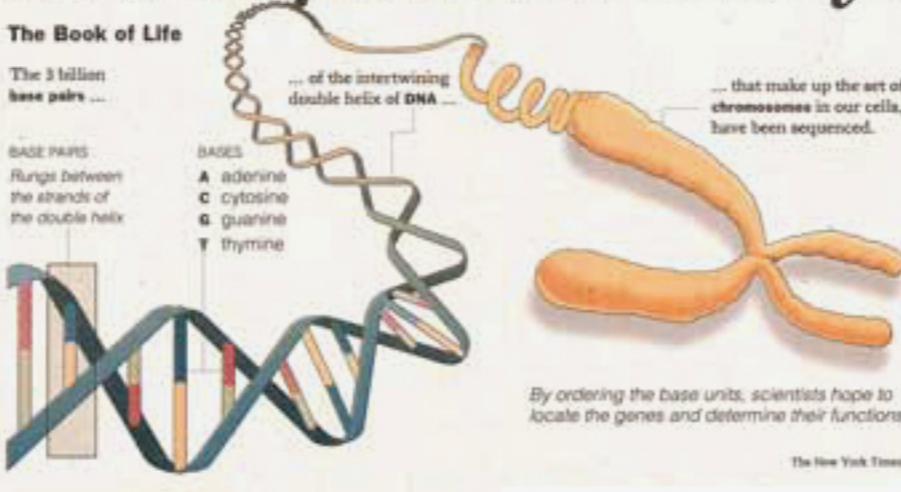
No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 Printed in America ONE DOLLAR

tic Code of Human Life Is Cracked by Scientists



become part ... that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntar

The Book of Life
The 3 billion base pairs ...
BASE PAIRS: Rungs between the strands of the double helix
BASES: A adenine C cytosine G guanine T thymine



... of the intertwining double helix of DNA ...
... that make up the set of chromosomes in our cells, have been sequenced.

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

National Edition
Arizona and New Mexico: M cloudy in New Mexico; thunderstorms in the mountains. Partly sunny elsewhere. Highs 80 mountains, over deserts. Weather map is on Page

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE
WASHINGTON, June 26 — The achievement that represents a milestone of human self-knowledge has been deciphered by rival groups of scientists said today that they had deciphered the human genetic script, the set of instructions that defines the human organism.

It took more than 10 years and \$3 billion

We need physical pieces of DNA to sequence



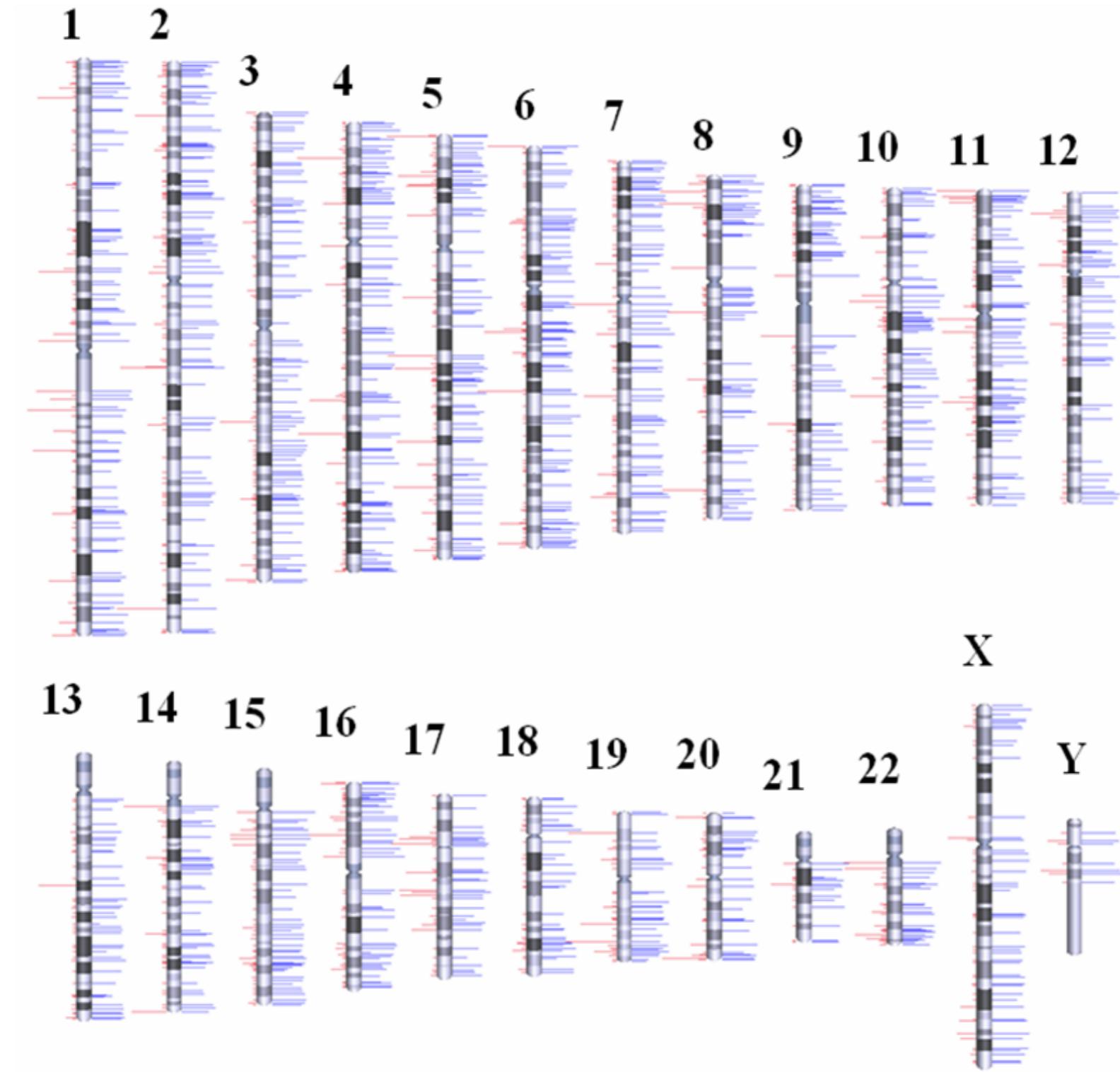
Who was sequenced?

We don't have one human genome



Nine humans had parts of their genomes sequenced to make the first draft.

