Survey of Genetics and Genomics

Christopher Keown, PhD

About Me

Bachelors in Computer Science

Masters in Computational Science

PhD in Cognitive Science

Computational Neuroepigenomics

About Me

WELBY HEALTH

Objectives

Familiarize you with:

- the biology of genetics and genomics
- how we measure and analyze them
- understanding your 23andMe better
- interesting data type with lots of publicly available data

Genetics

ge·net·ics

/jəˈnediks/

noun

the study of heredity and the variation of inherited characteristics.



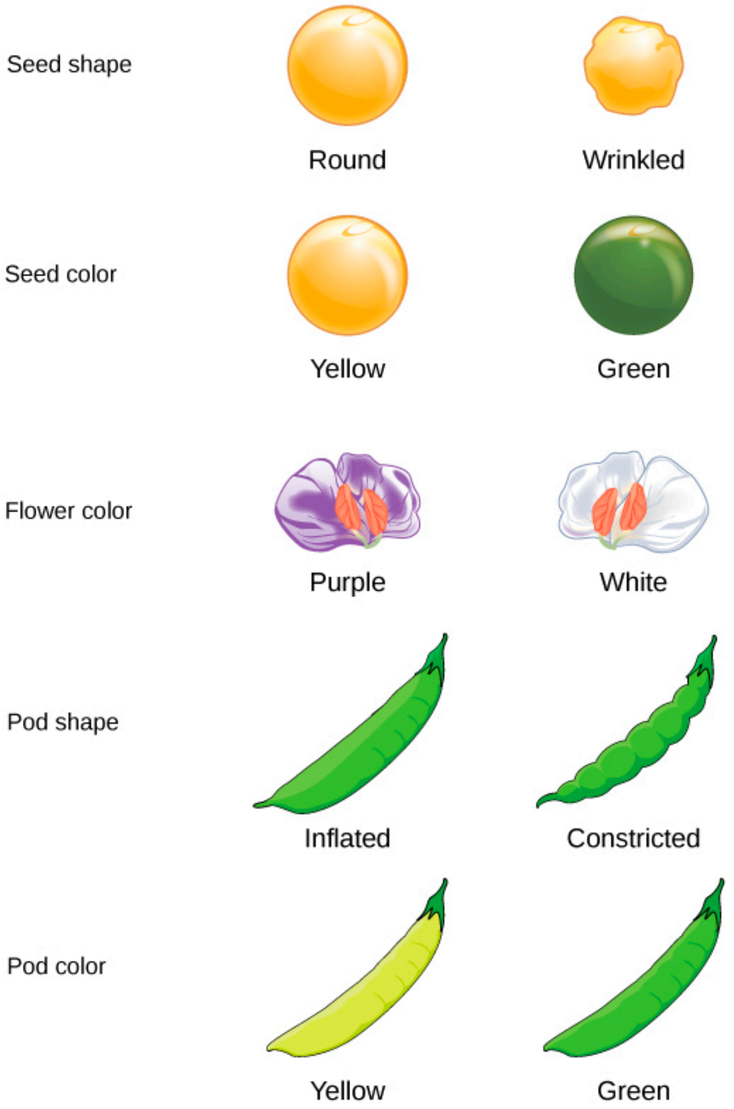

How is variation in DNA associated with variation in traits, e.g., specific diseases, height, cilantro tasting like soap?

Why does genetics matter?

- Understand the role of genetics versus environment in a given trait
- Diagnosis of diseases
- Compute risk scores for diseases like autism, schizophrenia, etc.
- Personalized medicine

Mendel's Peas

Gregor Mendel

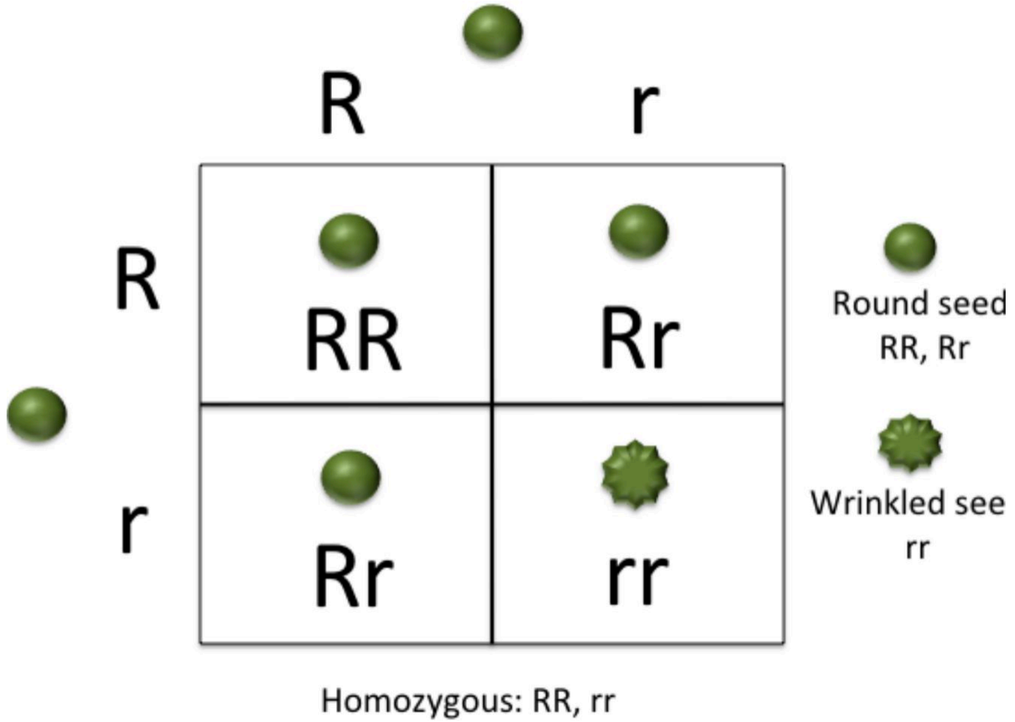


Theory of inheritance

Dominance and recessiveness

Each person has two copies of a trait (alleles)

Traits are passed randomly to offspring



What about quantitative phenotypes (e.g., height)?

William Bateson

Mendelians

Karl Pearson

Biometricians

Based on Galton's Ideas

"the pea vs. height debate"

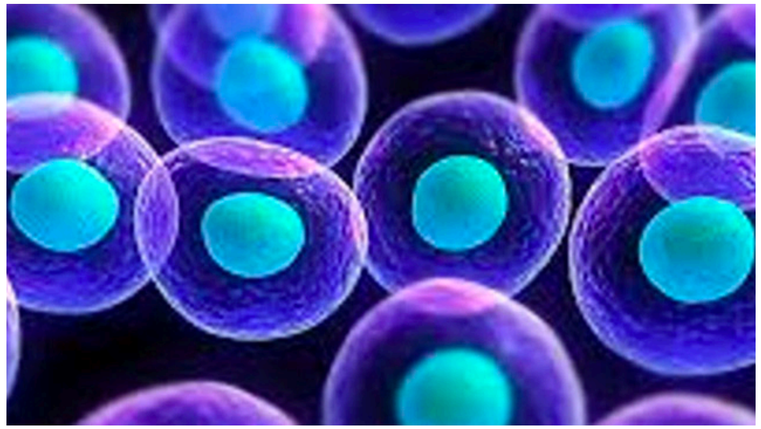
R. A. Fisher in 1918

Solution: multiple genes contribute to variation in the population, each of them obeying Mendelian rules

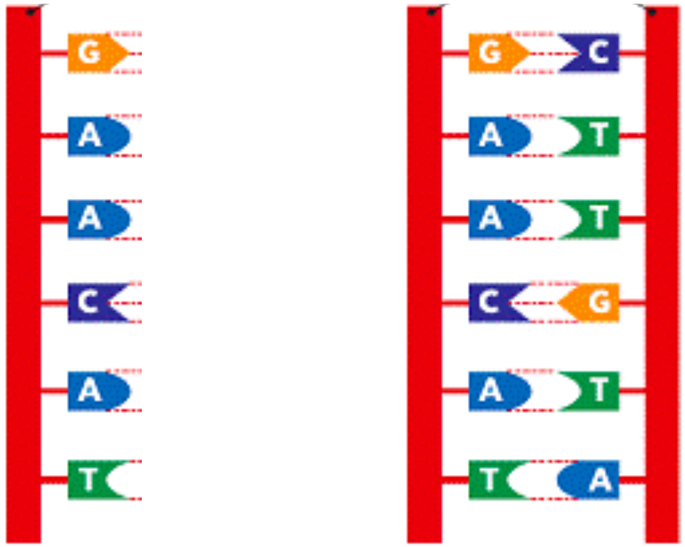
A bunch of genes control height

What is the biological mechanism?

The human body consists of ~37.2 trillion cells.

