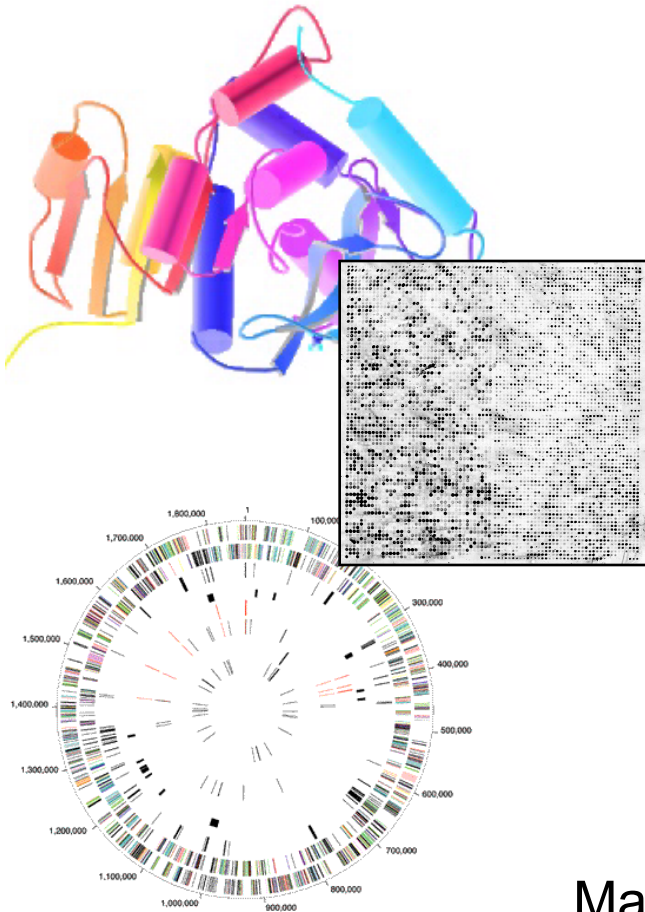


BIOINFORMATICS

Personal Genomes Intro.



Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring '19, MG lecture #2, final edit)

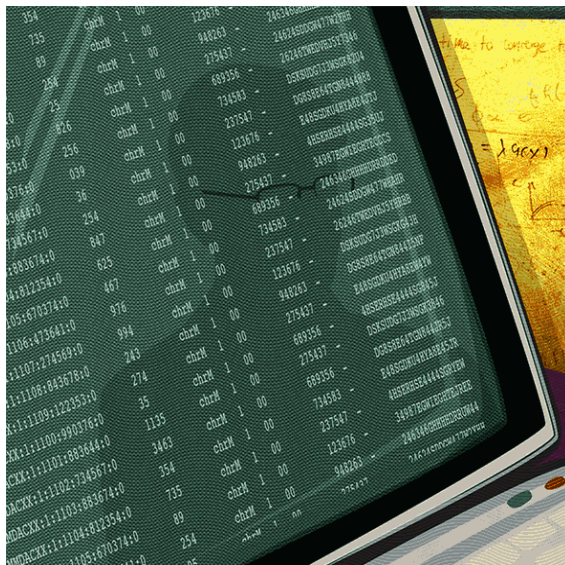
Personal Genomics

as an an organizing theme for this class

- A personal genome can reveal a lot about an individual.
 - Disease risks, ancestry, personal traits, etc.
- Personal genome annotation combined with multi-omic and longitudinal health data can inform new links between genotype and phenotype relevant to an individual and the larger population.
- Genomic privacy will become increasingly important as precision medicine becomes more common.
- In this class, we will look at how to identify key genomic variants with the most impact.
- We will also use analysis techniques including systems and network modeling as well as structural modeling to contextualize and interpret the mechanisms through which these variants impact health.

