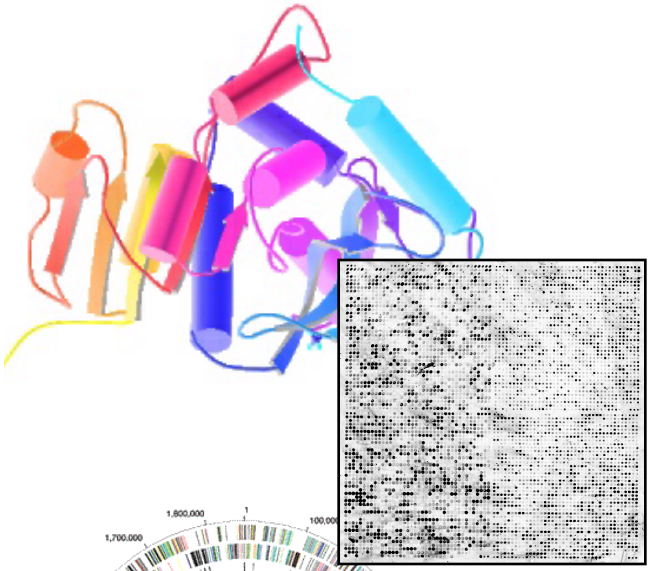
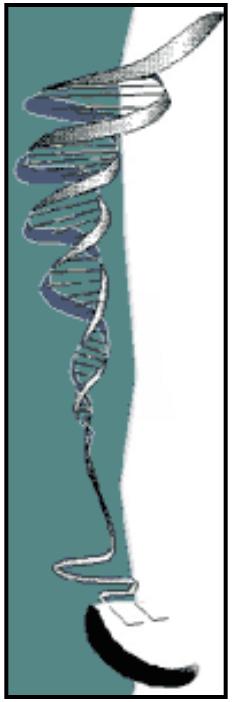
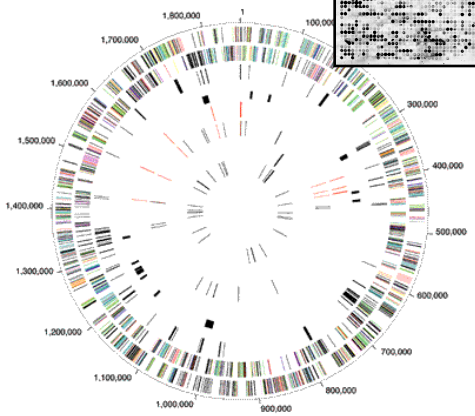
Biomed. Data Sci.
Personal Genomes Intro.

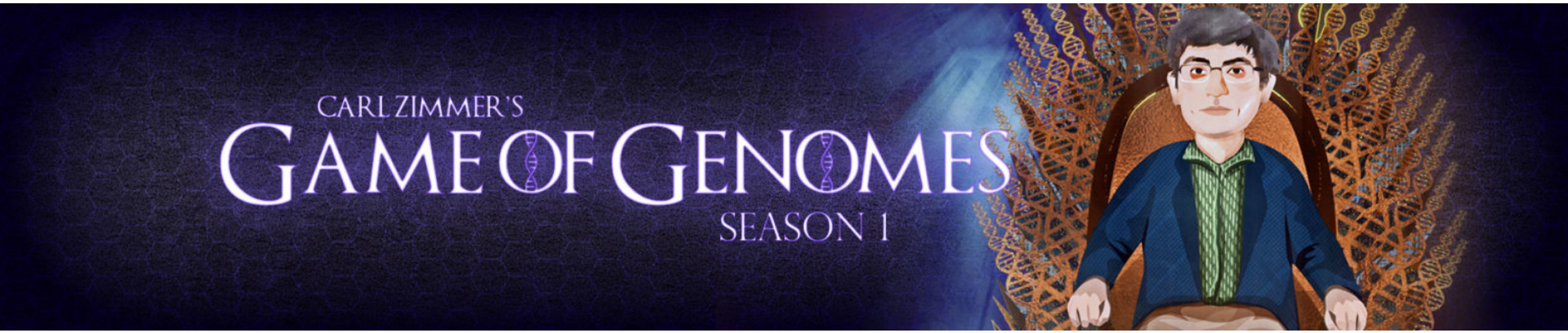


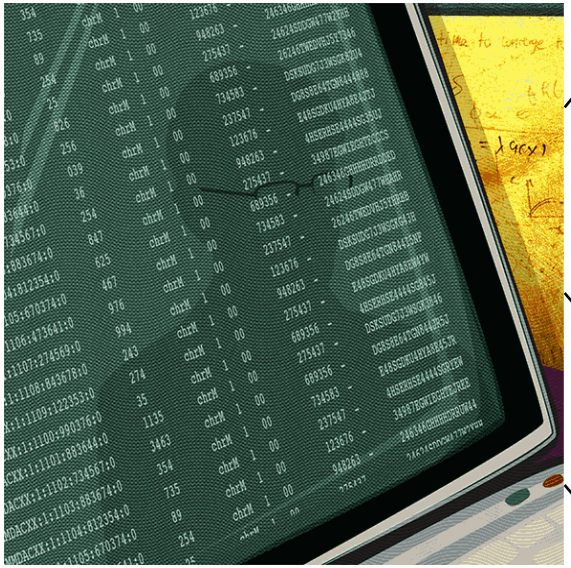




Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring ‘20)