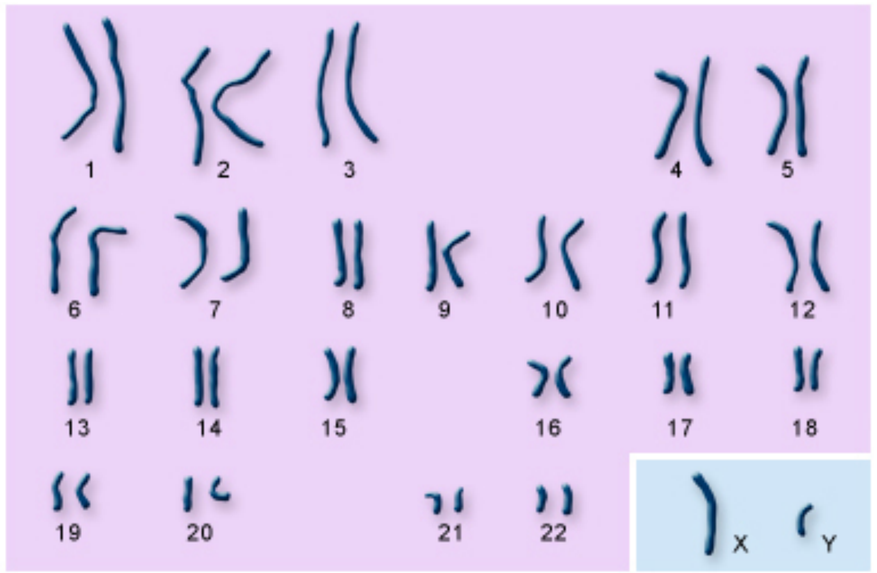


Each cell contains a copy of the hereditary information - DNA



Genotype is the sequence of DNA that encodes a particular trait: AA, AT, CG, etc.

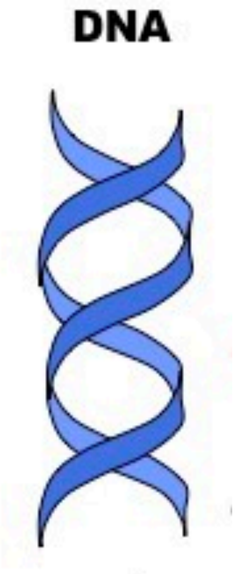
Phenotype is the physical manifestation of that trait: brown hair, height, autism.



How is DNA used to make genes?

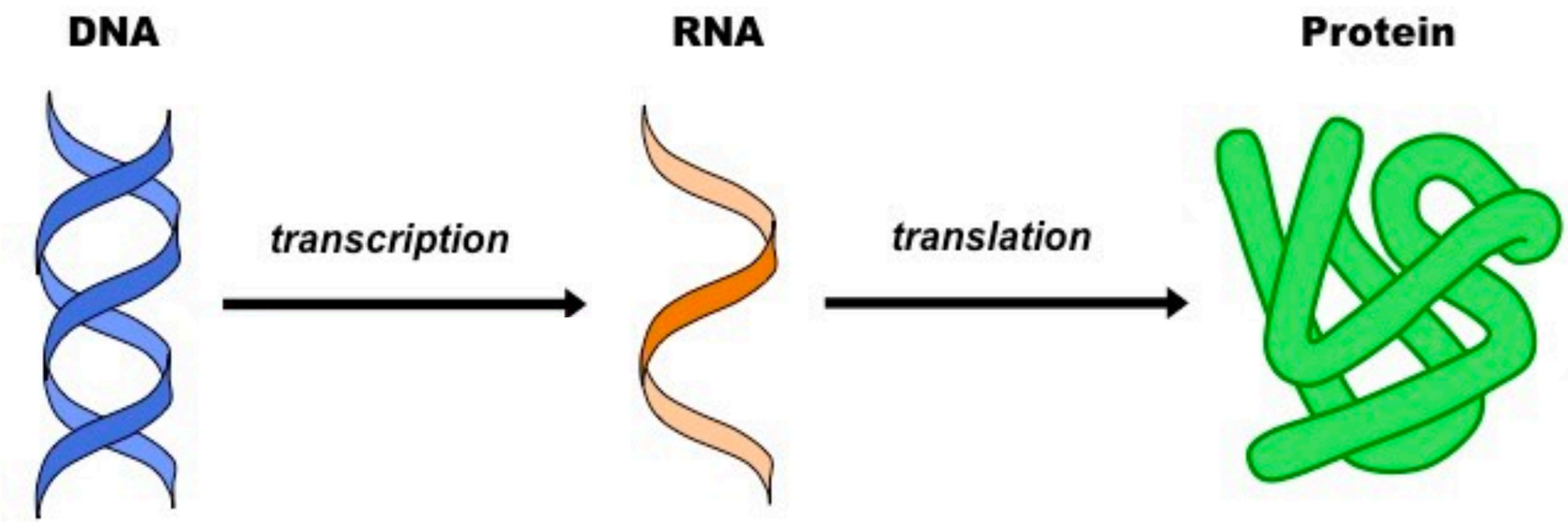
DNA has the recipes for proteins.
DNA is located in the nucleus,
but protein synthesis happens
outside of the nucleus.

Proteins are the actual
worker molecules that do
things in the body.



How is DNA used to make genes?

The recipe for a given gene is copied from DNA into RNA and shipped to the cell cytoplasm to make the protein.

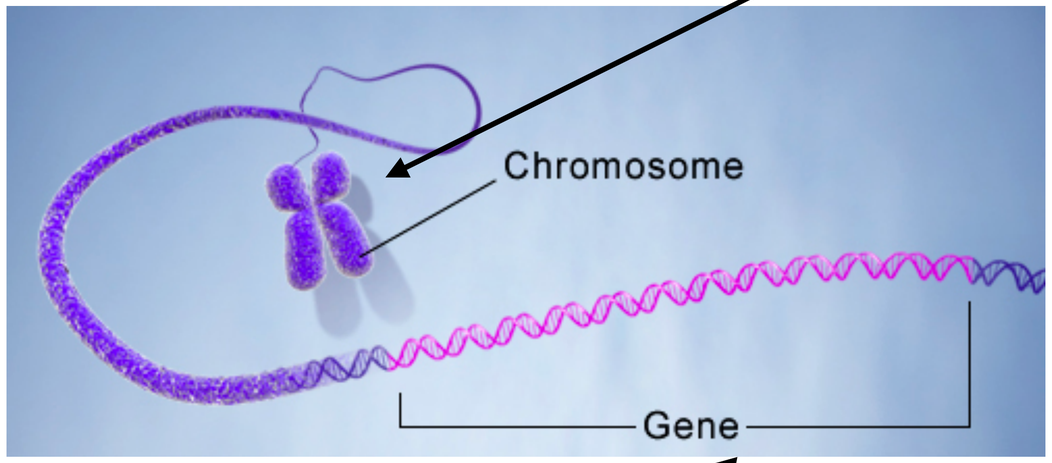


Formula for a genetic association analysis

1. Measure a phenotype of interest.
2. Measure the sequence of ACGTs for a given set of individuals.
3. Correlate the variance for each loci in the genome with the phenotype of interest.

Two types of analyses:

Genome-Wide Association Study (GWAS)



Hypothesis-driven

Example analysis in schizophrenia

1. Measure a phenotype of interest.
2. Measure the sequence of ACGTs for a given set of individuals.
3. Correlate the variance for each loci in the genome with the phenotype of interest.

ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

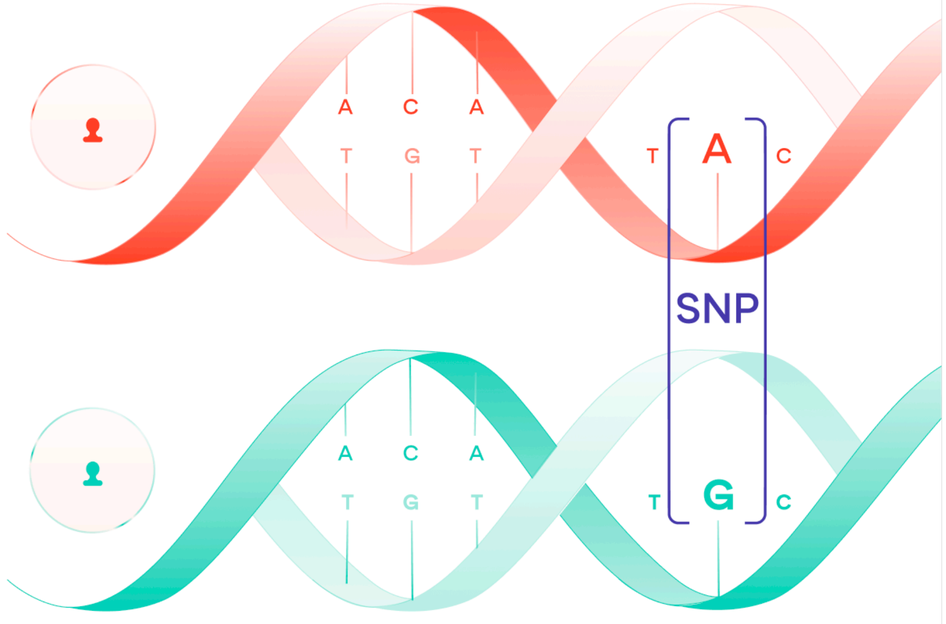
Cases: 36,989

Controls: 113,075

Idea behind a genetics analysis (recipe)

1. Measure a phenotype of interest—let's say schizophrenia.
2. Measure the sequence of ACGTs for a given set of individuals.
3. Correlate the variance for each loci in the genome with the phenotype of interest.