Analyzing Carl Zimmer's genome

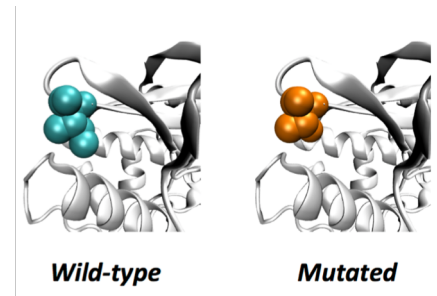
CARL ZIMMER'S GAME OF GENOMES SEASON 1



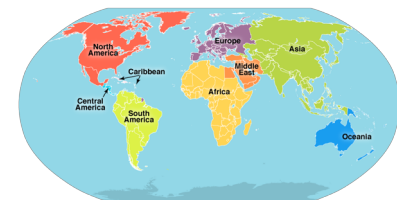
SNV

AAGCT → ACGCT

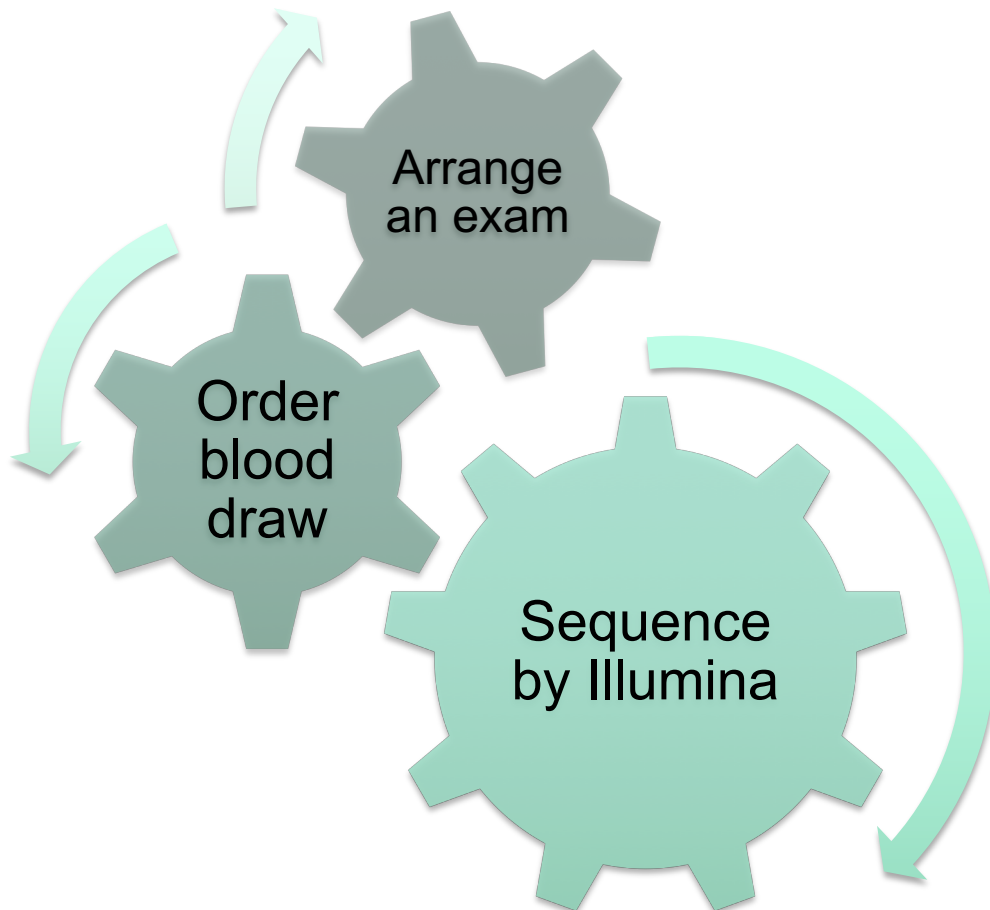
Protein
Structure



Ancestry

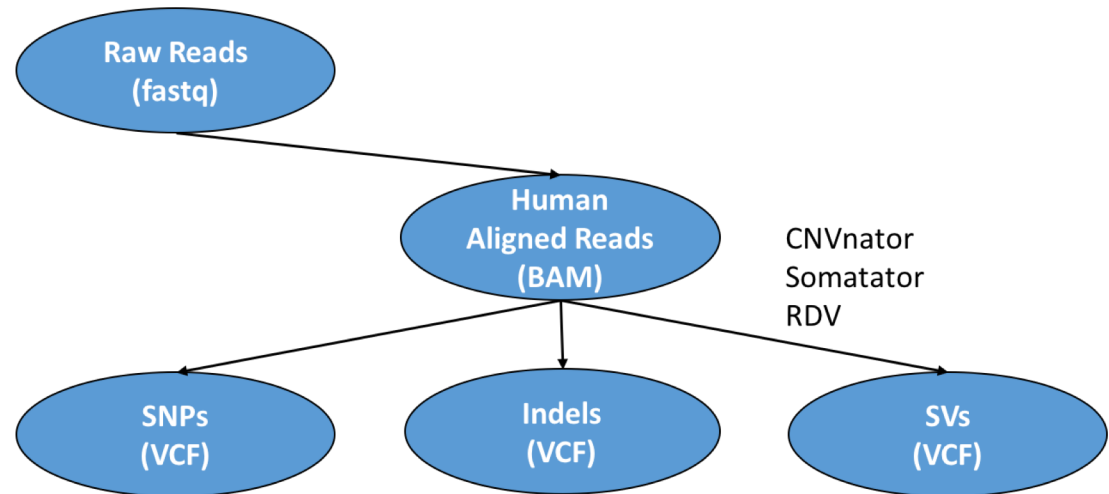


CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



- **Cost: \$3100**
- **Illumina briefly review the sequencing data, evaluating the risk for 1200 disorders, from familiar ones like lung cancer to obscure ones like cherubism**

CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



Genome Variation

TP53 Sequence:

...GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT...

Single Nucleotide Polymorphism (SNP) – 1nt:

...GGAGTCTTCCAGTGTGATGATGGT**G**AGGATGGGCCTCCGGTT...

T or A or C

Small Insertions and DEletions (INDEL) – 1-10nt:

...GGAGTCTTCCAGTGTGATGATGGT~~GAGGATG~~GGCCTCCGGTT...

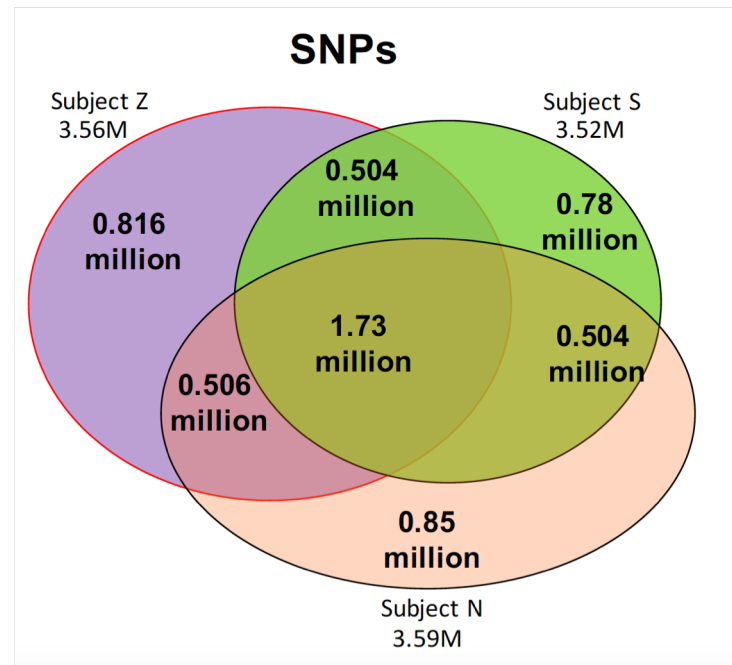
Large Structural Variations (SV) -- >100nt:

...GGAGTC~~TTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT~~...

