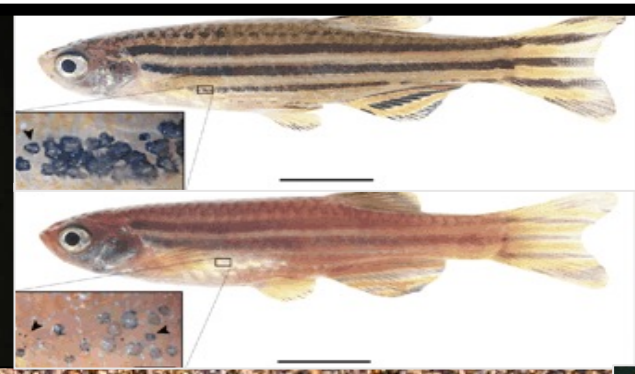
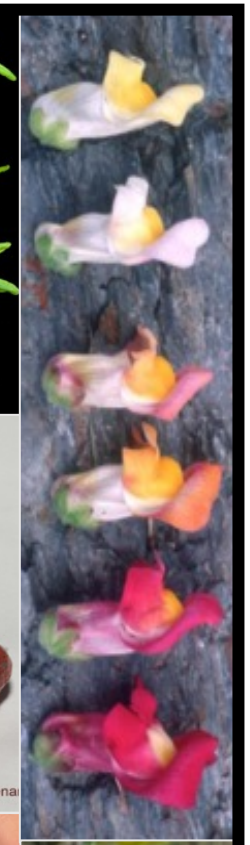
Genomes and their variation

Biol 5221

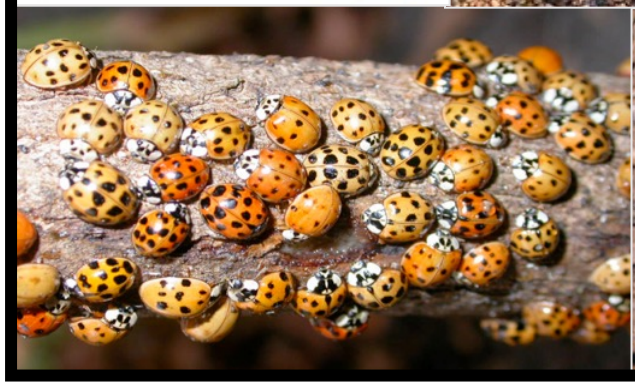
January 18, 2024

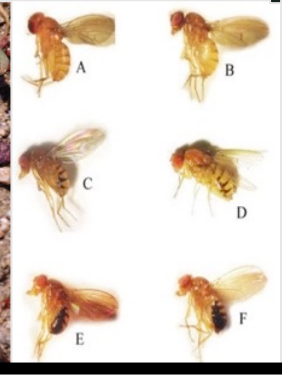
Hancock

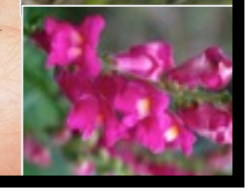






© 2011 Michael F. Bena



A
B
C
D
E
F

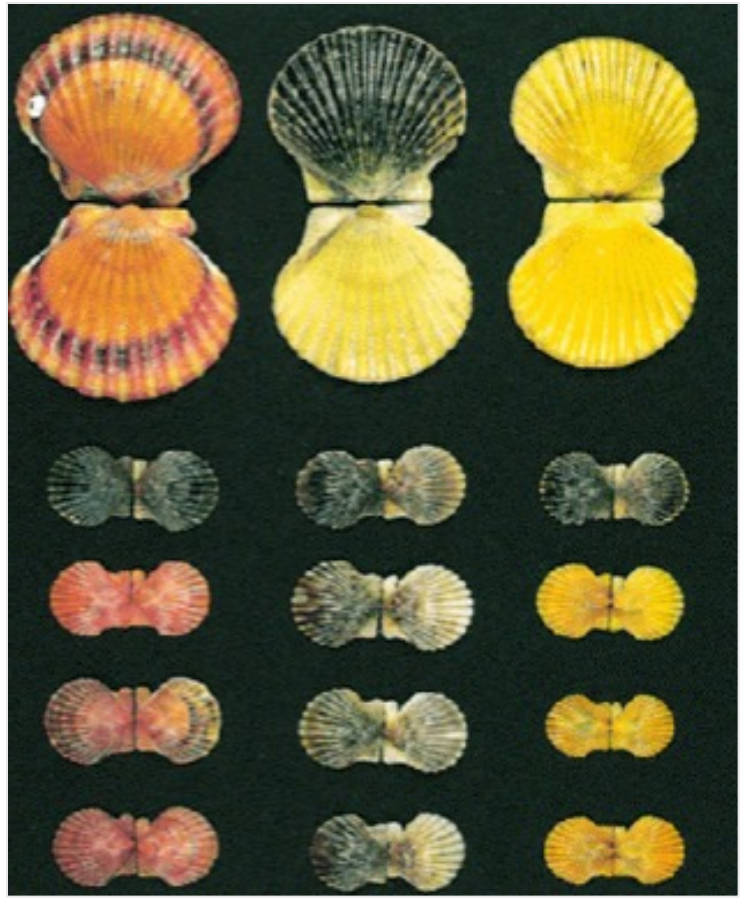


What is population genetics?

- Population genetics is the study of genetic variation within and between populations
- Studying genetic variation in present-day populations allows us to learn about the history of populations, genetic variants and traits and to predict the future
- We will often make simplifying assumptions, but simple models turn out to be powerful to make inferences, and future studies can build on these simple models

Some questions

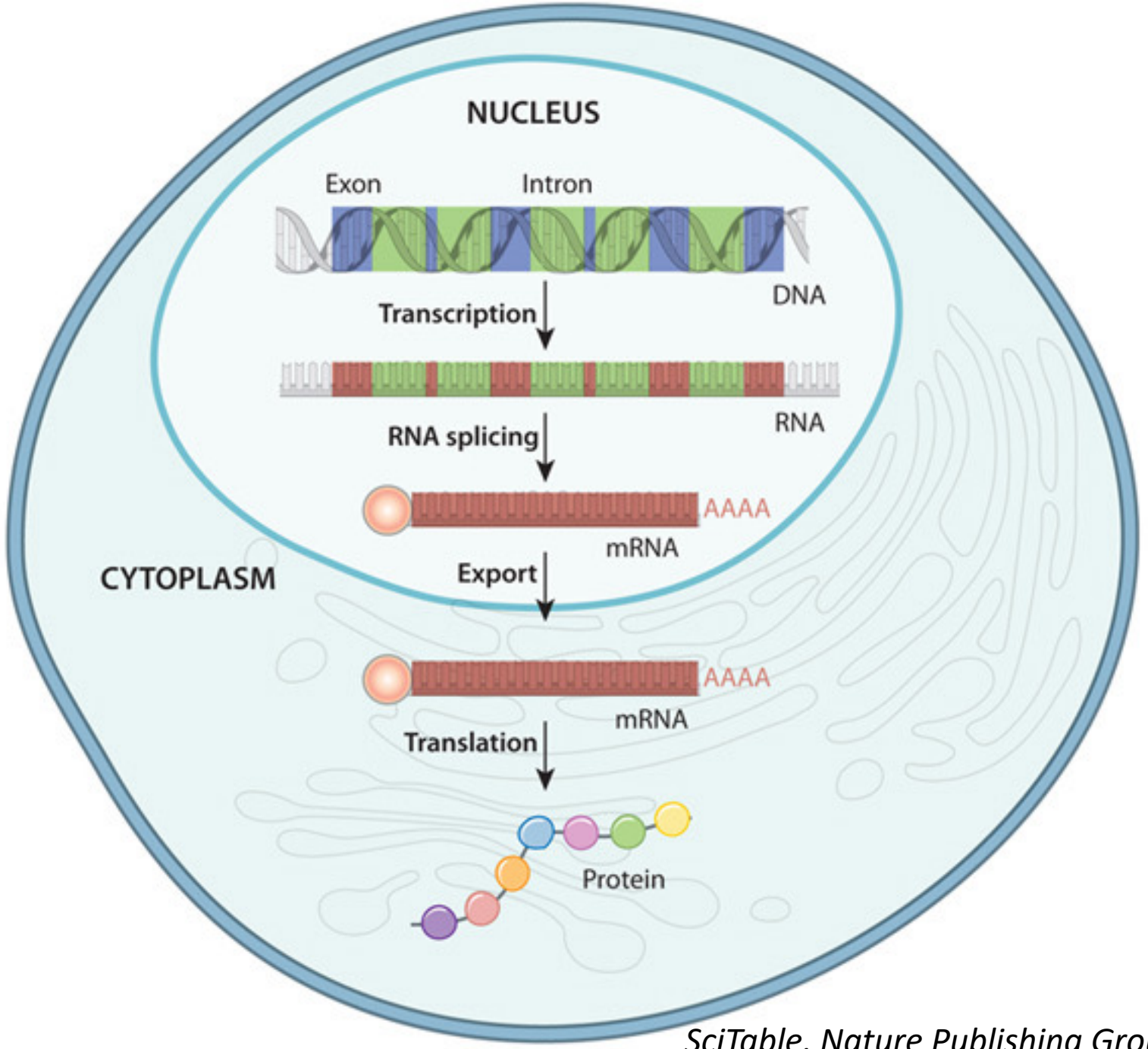


- *How can we assay genetic variation?*
- *What can we learn from genetic variation?*
- *What scientific and practical questions can we address using population genetic variation?*

Levels of regulation from DNA to protein

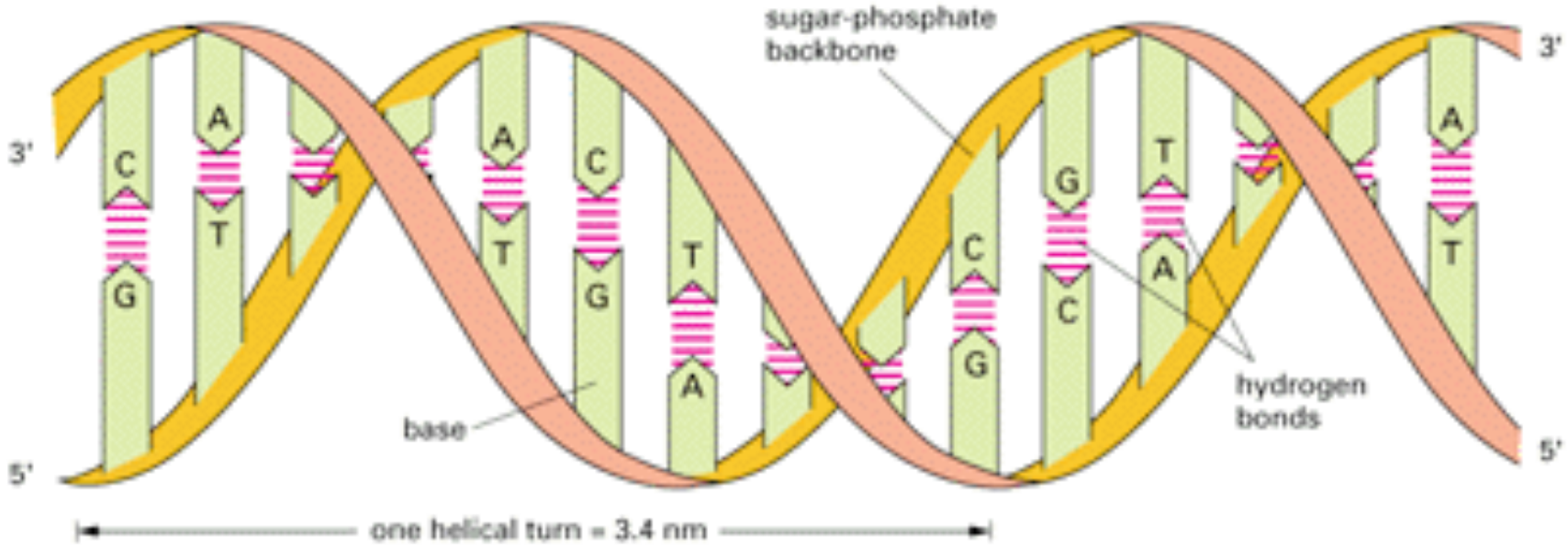
Not all DNA changes are equivalent!

Context matters!