Single Nucleotide Polymorphisms (SNPs)



rs6277

Organism	*Homo sapiens*
Position	chr11:113412737 (GRCh38.p12)
Alleles	G>A
Variation Type	SNV Single Nucleotide Variation
Frequency	A=0.411110 (103247/251142, GnomAD_exome) A=0.415771 (50453/121348, ExAC) A=0.37935 (11879/31314, GnomAD) (+ 18 more)

DB SNP: https://www.ncbi.nlm.nih.gov/snp/

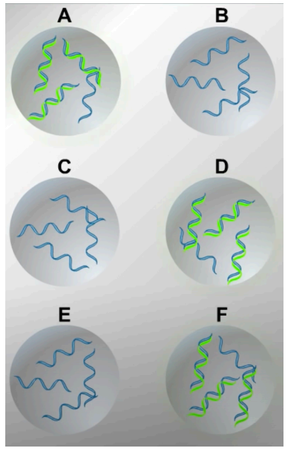
Result of the HAPMAP and 1000 Genome projects

How do we experimentally detect genetic variants?

rs6277

Organism	*Homo sapiens*
Position	chr11:113412737 (GRCh38.p12)
Alleles	G>A
Variation Type	SNV Single Nucleotide Variation
Frequency	A=0.411110 (103247/251142, GnomAD_exome) A=0.415771 (50453/121348, ExAC) A=0.37935 (11879/31314, GnomAD) (+ 18 more)

Genotyping



Most Common Sequence

T G T C G G G A

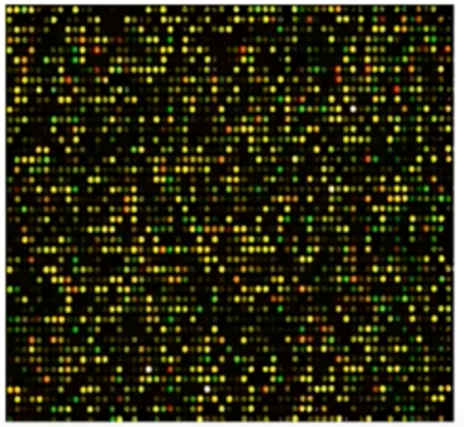
A C A G C C C T

Less Common Sequence

T G T C A G G A

A C A G T C C T

Use sequence complementarity to measure which allele someone has

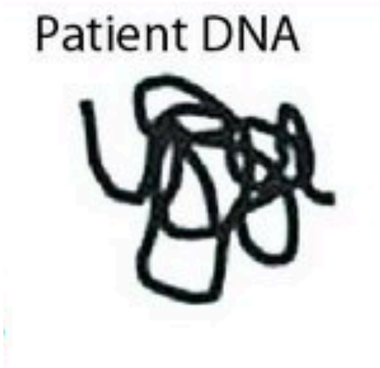


This is genotyping, not DNA sequencing.

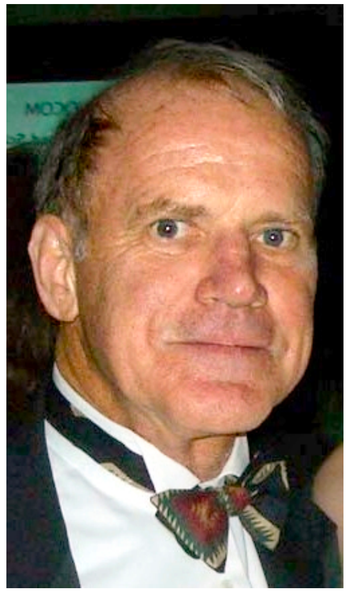
Usually measures ~ 1M SNPs —not the whole genome.

It's cheaper: ~$300 per person.

FUN FACT!



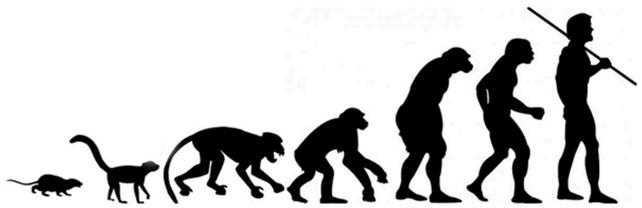
Genotyping and DNA sequencing require a decent amount of DNA to work



Kary Mullis

Polymerase chain reaction (PCR) for DNA amplification

Nobel Prize in Chemistry in 1993



Genetics - GWAS

1. Measure a phenotype of interest—let's say schizophrenia.
2. Measure the sequence of ACGTs for a given set of individuals.
3. Correlate the variance for each loci in the genome with the phenotype of interest.

GWAS regression analysis

Let Y_i be the phenotype for individual i

$Y_i = 0$ for controls

$Y_i = 1$ for cases

Let X_i be the genotype of individual i at a particular SNP

TT	$X_i = 0$
GT	$X_i = 1$
GG	$X_i = 2$

$$\text{logit}(p_i) \sim \beta_0 + \beta_1 X_i + \beta_2 C_i + \beta_3 D_i + \ldots$$

Correct for multiple comparisons: $p < 10^{-8}$

Results - The Manhattan Plot

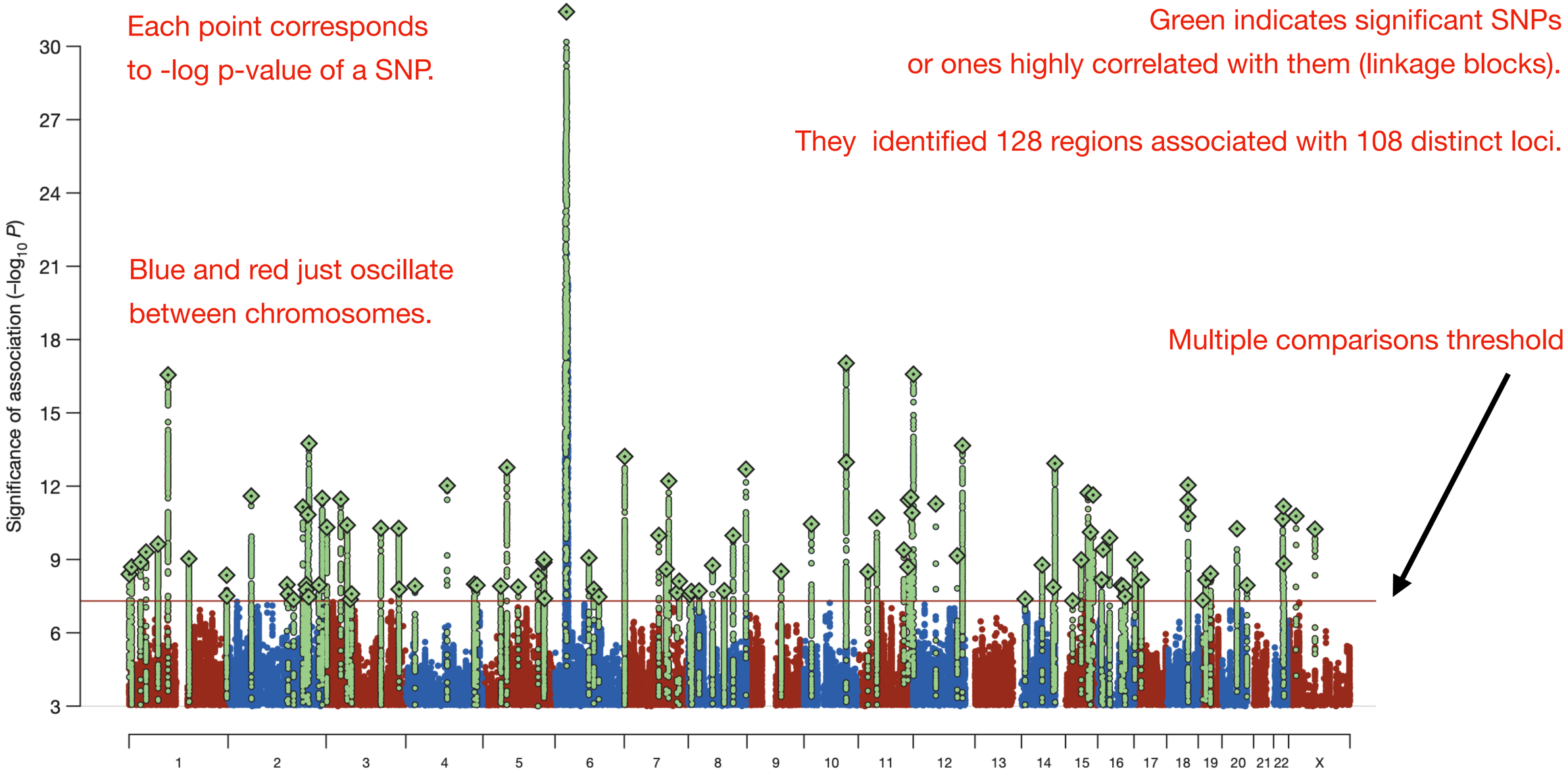
Each point corresponds to -log p-value of a SNP.

Green indicates significant SNPs or ones highly correlated with them (linkage blocks).

They identified 128 regions associated with 108 distinct loci.

Blue and red just oscillate between chromosomes.

Multiple comparisons threshold



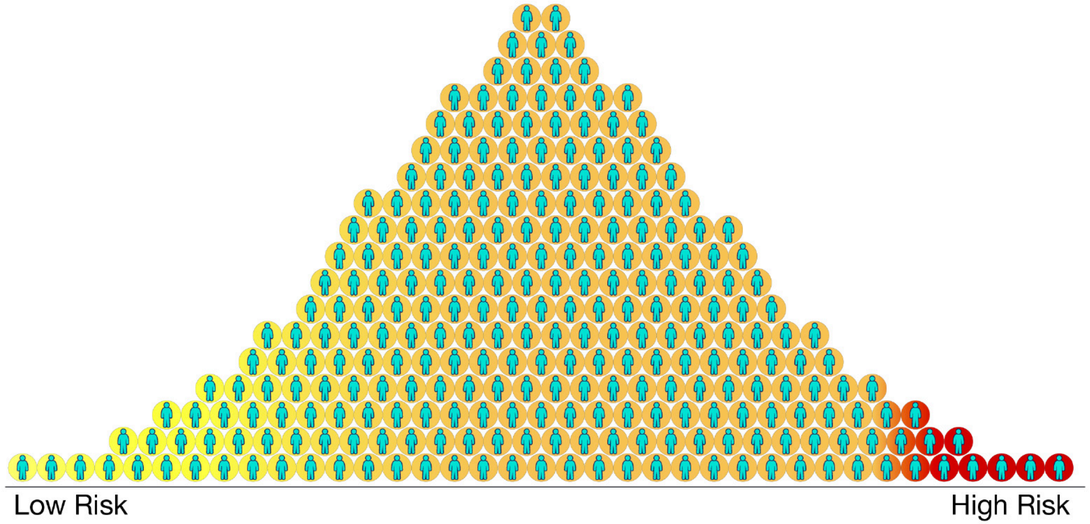
Learnings from GWAS analyses

- Very few diseases and phenotypes are Mendelian (single-gene) diseases: cystic fibrosis, sickle-cell anemia, Huntington's disease
- Most diseases and traits are associated with many loci, each accounting for a small amount of the overall variance.
- Most SNPs associated with schizophrenia and many other diseases do not code for proteins—they are noncoding.
- GWAS doesn't examine the interaction between genes (epistasis).