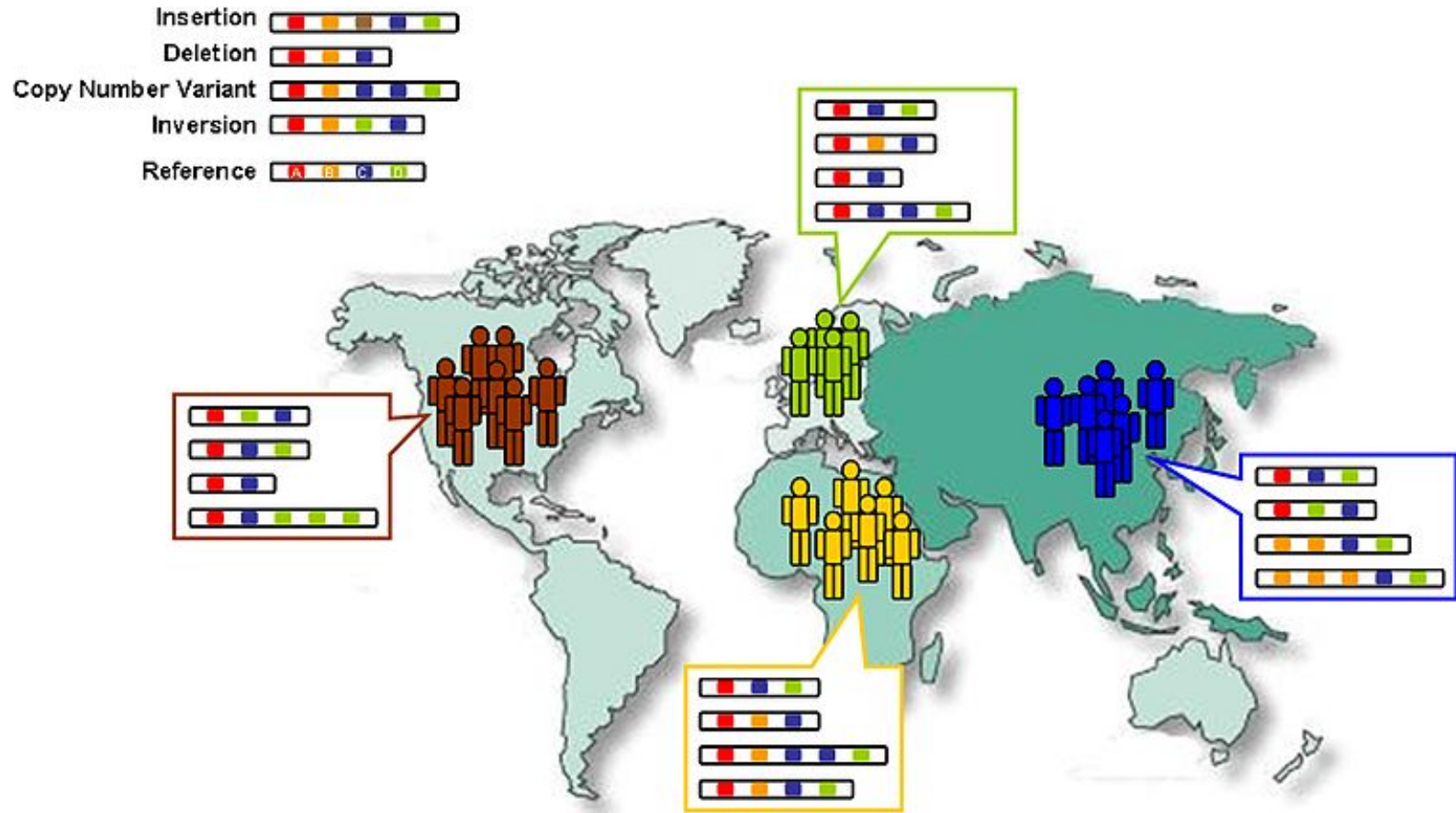
CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



- Normal range of number of SNPs
- Carl's case: more than 3M SNPs
- How do we know if the SNP is harmful?



- Thousand genome project
- Common SNP data base found in the population



CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



- **Got a variant in a gene for heart muscles, called DSG2**
- **DSG2 gene encodes a protein in humans called Desmoglein-2**
- **Mutations in desmoglein-2 have been associated with arrhythmogenic right ventricular cardiomyopathy**

1 in 200

People of European descent carry this variant

We're all different in our DNA. We're finally starting to understand when those differences matter ---- Carl Zimmer

Human Genetic Variation

A Cancer Genome



A Typical Genome



Population of 2,504 peoples



Origin of Variants

	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



Driver (~0.1%)

Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

Prevalence of Variants



Rare* (1-4%)

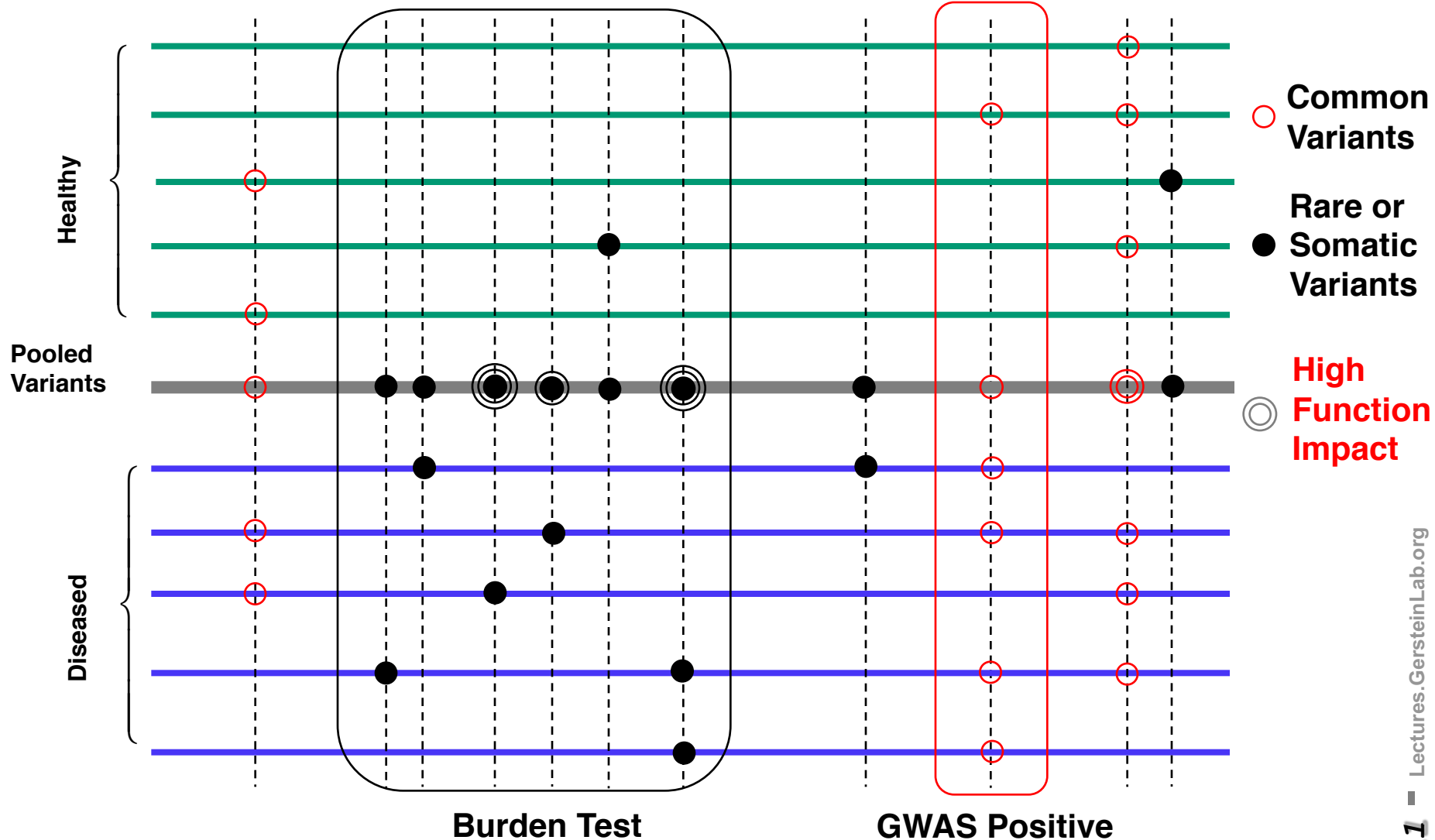
SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M