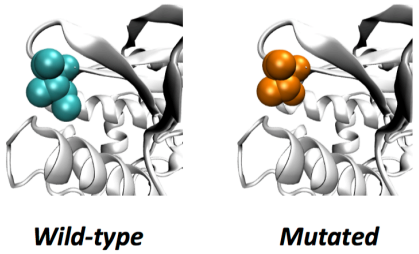
Analyzing Carl Zimmer's genome



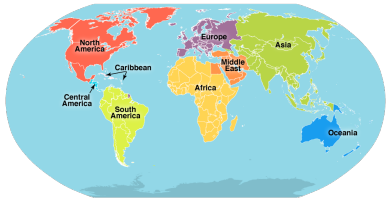


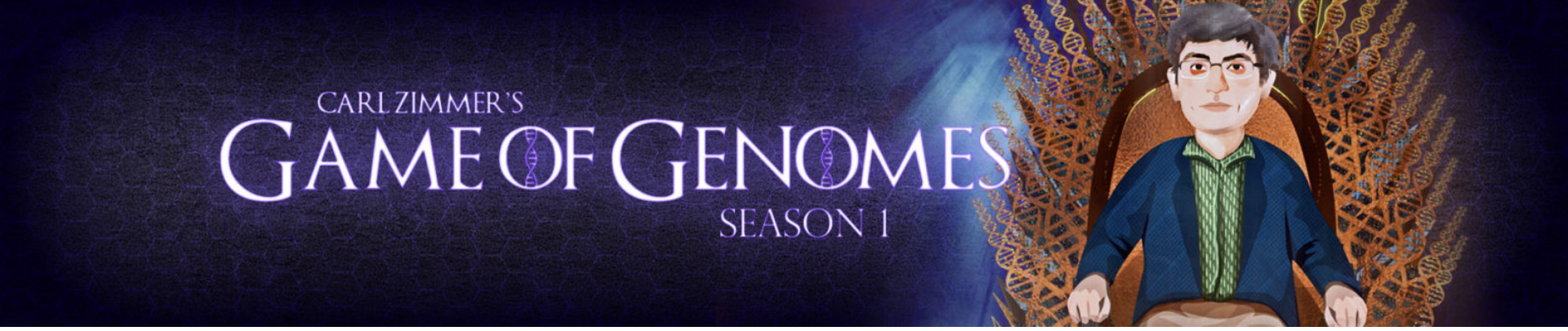
SNV AAGCT → ACGCT

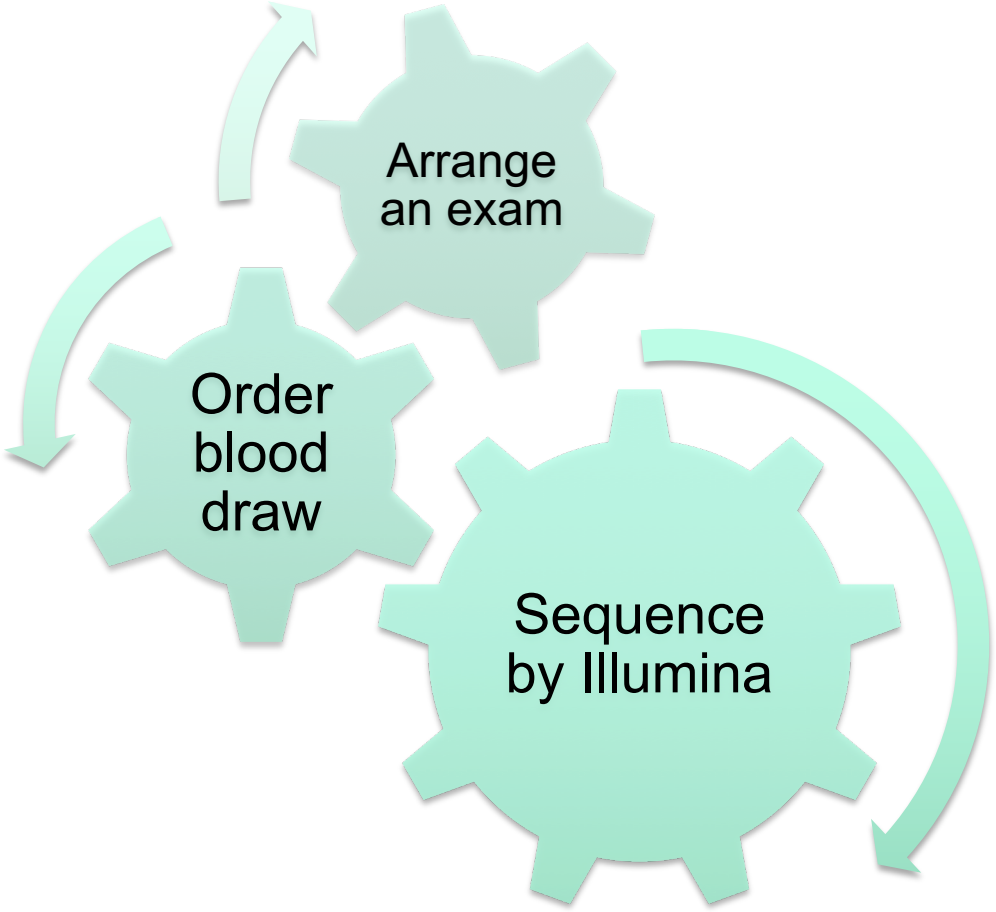
Protein Structure



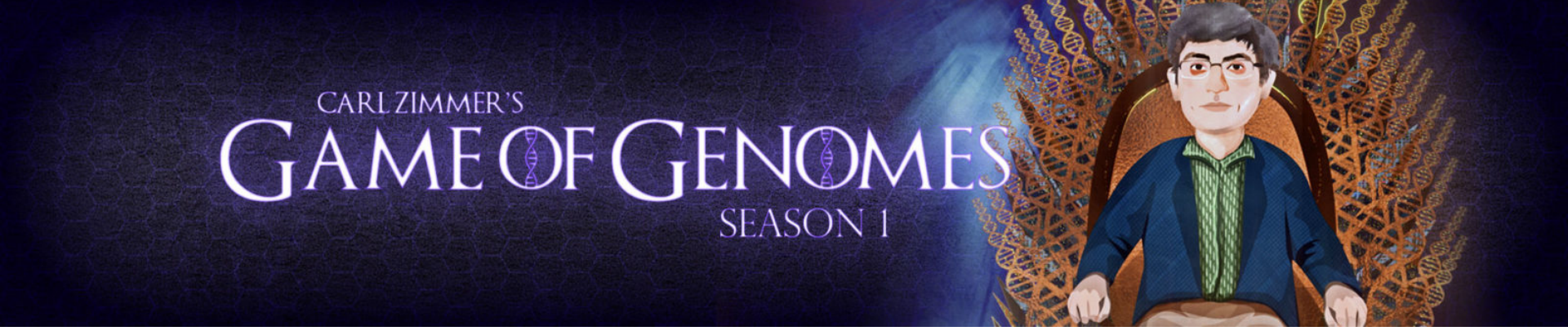
Ancestry

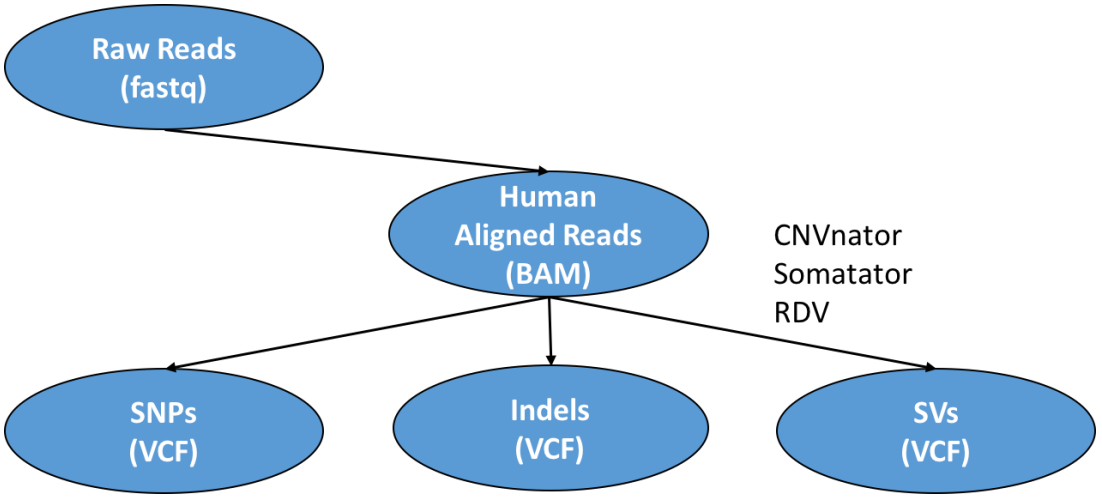






- **Cost: $3100**
- **Illumina briefly review the sequencing data, evaluating the risk for 1200 disorders, from familiar ones like lung cancer to obscure ones like cherubism**