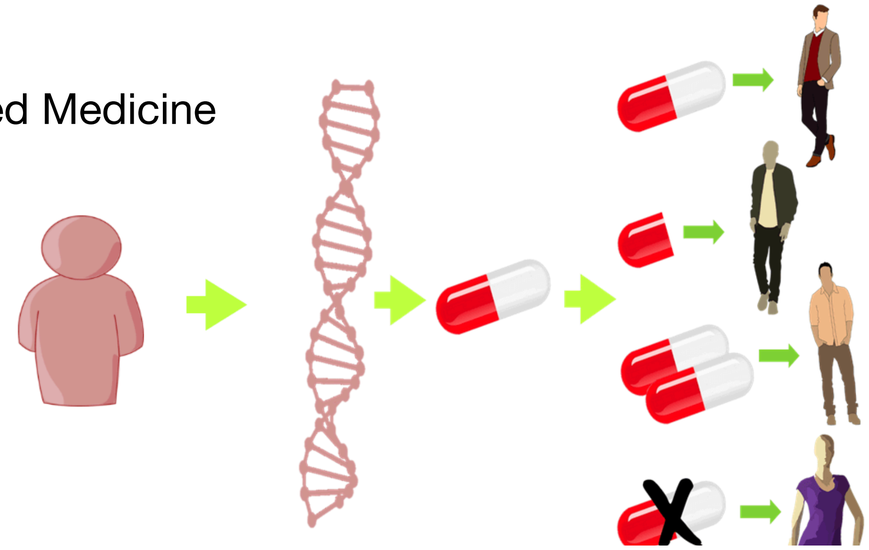
Applications for machine learning

We have a list of genes, but what do they do?
Mine the academic literature.

Polygenic Risk Scores



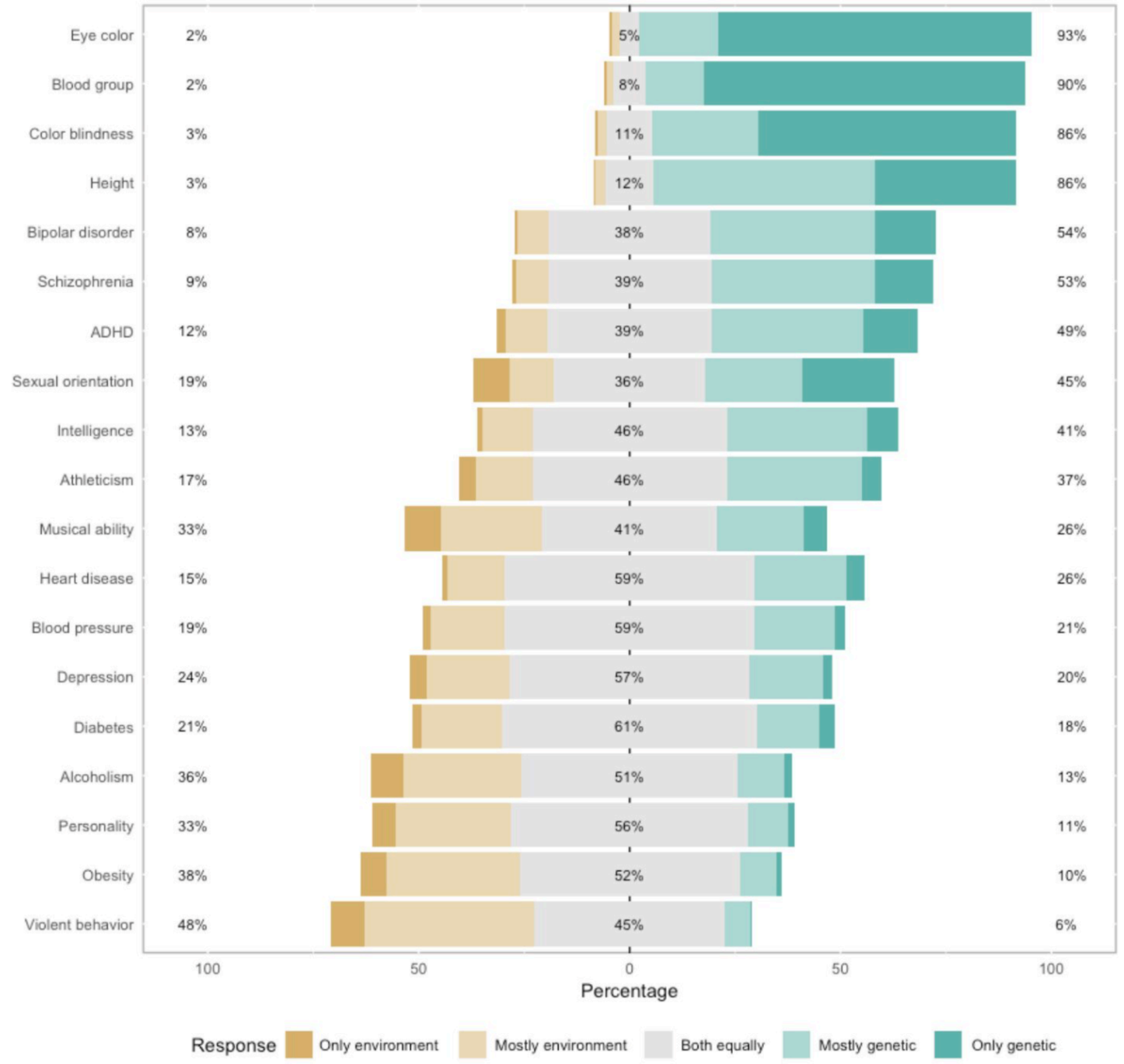
Personalized Medicine



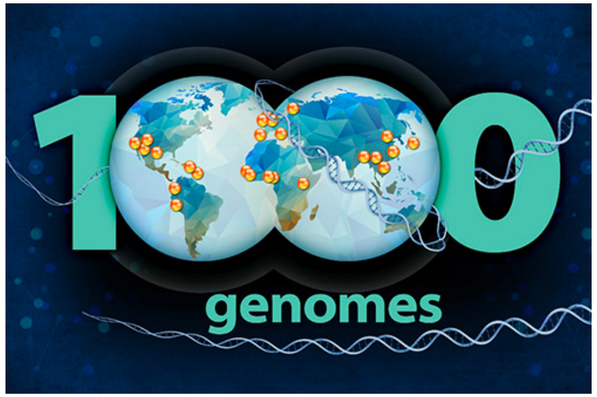
Heritability

A score for how much genetics play a role in the variability of a phenotypes, e.g, as opposed to the environment

Twin studies to measure the heritability: compare MZ twins with DZ twins.