Rare (~75%)

* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

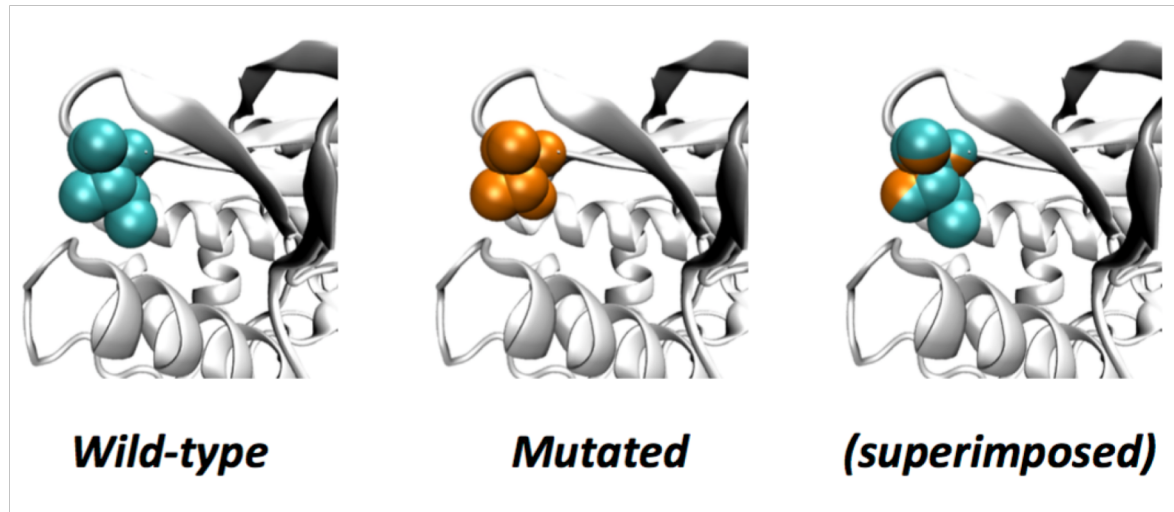
Association of Variants with Diseases



CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



SNP changing protein structure



- NAT2, an enzyme in the liver that breaks down caffeine and other toxins with a similar molecular structure.
- NAT2 helps break down certain medicines too. The variant puts people at risk of bad side effects from those drugs.

CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



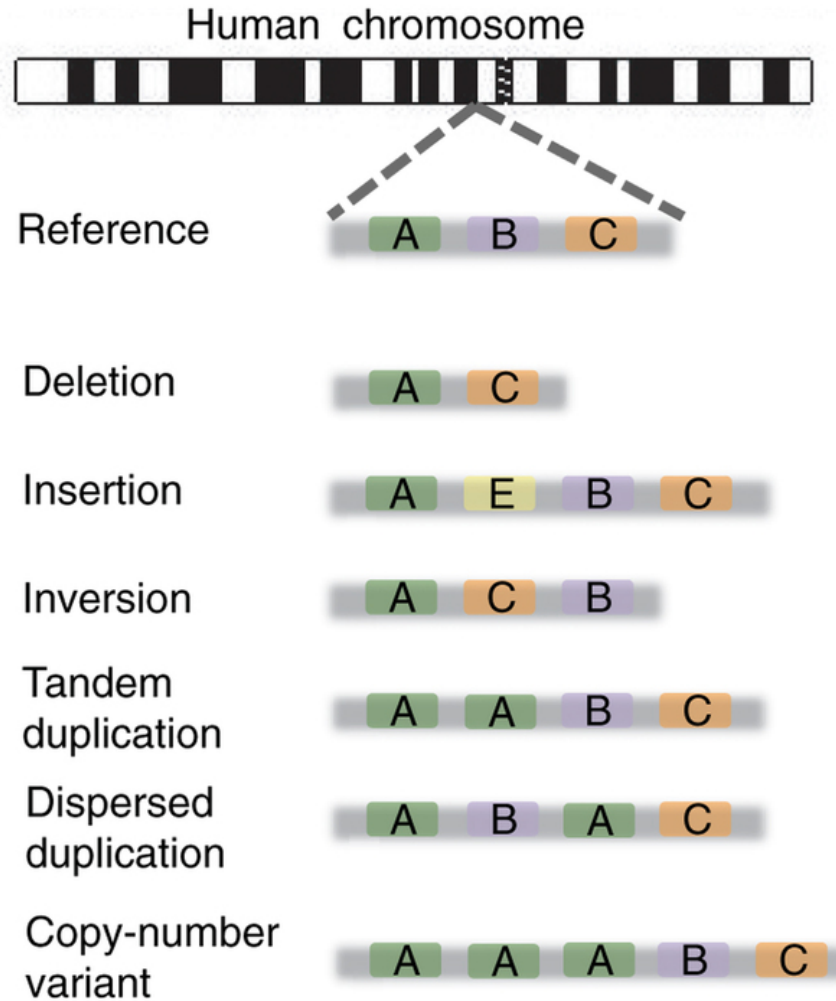
Indels (Insertions/deletions)

