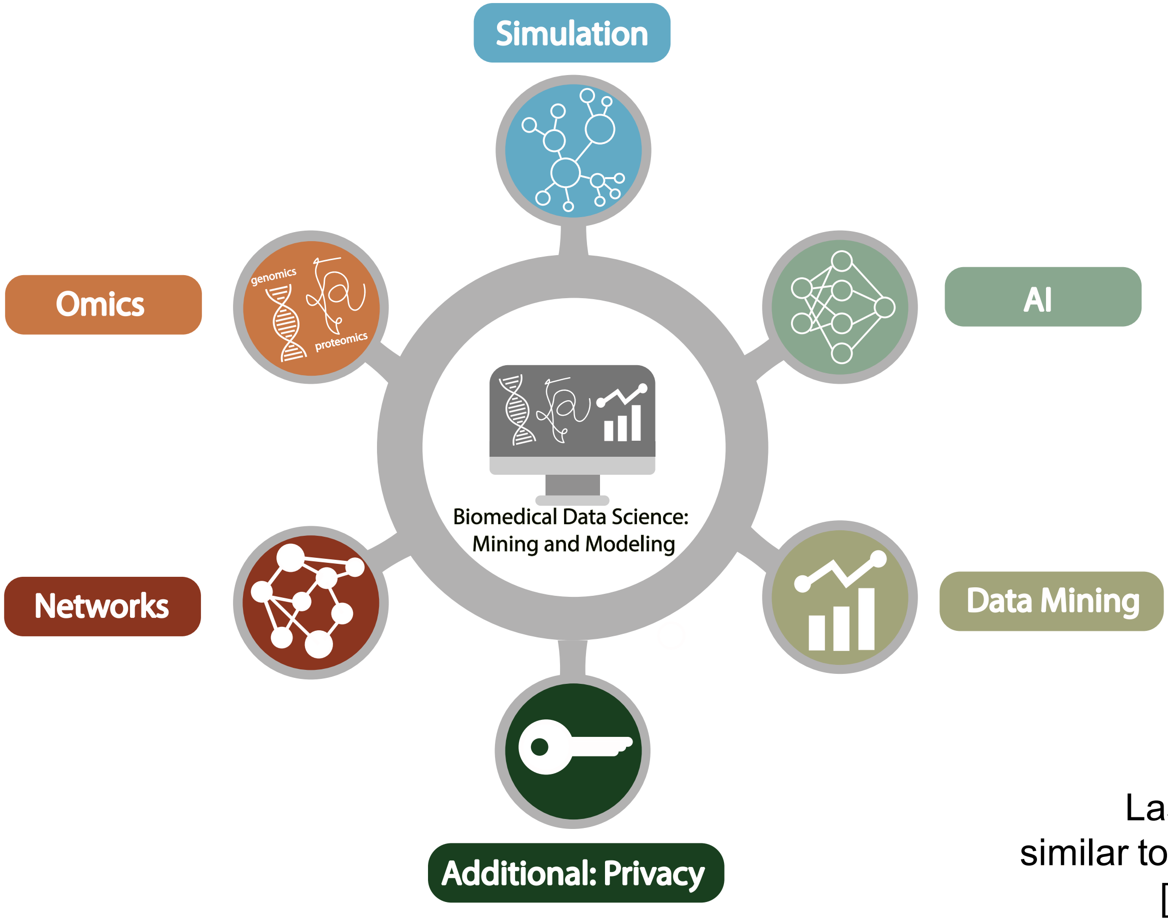
Biomedical Data Science (GersteinLab.org/courses/452)
Introduction to Personal Genomes (23i2a)



Simulation

AI

Omics

Biomedical Data Science: Mining and Modeling

Networks

Data Mining

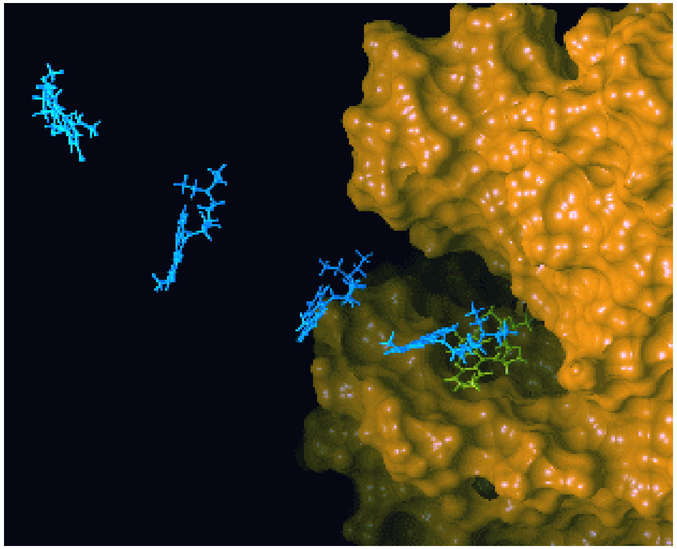
Additional: Privacy

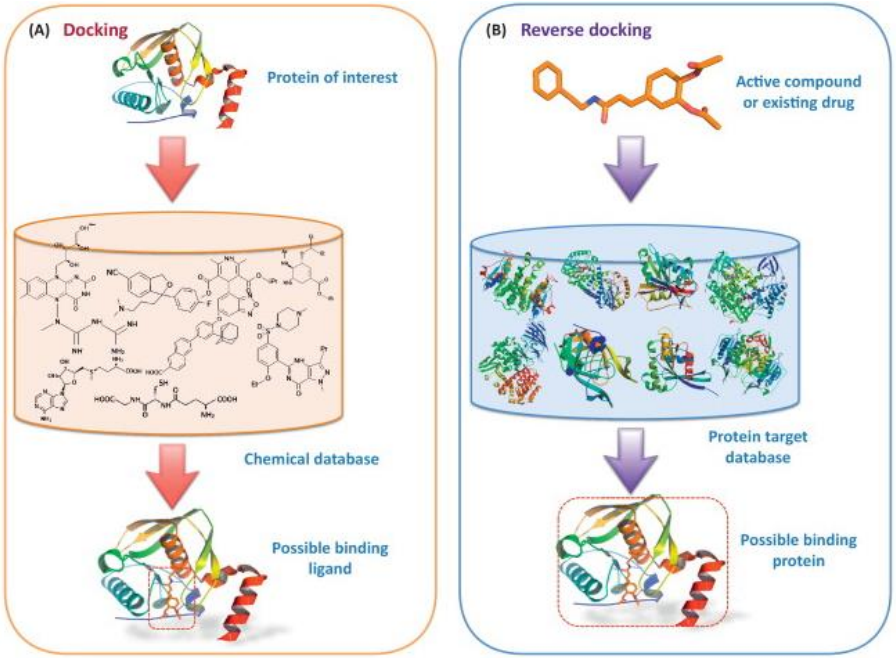
Last edit in spring ‘23,
similar to 22i2a & 2021’s I2a
[which has a video],
with additions beginning at slide 10, describing Zimmerome history & assignment. Slight modification at slide 5 too.