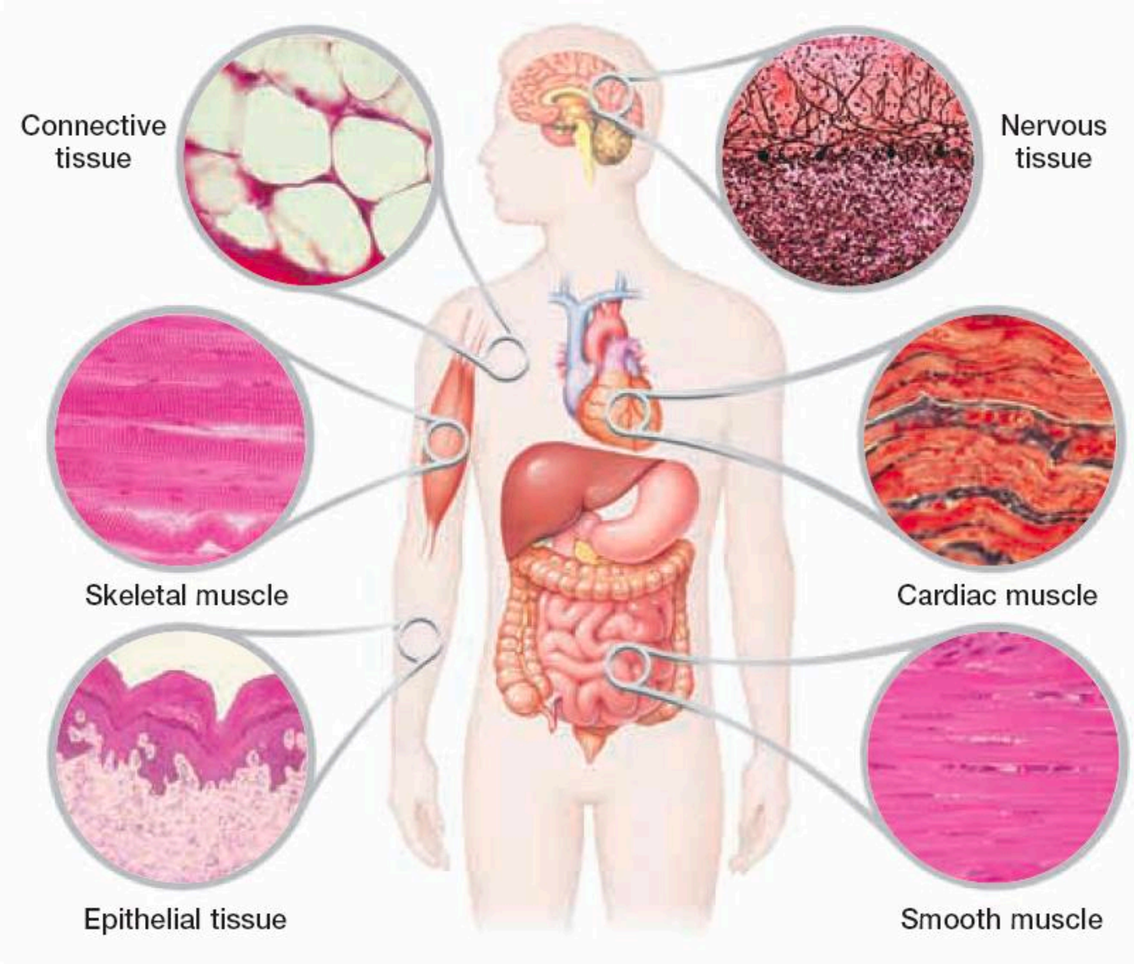
23andMe



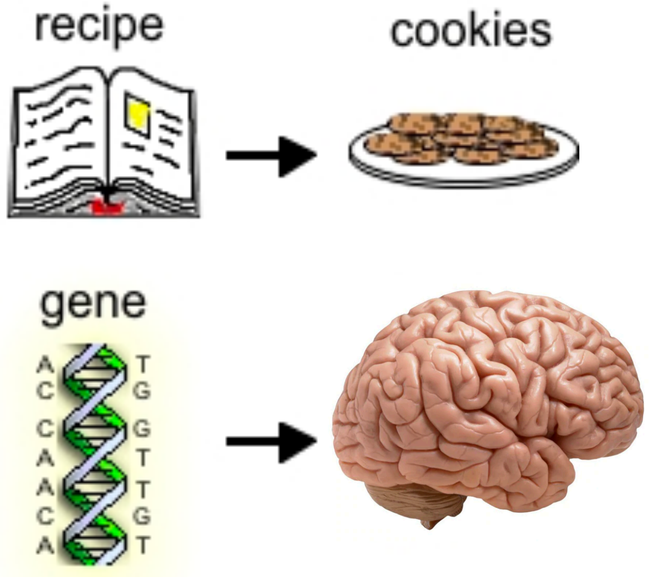
- What data does 23andMe collect?
- Haplotype - how much your SNPs match with certain ethnic groups in the world https://www.23andme.com/ancestry-composition-guide-pre-v5/
- Does cilantro taste like soap?
- Viewing your raw data
- The ethics of 23andMe

Genomics

Genomics: how are genes turned on and off in each cell to produce such cellular diversity—even though all cells have the same genome?



If DNA were a recipe book…

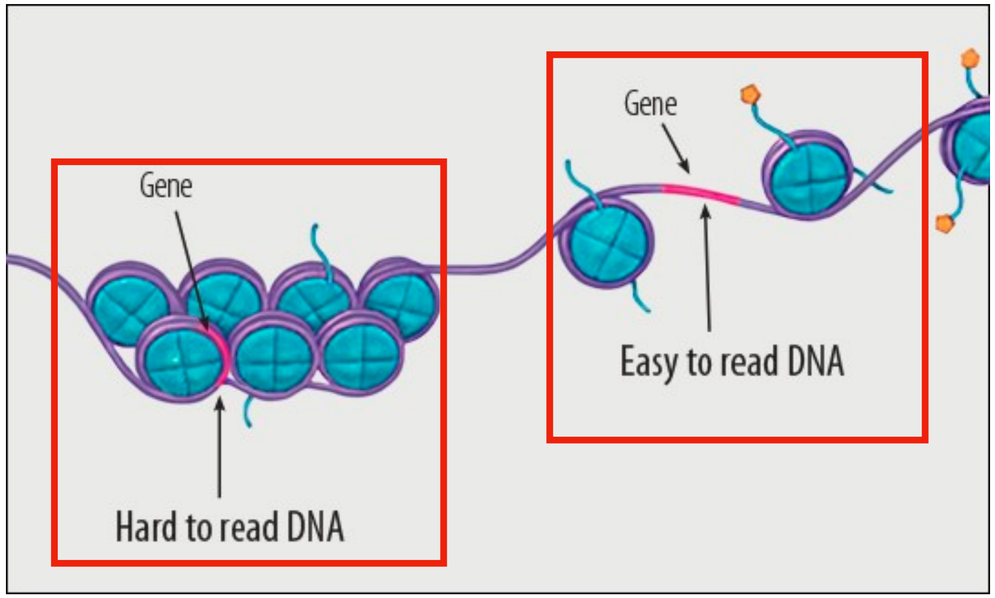


Put a single word on each page.

Now randomly shuffle all of the pages and break them into volumes.

Genomics study the mechanisms for switching to the right page at the right time to make chocolate chip cookies—or, a human brain.

Each cell is baking something different.



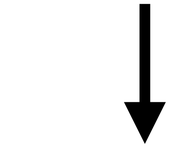
Manipulation of the accessibility of DNA is one way to turn genes on and off (chromatin accessibility)

There are many ways to modify DNA to turn genes on and off, or up and down:
methylation, chromatin accessibility, histone modifications, chromosome interactions, etc.

<u>The genome is spatially and temporally dynamic, and thus so is gene expression.</u>

Opportunities for machine learning

DNA sequence
+
Genomic markers



Protein

- Can we predict protein levels given genomic modifications?
- How do different mutations affect the expression of proteins?
- How does this regulation go awry in cancer?
- Massive amounts of publicly available data: https://www.ncbi.nlm.nih.gov/geo/

Identify cell types in the brain

Hypothesis: We can measure DNA methylation, a genomic marker, in individual cells and cluster them to identify neuronal cell types

RESEARCH

Science
AAAS

NEURAL EPIGENOMICS

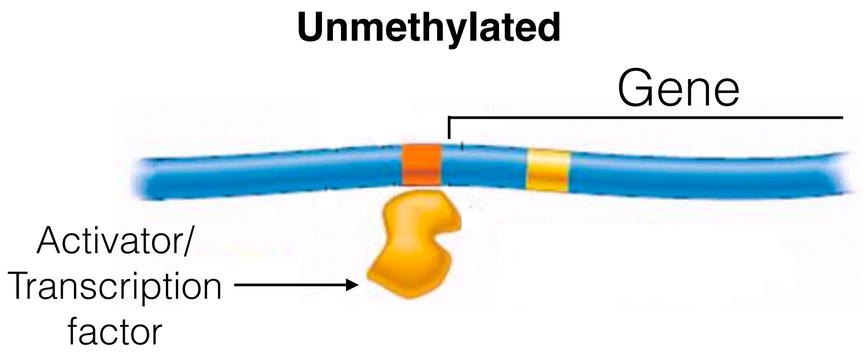
Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex

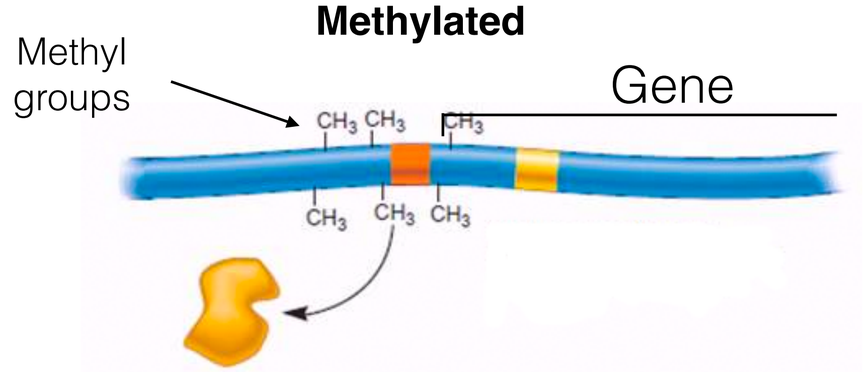
Chongyuan Luo,[1,2]* Christopher L. Keown,[3]* Laurie Kurihara,[4] Jingtian Zhou,[1,5] Yupeng He,[1,5] Junhao Li,[3] Rosa Castanon,[1] Jacinta Lucero,[6] Joseph R. Nery,[1] Justin P. Sandoval,[1] Brian Bui,[6] Terrence J. Sejnowski,[2,6,7] Timothy T. Harkins,[4] Eran A. Mukamel,[3]† M. Margarita Behrens,[6]† Joseph R. Ecker[1,2]†