SciTable, Nature Publishing Group

We will study variation at the DNA level

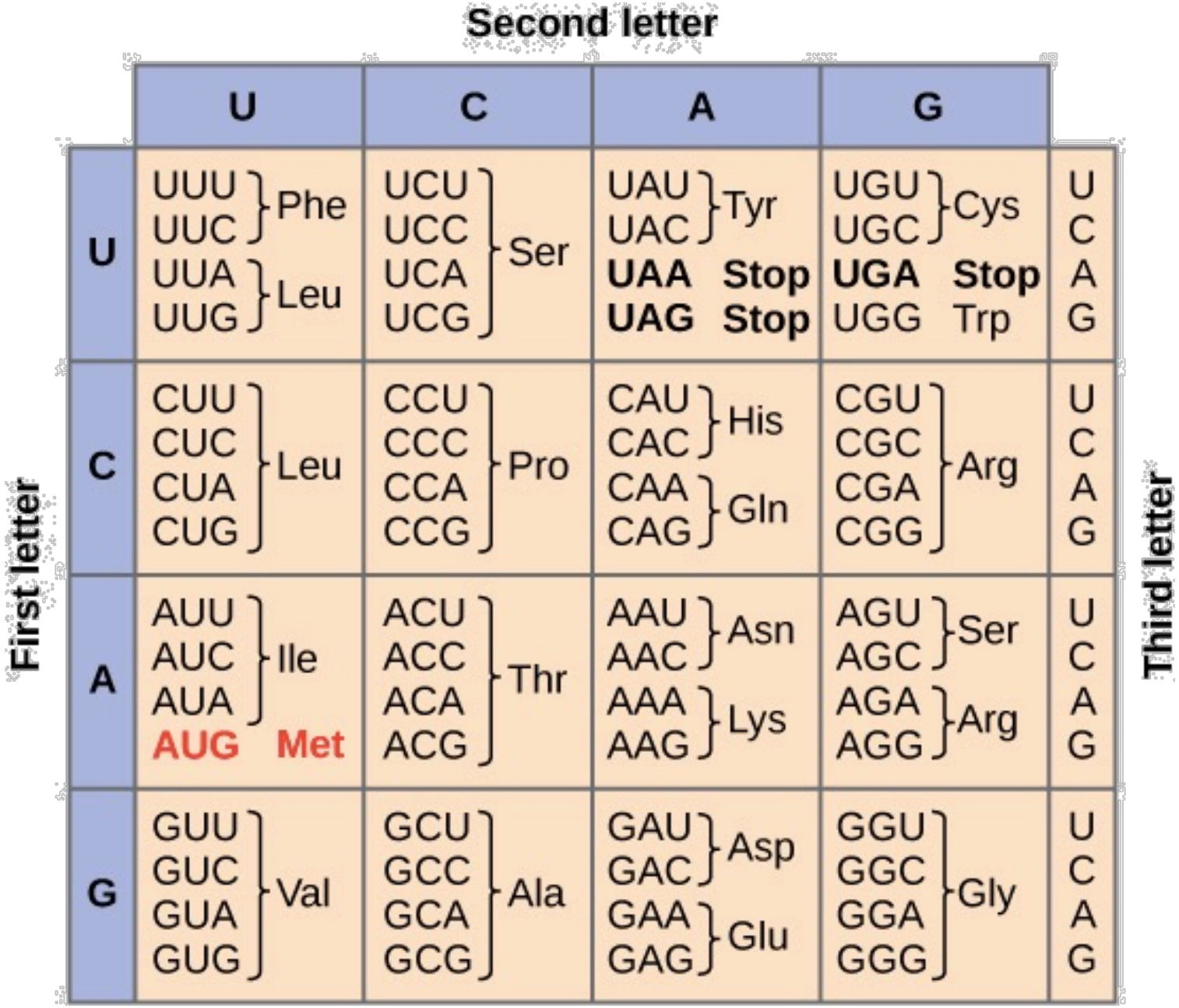


This variation may affect coding sequencing, regulatory sequences or it may have no effect on traits

The standard genetic code

Synonymous mutations are mutations that occur in a protein coding region that do not change the amino acid

Non-synonymous mutations are mutations that occur in a protein coding region that do change the amino acid



Source: OpenStax Biology

Types of genetic variation we study reflect available technology

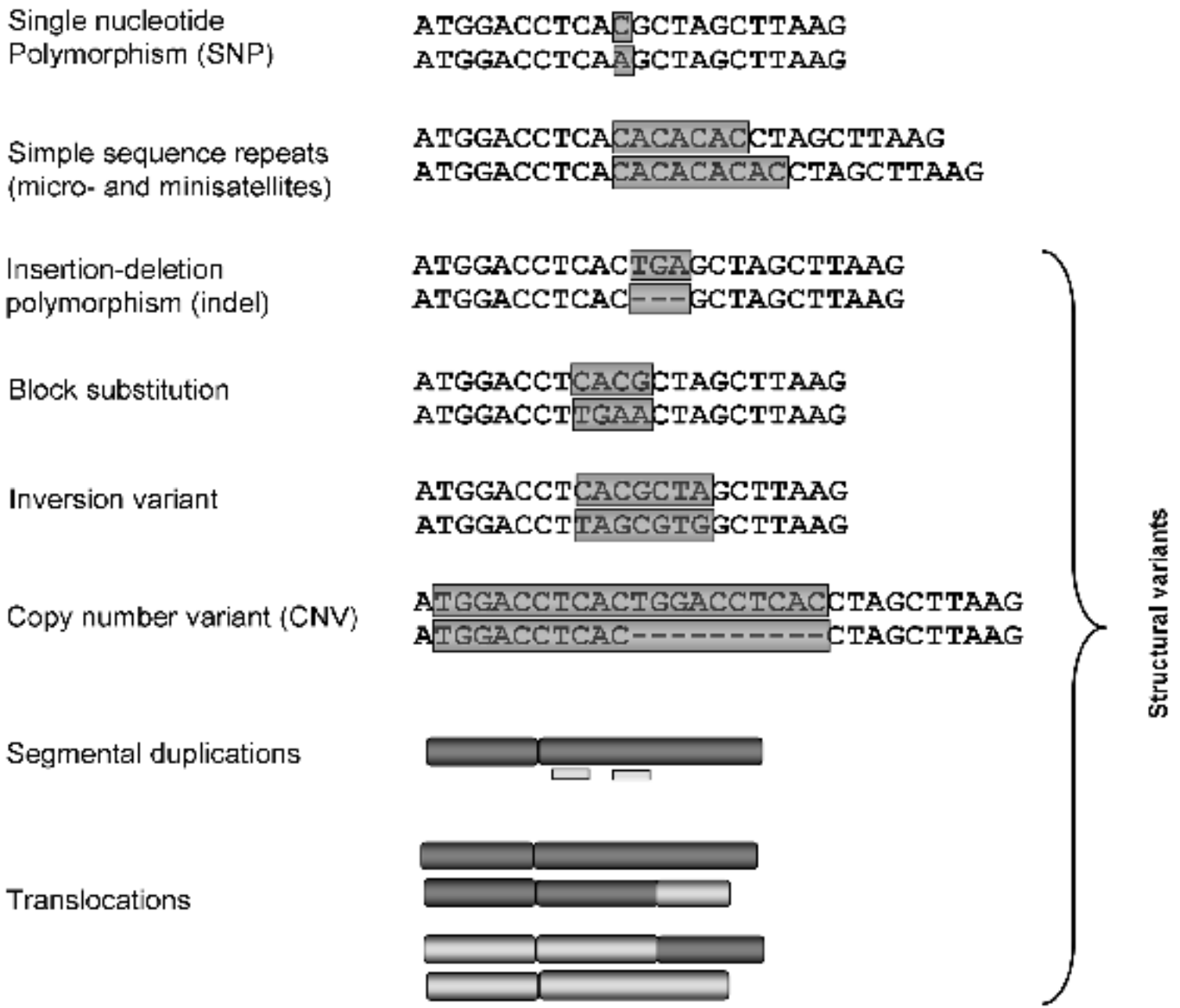
- Protein polymorphisms (using gel electrophoresis)
- RFLPs, microsatellites
- Single nucleotide variants (SNPs)
- Structural variants large and small
  - Indels
  - Large deletions
  - Tandem duplications
  - Inversions

Was the only method available before the DNA sequencing revolution

Most studies over the past 20 years have used these

SMRT-cell “long-read” data enables high quality genotyping of these

Types of variants



Major landmarks in population genetics and genomics

- Mendel's work showing how inheritance at the trait level could work through individual units, later termed "genes"
- Discovery of the structure of DNA
- Central Dogma: DNA-> RNA, RNA -> protein
- Assaying variation within a species using protein electrophoretic markers (Lewontin and Hubby, 1966)
- Assaying DNA sequence variation within a species (Kreitman 1983)
- Studies of variation using microsatellites
- Studies of DNA sequence of candidate genes and 'neutral' controls across genomes
- SNP-chips (genome-wide or partial genome-wide)
- Whole genome short-read sequencing
- Whole genome-long read sequencing (single-molecule real-time (SMRT) sequencing)

Timeline of sequencing technology

1977 2005 2010 present

First generation

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

Second generation
(next generation sequencing)

454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

Third generation

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

Tens of kb fragments, on average

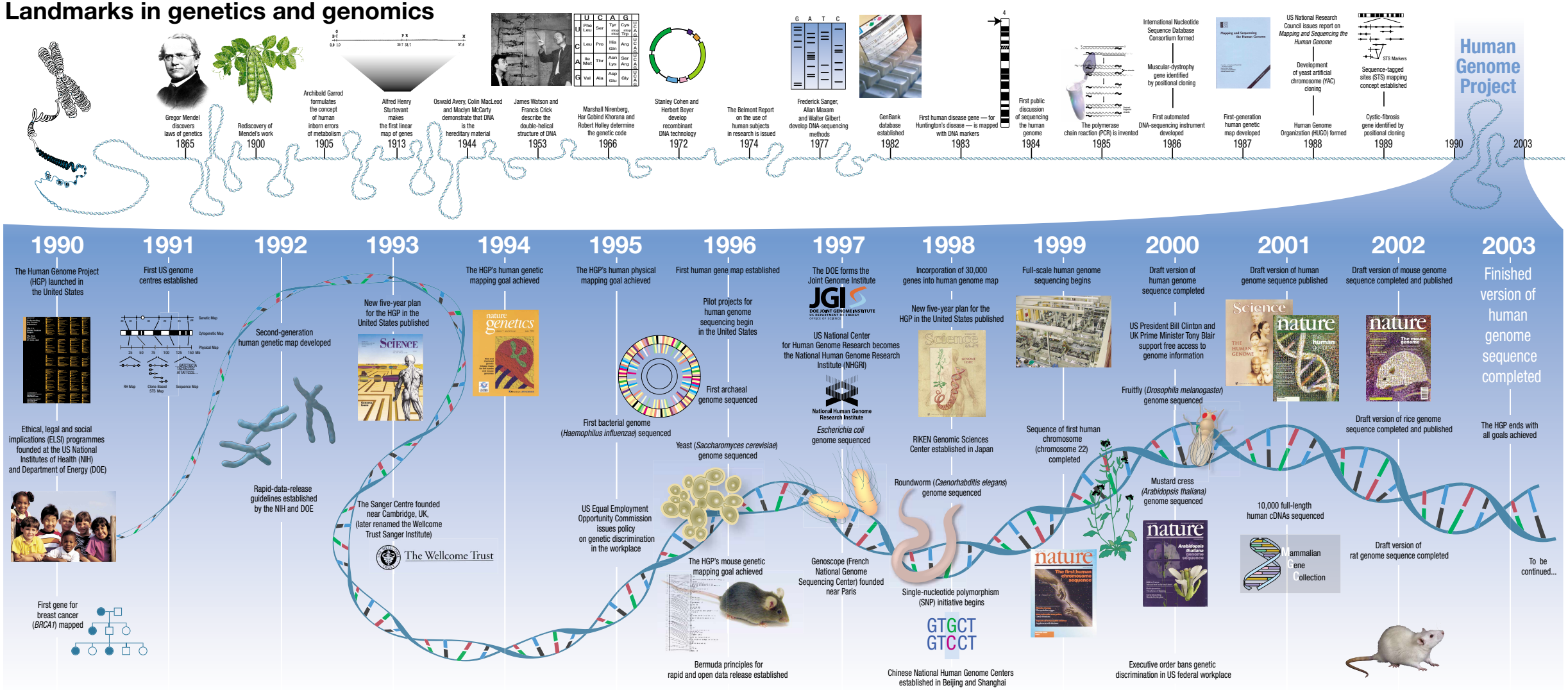
Short-read sequencing

Long-read sequencing

Source: https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

The Human Genome Project

Landmarks in genetics and genomics



1990
The Human Genome Project (HGP) launched in the United States

Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)

First gene for breast cancer (*BRCA1*) mapped

1991
First US genome centres established

1992
Second-generation human genetic map developed

Rapid-data-release guidelines established by the NIH and DOE

1993
New five-year plan for the HGP in the United States published

The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)

The Wellcome Trust

1994
The HGP's human genetic mapping goal achieved

1995
The HGP's human physical mapping goal achieved

First bacterial genome (*Haemophilus influenzae*) sequenced

US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996
First human gene map established

Pilot projects for human genome sequencing begin in the United States

First archaeal genome sequenced

Yeast (*Saccharomyces cerevisiae*) genome sequenced

The HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997
The DOE forms the Joint Genome Institute

US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)

Escherichia coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded near Paris

1998
Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the United States published

RIKEN Genomic Sciences Center established in Japan

Roundworm (*Caenorhabditis elegans*) genome sequenced

Single-nucleotide polymorphism (SNP) initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers established in Beijing and Shanghai

1999
Full-scale human genome sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000
Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced

Mustard cress (*Arabidopsis thaliana*) genome sequenced

Executive order bans genetic discrimination in US federal workplace

2001
Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002
Draft version of mouse genome sequence completed and published

Draft version of rice genome sequence completed and published

Draft version of rat genome sequence completed

2003
Finished version of human genome sequence completed

The HGP ends with all goals achieved

To be continued...

PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; *SCIENCE* COVERS COURTESY AAAS

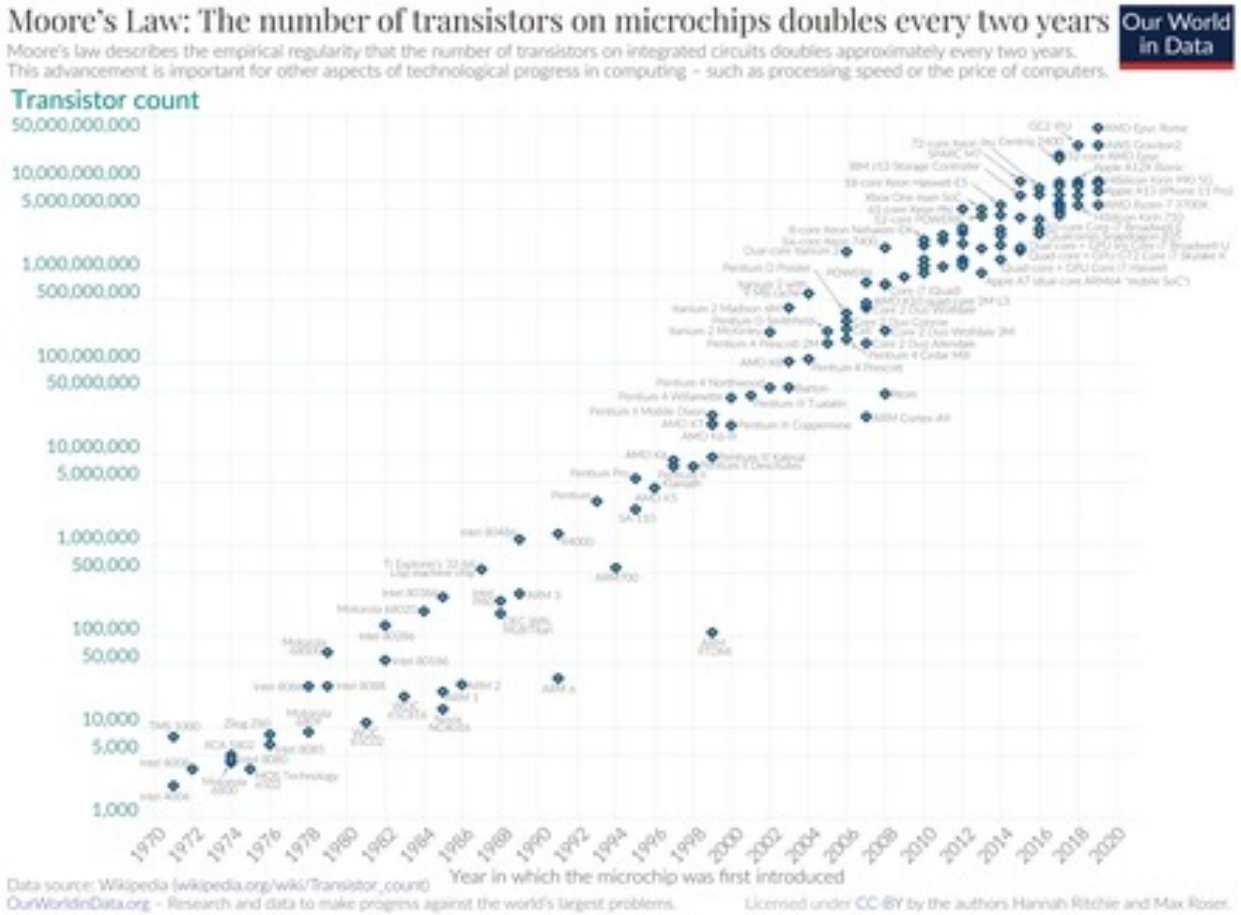
Source and more info: https://www.genome.gov/human-genome-project
https://geneticsunzipped.com/blog/2020/10/22/s322-the-past-present-and-future-of-the-human-genome-project

A summary of the structure of the human genome

- *Genome size: 3.1 Gb (haploid size).*
- *Number of chromosomes: 23 pairs*
- *Number of coding genes: ~20, 000*
- *Exons per gene: 8 (median)*
- *Number of genes per megabase: 6.5 (mean)*
- *Total in protein-coding exons: 1% of genome*
- *Total in genes (introns+exons): 40% of genome*
- *Active chromatin (per cell type): 1% of genome*
- *Active chromatin (all cell types): 13% of genome*

Source: Pritchard, J. An owner's guide to the human genome, Ch1.1, 2023

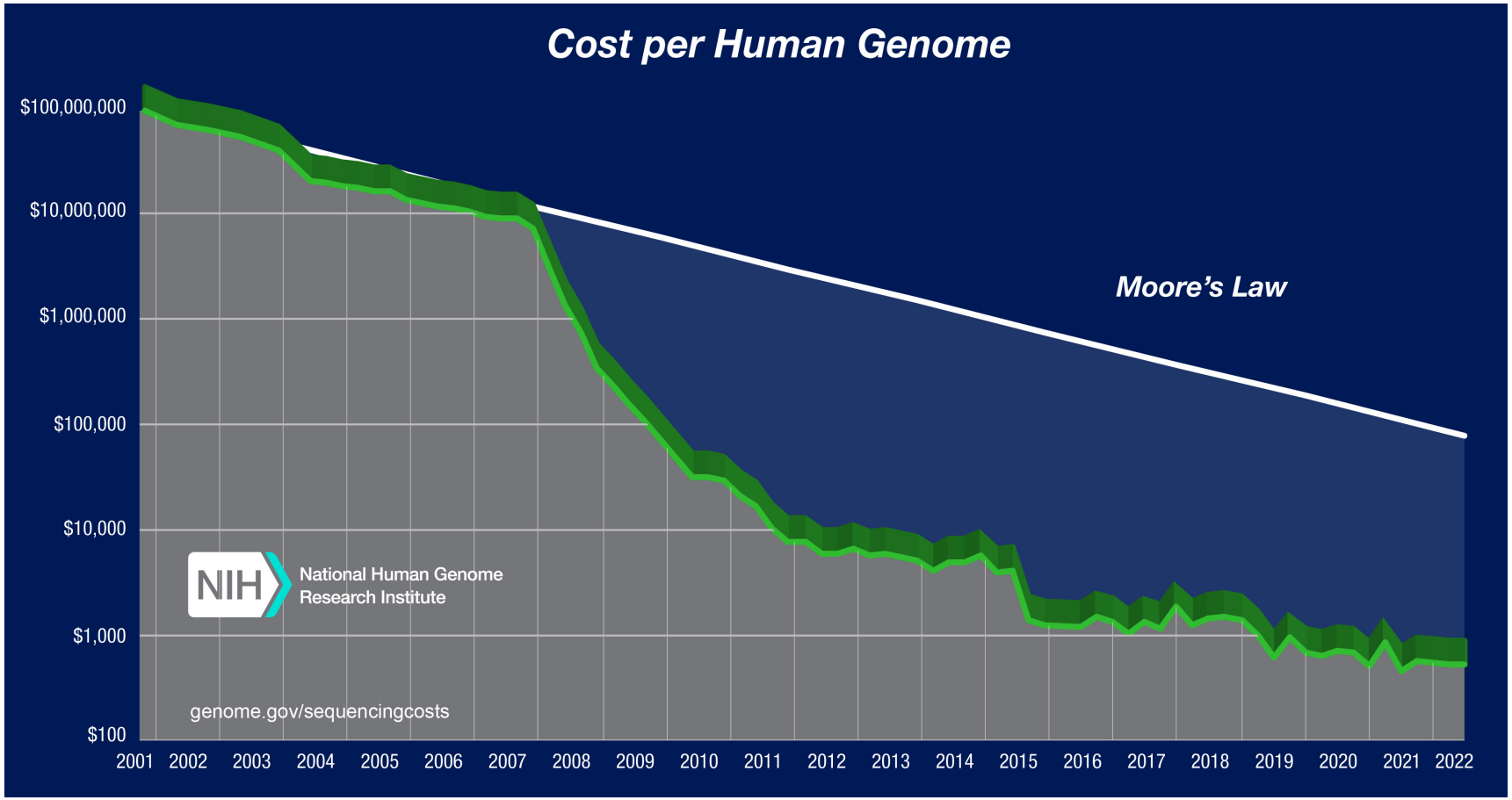
Moore’s law posits that manufacturing costs for semi-conductors should fall exponentially over time



Gordon Moore, projected that the ideal number of transistors per square inch on a microchip would double each year while the manufacturing cost per component would halve.

Moore’s law has since been widely applied to diverse technologies

Since 2007, with “next generation sequencing” the cost reduction per human genome has outpaced Moore’s law



Scale of sequencing in human populations

- Human Genome Project (1 reference sequence built from 4 individuals)
- Populations genomics aimed at assaying diversity across worldwide populations
  - HapMap Project (90 CEPH, 90 CHN, 90 YRI)
  - Human Genome Diversity Panel
  - 1000 Genomes Project assays diversity across worldwide populations
- Population genomics aimed at trait mapping
  - Wellcome Trust (WTCCC) 14K cases, 3K shared controls
  - UK Biobank 500K individuals
  - AllofUs Research Program -> goal: health + genomic data for 1M+ US citizens

Comparative genomics and comparative population genomics

The Human Genome Project also provided funding for sequencing other species. ***Why?***

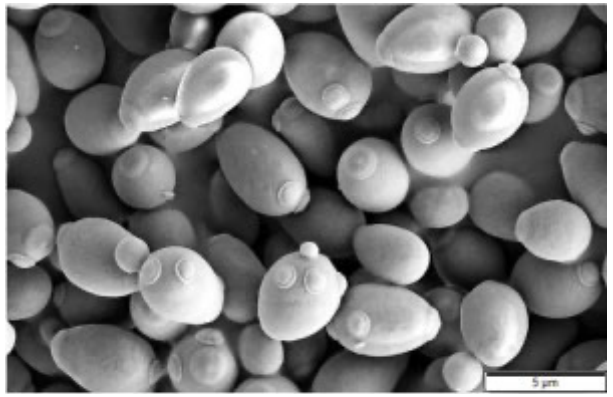
- Comparative genomics not only helps us to understand other genomes, but also our own
- Identifying non-coding loci that are deeply conserved is useful to annotate likely functional regions
- Regions that evolve relatively rapidly over phylogenetic time scales may be involved in adaptation
- Experiments in model organisms, which would not be possible in humans, can connect genetic variation to functional variation

https://www.genome.gov/about-genomics/fact-sheets/Comparative-Genomics-Fact-Sheet

Genomes projects in model organisms examine variation across the species distribution








Sources: NSF.gov/about/history/nifty50, WikiMedia Commons and Addgene.com (Anna Hendriks, Mogana Das Murtey and Patchamuthu Ramasamy)

10,000 plant genomes aims to assay diversity across species

- Plants make up the majority of biomass on Earth
- Plants are directly exposed to their environments, making adaptation especially important (***plants can't hide from the elements!***)
- Plants have high variation in genome size, chromosome number and diversity

The Zoonomia Project

Aims to identify the genetic basis for shared and distinct traits across animals

Publications:

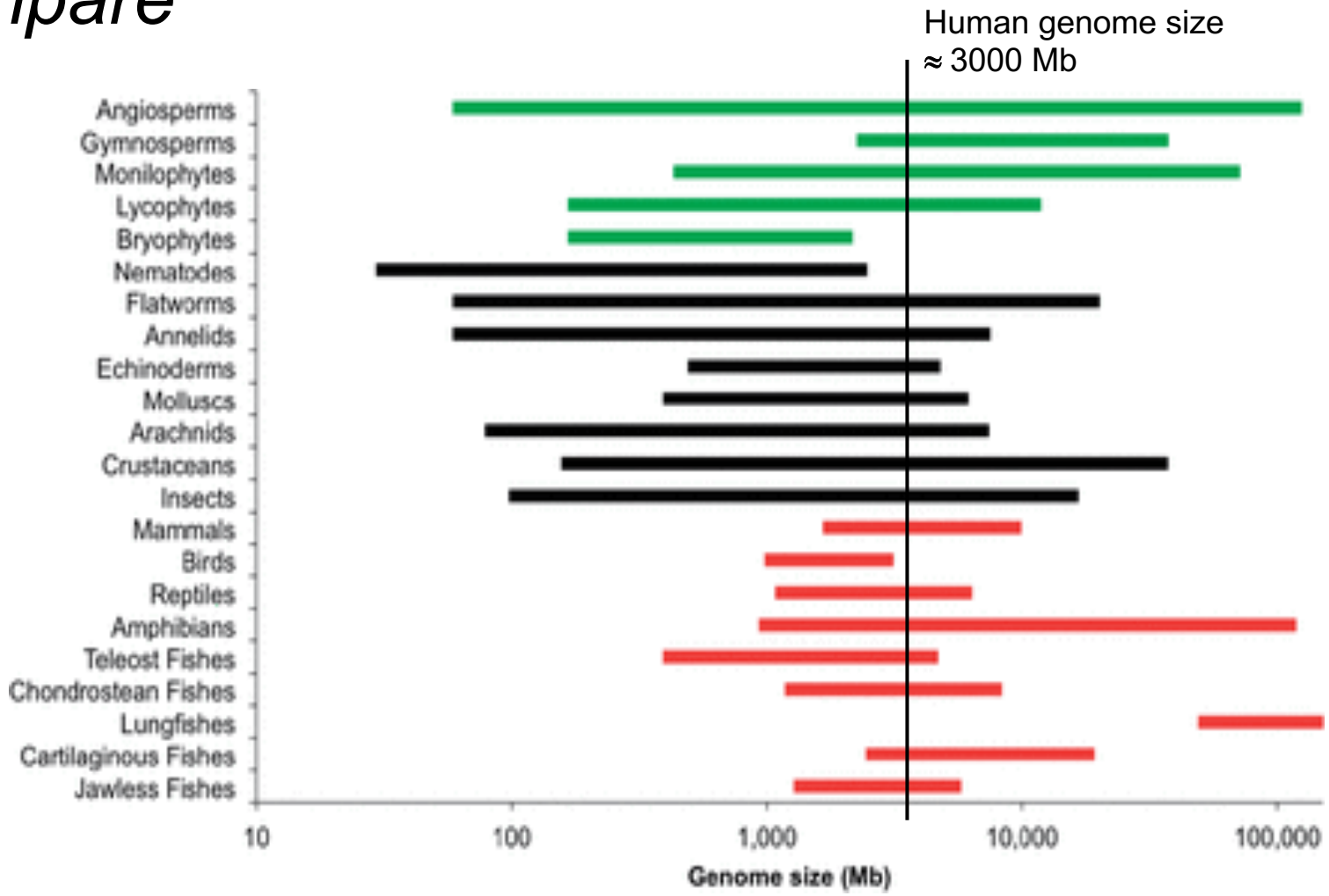
Christmas et al. Evolutionary constraint and innovation across hundreds of placental mammals
https://www.science.org/doi/10.1126/science.abn3943
Sullivan et al. Leveraging base pair mammalian constraint to understand genetic variation and human disease
https://www.science.org/doi/10.1126/science.abn2937
Andrews et al. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites
https://www.science.org/doi/10.1126/science.abn7930
Foley at al. A genomic timescale for placental mammal evolution
https://www.science.org/doi/10.1126/science.abl8189Kaplow et al. Relating enhancer genetic variation across mammals to complex phenotypes using machine learning
https://www.science.org/doi/10.1126/science.abm7993
Keough et al. Three-dimensional genome re-wiring in loci with human accelerated regions
https://www.science.org/doi/10.1126/science.abm1696
Kirilenko et al. Integrating gene annotation with orthology inference at scale
https://www.science.org/doi/10.1126/science.abn5887
Osmanski et al. Insights into mammalian TE diversity via the curation of 248 mammalian genome assemblies
https://www.science.org/doi/10.1126/science.abn1430
Wilder et al. The contribution of historical processes to contemporary extinction risk in placental mammals
https://www.science.org/doi/10.1126/science.abn5856
Xue et al. The functional and evolutionary impacts of human-specific deletions in conserved elements
https://www.science.org/doi/10.1126/science.abn2253

How does the human genome compare to others?

How does the size of the human genome compare with other species?

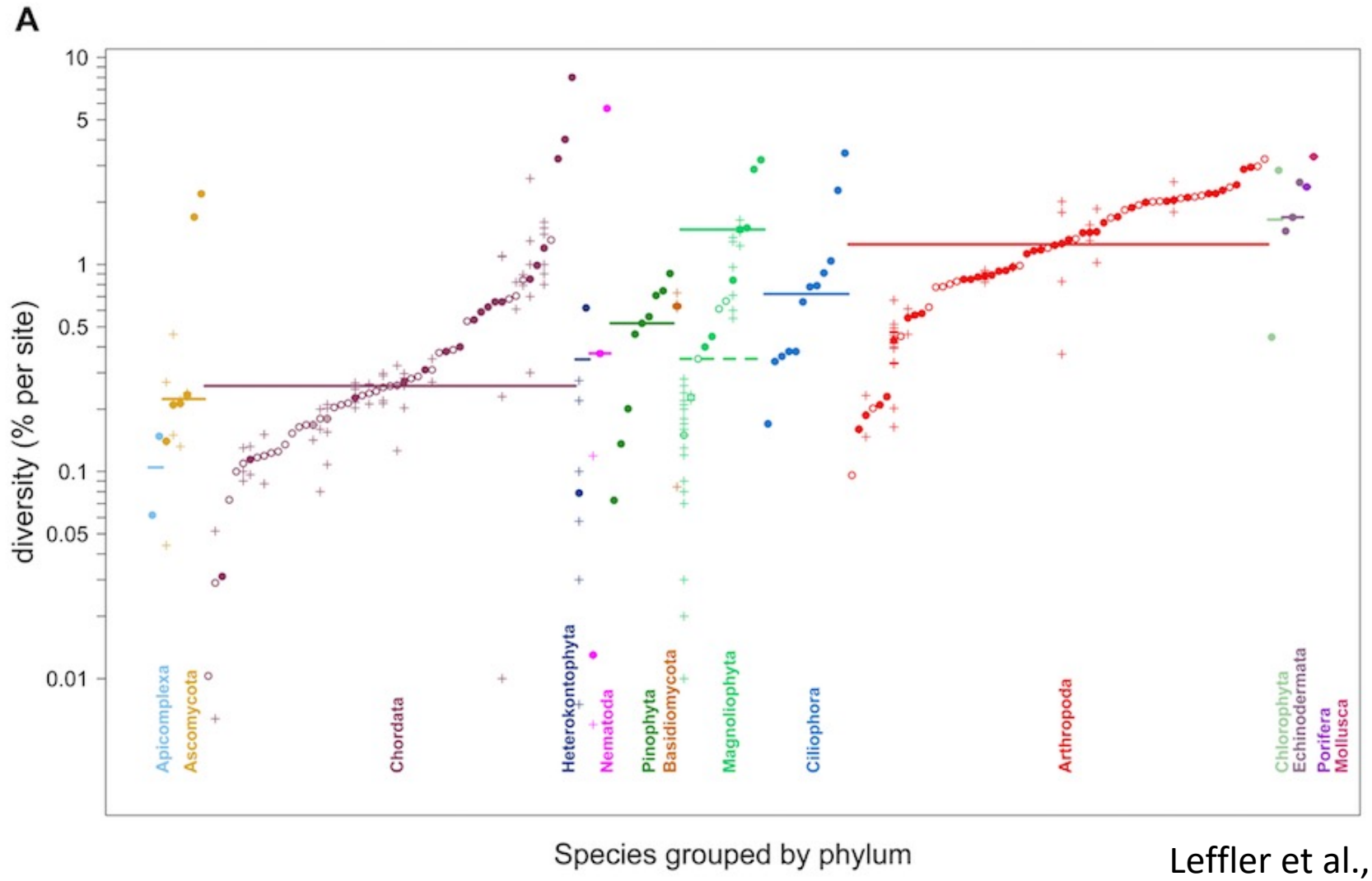
Four orders of magnitude of variation in genome size

Humans are not outliers for genome size!



Michael, Briefings in Functional Genomics, 2014

Diversity varies widely across species

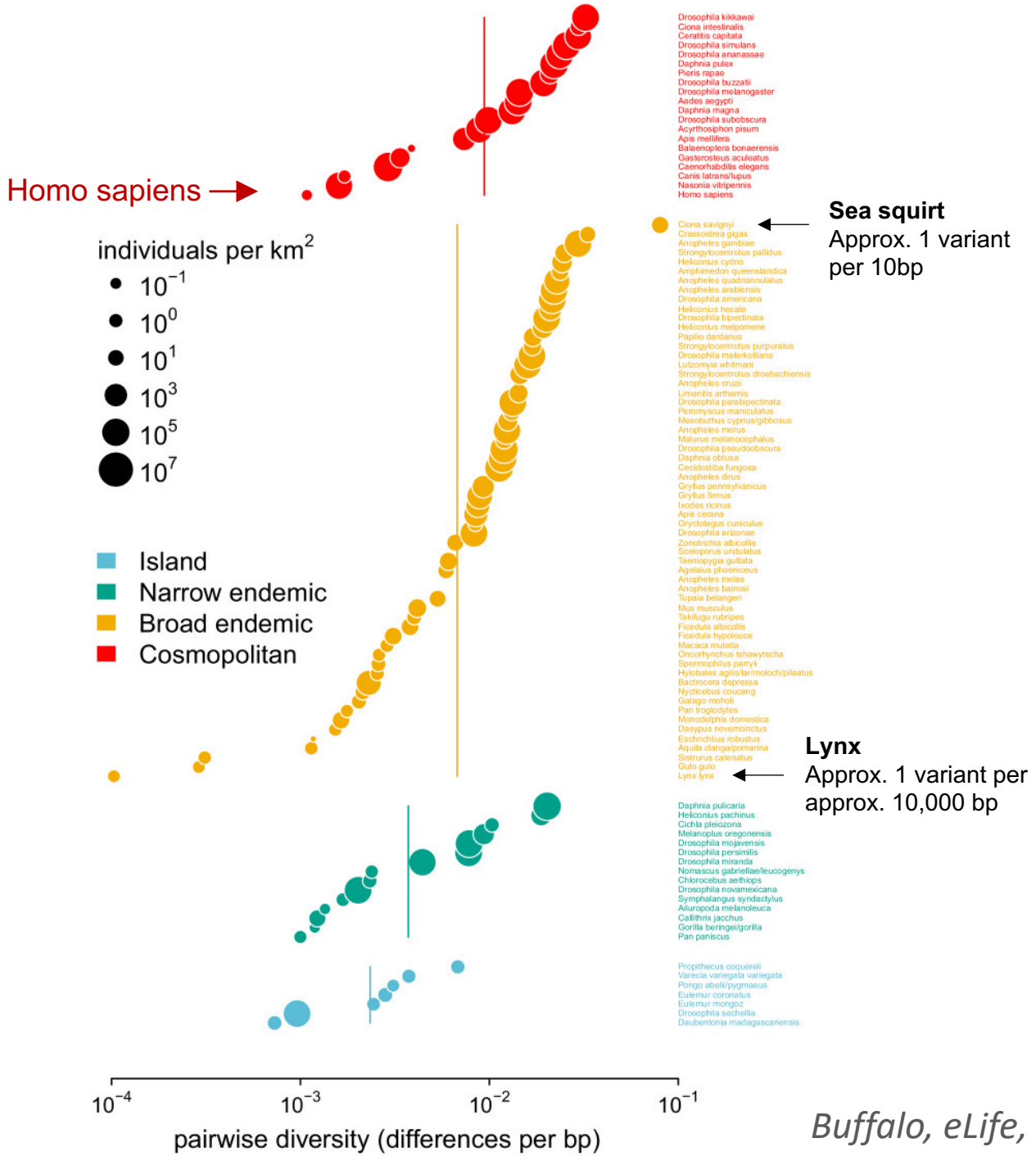


Leffler et al., Plos Biology, 2012

Can census size explain variation in diversity?