A genome sequence gives us the (incomplete) parts list



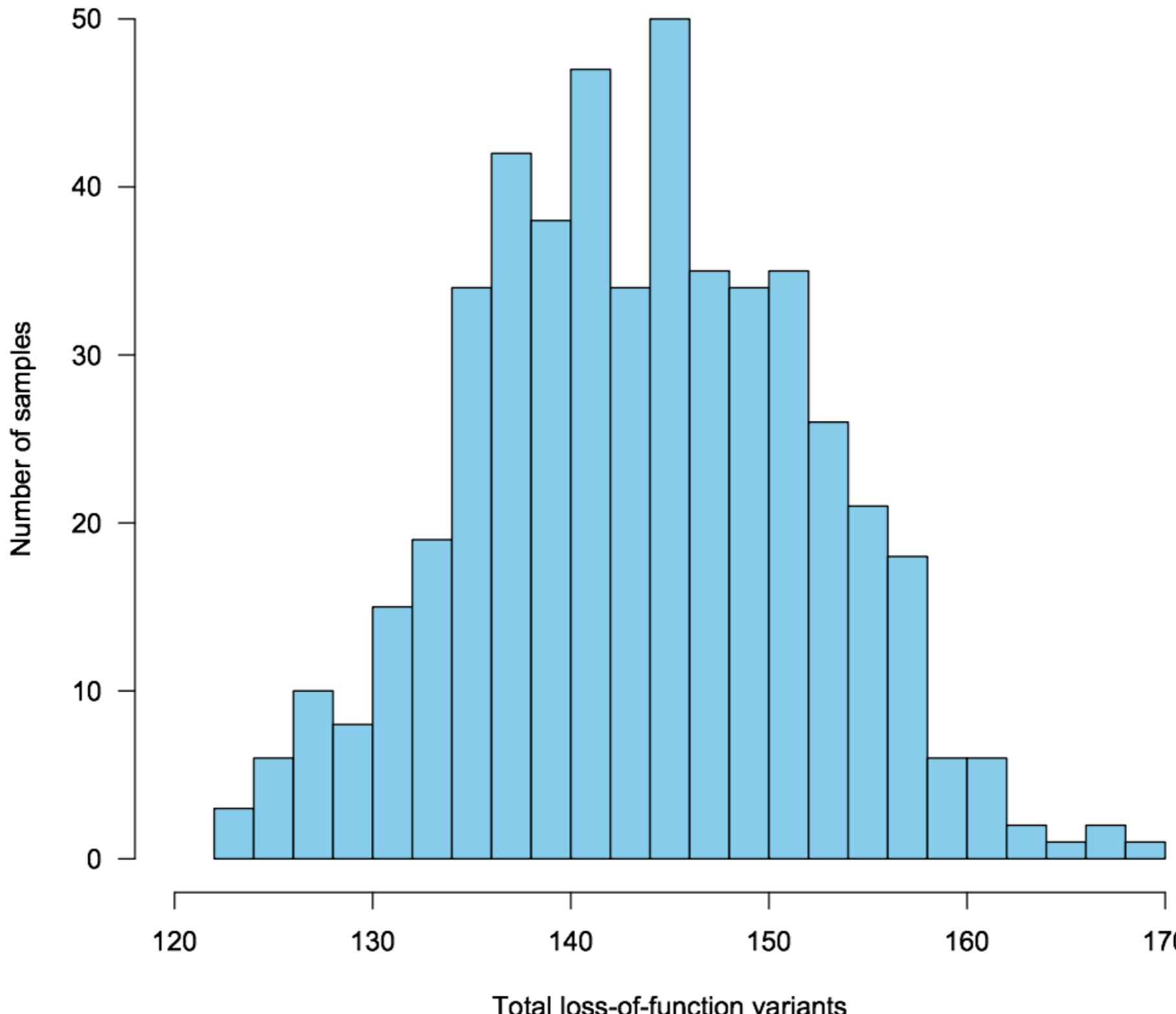
Types of variation



A large block of DNA sequence text, oriented vertically, consisting of multiple lines of nucleotide bases (A, T, C, G). The text is too long to be fully legible but represents a continuous stretch of genetic code.

Rare = variants found in less than 1% in population

We each have over 100 unique loss-of-function rare variants



Over 3,000 rare diseases have a known underlying genetic cause



One in twelve people have a rare disease

Compound heterozygosity underlies many diseases

Types of variation

A large block of DNA sequence text, oriented vertically, consisting of multiple lines of black text on a white background. The text represents a segment of a genome, showing various base pair sequences.

Rare = variants found in less than 1% in population

Common = variants found in more than 5% of the population

Intermediate = variants found in 1-5% of the population

Where does variation come from?



A large block of DNA sequence text, oriented vertically, representing a segment of genetic material. The sequence consists of four lines of text, each containing approximately 25 nucleotide bases (A, T, C, G). The text is color-coded by base: A is blue, T is orange, C is green, and G is red.

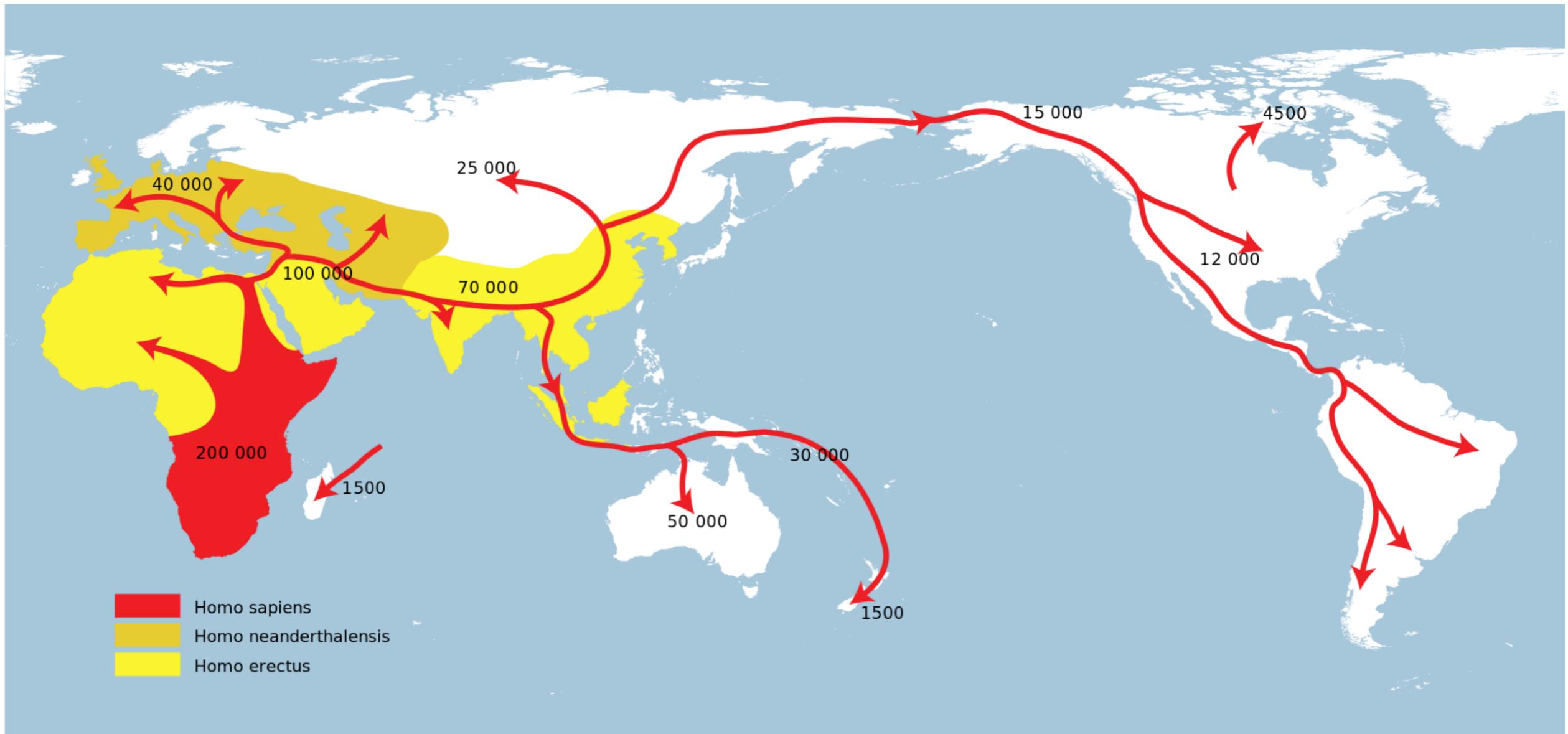
```
TTAGGGCCATGGCTGAA  
TGGCGTCAAAGTATC  
TAGCGACGTGACTCTGAATG  
CTTTCAAGATCATGCCGTAGGC  
CTACGCTGTACAGGTCCAGCTGATT  
AATCGCAGCAGGTTTCGAGCTTCAG  
TGAGCAATGCTCTGAAAATCATA  
TGCTTGACTGATAAGGCTCGAAC  
CTAGCAAACGTGCTCATGATGCC  
AGTGGCCGTACTTATGATGC  
ACGTGTGATATCCAGA  
GACAGTCCATGATGC
```

Random errors in replication, transcription, DNA repair, etc.