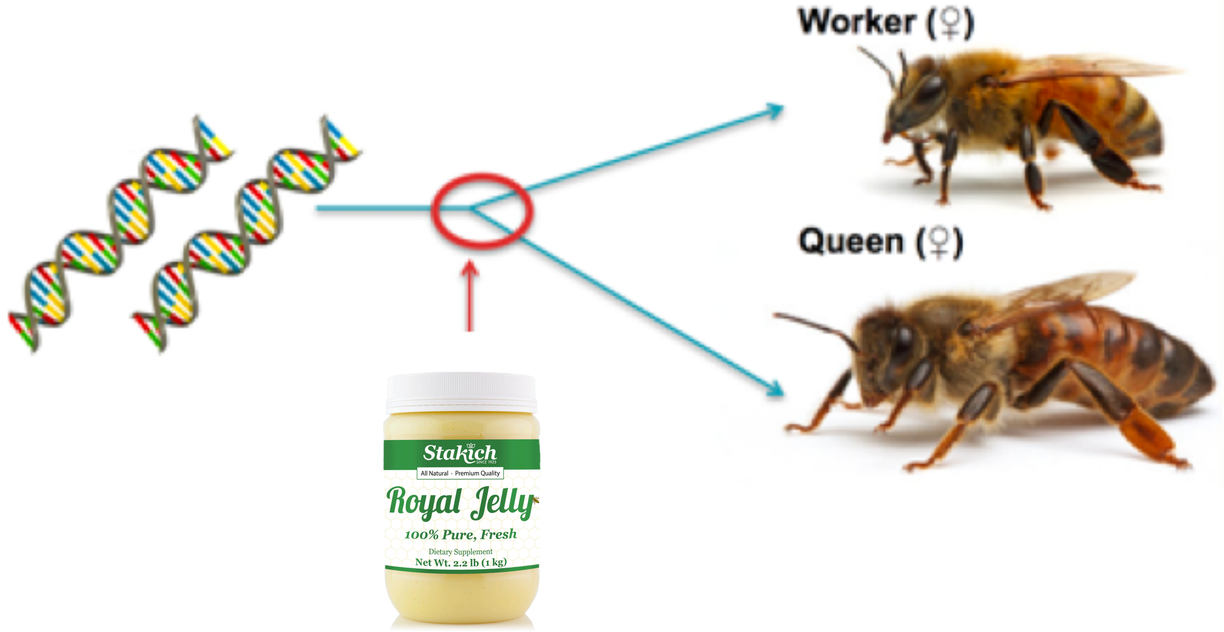
*contributed equally to this work

Luo *et al.*, *Science* **357**, 600–604 (2017) 11 August 2017

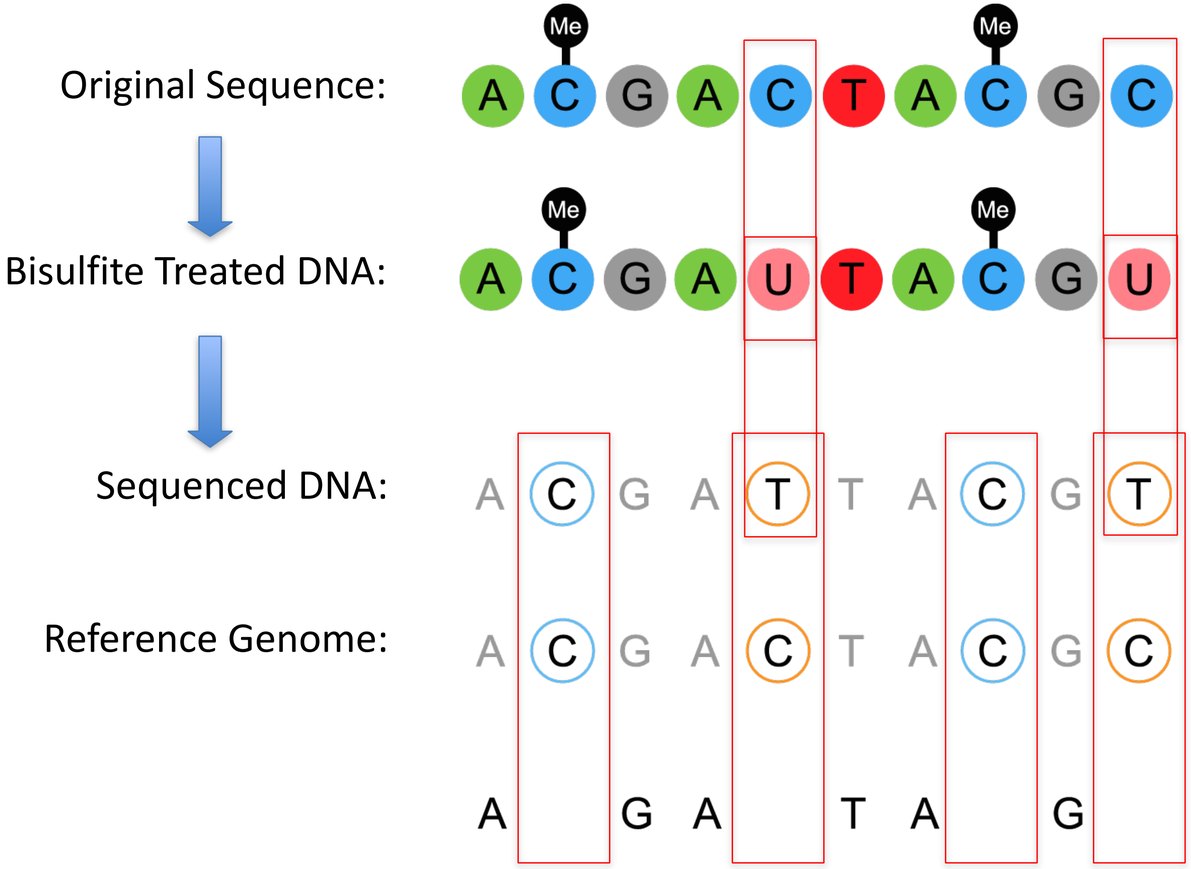
DNA Methylation







Whole genome measurement of DNA methylation with bisulfite sequencing



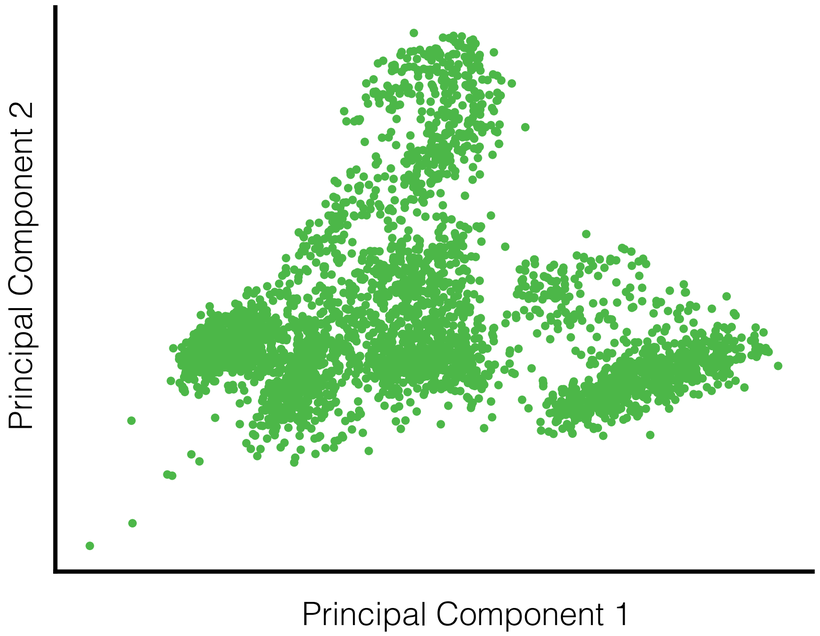
(Lister,Ecker 2009)