Buffalo, eLife, 2021

Differences in census population size explain some of the variation in diversity



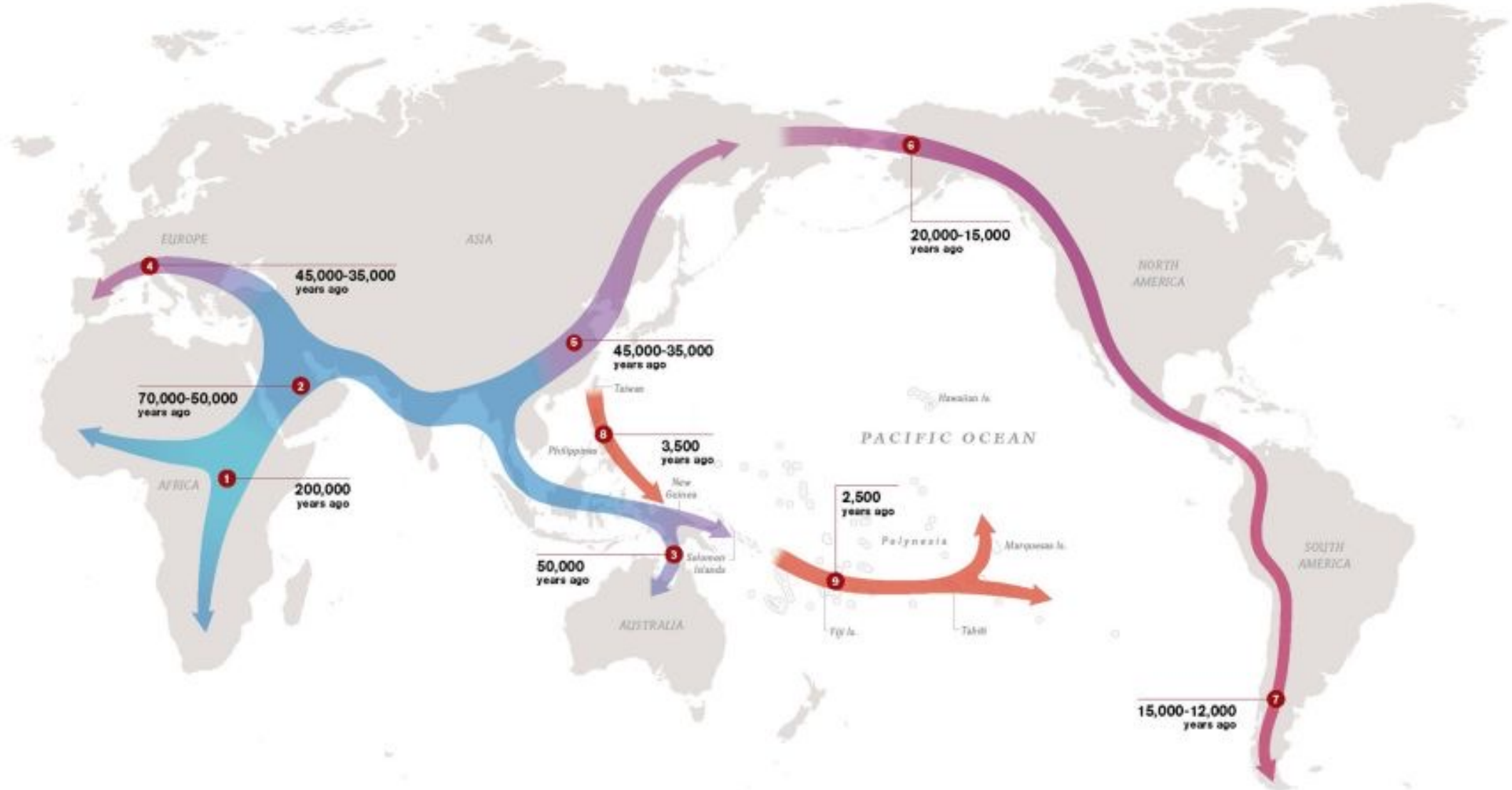
Buffalo, eLife, 2021

Some applications of population and quantitative genetics/genomics

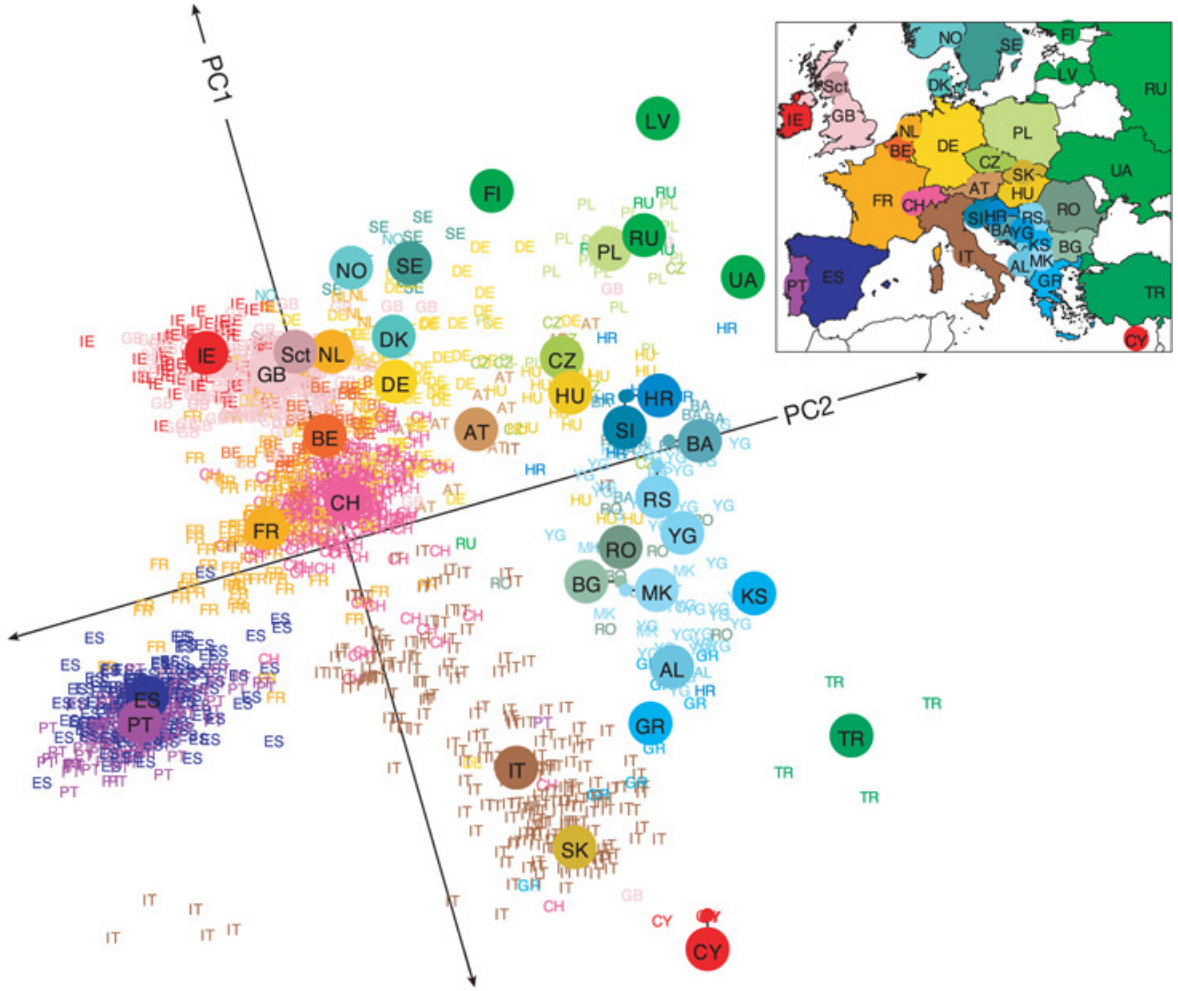
Applications of population genetics

- Estimate diversity in populations
- Reconstruct evolutionary history and predict future evolvability
- Understand the evolutionary mechanisms that act on variation
- Understand the processes of local adaptation and speciation
- Trait mapping and personal genomics
- DNA fingerprinting – tracking individuals/forensics
- Conservation

Reconstructing historical relationships



Genetics recapitulates geography in population samples from Europe



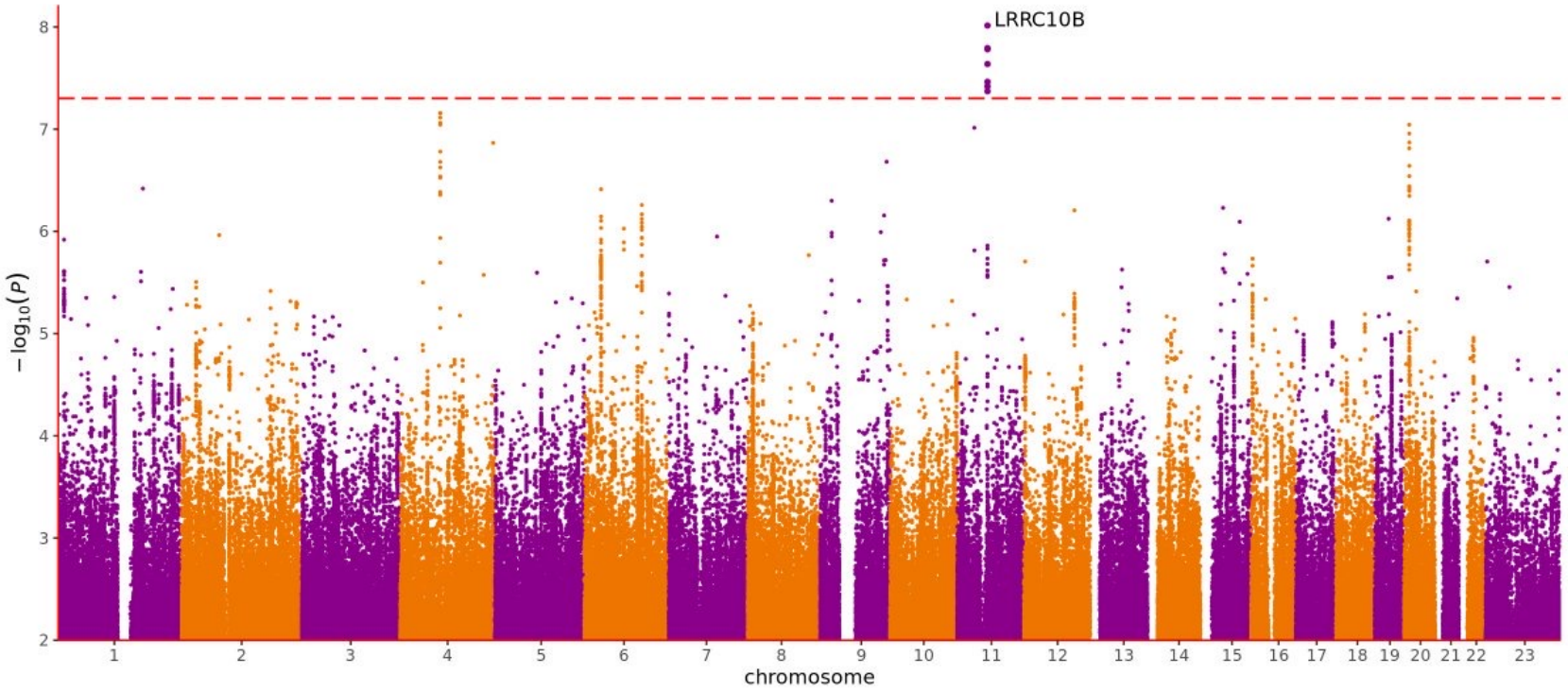
First two principal components derived from a matrix of SNP genotypes across a sample of Europeans

The resulting pattern is similar to the map of locations of origin

*Novembre et al. Nature **000**, 1-4 (2008) doi:10.1038/nature07331*

Trait mapping

Loci associated with diastolic blood pressure variation in the UK Biobank population sample