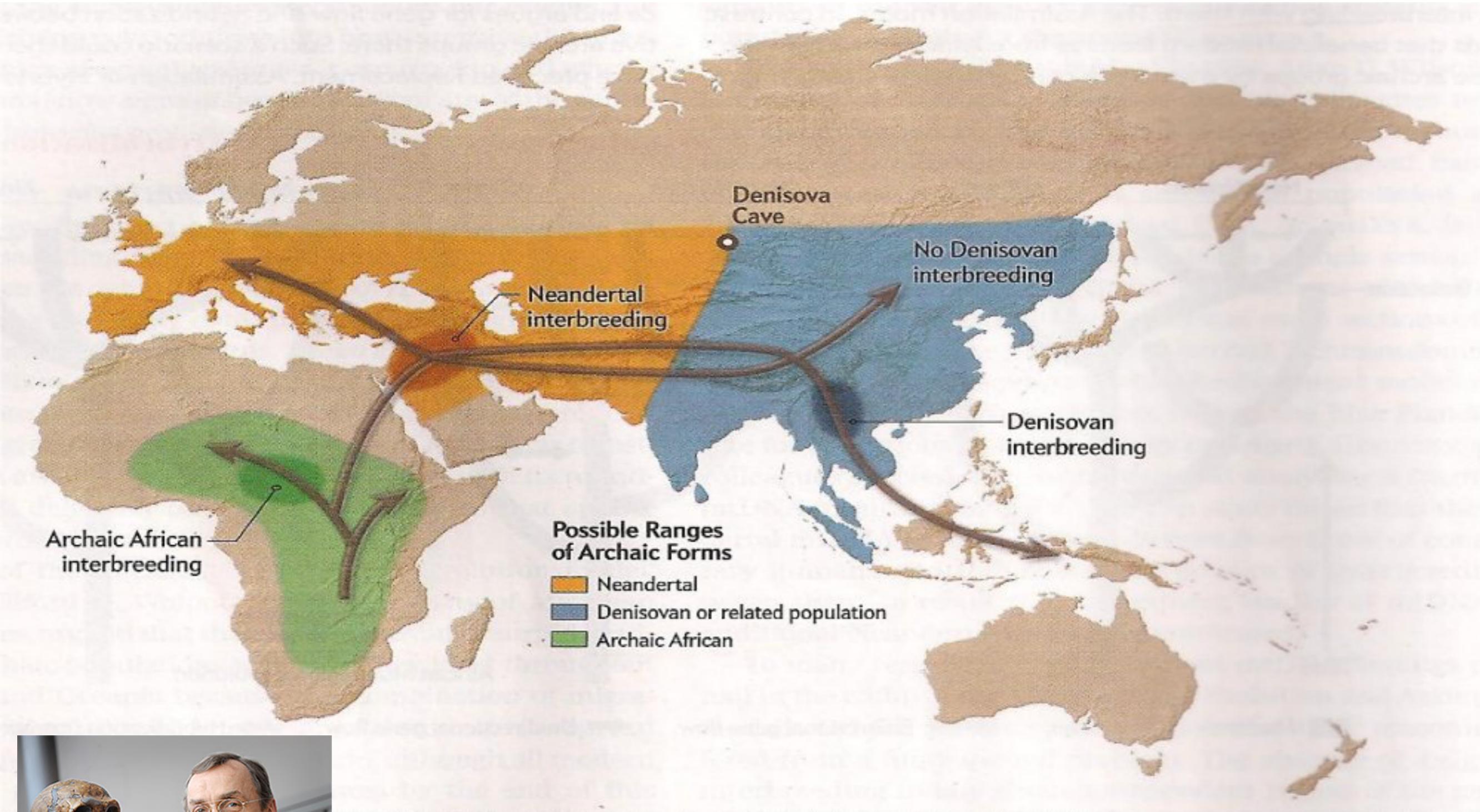
Somatic or germline errors

Once generated, germline variants are inherited

Human history drives our genetics



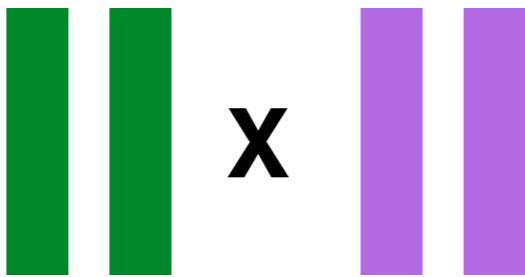
Human history drives our genetics



Svante Pääbo

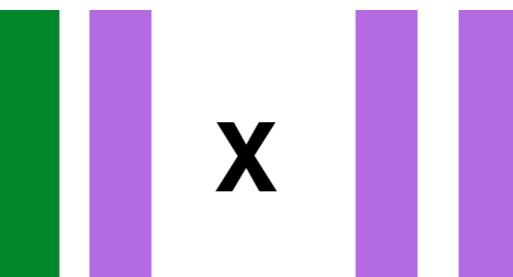


H. neanderthalensis

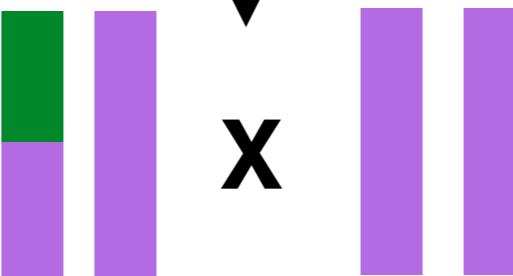


H. sapiens

50%



25%



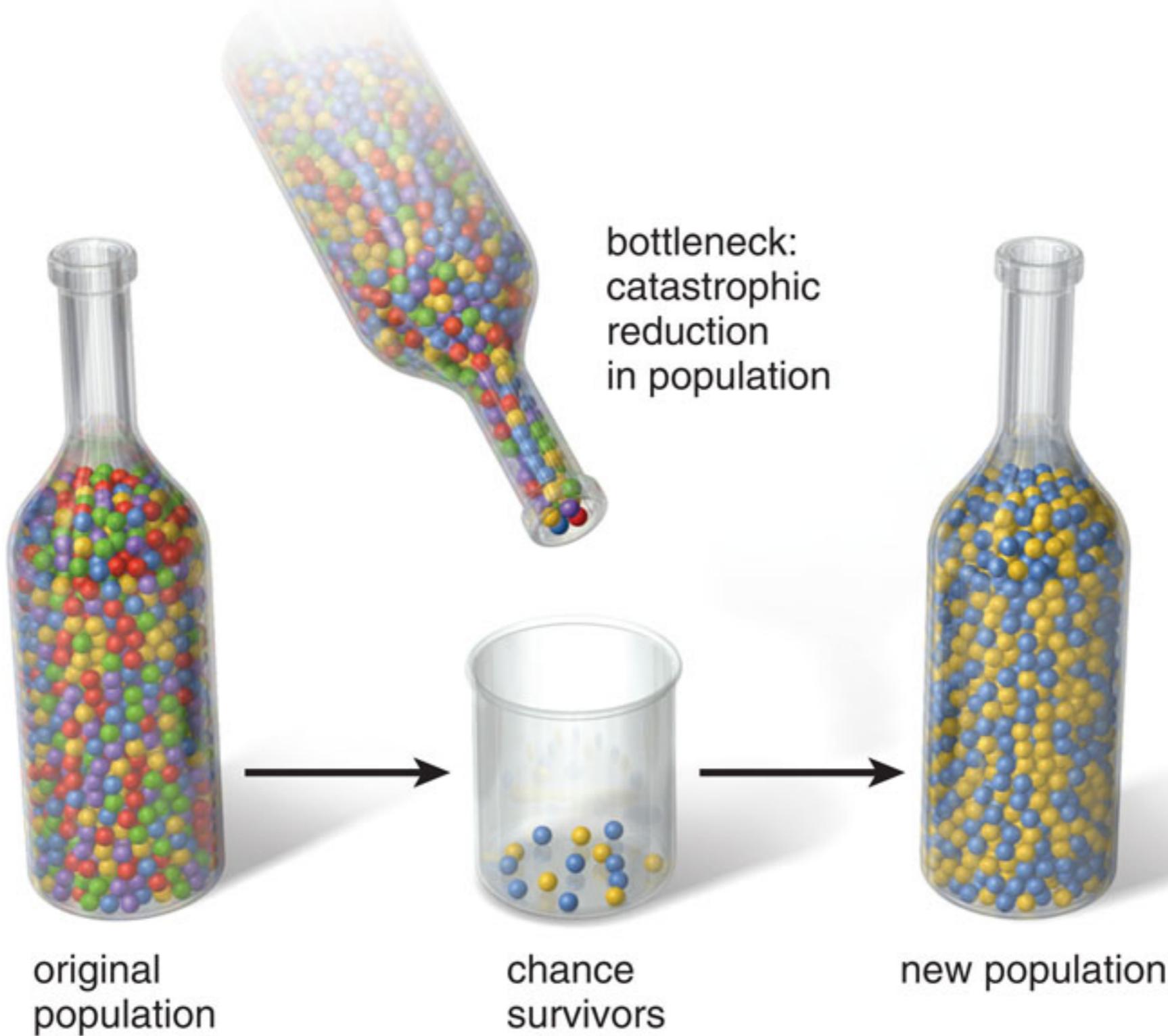
12.5%

6.25%

3.125%

Human history drives our genetics

Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.



The common disease - common variant (CD-CV) hypothesis



Diseases shared by lots of people
will be caused by variants shared by those same people

How do we find all these common variants?

To find common variants, we need markers shared by lots of people