CARL ZIMMER'S
GAME OF GENOMES
SEASON 1


Raw Reads
(fastq)
Human
Aligned Reads
(BAM)
CNVnator
Somatator
RDV
SNPs
(VCF)
Indels
(VCF)
SVs
(VCF)

Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) –1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

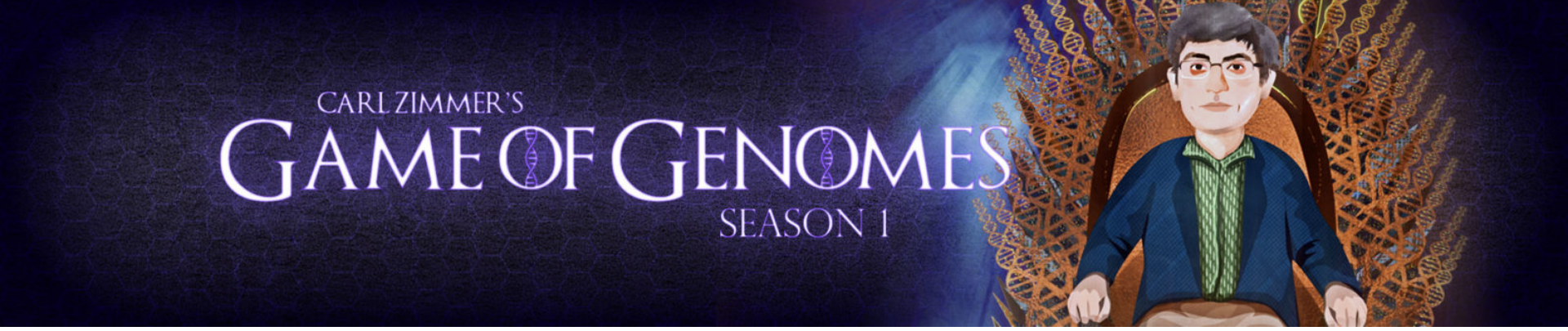
T or A or C

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

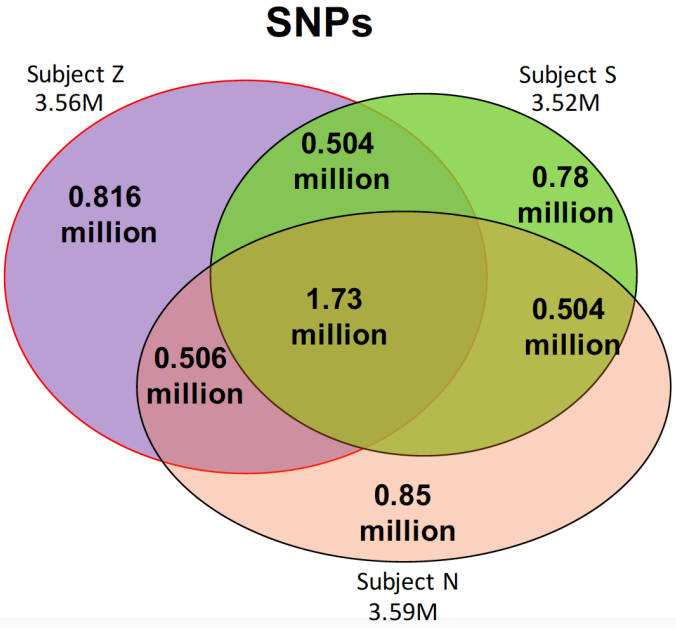
…GGAGTCTTCCAGTGTGATGATGGT~~GAGGATG~~GGCCTCCGGTT…

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

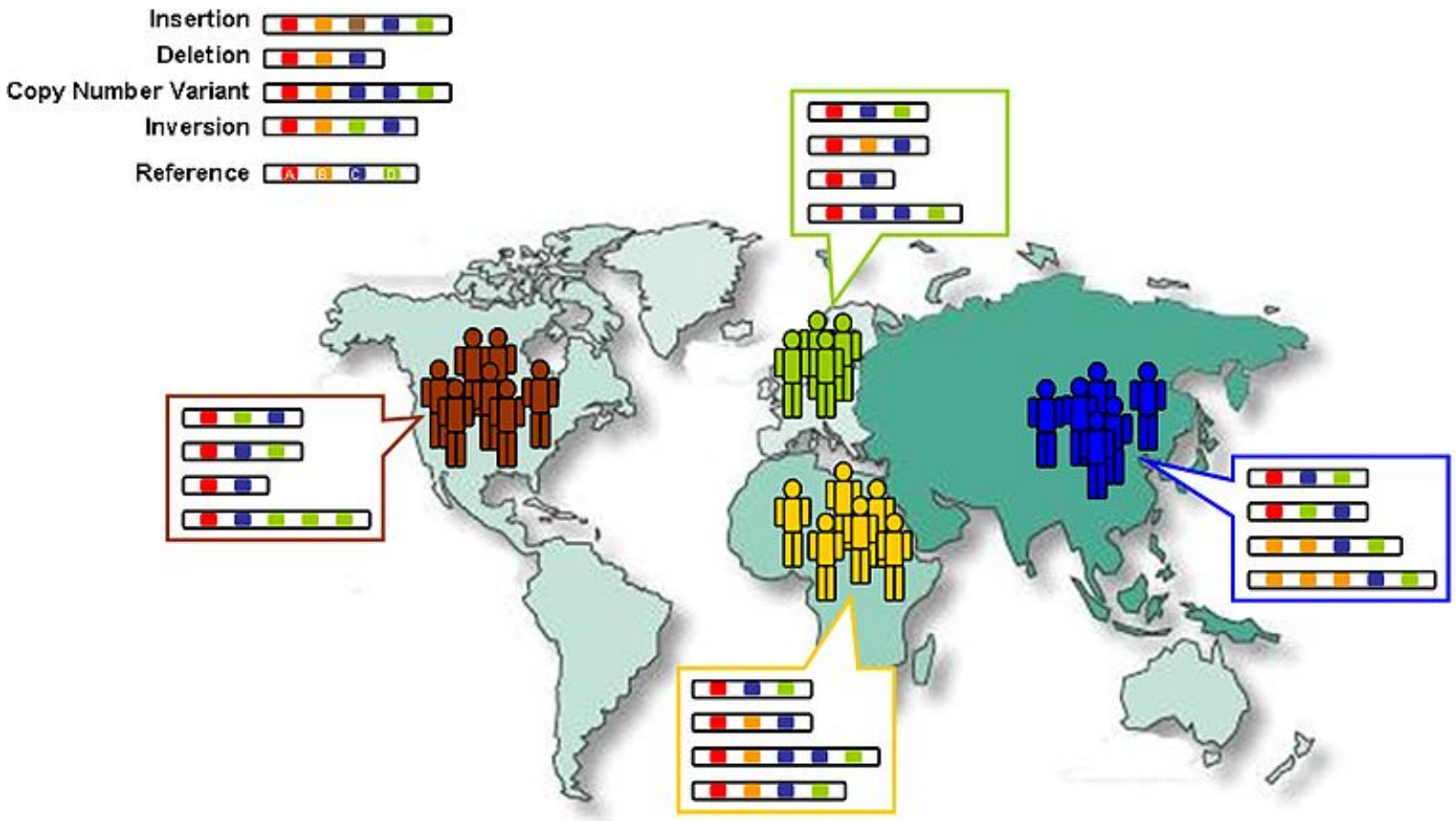
…GGAGTC~~TTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…~~