Sequencing by Synthesis / Next Generation Sequencing

https://www.youtube.com/watch?v=fCd6B5HRaZ8

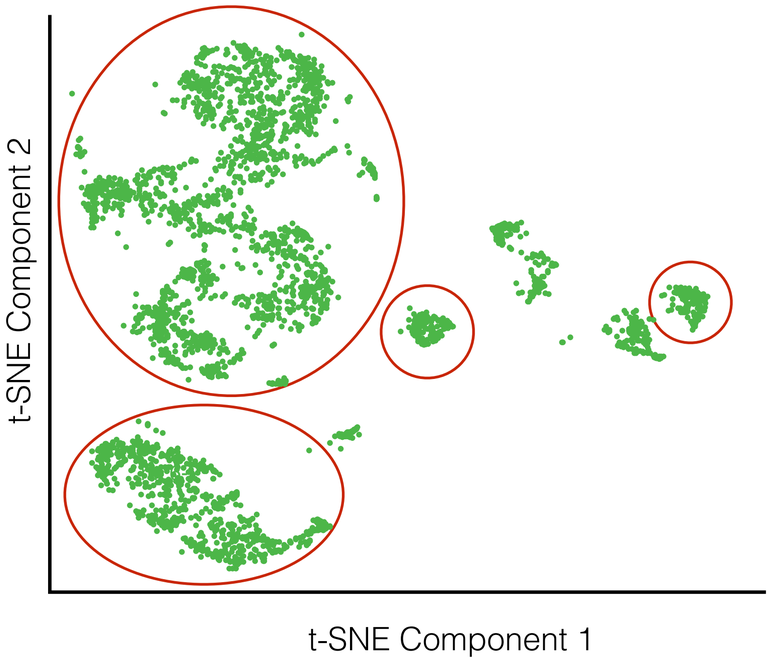
Data visualization

Principal components analysis

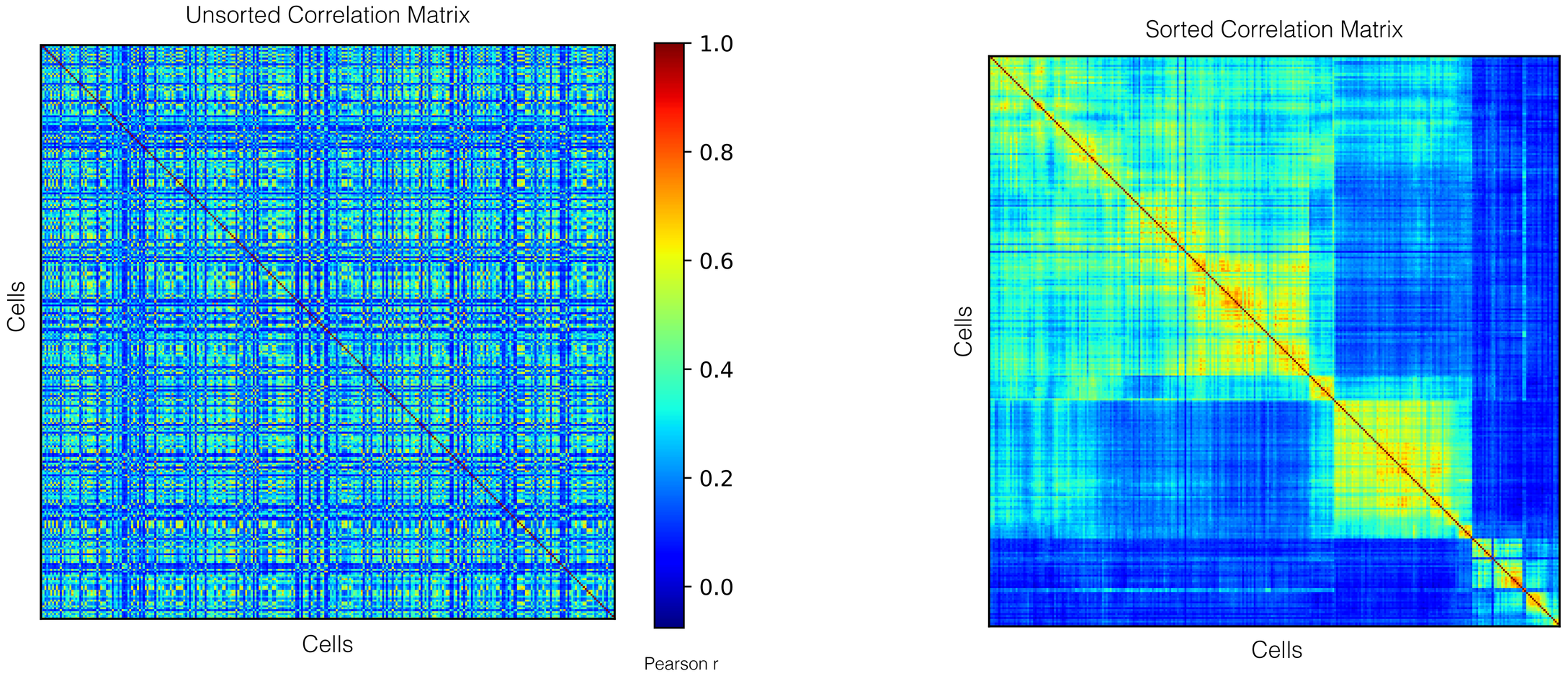


t-Distributed stochastic neighbor embedding

(van der Maaten, Hinton 2008)



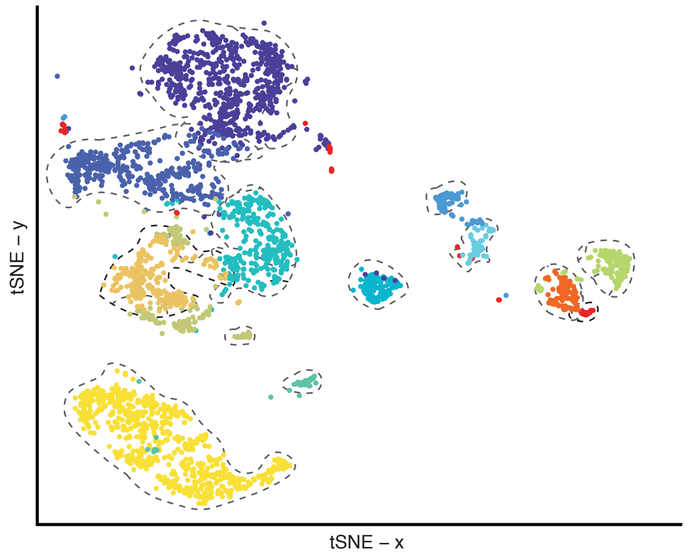
Clustering

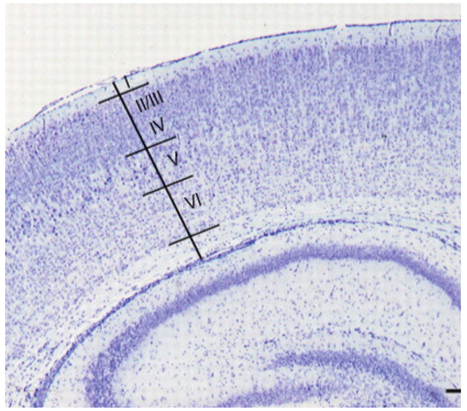


We used an algorithm call BackSPIN for clustering (Zeisel 2016)