Goal is to find all the common variants

After the HGP, the HapMap project was born.

All three types of variation can cause disease

A large block of DNA sequence text, oriented vertically, consisting of multiple lines of sequence data. The sequence is composed of the four bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G).

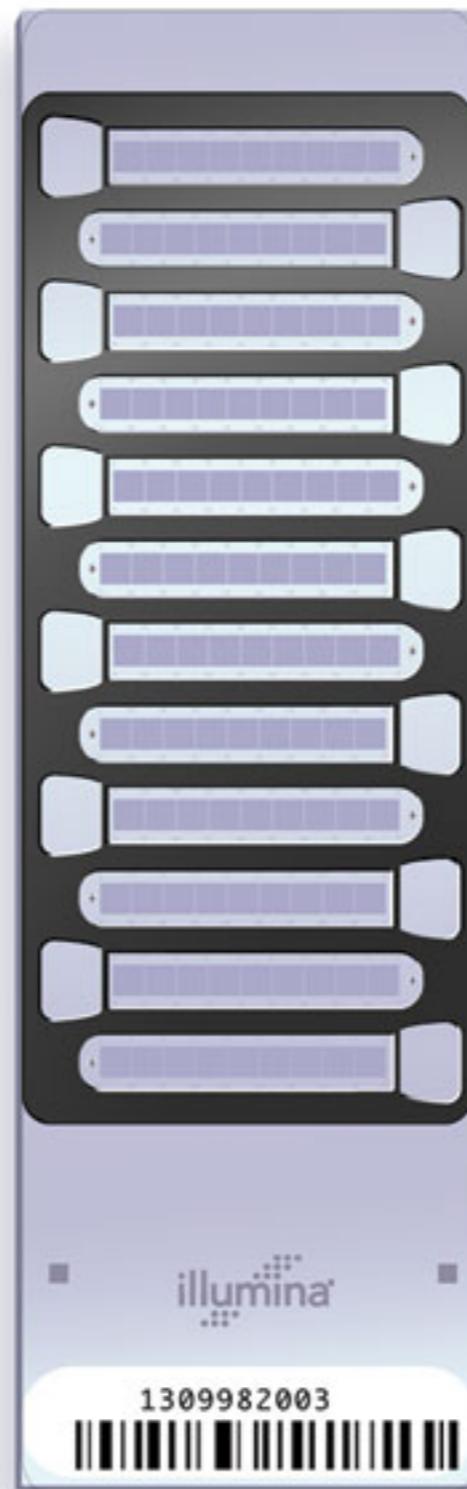
TAGGC
CCATGGCTGA
TAGCGACGTGACTCTGAATG
CTTTCAAGATCATGCCGTAGGTC
CTACGCTGTACAGGTCCAGCTGATT
AATCGCAGCAGGTTTGAGCTTCAAGGCT
TGAGCAATGCTCTGAAAATCATACG
TGCTTGACTGATAAGGCTCGAAC
CTAGCAAACGTGCTCATGATGCC
AGTGGCCGTACTTATGATGC
ACGTGTGATATCCAGA
GACAGTCCAT
CATGC

Rare = variants found in less than 1% in population

Common = variants found in more than 5% of the population

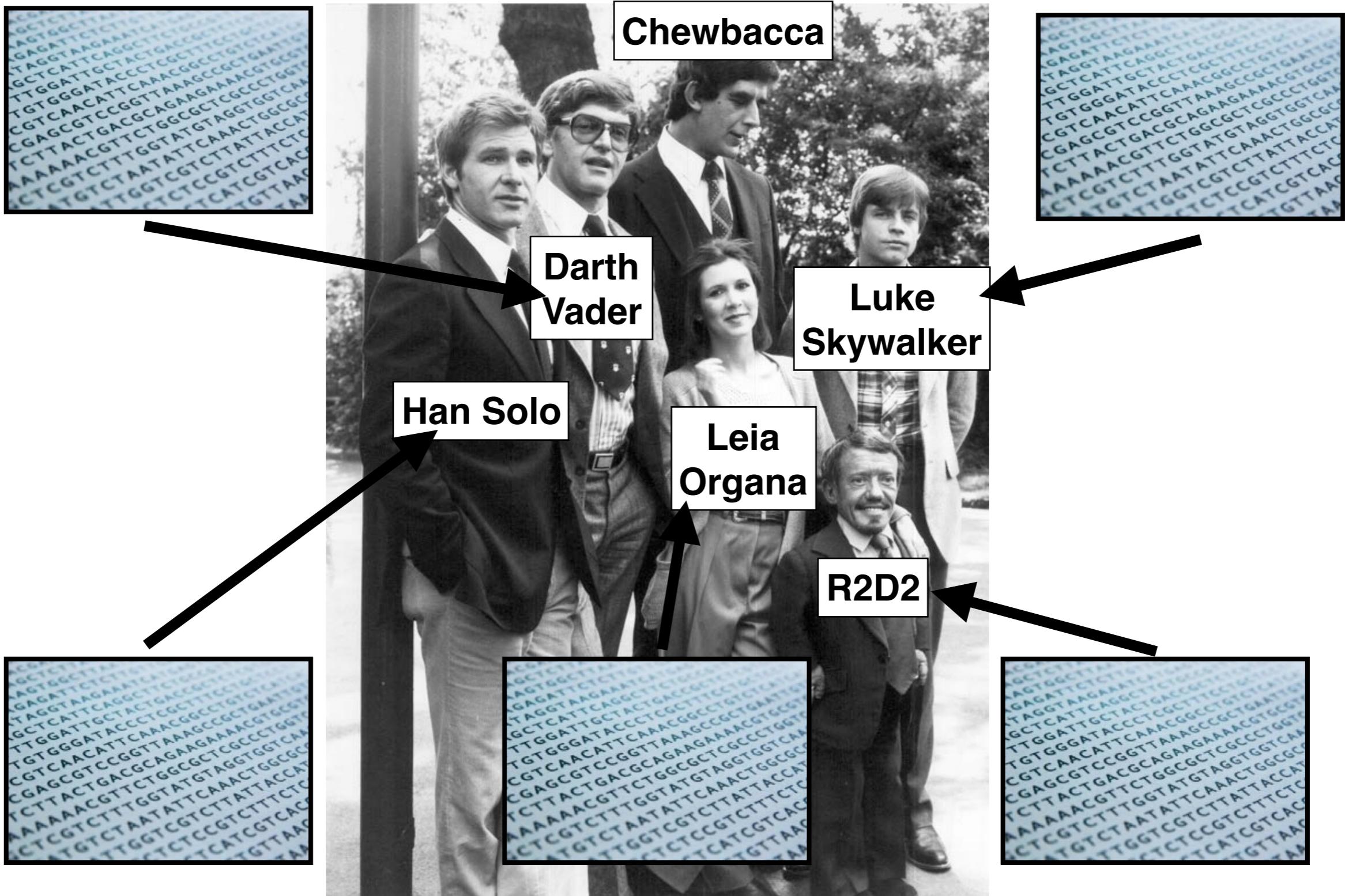
Intermediate = variants found in 1-5% of the population