Mark Gerstein
Yale U.

Major Application I:
Designing Drugs from Structural Targets

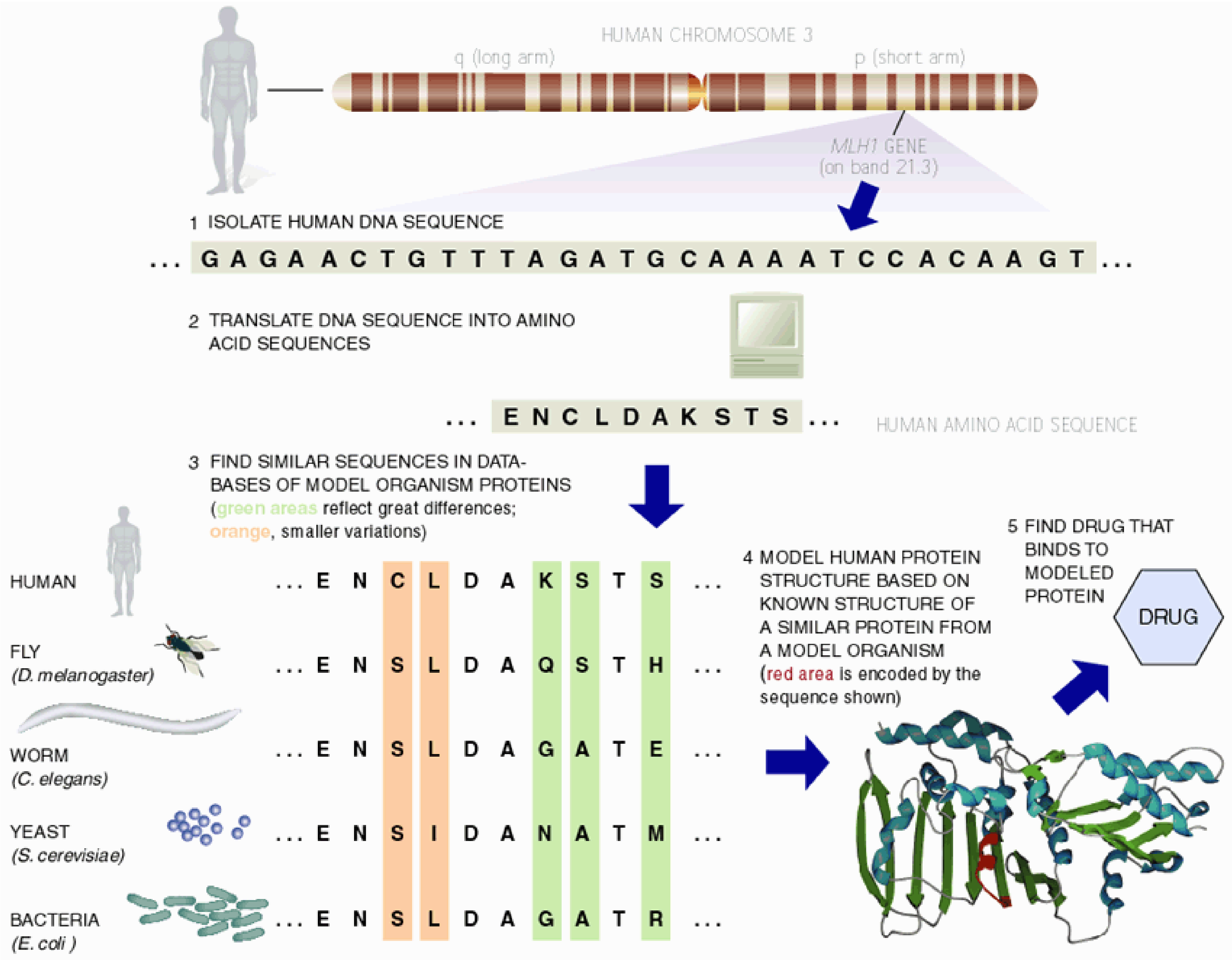
- Understanding how structures bind other molecules
- Designing inhibitors using docking, structure modeling
- *In silico* screens of chemical and protein databases





(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Zheng et al. Trends in Pharmacological Sciences 2013)

Major Application II: Finding Homologs, to Find Experimentally Tractable Gene Targets



1 ISOLATE HUMAN DNA SEQUENCE

... G A G A A C T G T T T A G A T G C A A A A T C C A C A A G T ...

2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES

... E N C L D A K S T S ...

3 FIND SIMILAR SEQUENCES IN DATABASES OF MODEL ORGANISM PROTEINS (green areas reflect great differences; orange, smaller variations)

HUMAN	...	E	N	C	L	D	A	K	S	T	S	...
FLY *(D. melanogaster)*	...	E	N	S	L	D	A	Q	S	T	H	...
WORM *(C. elegans)*	...	E	N	S	L	D	A	G	A	T	E	...
YEAST *(S. cerevisiae)*	...	E	N	S	I	D	A	N	A	T	M	...
BACTERIA *(E. coli)*	...	E	N	S	L	D	A	G	A	T	R	...

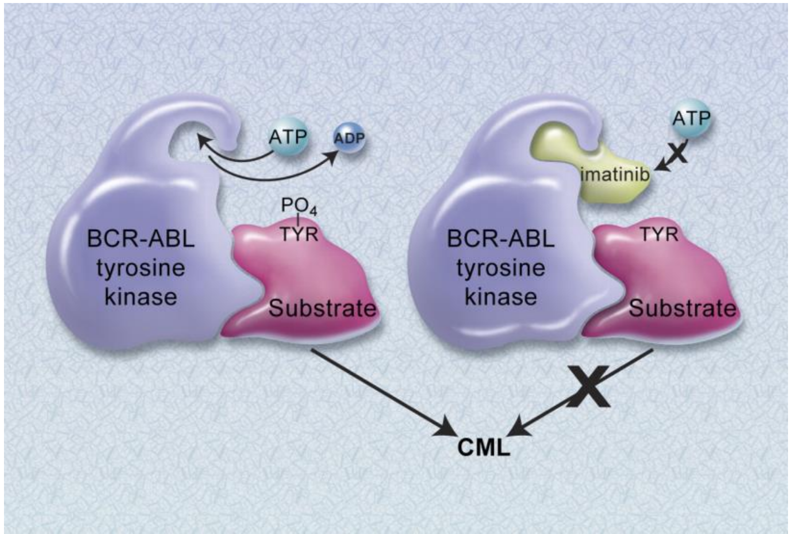
4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)