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

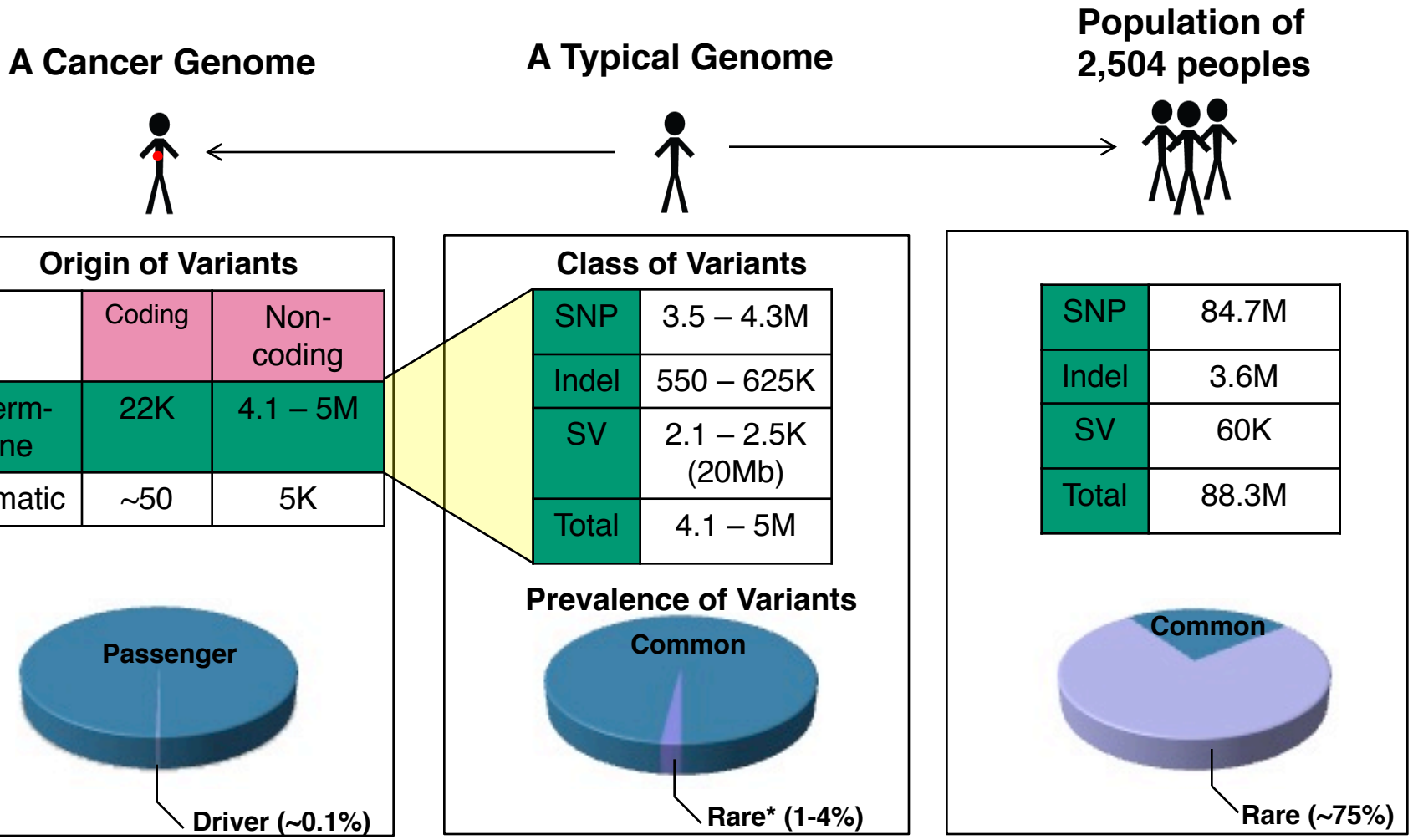


- Thousand genome project
- Common SNP data base found in the population



https://en.wikipedia.org/wiki/1000_Genomes_Project

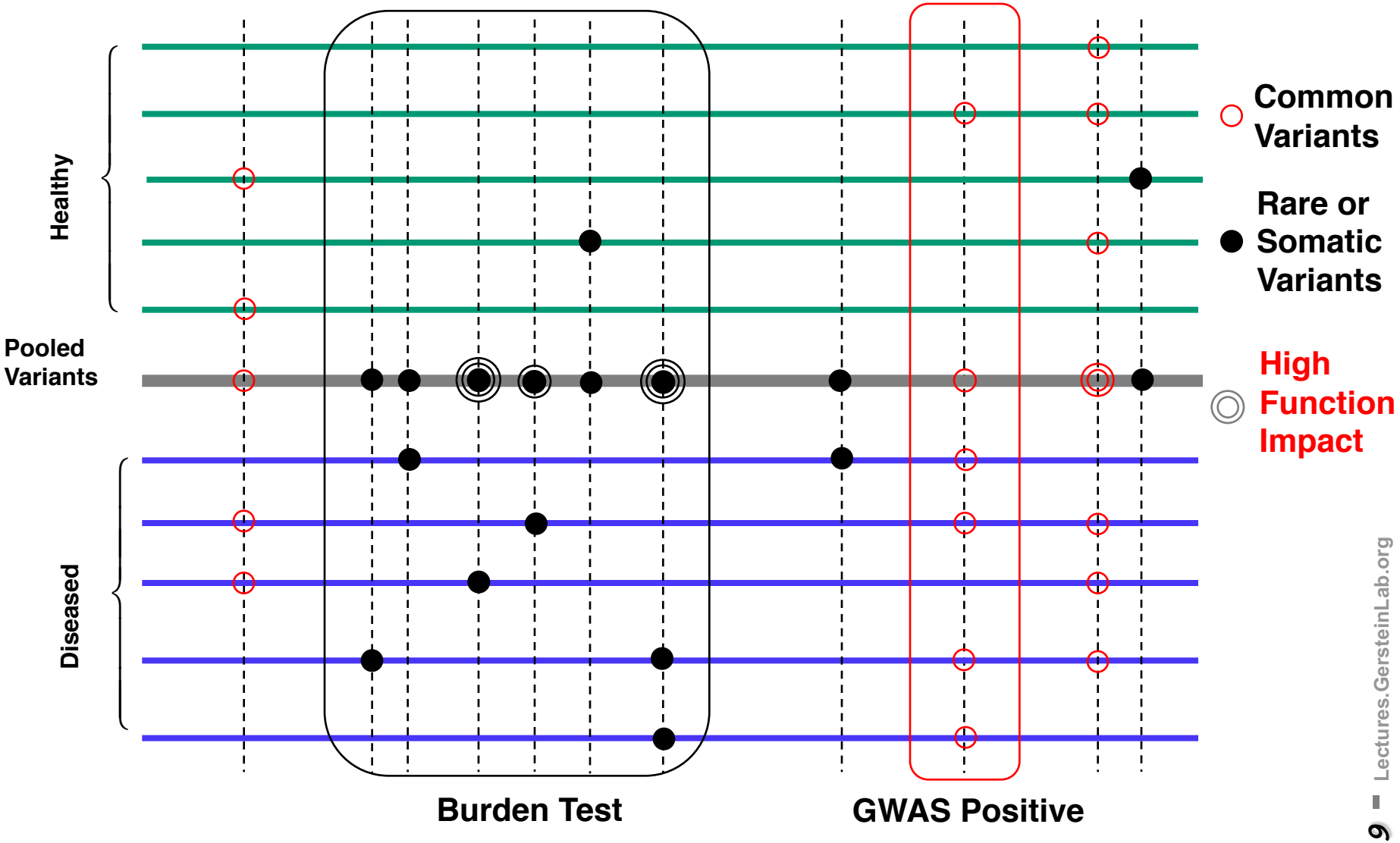
Human Genetic Variation

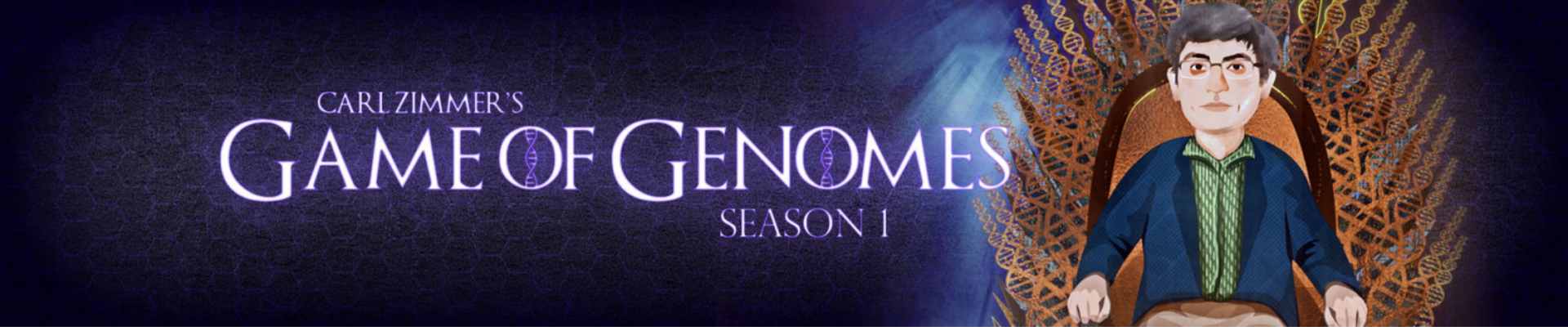


* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Association of Variants with Diseases



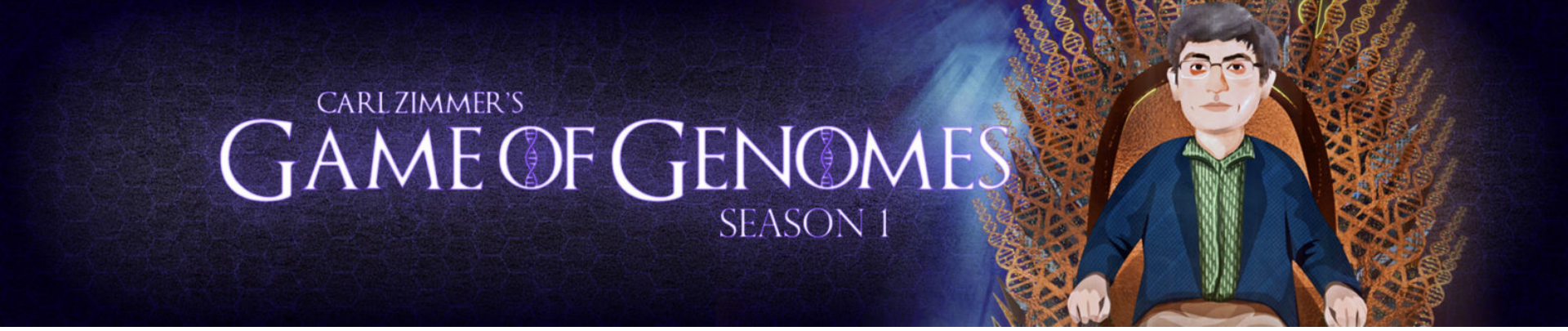


- **Got a variant in a gene for heart muscles, called DSG2**

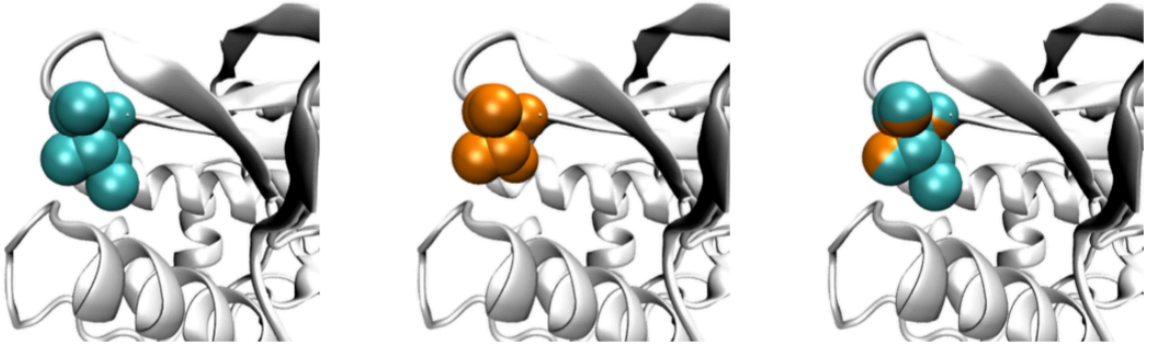
- **DSG2 gene encodes a protein in humans called Desmoglein-2**
- **Mutations in desmoglein-2 have been associated with arrhythmogenic right ventricular cardiomyopathy**

