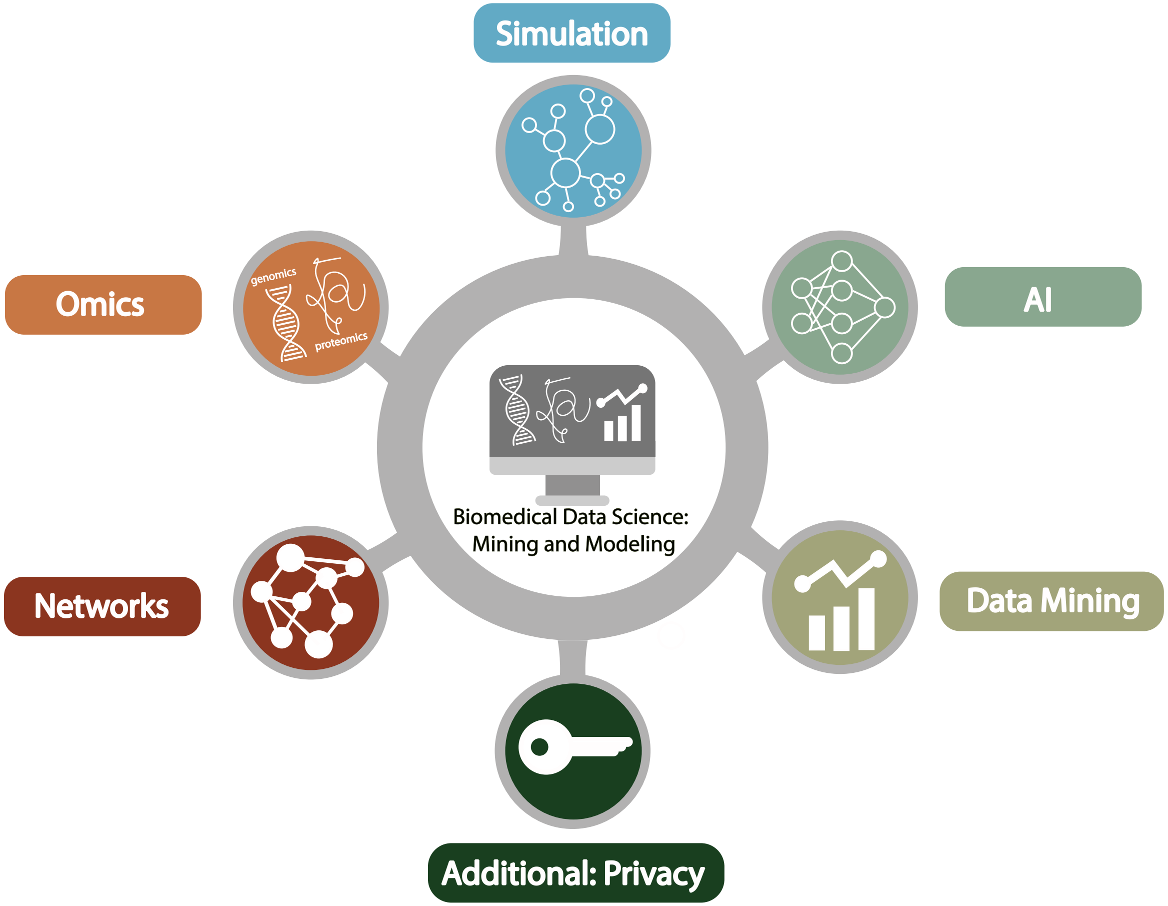
Biomedical Data Science (GersteinLab.org/courses/452)

Introduction (25i1+25i2a)



Last edit in spring ’25.
Combines & integrates i1 [which has a video] & i2a from previous years.

Takes ~50’ with rest of class going over website syllabus

Mark Gerstein
Yale U.

Please Fill Out Course Web Forms – Right now, if you haven’t already!