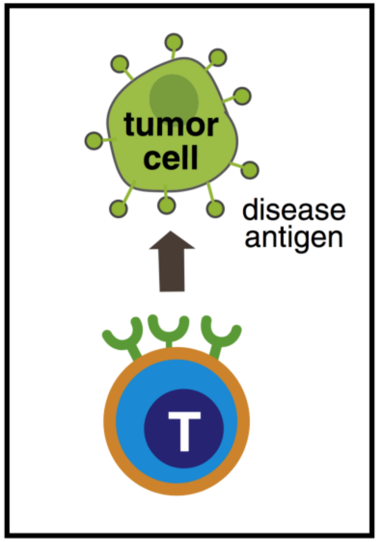
5 FIND DRUG THAT BINDS TO MODELED PROTEIN

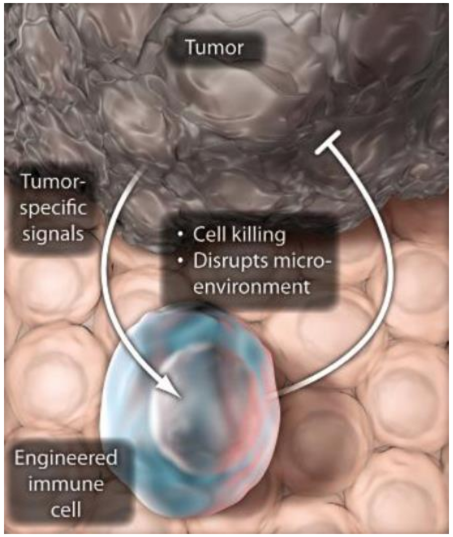
[Adapted from Sci. Am.]

Major Application III: Customizing treatment in oncology

- Identifying disease causing mutations in individual patients
- Designing targeted therapeutics
  - e.g. BCR-abl and Gleevec
  - Cancer immunotherapies targeting neoantigens



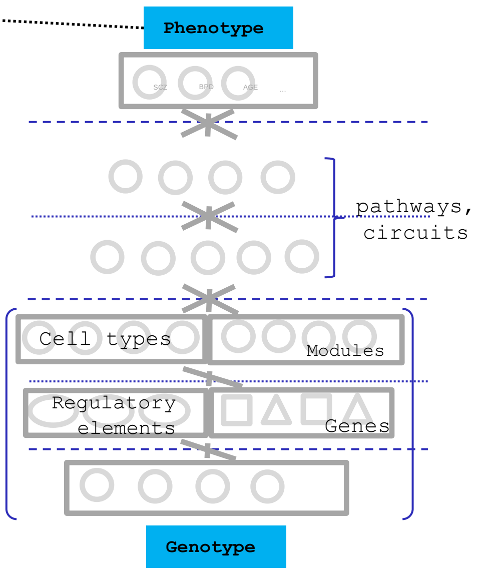




(From left to right, figures adapted from Druker BJ. Blood 2008 and the Lim Lab at UCSF)

Major Application IV: Finding molecular mechanisms & drug targets for diseases we know little about (Neuro-psychiatic Diseases)

Disease	Heritability*	Molecular **Mechanisms**
Schizophrenia	**81%**	-
Bipolar disorder	70%	-
Alzheimer's disease	58 - 79%	Apolipoprotein E (APOE), Tau
Hypertension	30%	Renin–angiotensin–aldosterone
Heart disease	34-53%	Atherosclerosis, VCAM-1
Stroke	32%	Reactive oxygen species (ROS), Ischemia
Type-2 diabetes	26%	Insulin resistance
Breast Cancer	25-56%	BRCA, PTEN



Many psychiatric conditions are highly heritable

Schizophrenia: up to 80%

But we don't understand basic molecular mechanisms underpinning this association

(in contrast to many other diseases such as cancer & heart disease)

Moreover, current models substantially underestimate heritability using genetic data

Schizophrenia : ~25%

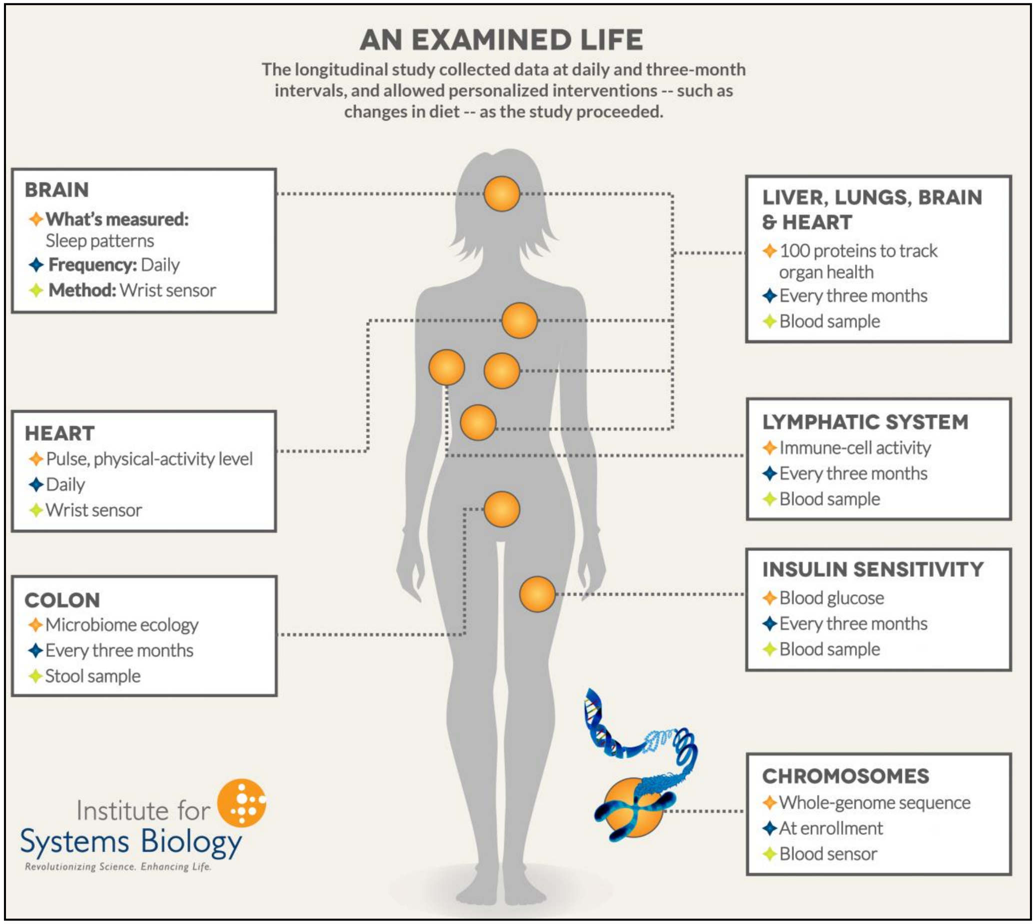
Thus, interested in developing predictive models of psychiatric traits which:

Use observations at intermediate (molecular levels) levels to inform latent structure.