- In coding regions, unless the length of an indel is a multiple of 3, it will produce frameshift mutation
- Likely to disrupt genes (loss-of-function variant)

Example: Non-functional F8 gene

- Can't make essential clotting protein
- Get hemophilia and can bleed to death from a little cut

Structural Variation



CARL ZIMMER'S
GAME OF GENOMES
SEASON 1



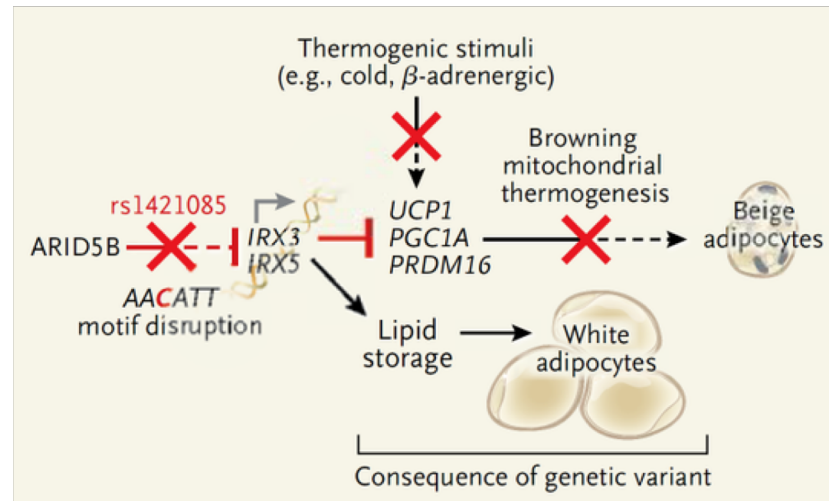
- Structural variation
- Example: HTT
- Certain mutations in HTT cause Huntington's disease.
- Healthy people have a wide range of CAG repeats. It's only when people get 37 or more CAG repeats in HTT that they are at risk of developing Huntington's disease.
- The reference genome has 19 CAG repeats. Carl has 17.

CARL ZIMMER'S
GAME OF GENOMES
SEASON 2



Non-coding variant

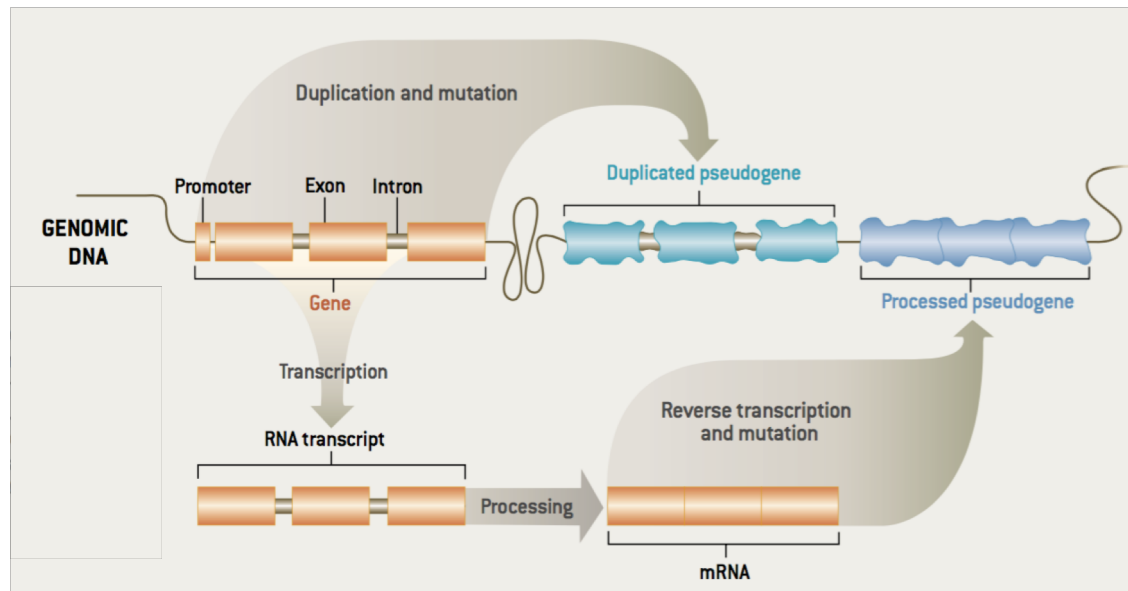
- Variant rs1421085
- Located in a genetic switch that activates several genes in fat cells
- The variant causes people to put on an average of 7 pounds



CARL ZIMMER'S
GAME OF GENOMES
SEASON 1

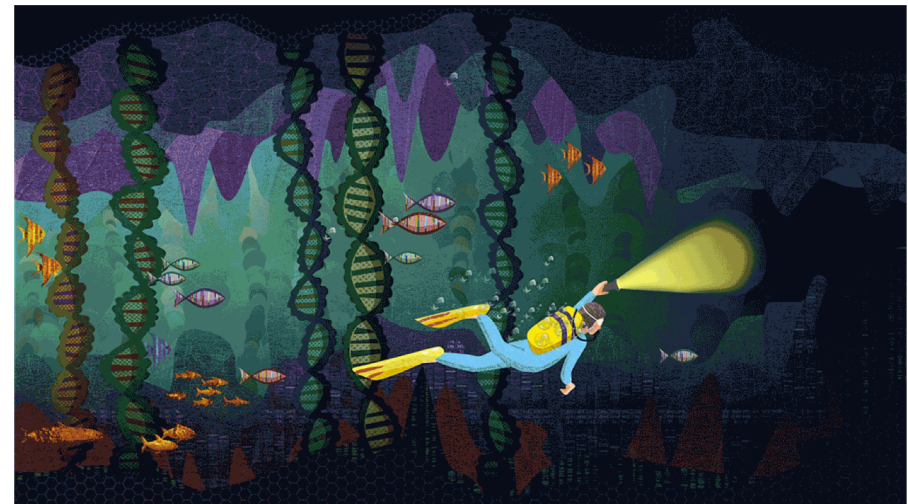


- What else are in the genome?
- Pseudogenes
- About 14000 pseudogenes carried in our genome