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



SNP changing protein structure

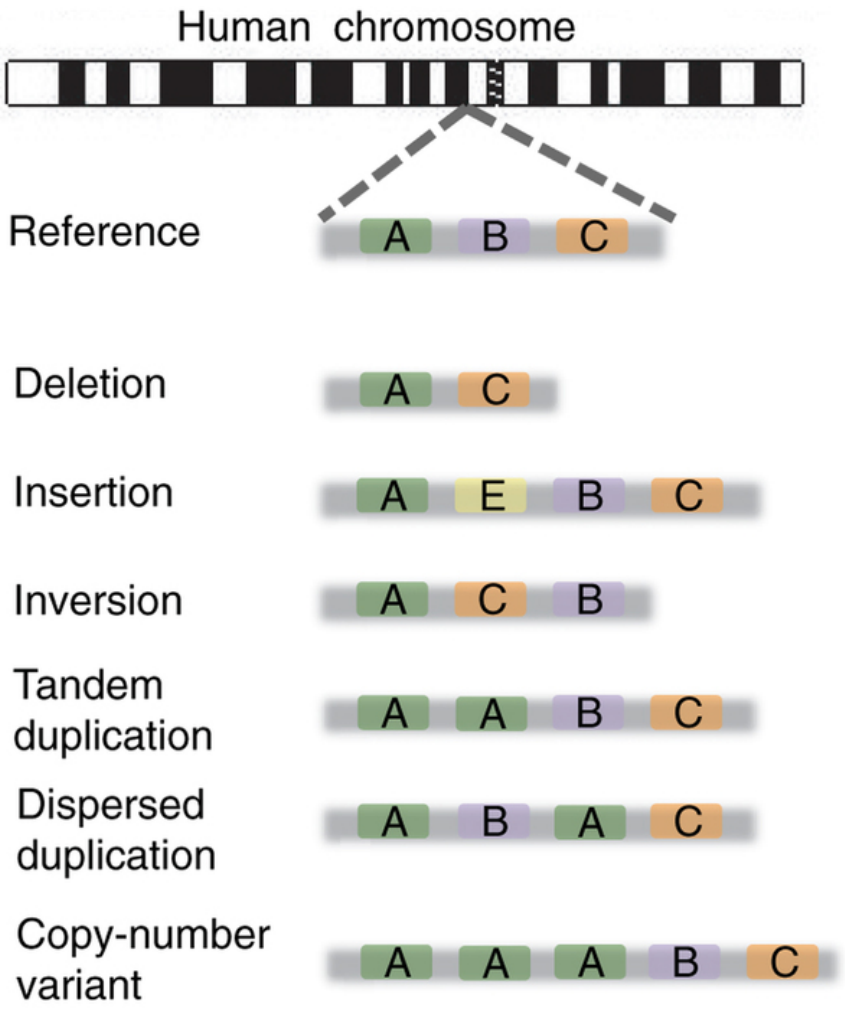


Wild-type *Mutated* *(superimposed)*

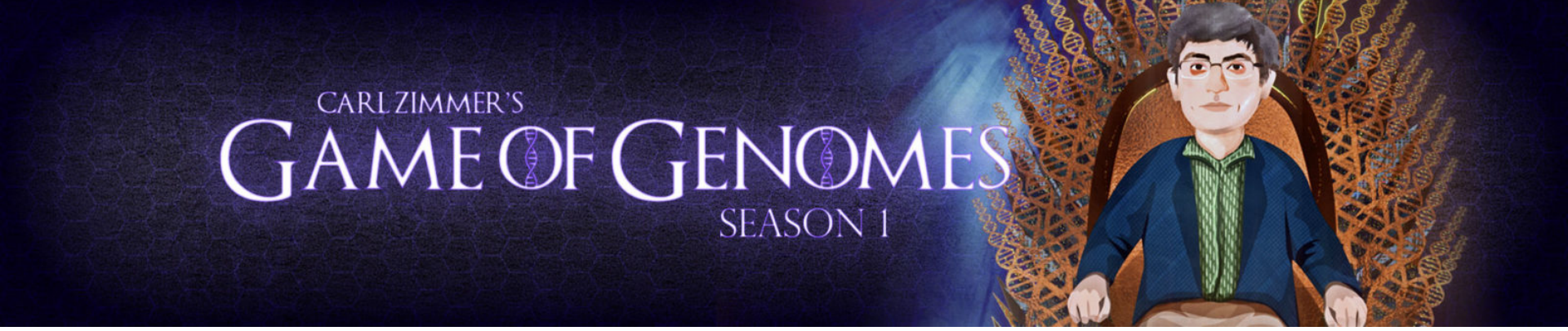
114: I->T

- 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.

Structural Variation

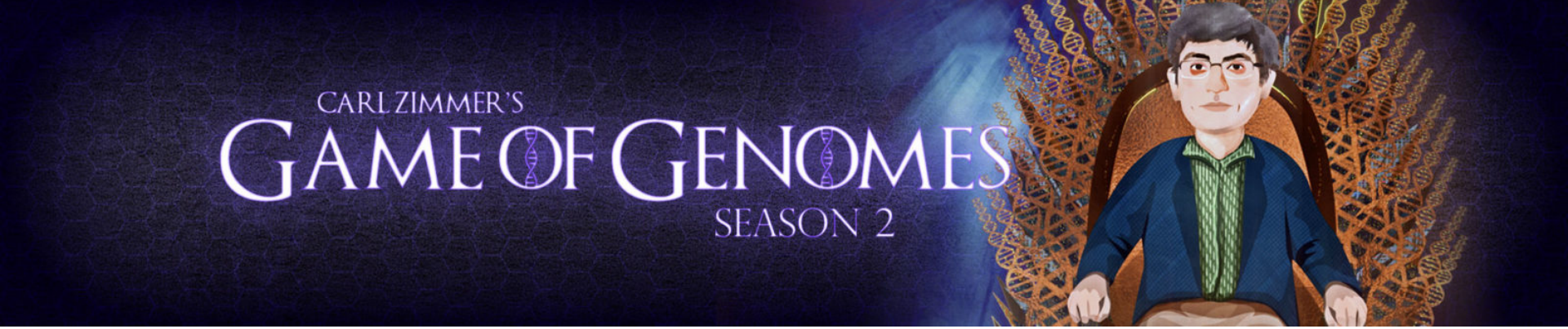


Baker, M., 2012. *Nature methods*, *9*(2), pp.133-137.