Use the predictive features of these "molecular endo phenotypes" to begin to suggest actors involved in mechanism

*https://www.snpedia.com/index.php/Heritability

Major Application V: Holistic Personal Genome Characterization, in Normal Individuals



AN EXAMINED LIFE

The longitudinal study collected data at daily and three-month intervals, and allowed personalized interventions -- such as changes in diet -- as the study proceeded.

BRAIN
- **What's measured:** Sleep patterns
- **Frequency:** Daily
- **Method:** Wrist sensor

HEART
- Pulse, physical-activity level
- Daily
- Wrist sensor

COLON
- Microbiome ecology
- Every three months
- Stool sample

LIVER, LUNGS, BRAIN & HEART
- 100 proteins to track organ health
- Every three months
- Blood sample

LYMPHATIC SYSTEM
- Immune-cell activity
- Every three months
- Blood sample

INSULIN SENSITIVITY
- Blood glucose
- Every three months
- Blood sample

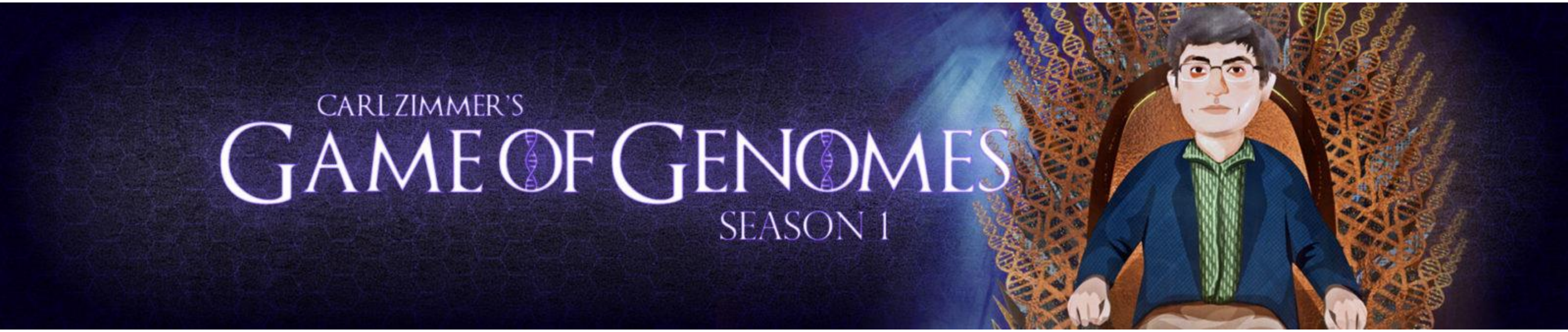
CHROMOSOMES
- Whole-genome sequence
- At enrollment
- Blood sensor

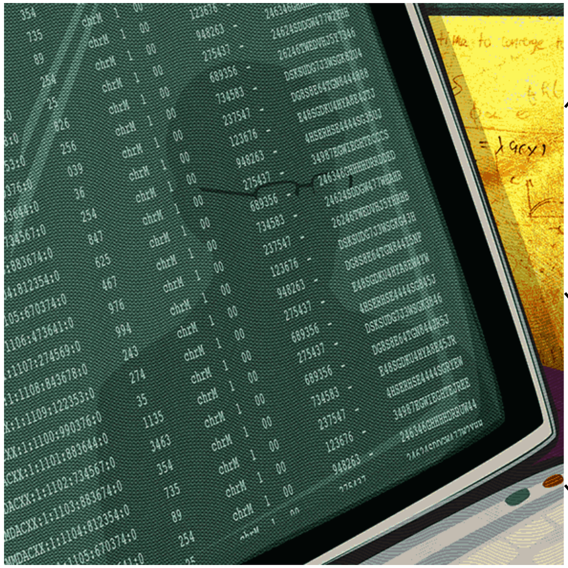
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

(Figure from Institute for Systems Biology)

- Mental disease & cancer are two extremes with respect to genomics (CEN, 92: 26)
  - Many other conditions in between, often involving interaction with the environment
- Pers. Genome Characterization
  - Identify mutations in personal genomes (SNPs, SVs, &c)
  - Estimate phenotypic (deleterious or protective) impact of variants.
  - Compare one person to wider population.
- Track changes over time & consider interaction w/ environment
  - Transcriptome studies
  - Longitudinal health studies (e.g. 100K wellness project, Framingham Heart Study)

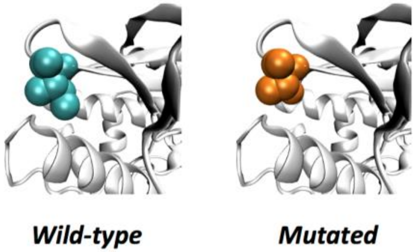
Analyzing Carl Zimmer’s genome





SNV AAGCT → ACGCT

Protein Structure



Ancestry

Expanding personalized medicine beyond the genome.