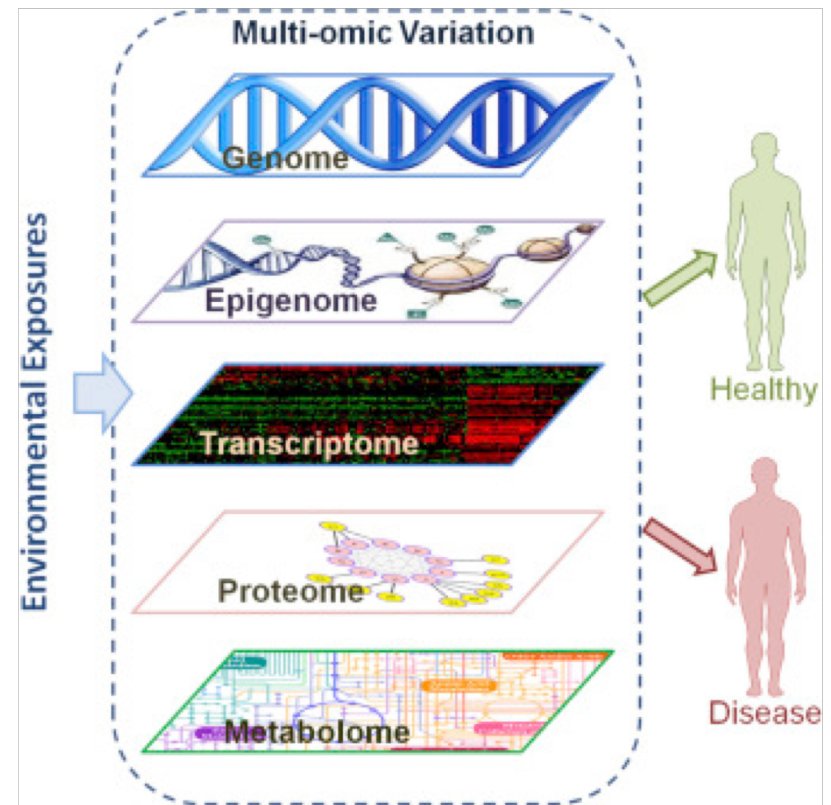
Looking beyond the genome

- In the Game of Genomes Carl Zimmer explored his genomic sequence.
- The genome provides a mostly static view.
- Misses the active regulation, transcription, and translation



Integrating environmental factors, genetic background, and large scale datasets

- Difference between health and disease depends on many factors.
- Environment, genome, cellular contents, etc. all play a role.
- Important to integrate information from multiple large-scale datasets.



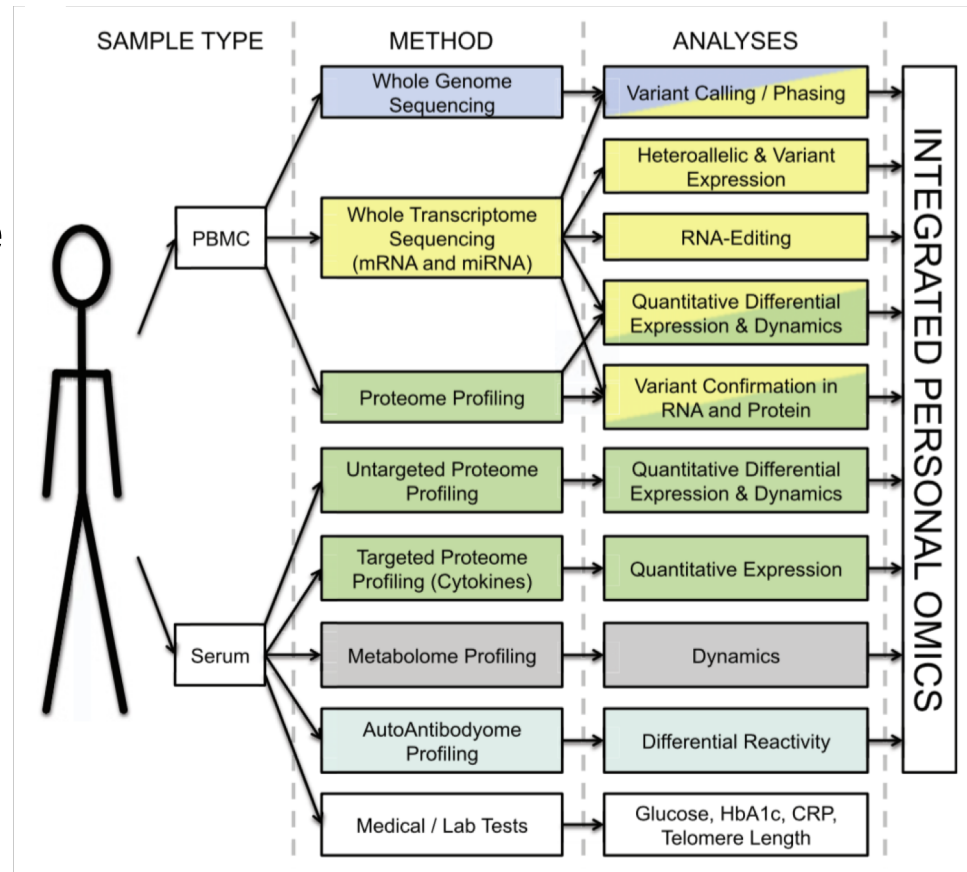
Expanding personalized medicine beyond the genome.

- An integrated personal omics profile (iPOP) is an example of a more comprehensive version of personalized medicine.
- Michael Snyder had his genome sequenced and collected many other large scale datasets over an extended period of time.



Integrated personal omics profile (iPOP)

- Numerous types of data were collected, primarily from blood samples. The datasets include:
 - Transcriptomic
 - Proteomic
 - Metabolomic
 - Cytokine profiling
 - Autoantibody profiling
 - Medical exams



Prioritizing variants by leveraging mutation databases

- Using existing databases of population level genetic variation, rare and disease associated variants could be identified.
- Helped prioritize medical conditions for monitoring (e.g. glucose for diabetes)