An array to genotype at >4.3 million sites



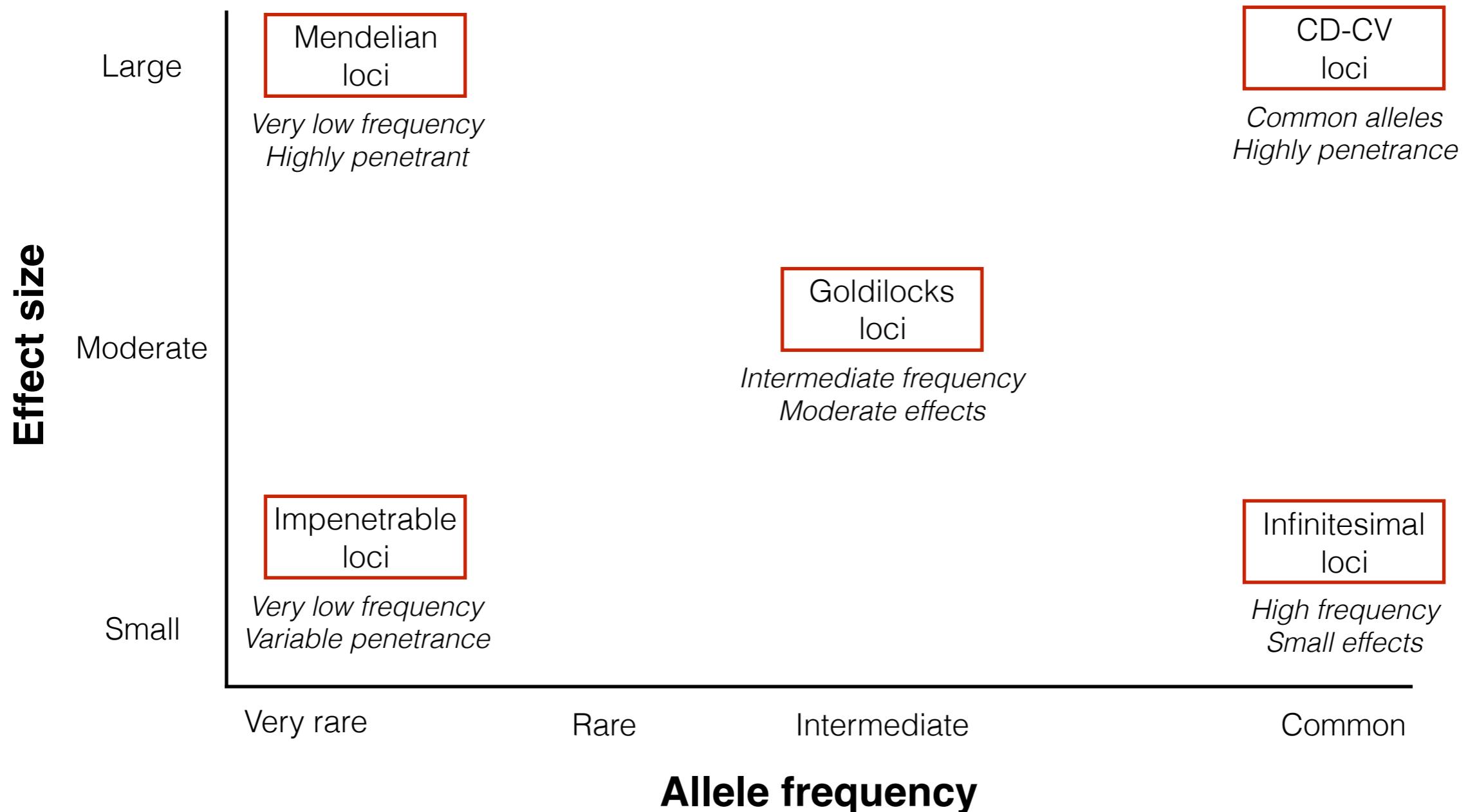
Tool to genotype intermediate and common variation

We want to be able to read genomes and make predictions



The cast of the original *Star Wars*

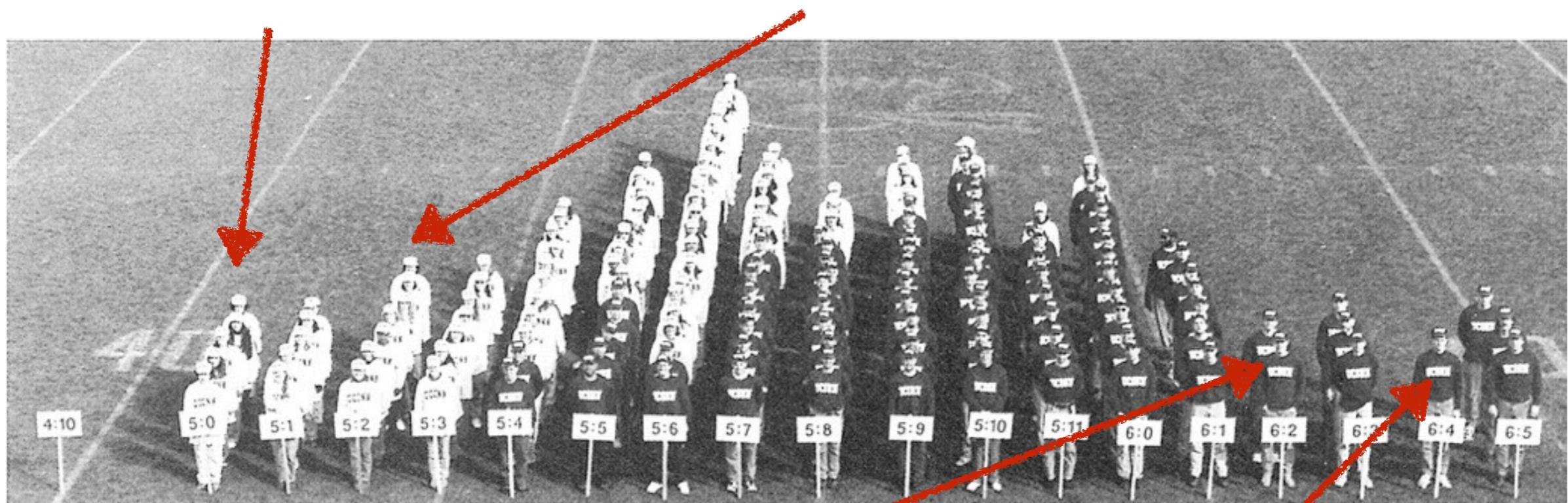
The spectrum of how variation contributes to disease



How do we find the variants that cause common disease?