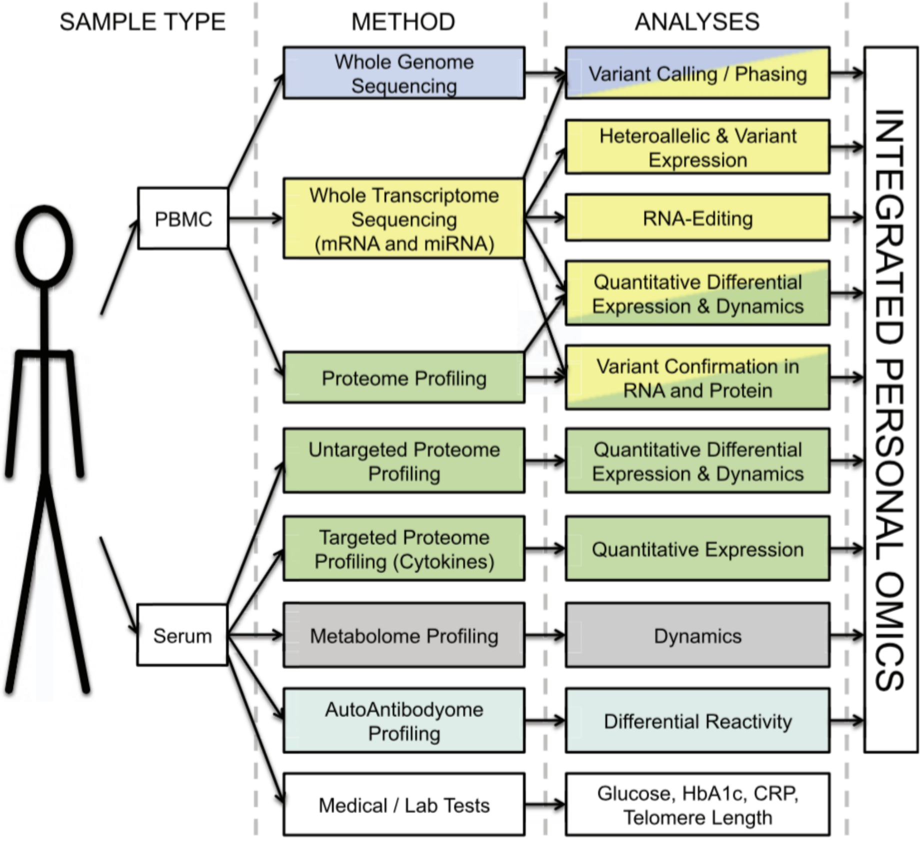
- An integrated personal omics profile (iPOP) is an example of a more comprehensive version of personalized medicine.

- 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


SAMPLE TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
AutoAntibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA-Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dynamics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

Chen et al. Cell 2012

History of the Analysis of the “Zimmerome” in the Class

2017

- Each group created a GitHub page detailing the work of each team
  - Additionally, each group has a power point presentation:
- Topics of projects include:
  - Comparative analysis of personal genomes
  - Personal genomes and personalized medicine (CRISPR)
  - Network analysis of personal genomes
  - Structure analysis

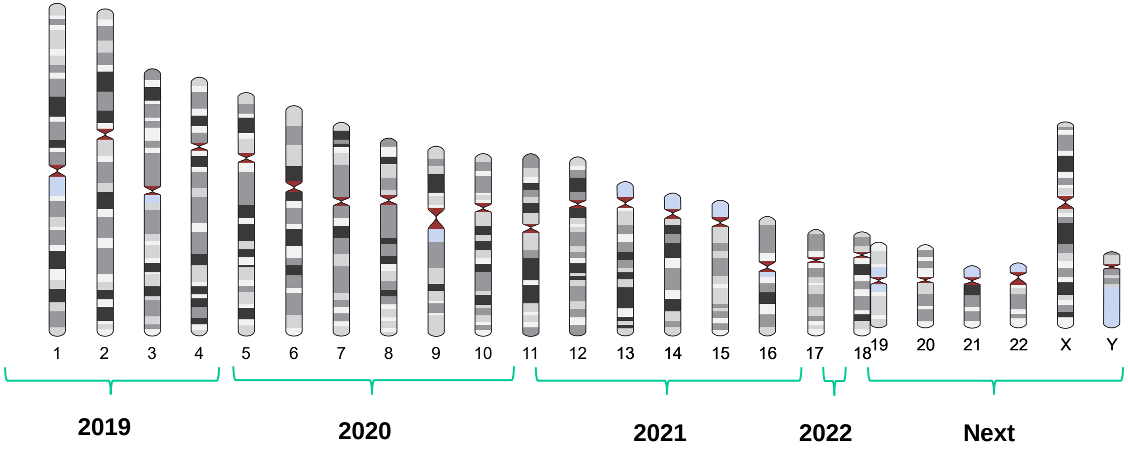
2018

- Each group had a power point presentation and a writeup
- Topics of projects include:
  - Finding how much of your genetic material comes from the Neanderthals
  - Using carl's genome to predict differences in gene expression from the average human and infer possible changes in physiology from these differences (GTEx analysis)
  - Predicting gene expression values from Carl’s SNP information
  - Finding a common variant associated with inflammatory response in Carl
  - Calculating Zimmer's risk for Alzheimer's disease
  - Identifying significant protein-coding mutations in Carl's genome
  - polygenic risk score prediction in coronary artery disease, type II diabetes, and schizophrenia for Carl

History of the Analysis of the "Zimmerome" in the Class

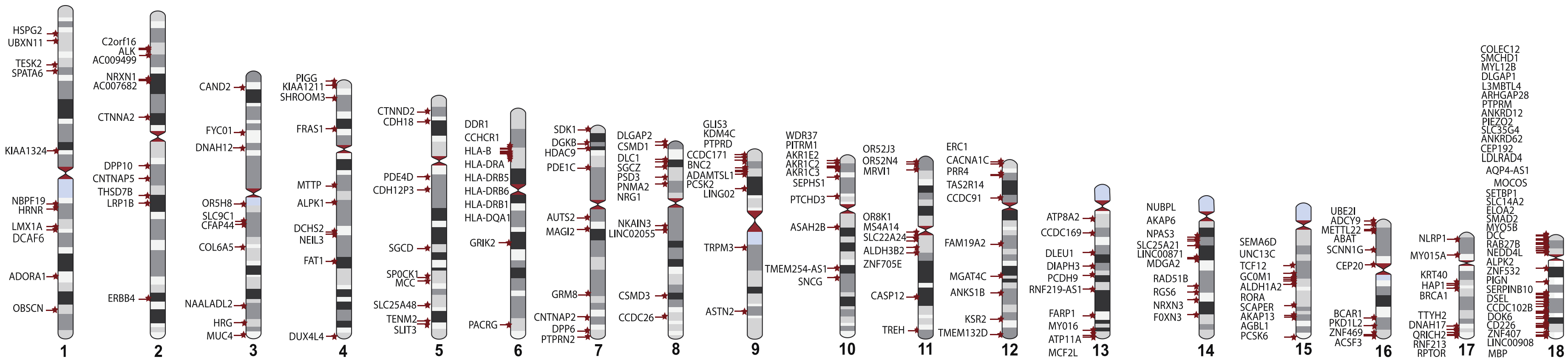
2019 - 2022

- Each group had a power point presentation and a write-up
- Started analyzing Carl's gene chromosome by chromosome
  - Part 1: Prioritization of 10 genes
  - Part 2: In-depth Analysis of prioritized genes:
    - Gene expression analysis
    - Network analysis
    - Protein structure analysis
    - Text mining analysis