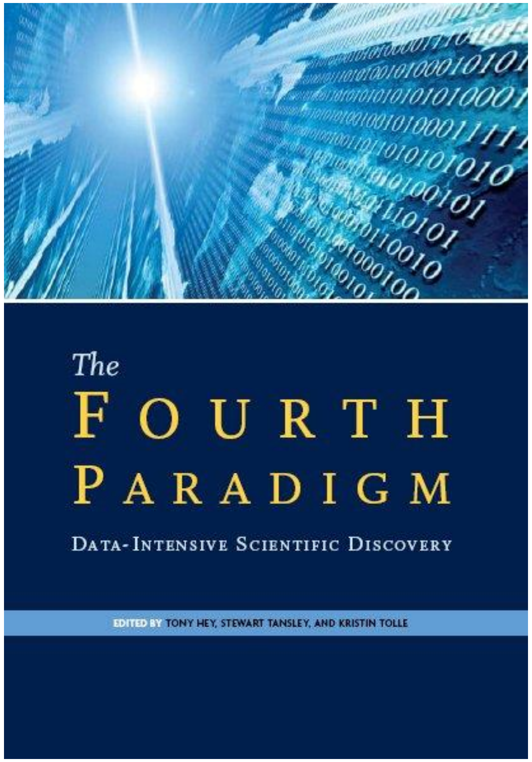
Course Web Form
Due Today（1/13）

Link is also available from class website:
GersteinLab.org/courses/452

Overview: what is Biomed. Data science?

(in the context of Data Science, in general)

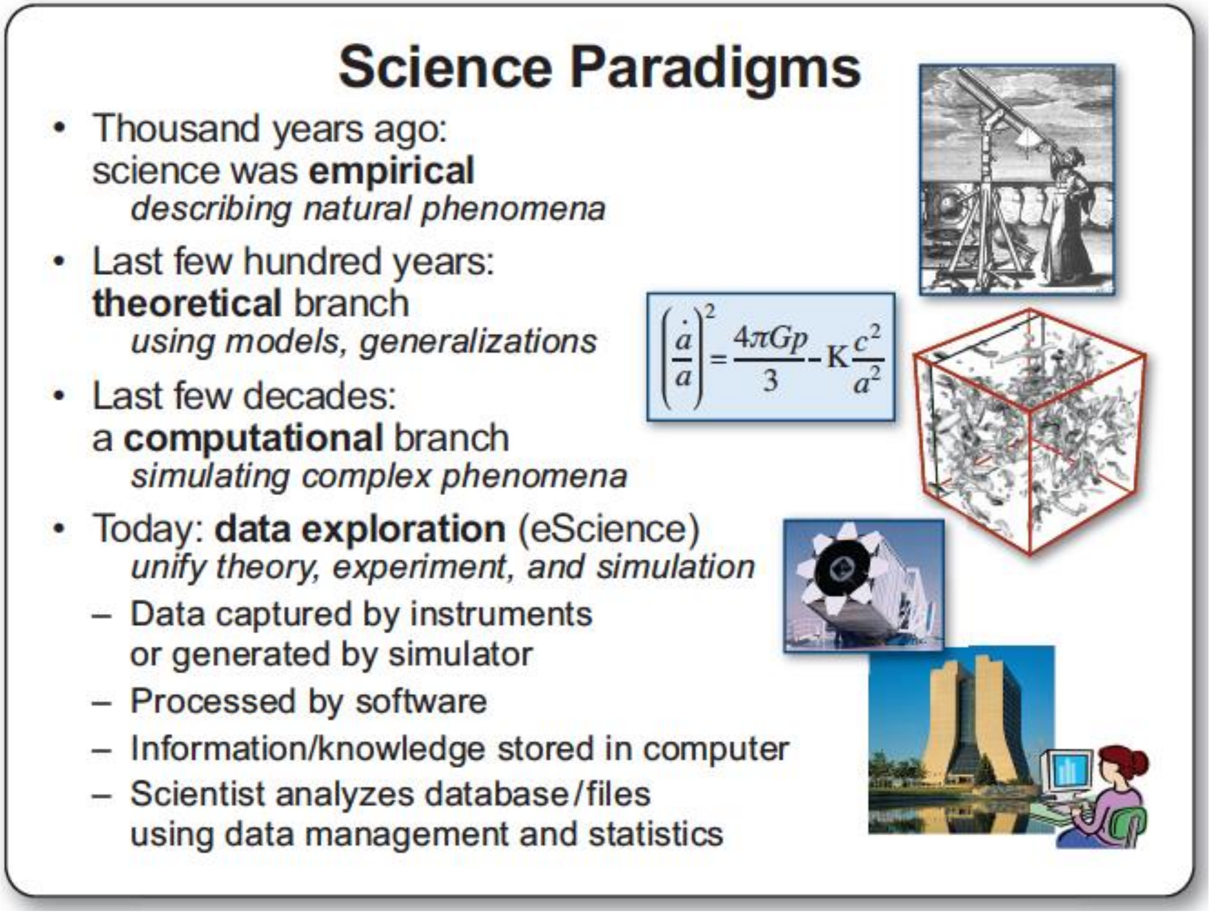
Jim Gray's 4th Paradigm



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files
  using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$