To find genes in humans, we must correlate genotype with phenotype

CAGCGATAGGCTTAATGTT	CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>T</ins> ATGACCAACG	AGCCC GTTT <ins>T</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT	GGGTTCACAGTGAGCTGTGT

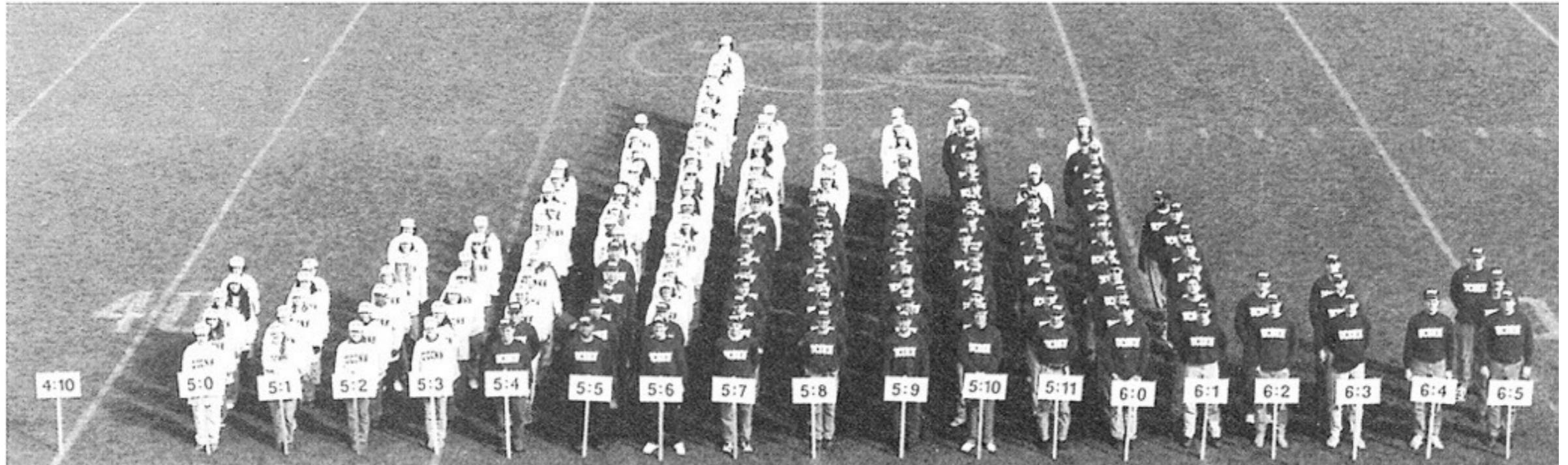


University of Connecticut, 1997

CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>G</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT

CAGCGATAGGCTTAATGTT
AGCCC GTTT <ins>G</ins> ATGACCAACG
GGGTTCACAGTGAGCTGTGT

For traits controlled by many genes, we need many, many people

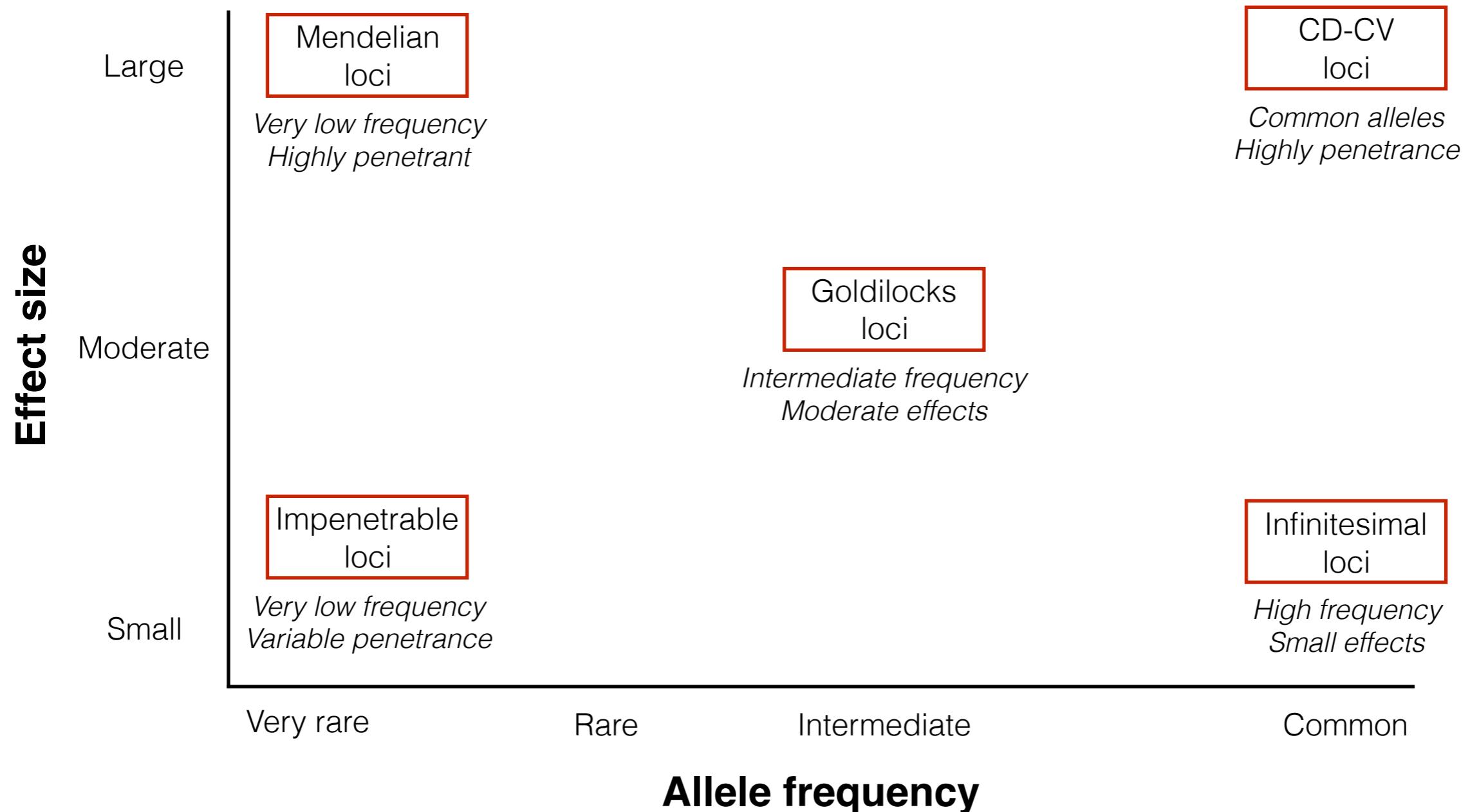


University of Connecticut, 1997

Variation shared by lots of tall people
and not shared by lots of short people

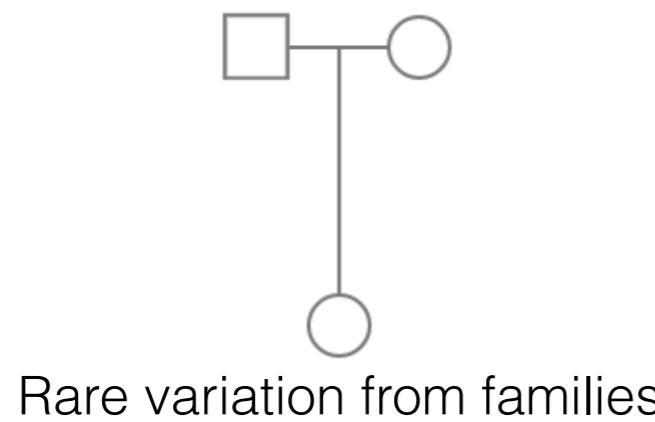
~700,000 people genotyped led to 60%
of height differences explained