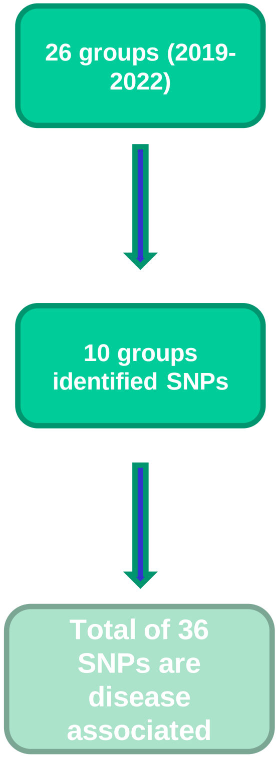


History of the Analysis of the “Zimmerome” in the Class

Genes prioritized



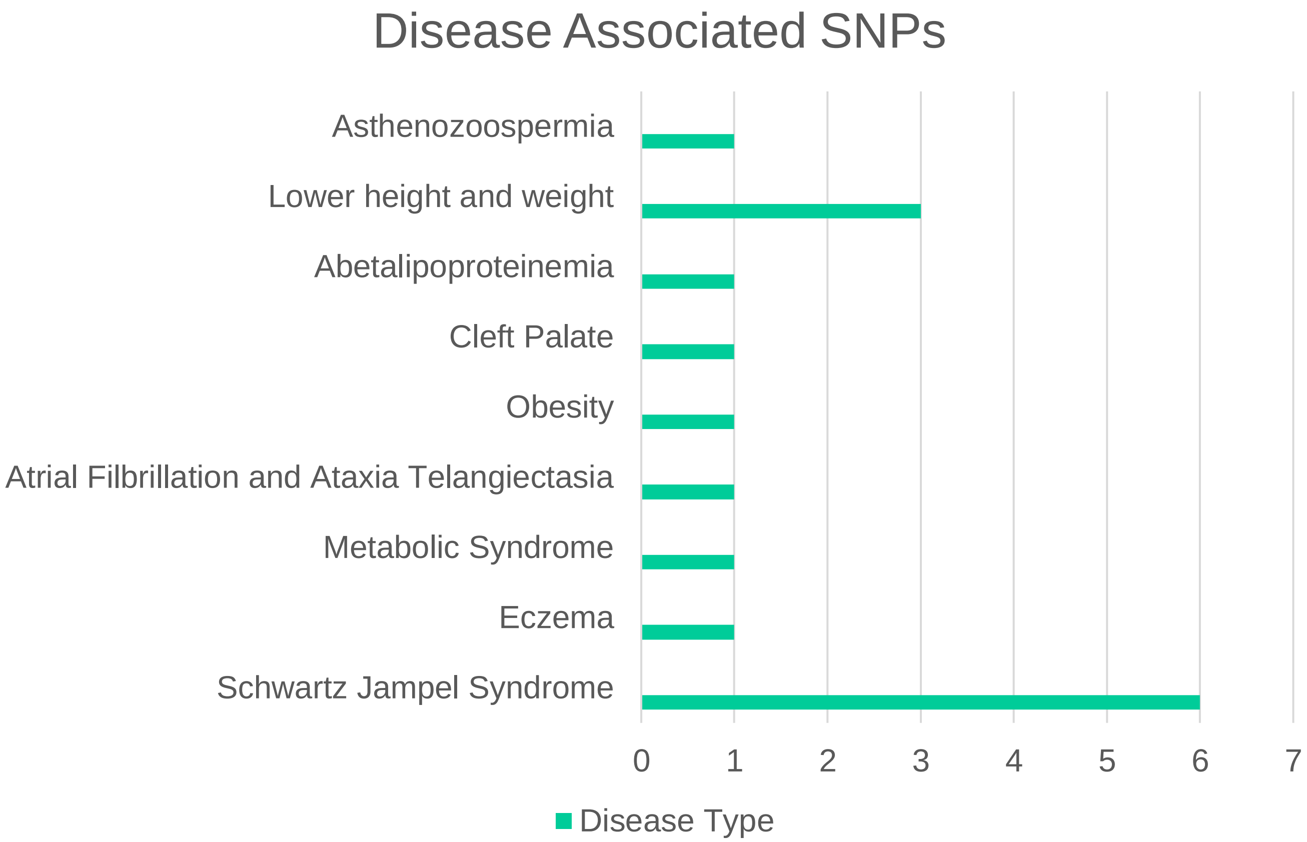
History of the Analysis of the “Zimmerome” in the Class



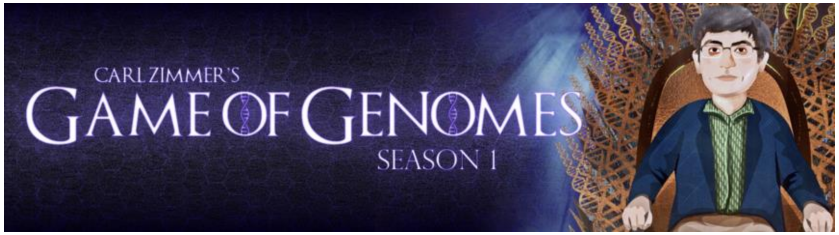
26 groups (2019-2022)

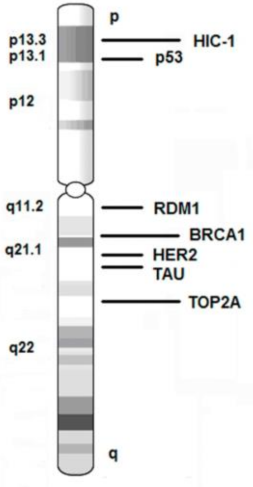
10 groups identified SNPs

Total of 36 SNPs are disease associated



This year’s Zimmerome Assignment: Investigate and Analyze a Personal Genome Using Bioinformatic/Biomedical Tools





Team based approach

- Assigned Teams (4-5 people in your section, assigned by TFs)
- Each team focuses on a single chromosome
- Cross-disciplinary