Single neurons are distributed according to cortical layer

Clustering with BackSPIN

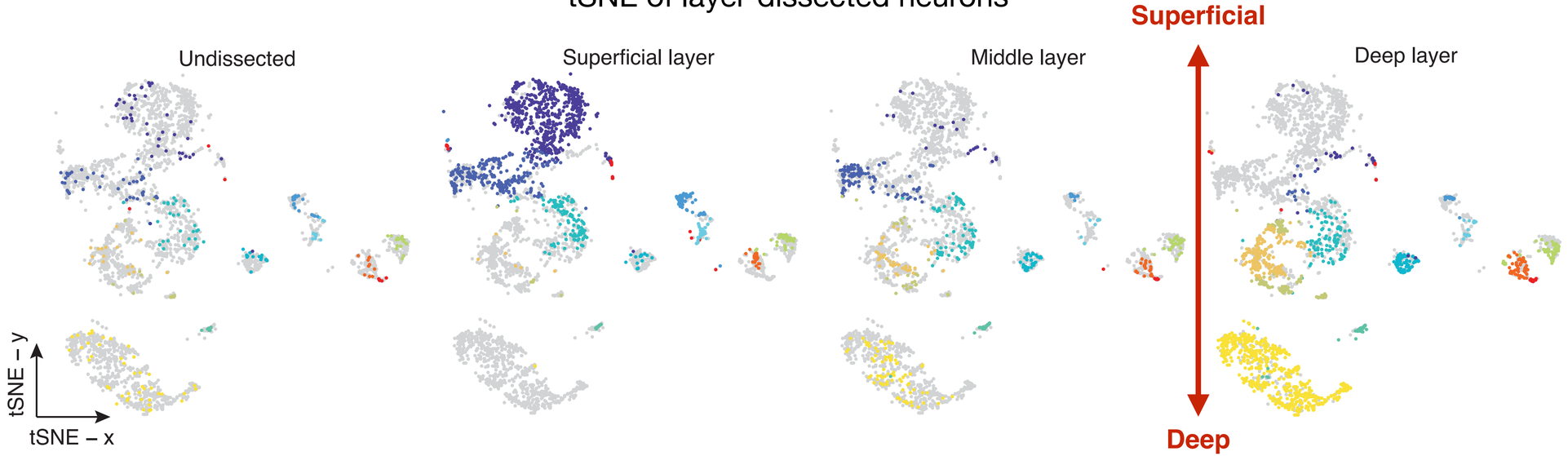




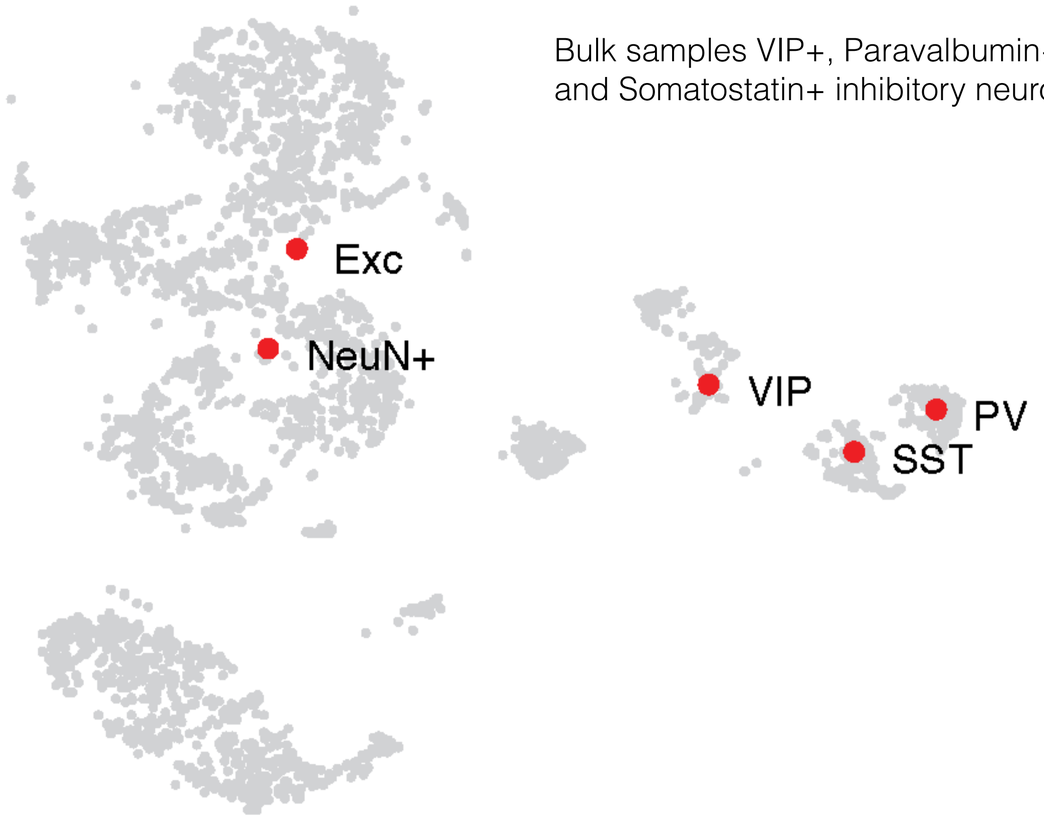
Coronal section

Margarita Behrens, Salk

tSNE of layer-dissected neurons

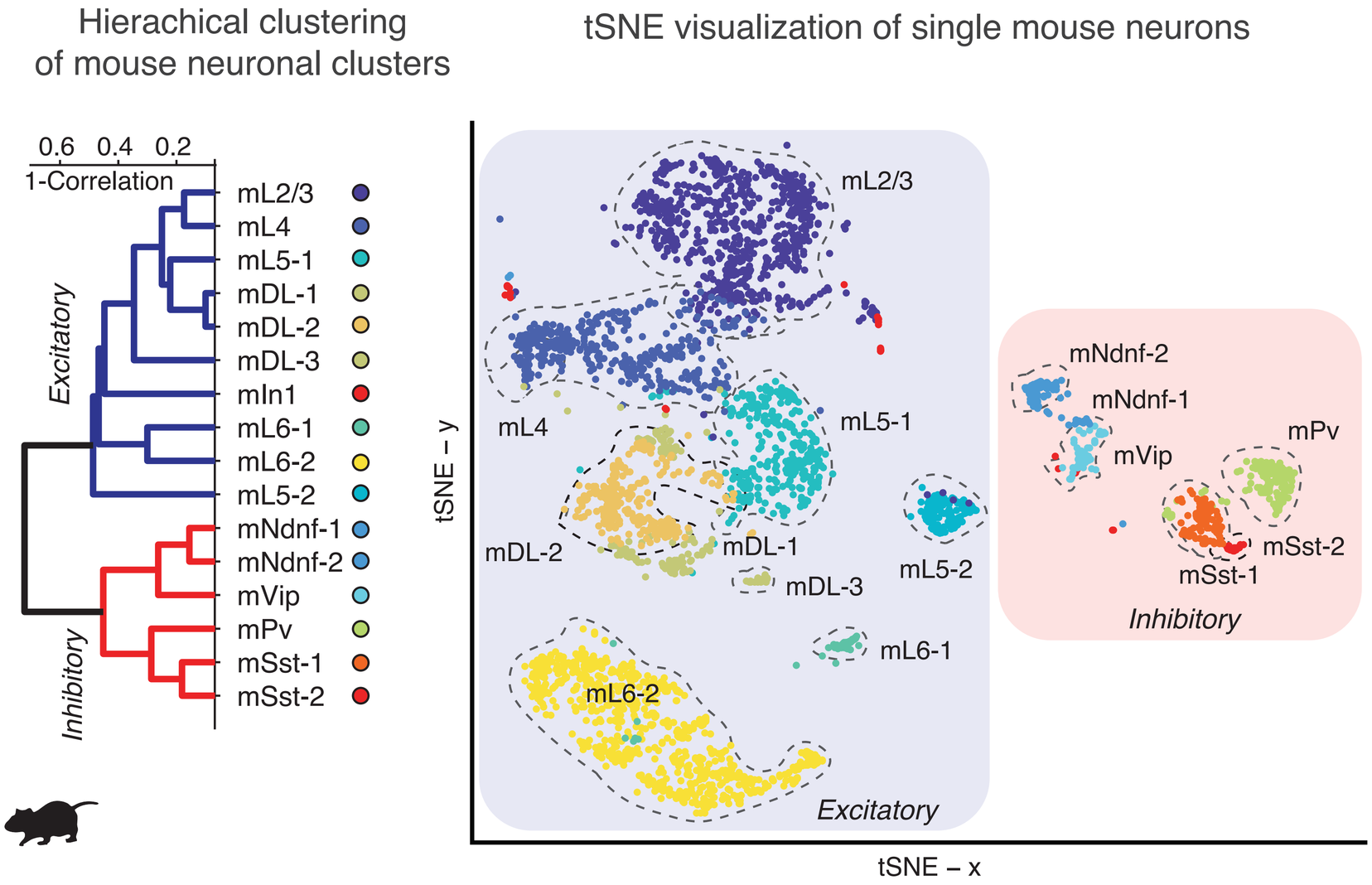


Co-clustering of purified bulk cells reveals cluster identity

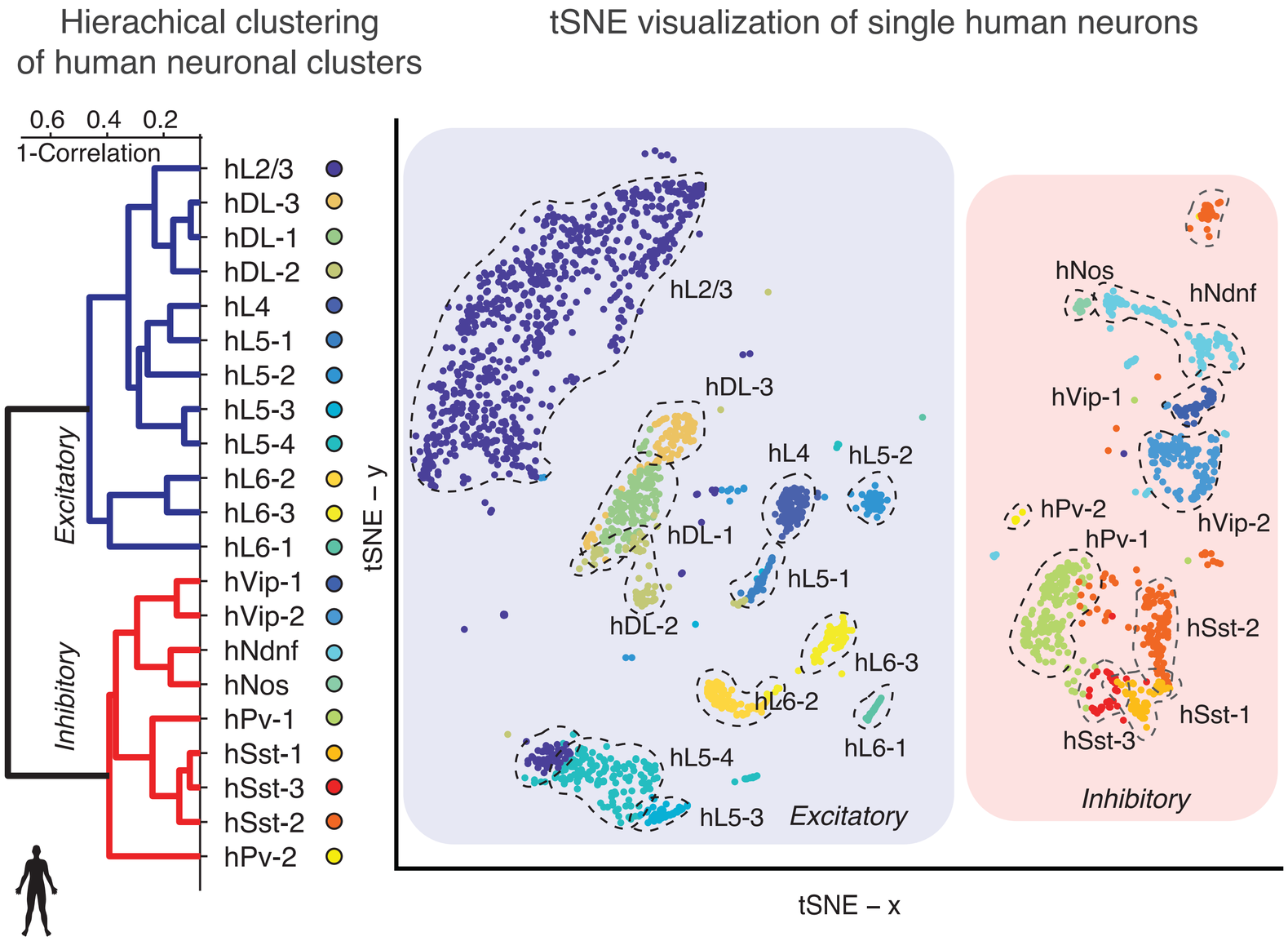


Bulk samples VIP+, Paravalbumin+
and Somatostatin+ inhibitory neurons (Mo 2015)

Identified 16 clusters in mouse with greater diversity in deep layers.



Identified 21 clusters in human, showing increased diversity over mouse.



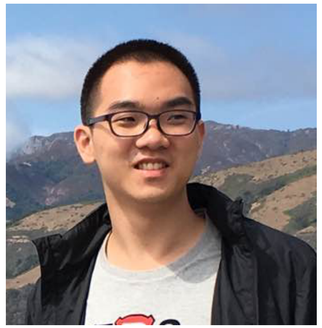
Single cell project team

Eran Mukamel

Joseph Ecker

M. Margarita Behrens

Chongyuan Luo



Jingtian Zhou

Junhao Li

Jacinta Lucero

Joe Nery

Rosa Castanon

Laurie Kurihara
Yupeng He
Justin P. Sandoval
Brian Bui
Terrence J. Sejnowski
Timothy T. Harkins

UC San Diego

salk®
Where cures begin.

Swift
BIOSCIENCES™

In summary

- Genomics is an exciting area with many open questions and lots of public data.
- San Diego and the Bay Area are huge in biotech.
- Evolution and development are computational processes that produce intelligence.
- Hopefully you understand your 23andMe better now.
- Contact me if you have questions: christopher.keown@gmail.com

Thank you.