The spectrum of how variation contributes to disease

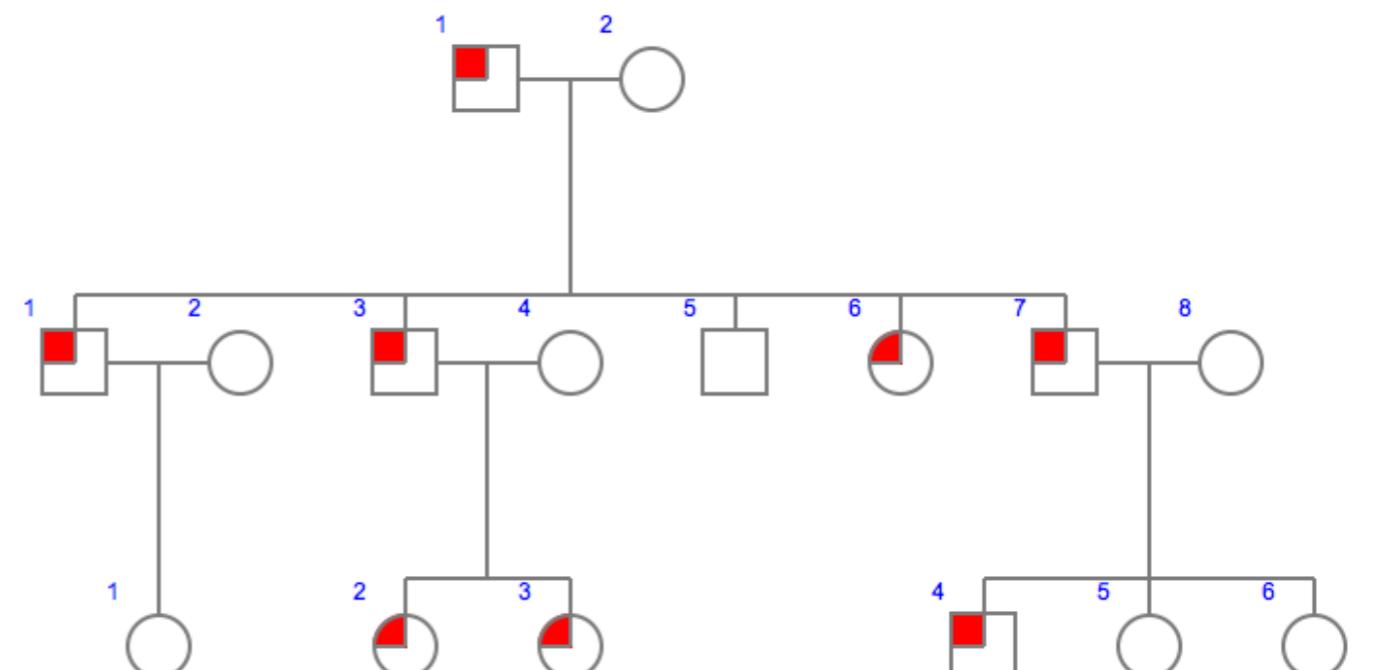


How do we find the variants that cause rare disease?

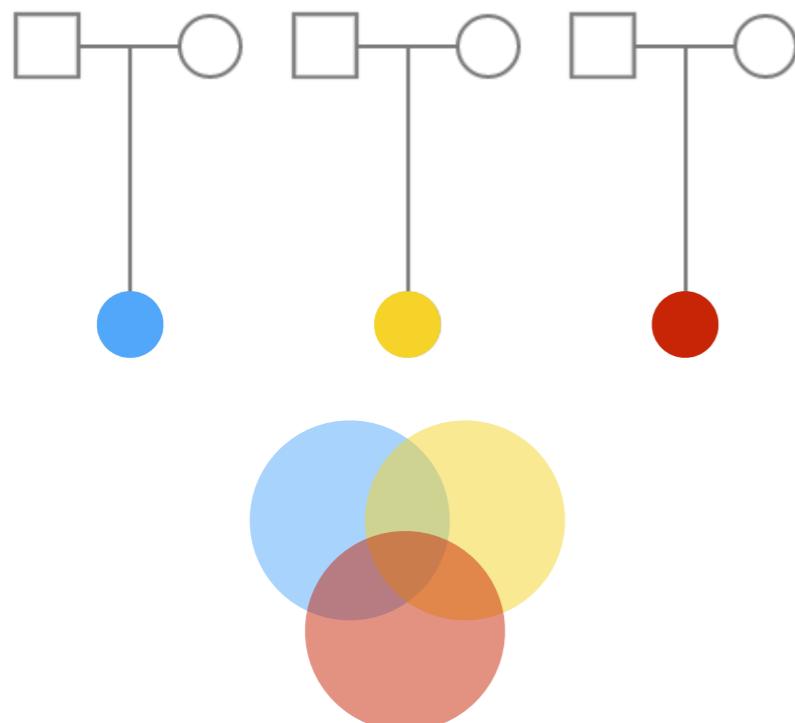
Strategies to identify disease-causing rare variants