1. Computational

- Leveraging tools to prioritize genes or variants
- Pipeline Development

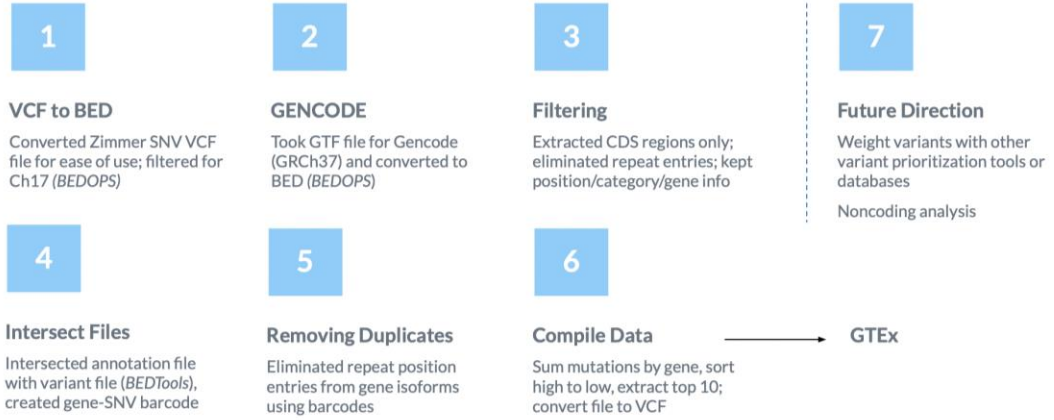
2. Biological/Biomedical

- Interpretation of prioritized genes or loci

3. Written and Oral

- Communication of project and results through written report

1. Computational Pipeline Development



1

VCF to BED

Converted Zimmer SNV VCF file for ease of use; filtered for Ch17 *(BEDOPS)*

2

GENCODE

Took GTF file for Gencode (GRCh37) and converted to BED *(BEDOPS)*

3

Filtering

Extracted CDS regions only; eliminated repeat entries; kept position/category/gene info

7

Future Direction

Weight variants with other variant prioritization tools or databases

Noncoding analysis

4

Intersect Files

Intersected annotation file with variant file (*BEDTools*), created gene-SNV barcode

5

Removing Duplicates

Eliminated repeat position entries from gene isoforms using barcodes

6

Compile Data

Sum mutations by gene, sort high to low, extract top 10; convert file to VCF

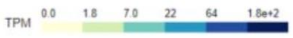
→ **GTEx**

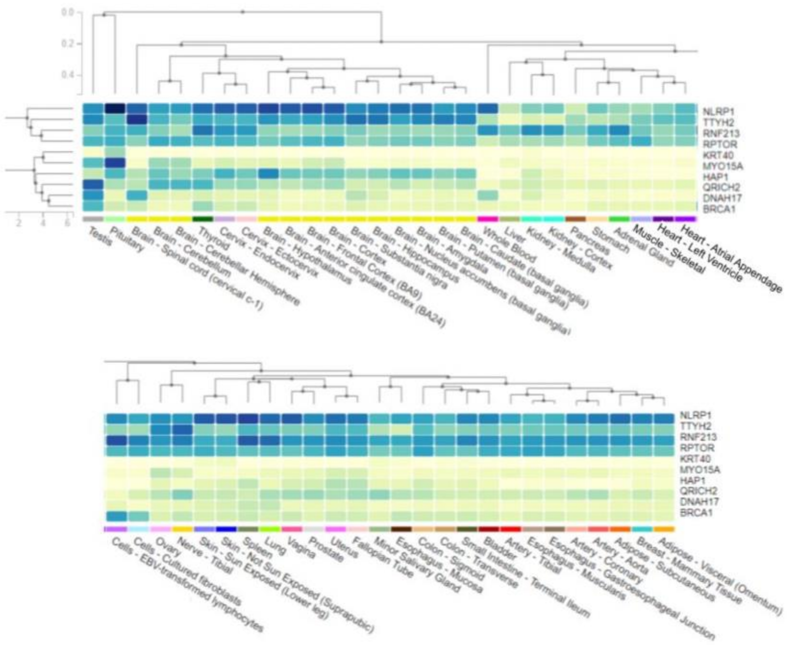
Computational Pipeline

- **Full code/software/script package**
- **GitHub**
- **Data files**
- **readme**

2. Biological/Biomedical Interpretation

Tissue Specific Expression Extracted from GTEx

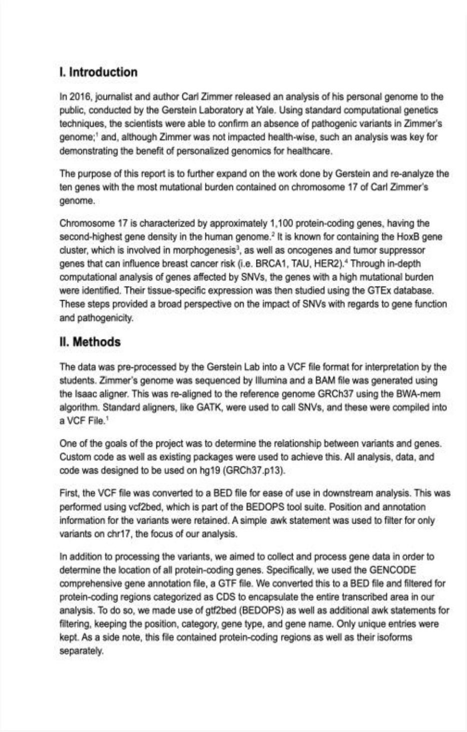




Interpretation of Results

- Biological interpretation of prioritized genes or loci
- Leveraging public omics or biomedical data
- Further discussion of results

3. Oral Presentation and Written Report



I. Introduction

In 2016, journalist and author Carl Zimmer released an analysis of his personal genome to the public, conducted by the Gerstein Laboratory at Yale. Using standard computational genetics techniques, the scientists were able to confirm an absence of pathogenic variants in Zimmer's genome;[1] and, although Zimmer was not impacted health-wise, such an analysis was key for demonstrating the benefit of personalized genomics for healthcare.

The purpose of this report is to further expand on the work done by Gerstein and re-analyze the ten genes with the most mutational burden contained on chromosome 17 of Carl Zimmer's genome.

Chromosome 17 is characterized by approximately 1,100 protein-coding genes, having the second-highest gene density in the human genome.[2] It is known for containing the HoxB gene cluster, which is involved in morphogenesis[3], as well as oncogenes and tumor suppressor genes that can influence breast cancer risk (i.e. BRCA1, TAU, HER2).[4] Through in-depth computational analysis of genes affected by SNVs, the genes with a high mutational burden were identified. Their tissue-specific expression was then studied using the GTEx database. These steps provided a broad perspective on the impact of SNVs with regards to gene function and pathogenicity.