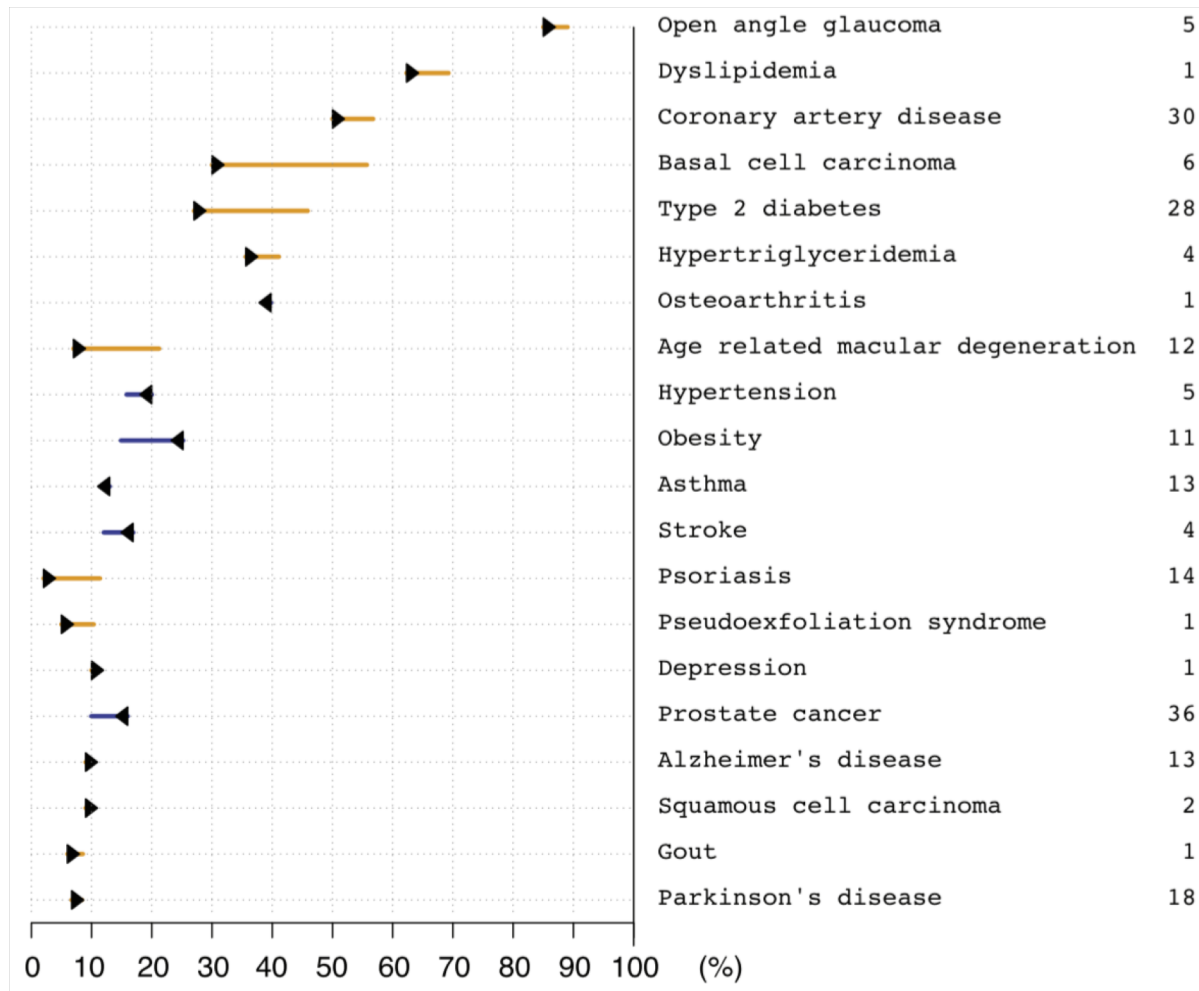
High Interest Disease-Associated Rare Variants.

Gene	Position	Genotype	OMIM
SERPINA1	14:94844947	C/T	Emphysema due to AAT deficiency
TERT	5:1294397	C/T	Aplastic anemia
KCNJ11	11:17409571	T/T	Type 2 diabetes
GCKR	2:27730939	T/T	Hypertriglyceridemia
NUP54	4:77055431	G/A	Nuclear Pore Complex Protein

High Interest Drug-Related Variants.

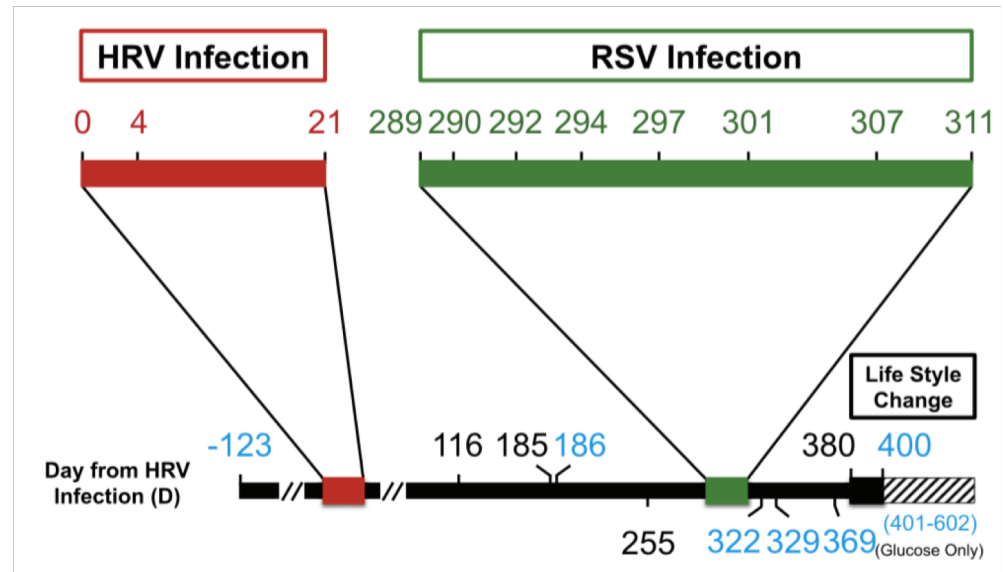
Gene	rsID	Genotype	Drug Response Affected
	rs10811661	C/T	Troglitazone (Increased Beta-Cell Function)
CYP2C19	rs12248560	C/T	Clopidogrel (Increased Activation)
LPIN1	rs10192566	G/G	Rosiglitazone (Increased Effect)
SLC22A1	rs622342	A/A	Metformin (Increased Effect)
VKORC1	rs9923231	C/T	Warfarin (Lower Dose Required)