Precision medicine

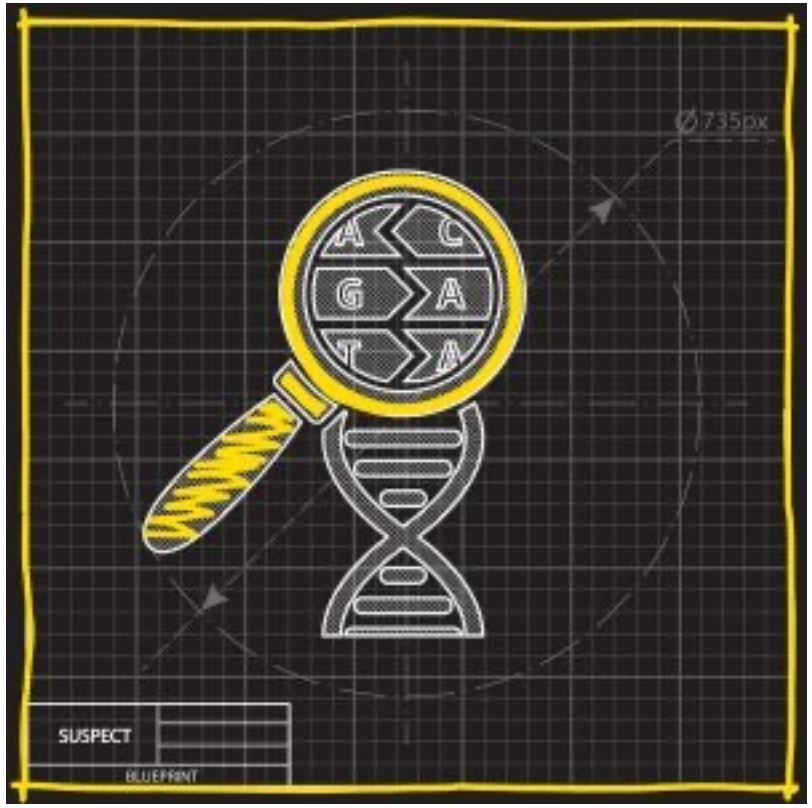
Goals:

- To understand genetic and environmental risk factors for human disease
- To learn which treatments work best for people of different backgrounds
- To precision medicine aims to improve treatment to take into account a person's specific genetic and environmental differences

Precision agriculture and resilience

Understanding how plant populations have adapted to extreme climates can provide insights into how to improve resilience in crops

Forensics



Source: HudsonAlpha.org

Genomics is establishing more robust methods for DNA-based forensic analyses

For more see https://www.genome.gov/dna-day/15-ways/enhanced-forensics

Conservation

Understanding the amount and spatial distribution of variation within threatened species can inform conservation strategies

Timeline of population genetic/genomics

- Mendel's work showing how inheritance at the trait level could work; connected to "genes", which were abstract constructs
- Discovery of the structure of DNA: Watson, Crick and Wilkens received the Nobel prize from work that was inspired by a photo by Rosalind Franklin
- DNA-> RNA -> protein
- Assaying variation using protein electrophoretic markers
- **First study of DNA sequence variation within a species (Kreitman 1983)**
- Studies of variation using microsatellites
- Studies of DNA sequence of candidate genes and 'neutral' controls across genomes
- SNP-chips (genome-wide or partial genome-wide)
- Whole genome short-read sequencing
- Whole genome-long read sequencing (single-molecule real-time (SMRT) sequencing)

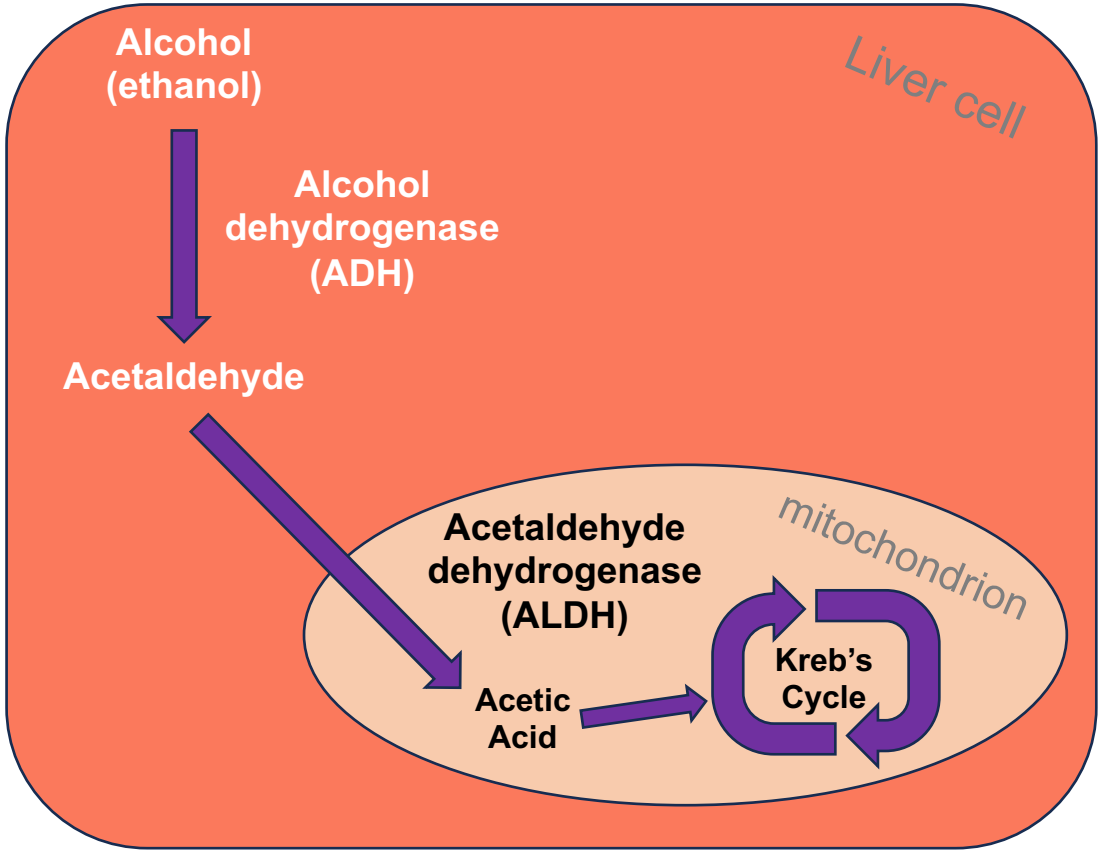
*Time flies like an arrow, a fruit fly likes a banana**

*attributed to Groucho Marx, but this is questionable

Alcohol is metabolized using two enzymes: ADH and ALDH

ADH breaks alcohol down into acetaldehyde, representing the first step of alcohol metabolism

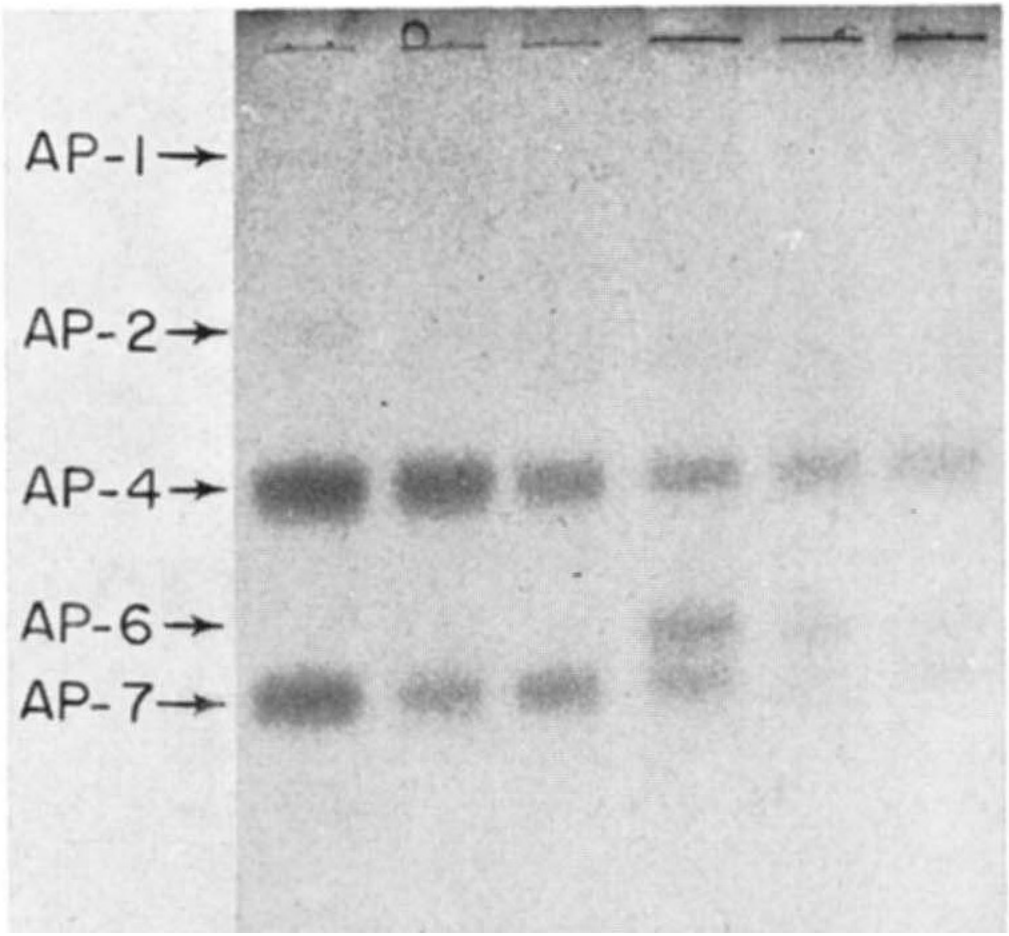
In a second step, ALDH breaks acetaldehyde into acetic acid



Two alleles segregate in *Drosophila melanogaster*: a *fast* (*Adh-f*) and a *slow* (*Adh-s*) metabolizing allele

The fast allele is encoded by a Thr -> Lys amino acid replacement

The standard way to assay variation was using protein electrophoresis gels



The rate at which a protein moves through the gel depends on its electrostatic charge.

Protein variation could be assayed across a set of individuals, as in this gel showing different alkaline phophatase proteins in *Drosophila pseudoobscura* samples.

Lewontin and Hubby, Genetics, 1966

In the 1980's Church and Gilbert were working out an approach to sequence DNA

Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 1991–1995, April 1984
Biochemistry

Genomic sequencing

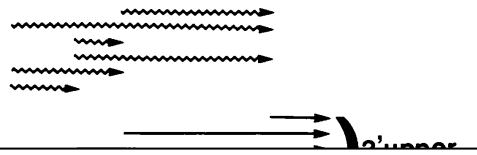
(DNA methylation/UV crosslinking/filter hybridization/immunoglobulin genes)

GEORGE M. CHURCH* AND WALTER GILBERT*†

*Biological Laboratories, Harvard University, Cambridge, MA 02138; and †Biogen, Inc., 14 Cambridge Center, Cambridge, MA 02142

Contributed by Walter Gilbert, December 19, 1983

ABSTRACT **Unique DNA sequences can be determined directly from mouse genomic DNA. A denaturing gel separates by size mixtures of unlabeled DNA fragments from complete restriction and partial chemical cleavages of the entire genome.**