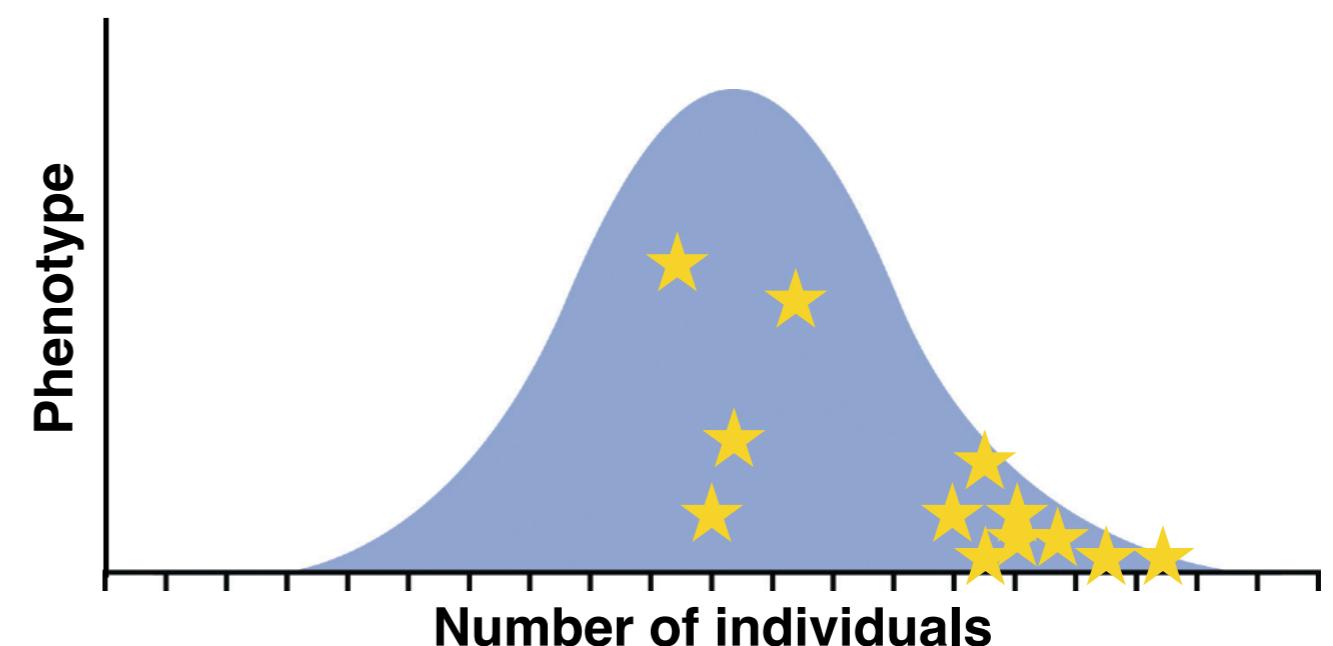
Rare variation from families



Shared variants from affected individuals in large families

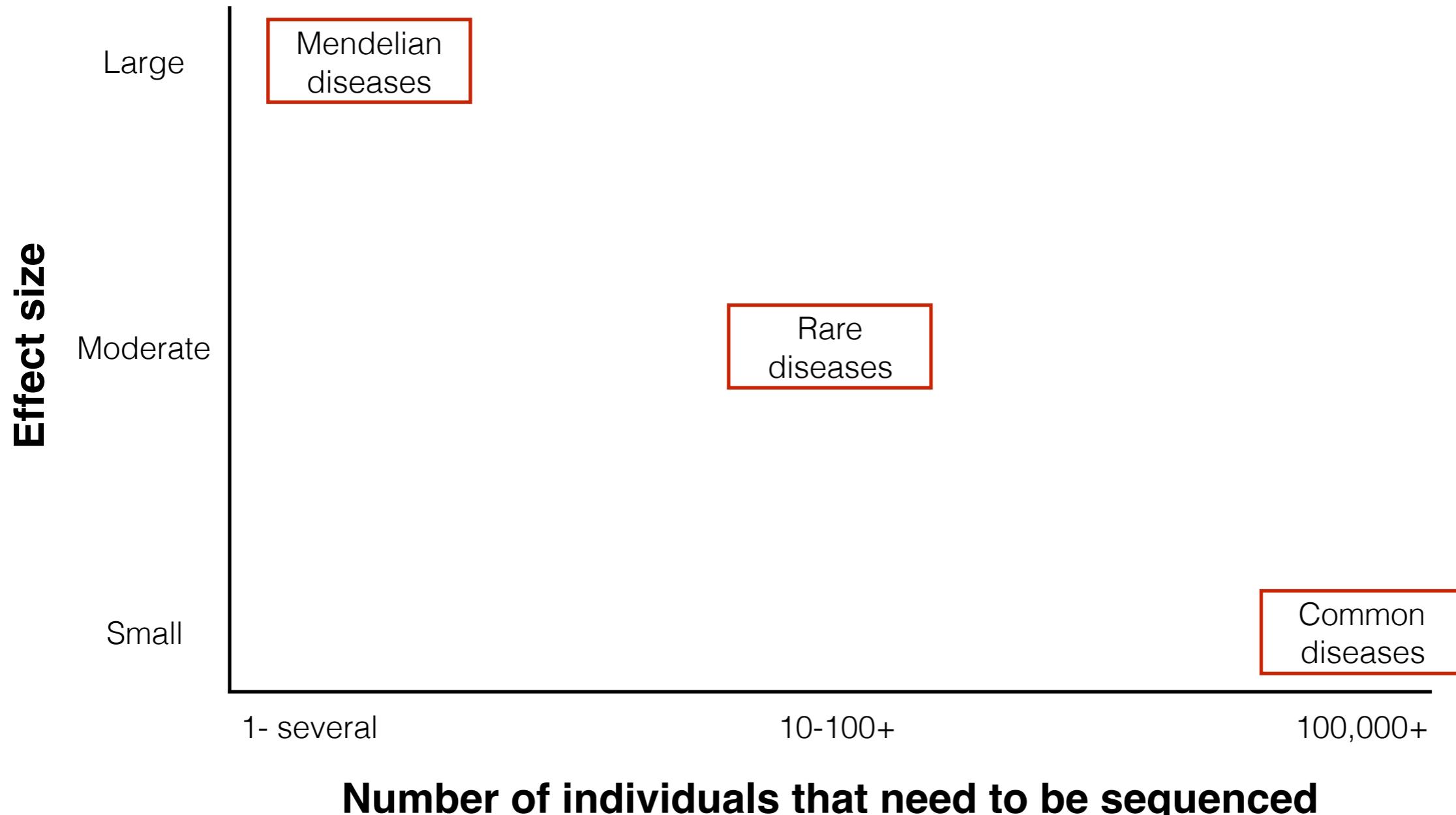


Shared variation from trios



Shared variants from many people

How can sequencing help us to identify these variants?



Why can't we read the genome?



We don't know all the variants.

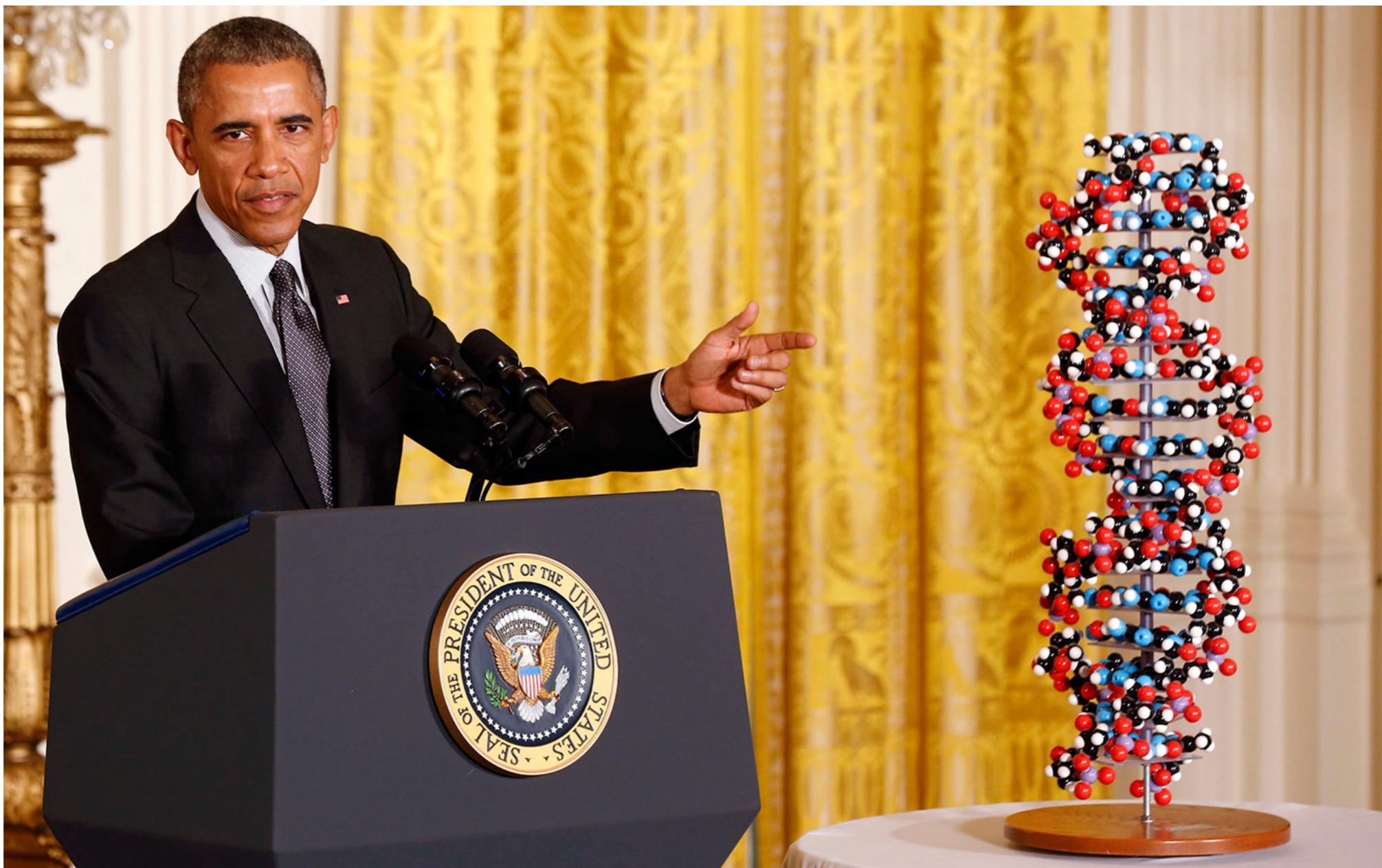
We don't know which ones affect phenotype.

Single genes don't cause most disease or control most traits

The human genome is big.

Phenotypes are highly variable.

What is precision medicine?



We are living in the human genetics renaissance



Under \$1000 genome
Rare disease sequencing for Mendelian disorders
Family genetics
Fetal testing from sequence
Disease outbreaks and diagnosis
Drug response prediction
Cancer genome sequencing