Genomic information helps refine disease risk estimates

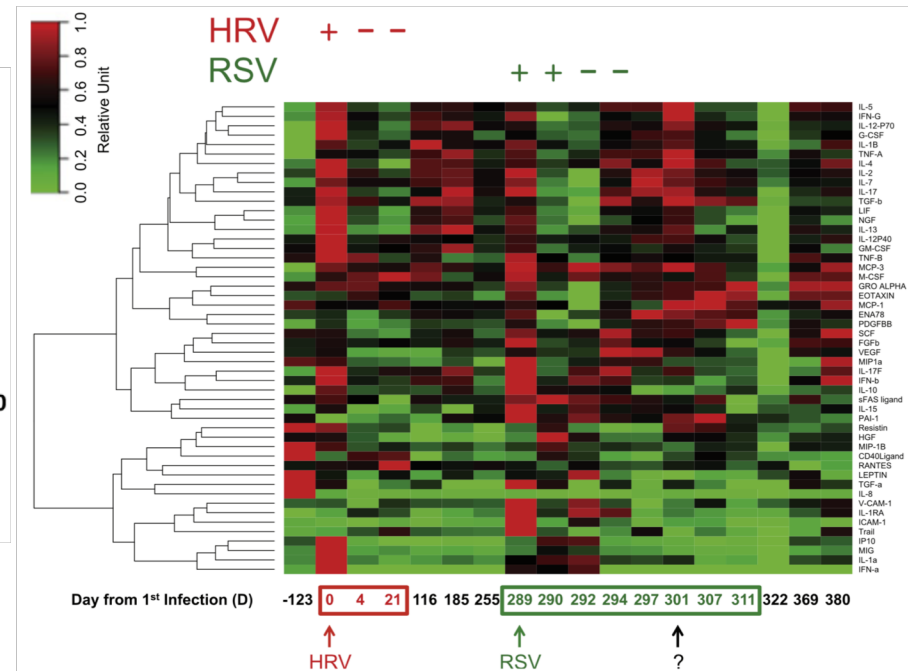
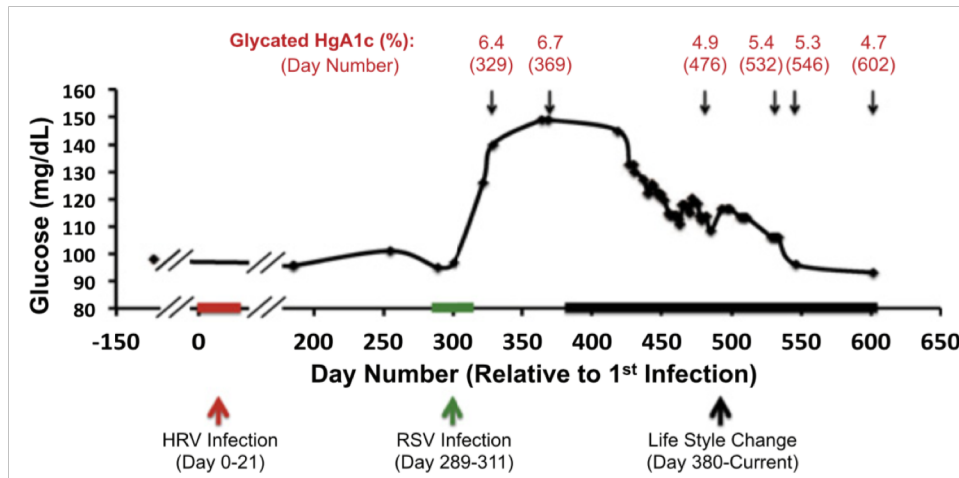


Personal omics profiling time course

- Changing cellular state and environmental perturbations impact the genome.
- Longitudinal data collection tracks the dynamic regulation of the genome.



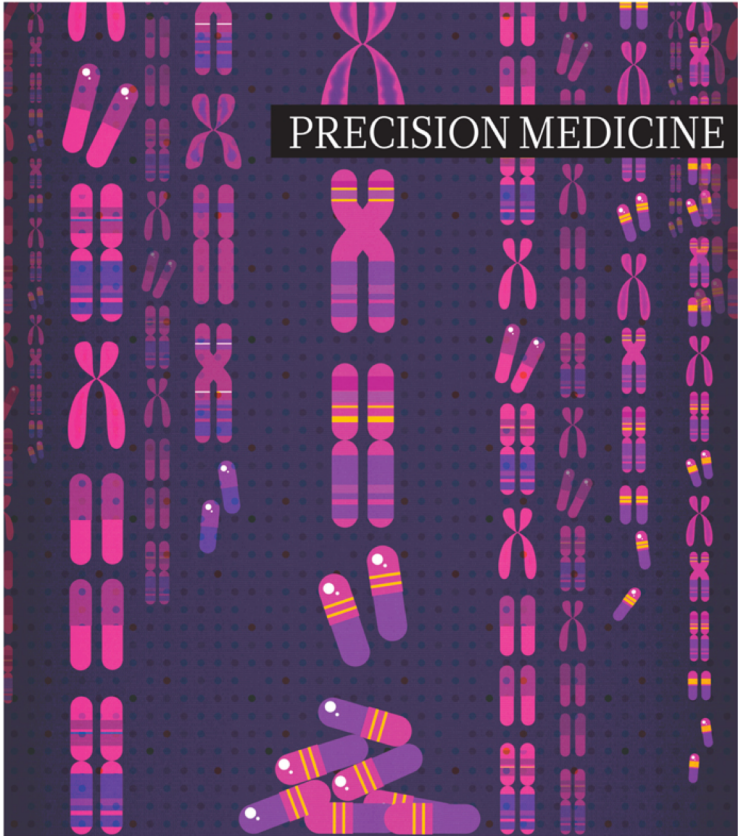
Longitudinal medical data



- Tracking relevant medical (e.g. blood glucose) data over time helps link phenotypic changes with changes at the molecular level.

Precision medicine in the clinic

- Precision medicine is leading to better defining and treating disease at the molecular level.
- It is both changing the prescription of existing medications and inspiring new targeted therapies.
- Precision medicine requires high quality patient genome sequences be obtained at reasonable cost.



The graphic features a dark blue background with a pattern of colorful chromosomes in shades of pink, purple, and yellow. A black banner with the text "PRECISION MEDICINE" is positioned in the upper right. The "nature" part of the logo is in a serif font, and "OUTLOOK" is in a bold, sans-serif font.

natureOUTLOOK

PRECISION MEDICINE

Produced with support from:

illumina

A personal approach
to health care