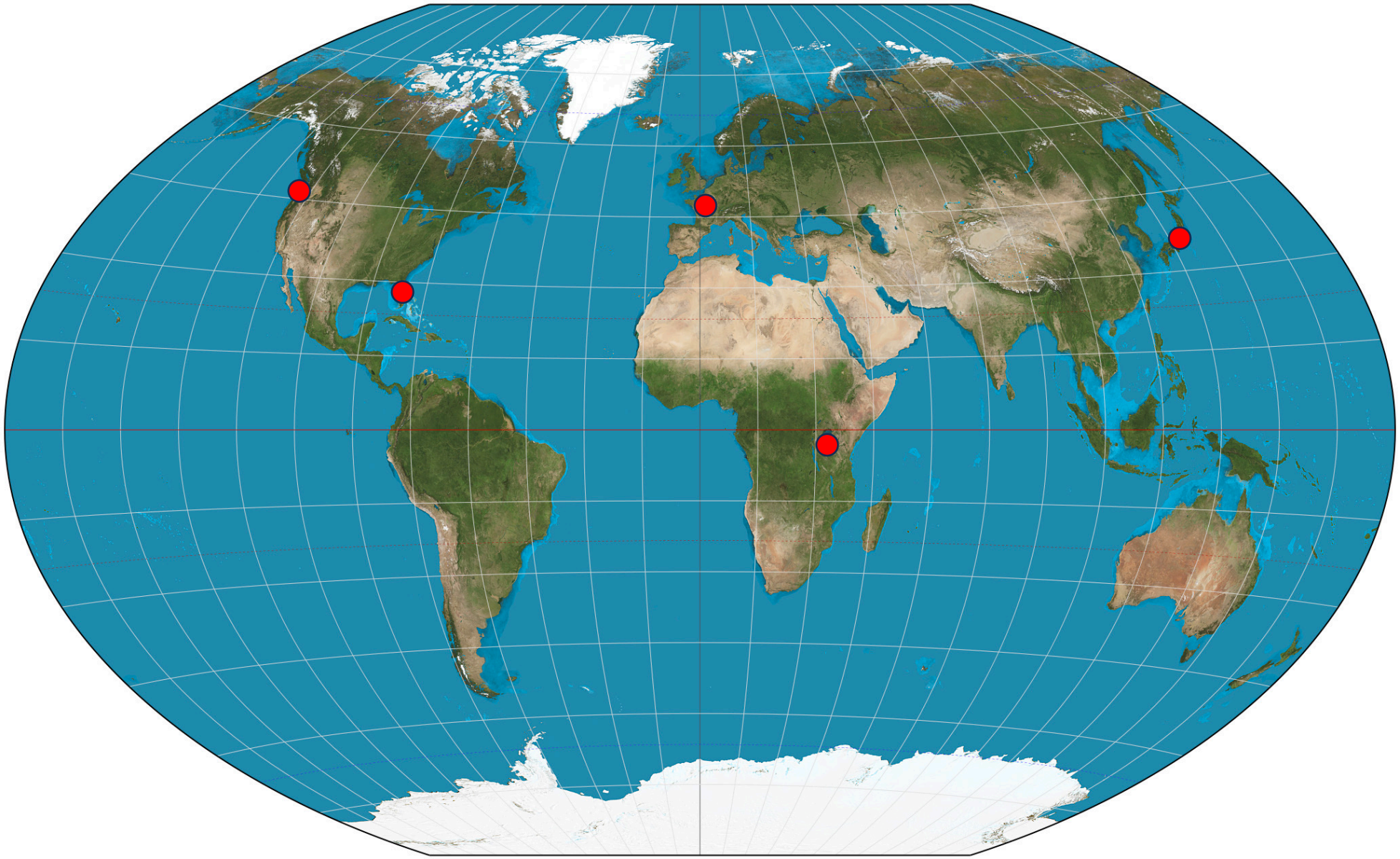
Marty Kreitman, a student in Dick Lewontin's lab thought it would be interesting to look at variation within a species at the DNA sequence level

Adh gene variation in *D. melanogaster*

Abstract

The sequencing of eleven cloned Drosophila melanogaster *alcohol dehydrogenase* (Adh) *genes from five natural populations has revealed a large number of previously hidden polymorphisms. Only one of the 43 polymorphisms results in an amino acid change, the one responsible for the two electrophoretic variants (fast, Adh-f, and slow, Adh-s) found in nearly all natural populations. The implication is that most amino acid changes in Adh would be selectively deleterious.*

The study used 11 *D. melanogaster* strains derived from 5 locations



Structure of the *Adh* gene in *Drosophila melanogaster*

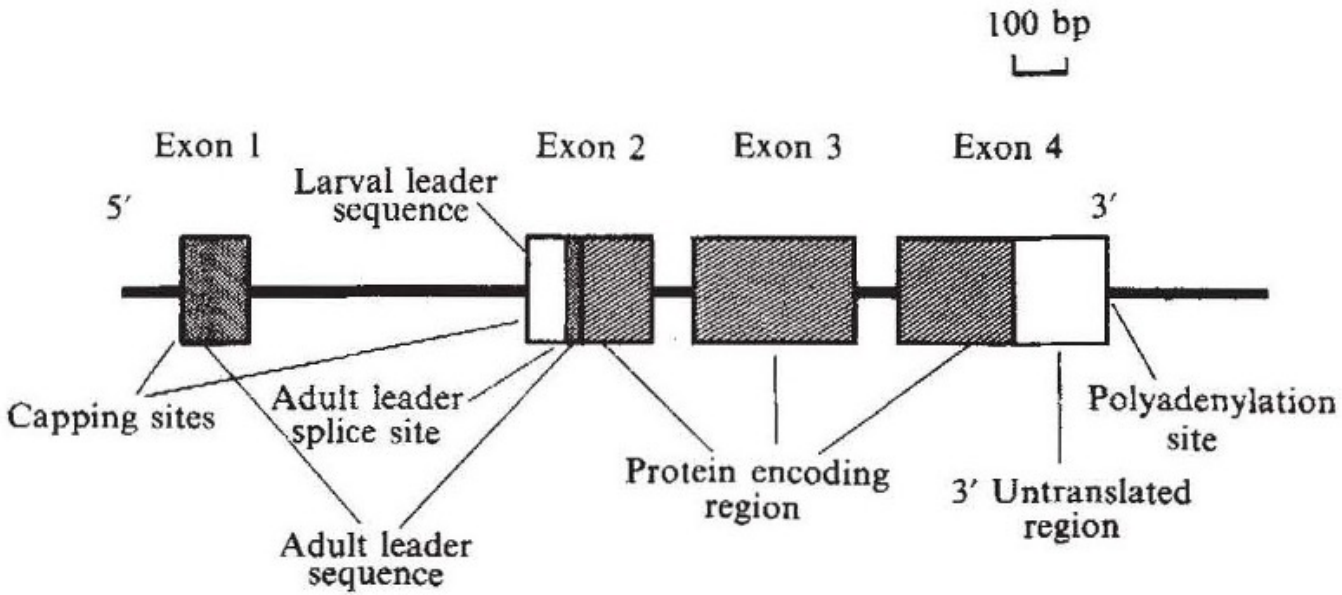


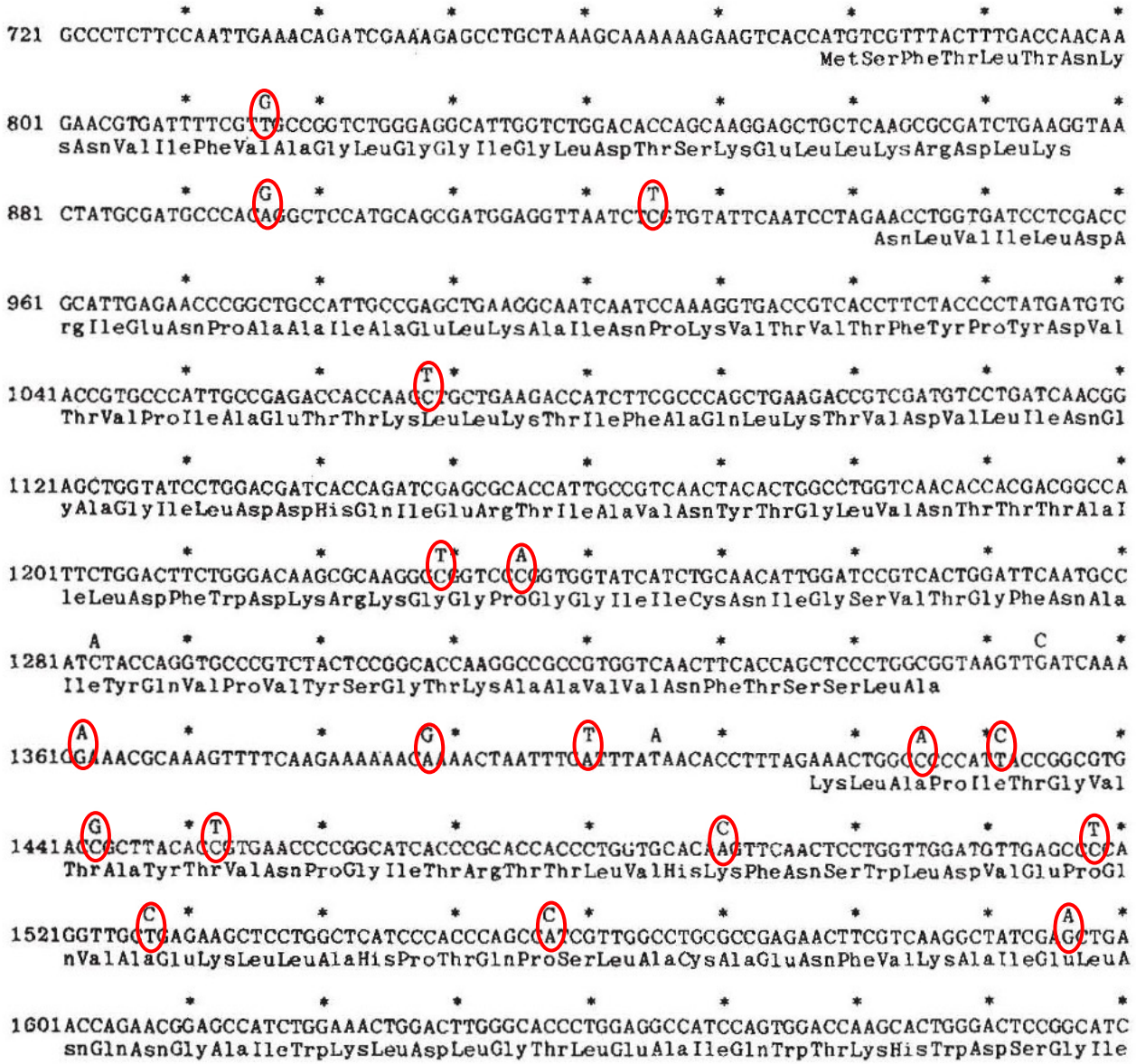
Figure 1 from Kreitman 1983

Kreitman 1983, *Nature*

DNA sequence variation at *Adh*

Polymorphic sites are sites that vary across individuals in the population

Consensus sequence showing polymorphic sites



Kreitman 1983, *Nature*

DNA sequence variation at *Adh*

Consensus sequence showing polymorphic sites

A polymorphism at site 1490 encodes a Lys -> Thr amino acid ***nonsynonymous substitution***, which results in a change from the 'slow' metabolising to 'fast' metabolizing allele

Kreitman 1983, *Nature*

Key results from Kreitman 1983

- Variation was higher than expected: 43 out of 2721 sites varied within the sample
- Most variation involved ***single nucleotide polymorphisms*** (**SNPs**)
- Only one polymorphism was ***non-synonymous****, i.e., affected the amino acid sequence, while 14 polymorphisms were ***synonymous*** (i.e., affected the coding sequence but did not change the amino acid sequence). The others affected **non-coding DNA**.

What was novel about this study?

- Previous studies focused on protein variation assayed using rate of movement through an electrophoretic gel. ***This was the first study to examine sequence variation within a species.***
- This study showed that contrary to the limited variation that could be assayed using gel electrophoresis, ***variation at the DNA sequence level was relatively common.***
- However, ***no new variation was detected that affected the amino acid sequence*** of the *Adh* protein.
- These results imply ***that strong purifying selection likely acted to limit changes in the Adh protein sequence.***

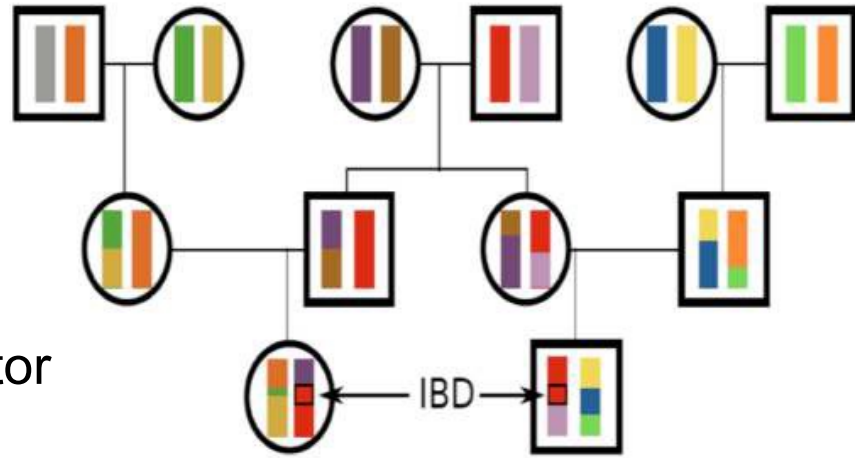
Describing population genetic variation (part 0)

Some terminology

- A *locus* is a genomic location
- Sites that vary with a population are called *segregating sites* or *polymorphic sites*
- In the protein-coding or gene regions
  - Coding variants that change the amino acid sequence are called *nonsynonymous* or *replacement* polymorphisms
  - Coding variants that do not change the amino acid sequence are called *synonymous* or *silent* polymorphisms
- An *allele* is the physical copy of DNA at a locus on a chromosome; usually used in the context of a polymorphic locus
- A *heterozygote* at a given locus carries two different alleles on its two chromosomes
- A *homozygote* at a given locus carries the same allele on both chromosomes