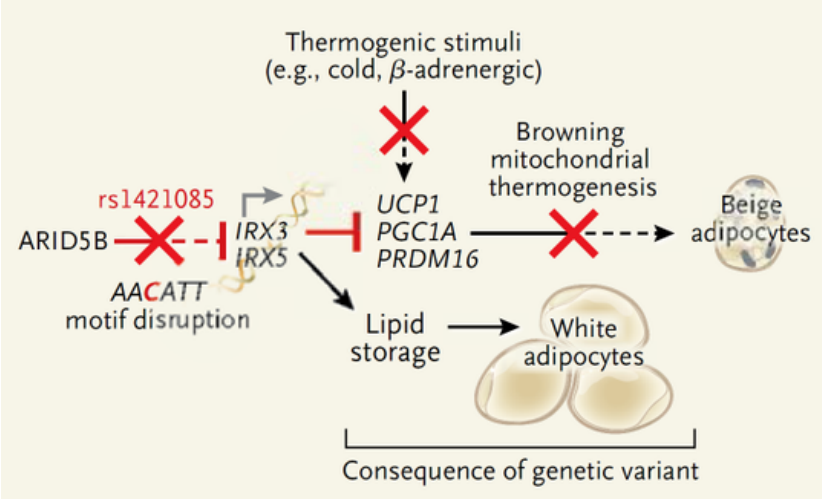
- Structural variation
- Example: HTT
- Certain mutations in HTT cause Huntington's disease.
- Healthy people have a wide range of CAG repeats. It's only when people get 37 or more CAG repeats in HTT that they are at risk of developing Huntington's disease.

- The reference genome has 19 CAG repeats. Carl has 17.



Non-coding variant

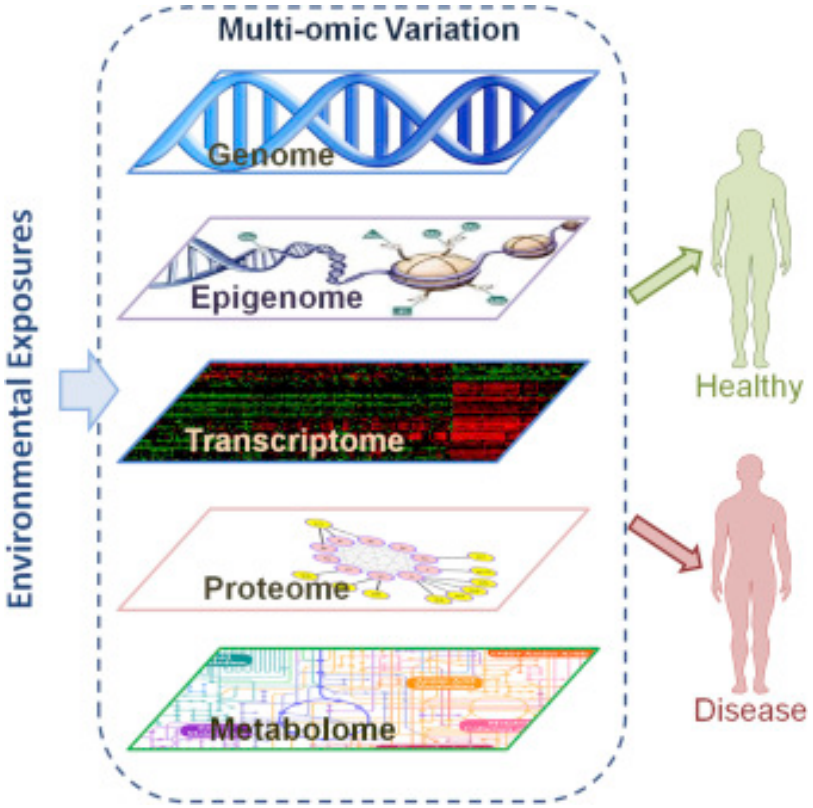
- Variant rs1421085
- Located in a genetic switch that activates several genes in fat cells
- The variant causes people to put on an average of 7 pounds



Claussnitzer, Melina, et al. *N Engl J Med* 2015.373 (2015): 895-907.

Integrating environmental factors, genetic background, and large-scale datasets

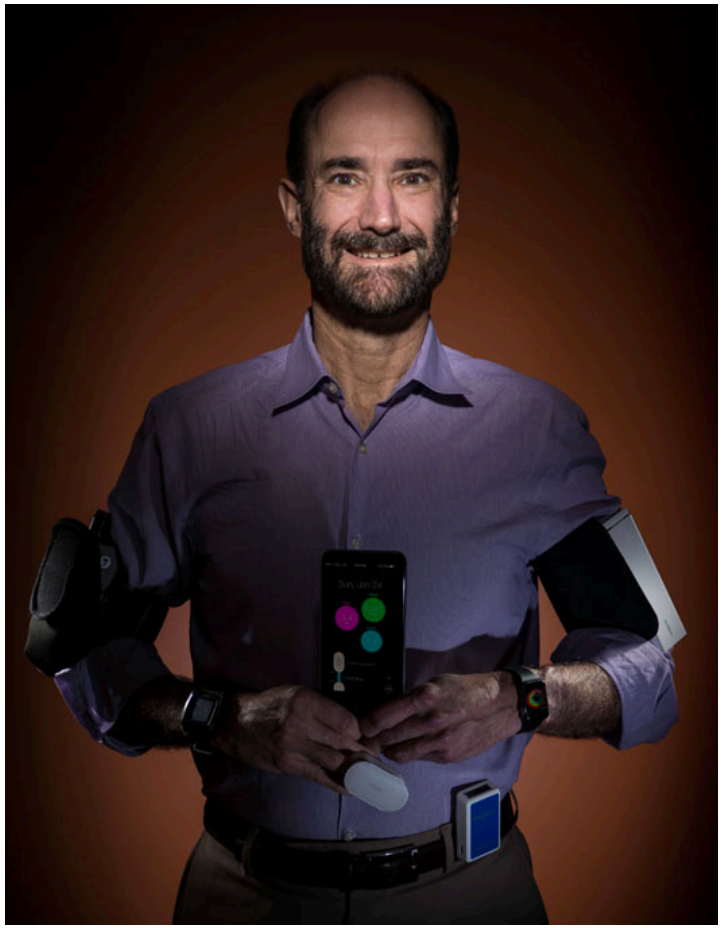
- Difference between health and disease depends on many factors.
- Environment, genome, cellular contents, etc. all play a a role.
- Important to integrate information from multiple large-scale datasets.



Sun et al. Advances in Genetics 2016

Expanding personalized medicine beyond the genome.