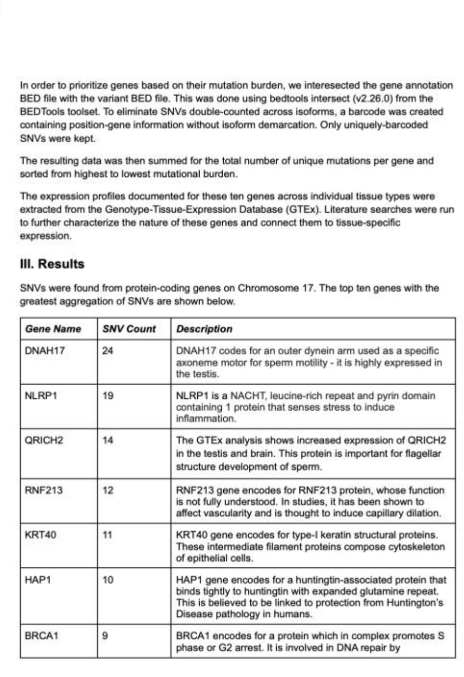
II. Methods

The data was pre-processed by the Gerstein Lab into a VCF file format for interpretation by the students. Zimmer's genome was sequenced by Illumina and a BAM file was generated using the Isaac aligner. This was re-aligned to the reference genome GRCh37 using the BWA-mem algorithm. Standard aligners, like GATK, were used to call SNVs, and these were compiled into a VCF File.[1]

One of the goals of the project was to determine the relationship between variants and genes. Custom code as well as existing packages were used to achieve this. All analysis, data, and code was designed to be used on hg19 (GRCh37.p13).

First, the VCF file was converted to a BED file for ease of use in downstream analysis. This was performed using vcf2bed, which is part of the BEDOPS tool suite. Position and annotation information for the variants were retained. A simple awk statement was used to filter for only variants on chr17, the focus of our analysis.

In addition to processing the variants, we aimed to collect and process gene data in order to determine the location of all protein-coding genes. Specifically, we used the GENCODE comprehensive gene annotation file, a GTF file. We converted this to a BED file and filtered for protein-coding regions categorized as CDS to encapsulate the entire transcribed area in our analysis. To do so, we made use of gtf2bed (BEDOPS) as well as additional awk statements for filtering, keeping the position, category, gene type, and gene name. Only unique entries were kept. As a side note, this file contained protein-coding regions as well as their isoforms separately.



In order to prioritize genes based on their mutation burden, we interesected the gene annotation BED file with the variant BED file. This was done using bedtools intersect (v2.26.0) from the BEDTools toolset. To eliminate SNVs double-counted across isoforms, a barcode was created containing position-gene information without isoform demarcation. Only uniquely-barcoded SNVs were kept.

The resulting data was then summed for the total number of unique mutations per gene and sorted from highest to lowest mutational burden.

The expression profiles documented for these ten genes across individual tissue types were extracted from the Genotype-Tissue-Expression Database (GTEx). Literature searches were run to further characterize the nature of these genes and connect them to tissue-specific expression.

III. Results

SNVs were found from protein-coding genes on Chromosome 17. The top ten genes with the greatest aggregation of SNVs are shown below.

Gene Name	*SNV Count*	*Description*
DNAH17	24	DNAH17 codes for an outer dynein arm used as a specific axoneme motor for sperm motility - it is highly expressed in the testis.
NLRP1	19	NLRP1 is a NACHT, leucine-rich repeat and pyrin domain containing 1 protein that senses stress to induce inflammation.
QRICH2	14	The GTEx analysis shows increased expression of QRICH2 in the testis and brain. This protein is important for flagellar structure development of sperm.
RNF213	12	RNF213 gene encodes for RNF213 protein, whose function is not fully understood. In studies, it has been shown to affect vascularity and is thought to induce capillary dilation.
KRT40	11	KRT40 gene encodes for type-I keratin structural proteins. These intermediate filament proteins compose cytoskeleton of epithelial cells.
HAP1	10	HAP1 gene encodes for a huntingtin-associated protein that binds tightly to huntingtin with expanded glutamine repeat. This is believed to be linked to protection from Huntington's Disease pathology in humans.
BRCA1	9	BRCA1 encodes for a protein which in complex promotes S phase or G2 arrest. It is involved in DNA repair by

Oral Presentation

- **April 26**
  - **2 minute mp4 recording per group (nominate 1 person to make the recording)**
  - **We will play these recordings in class on 4/26**
- **April 27 or 28 in your Section**
  - **10 minute presentation by other members of the group**

Written Report

- **Due: May 10, 2023**
- **At least 1000 words**

Summary Slide

- **1 summary slide giving an overview of your project**

Summary Metadata File

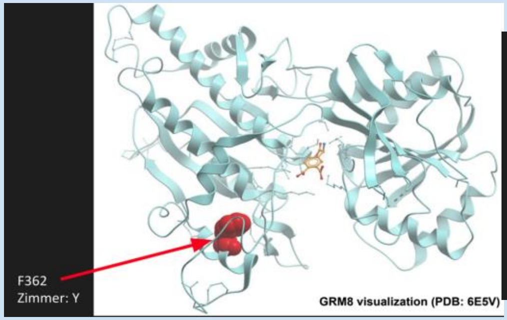
- **A single text file containing relevant information**
- **More description in assignment file**

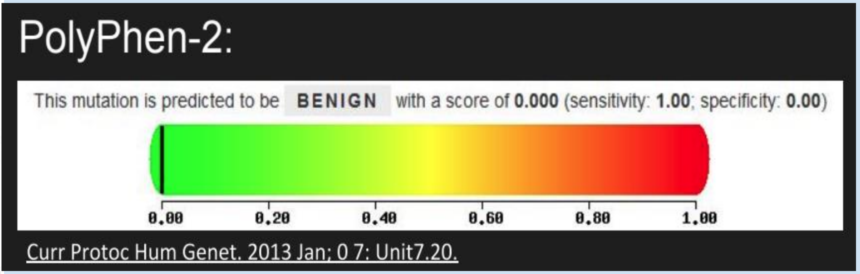
2020 group 2 (chr 7)

Top 10 Prioritized Genes

1. CNTNAP2
2. MAGI2
3. PTPRN2
4. DPP6
5. SDK1
6. DGKB
7. AUTS2
8. HDAC9
9. GRM8
10. PDE1C

Summary Figure:





PolyPhen-2:

This mutation is predicted to be **BENIGN** with a score of **0.000** (sensitivity: **1.00**; specificity: **0.00**)

Curr Protoc Hum Genet. 2013 Jan; 0 7: Unit7.20.

Summary:

1. Prioritization approach: mutational burden
2. Downstream analysis: PDB structural analysis
3. Findings:
   a. Among the top 10 most mutated genes on chromosome 7, there are 6 missense variants within 4 genes
   b. Only one variant, conferring a protein, is characterized: GRM8
   c. PolyPhen analysis shows that substitution at pos 362 from F to Y is predicted to be tolerated

2023 group N (chr Z)

Summary Figure:

Top 10 Prioritized Genes

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

Summary:

1. Prioritization approach:
2. Downstream analysis::
3. Findings:
   a.
   b.