Evolution involves descent with modification

Identity by state (IBS) occurs when two alleles at a given locus are the same

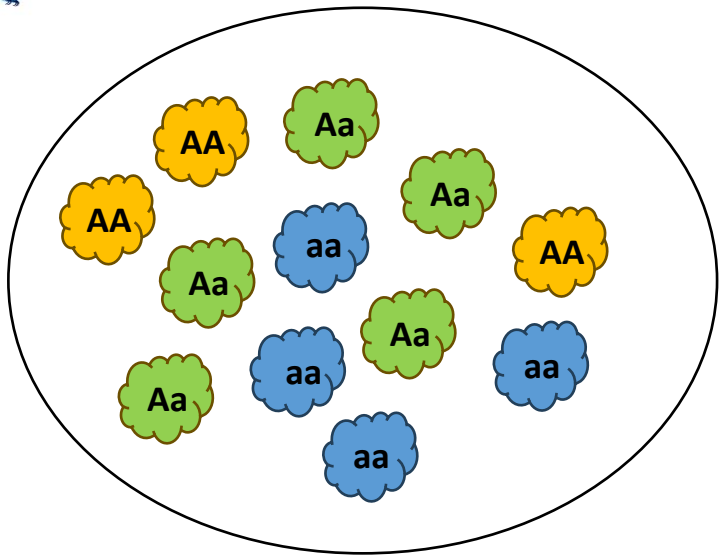


Two alleles derived from the same ancestor are shared identical by descent (IBD)

** Note that two alleles could be identical by state but differ by descent due to mutation

Figure source: Coop, 2022

Genotype and allele frequencies



Since we normally cannot sample an entire population, we are generally working with a sample from the populations we are studying.

Therefore, the genotype and allele frequencies we calculate from our sample are only estimates of the actual value in the population.

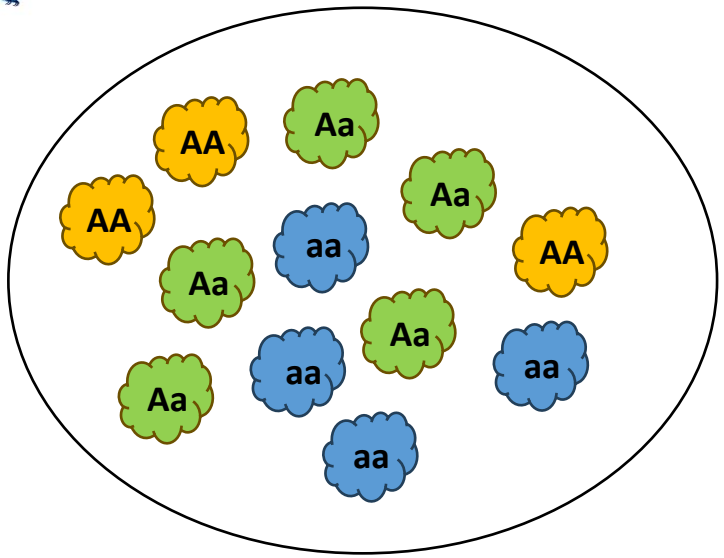
The confidence we have in our allele frequency estimates depends on the allele frequency and on how deeply we sampled from the population

95% confidence interval on estimate of the allele frequency

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

As sample size increases, the confidence interval becomes tighter

Genotype frequencies

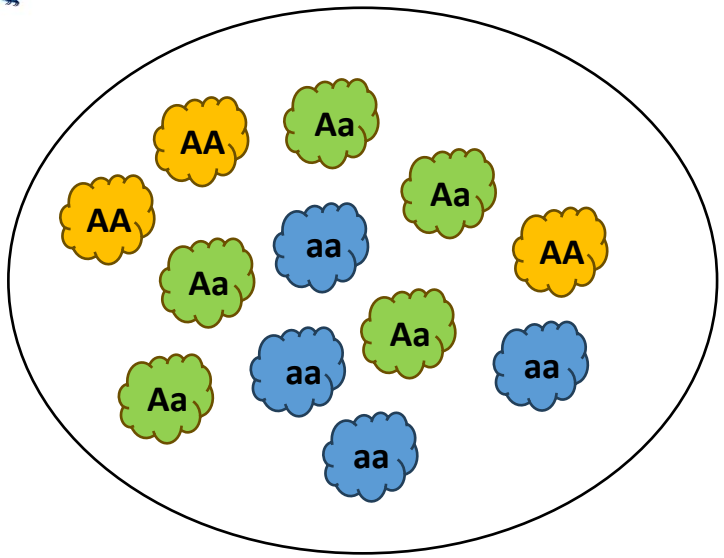


Genotype frequencies sum up to 1:

$$x_{11} + x_{12} + x_{22} = 1$$

Or, alternatively: $x_{AA} + x_{Aa} + x_{aa} = 1$

Genotype frequencies



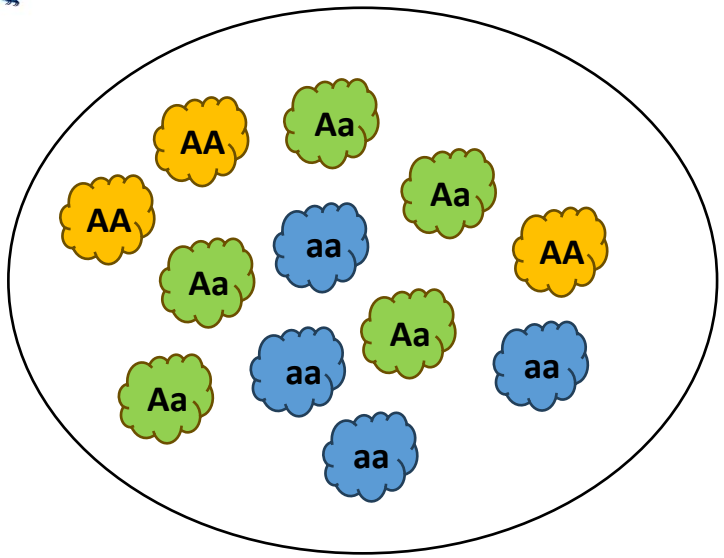
Genotype frequencies sum up to 1:

$$x_{11} + x_{12} + x_{22} = 1$$
$$x_{AA} + x_{Aa} + x_{aa} = 1$$

The frequency of each genotype is the number of that genotype divided by the total number of individuals:

$$x_{AA} = N_{AA}/N$$
$$x_{Aa} = N_{Aa}/N$$
$$x_{aa} = N_{aa}/N$$

Genotype frequencies



Genotype frequencies sum up to 1:

$$x_{11} + x_{12} + x_{22} = 1$$
$$x_{AA} + x_{Aa} + x_{aa} = 1$$

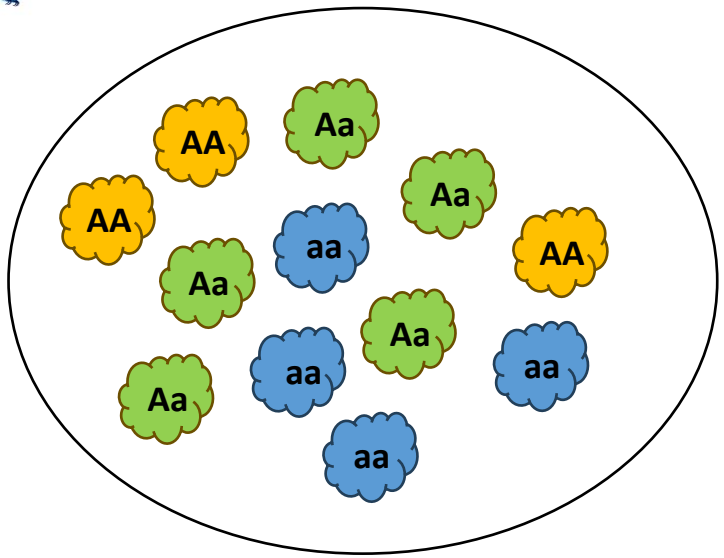
The frequency of each genotype is the number of that genotype divided by the total number of individuals:

$$x_{AA} = \frac{N_{AA}}{N} = \frac{3}{12} = 0.25$$

$$x_{Aa} = \frac{N_{Aa}}{N} = \frac{5}{12} = 0.42$$

$$x_{aa} = \frac{N_{aa}}{N} = \frac{4}{12} = 0.33$$

Genotype and allele frequencies



Genotype frequencies

$$x_{AA} = 0.25$$

$$x_{Aa} = 0.42$$

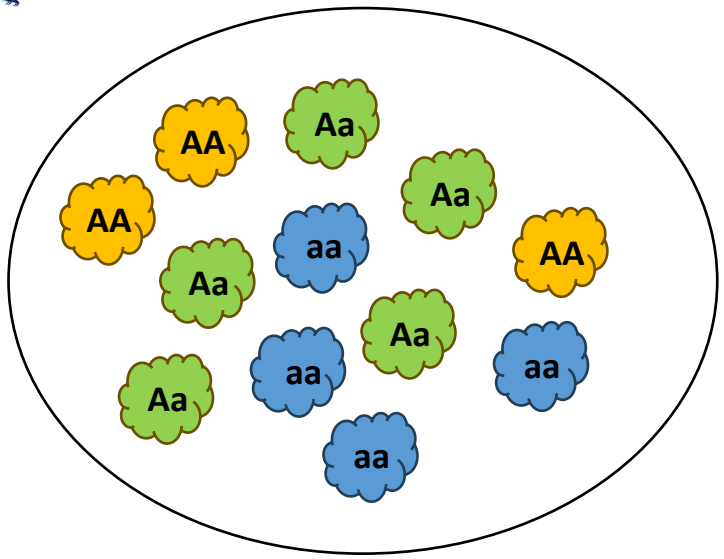
$$x_{aa} = 0.33$$

Allele frequencies

$$p = x_{AA} + \frac{1}{2}x_{Aa}$$

$$q = 1 - p = x_{aa} + \frac{1}{2}x_{Aa}$$

Genotype and allele frequencies



Genotype frequencies

$$x_{AA} = 0.25$$

$$x_{Aa} = 0.42$$

$$x_{aa} = 0.33$$

Allele frequencies

$$p = x_{AA} + \frac{1}{2}x_{Aa} = 0.25 + 0.21 = 0.46$$

$$q = 1 - p = x_{aa} + \frac{1}{2}x_{Aa} = 0.33 + 0.21 = 0.54$$

Generalization to K-allelic loci

What about the case(s) where a locus is not bi-allelic?

For example, microsatellite loci tend to have more than two alleles.

Genotype frequencies still sum up to 1:

$$\mathbf{1} = \boldsymbol{x_{11}} + \boldsymbol{x_{22}} + \ \ldots + \ \boldsymbol{x_{nn}} + \boldsymbol{x_{12}} + \boldsymbol{x_{13}} + \cdots + \boldsymbol{x_{(n-1)n}}$$

$$= \sum_{i=1}^{n} \sum_{j \geq 1}^{n} x_{ij}$$

And the frequency of the ith allele is:

$$p_i = x_{ii} + \frac{1}{2} \sum_{j=1}^{i-1} x_{ji} + \frac{1}{2} \sum_{j=i+1}^{n} x_{ij}$$

Evolution is the change in allele frequencies over time.

What causes allele frequencies to change?

- Mutation
- Genetic drift
- Migration
- Selection

We will explore each of these forces of evolution in more detail in future lectures

Summary

- Population genetics can help to address a wide variety of basic and applied questions
- The genomic revolution is opening up new opportunities to study variation within and between populations
- Not all mutations are created equal; some affect non-coding DNA and some affect coding regions. Those that impact coding DNA may be synonymous or non-synonymous
- Evolution involves a process of descent with modification
- We can estimate population allele frequencies from a sample. The size of the sample and frequency of the allele in the population determines the precision of our estimate

More resources

- https://www.genome.gov/leadership-initiatives/History-of-Genomics-Program
- https://www.genome.gov/about-genomics/policy-issues
- Genetics unzipped podcasts from The Genetics Society
- https://geneticsunzipped.com/blog/2019/4/11/011-darwin-vs-mendel
- https://geneticsunzipped.com/blog/2024/1/11/the-battle-for-biology-mendel
- https://geneticsunzipped.com/blog/2020/10/22/s322-the-past-present-and-future-of-the-human-genome-project
- https://www.nature.com/immersive/d42859-020-00099-0/index.html