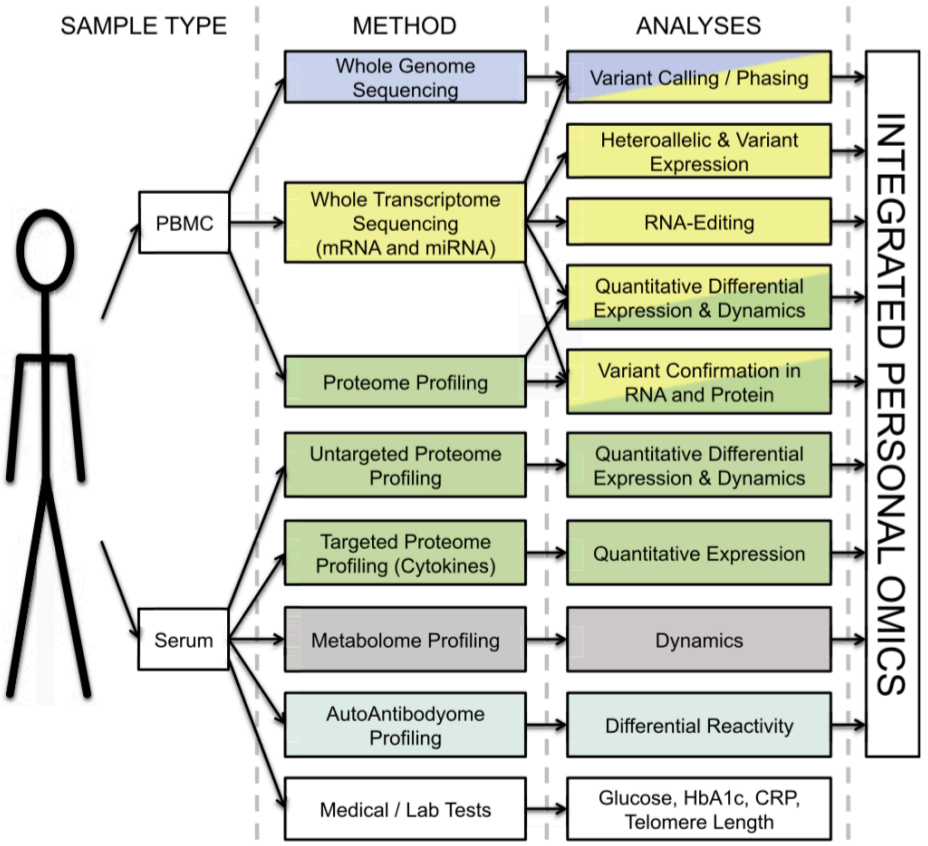
- An integrated personal omics profile (iPOP) is an example of a more comprehensive version of personalized medicine.
- Michael Snyder had his genome sequenced and collected many other large scale datasets over an extended period of time.



https://med.stanford.edu/news/all-news/2017/01/wearable-sensors-can-tell-when-you-are-getting-sick.html

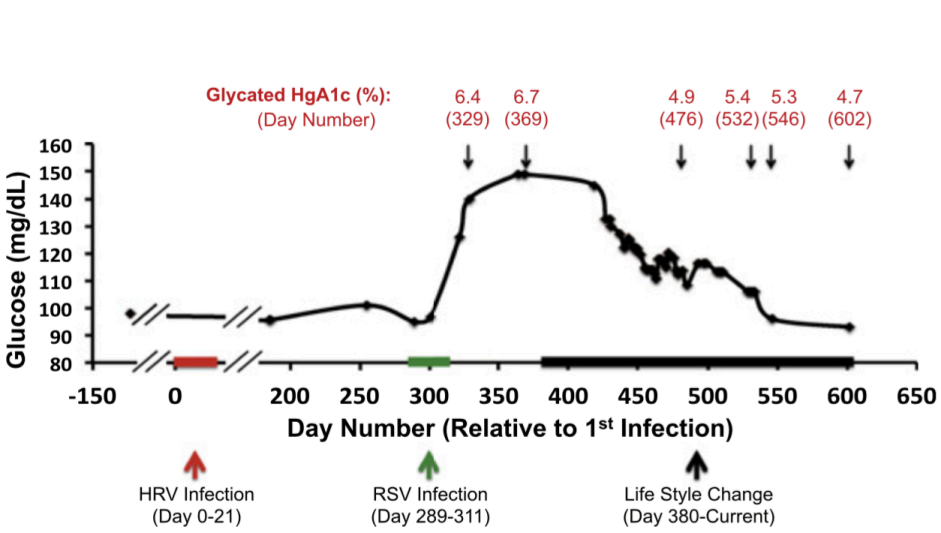
Integrated personal omics profile (iPOP)

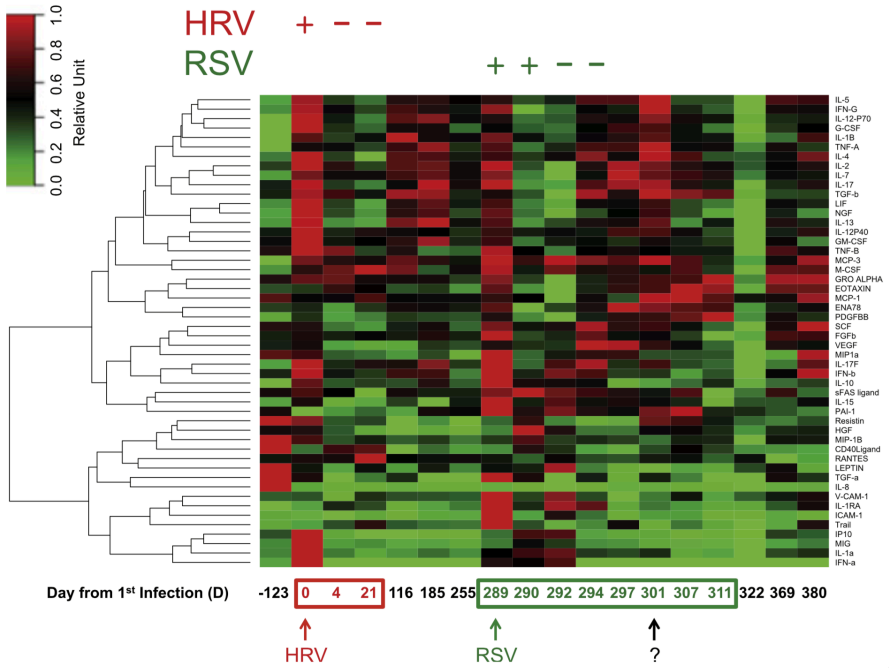
- Numerous types of data were collected, primarily from blood samples. The datasets include:
  - Transcriptomic
  - Proteomic
  - Metabolomic
  - Cytokine profiling
  - Autoantibody profiling
  - Medical exams



Chen et al. Cell 2012

Longitudinal medical data





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

Chen et al. Cell 2012