Jim Gray’s 4th Paradigm

#3 - Simulation

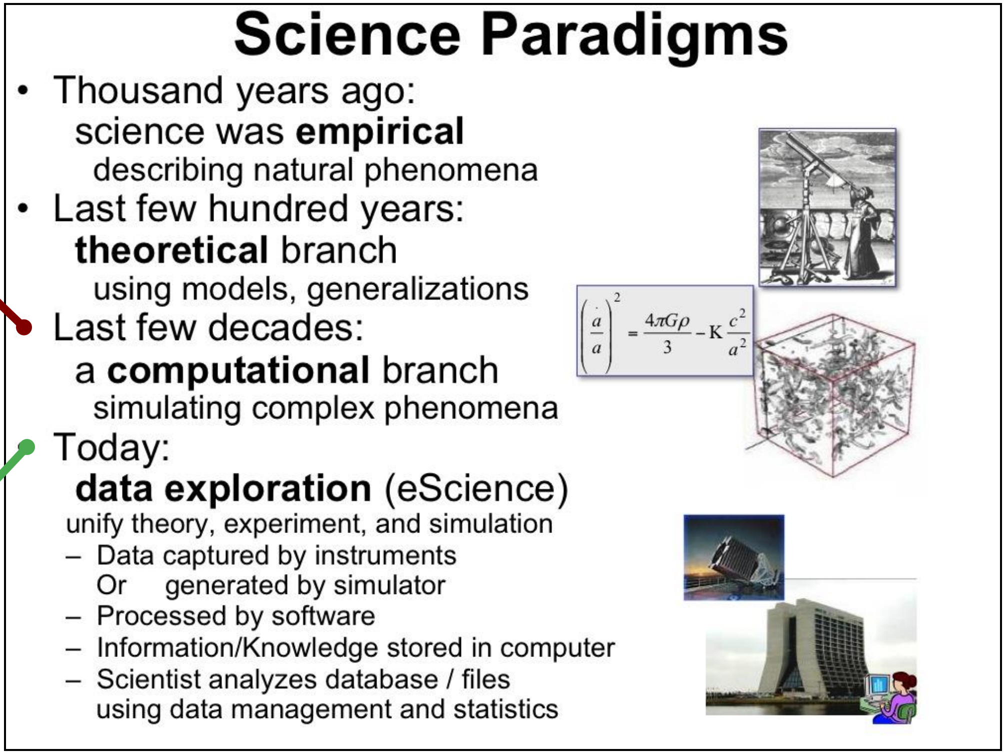
Prediction based on physical principles (eg Exact Determination of Rocket Trajectory)

Emphasis on: Supercomputers

#4 - Data Mining

Classifying information & discovering unexpected relationships

Emphasis: networks, “federated” DBs



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
    Or generated by simulator
  – Processed by software
  – Information/Knowledge stored in computer
  – Scientist analyzes database / files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

Gray died in ’07.
Book about his ideas came out in ‘09…..

[Slide from : http://research.microsoft.com/en-us/um/people/gray/talks/stanford%2520symbolic%2520systems%2520seminar.ppt]

What is Data Science? An overall, bland definition…

- Data Science encompasses the study of the entire **lifecycle of data**
  - Understanding of how data are **gathered** & the issues that arise in its collection
  - Knowledge of what data sources are available & how they may be synthesized to solve problems
  - The **storage**, access, annotation, management, & transformation of data
- Data Science encompasses many aspects of **data analysis**
  - Statistical inference, machine learning, & the design of algorithms and computing systems that enable **data mining**
  - Connecting this mining where possible with **physical modeling**
  - The presentation and **visualization** of data analysis
  - The use of data analysis to make **practical decisions** & policy
- Secondary aspects of data, not its intended use – eg the **data exhaust**
  - The appropriate protection of **privacy**
  - Creative **secondary uses** of data – eg for Science of science
  - The elimination of inappropriate bias in the entire process

Data Science in the wider world: a buzz-word for successful Ads

- Ads, media, product placement, supply optimization,
- Integral to success of GOOG, FB, AMZN, WMT…

The Economist

Obama the warrior

Misgoverning Argentina

The economic shift from West to East

Genetically modified crops blossom

The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

CIO Network

INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.

+ Follow (469)

TECH | 12/12/2012 @ 1:57AM | 3,289 views

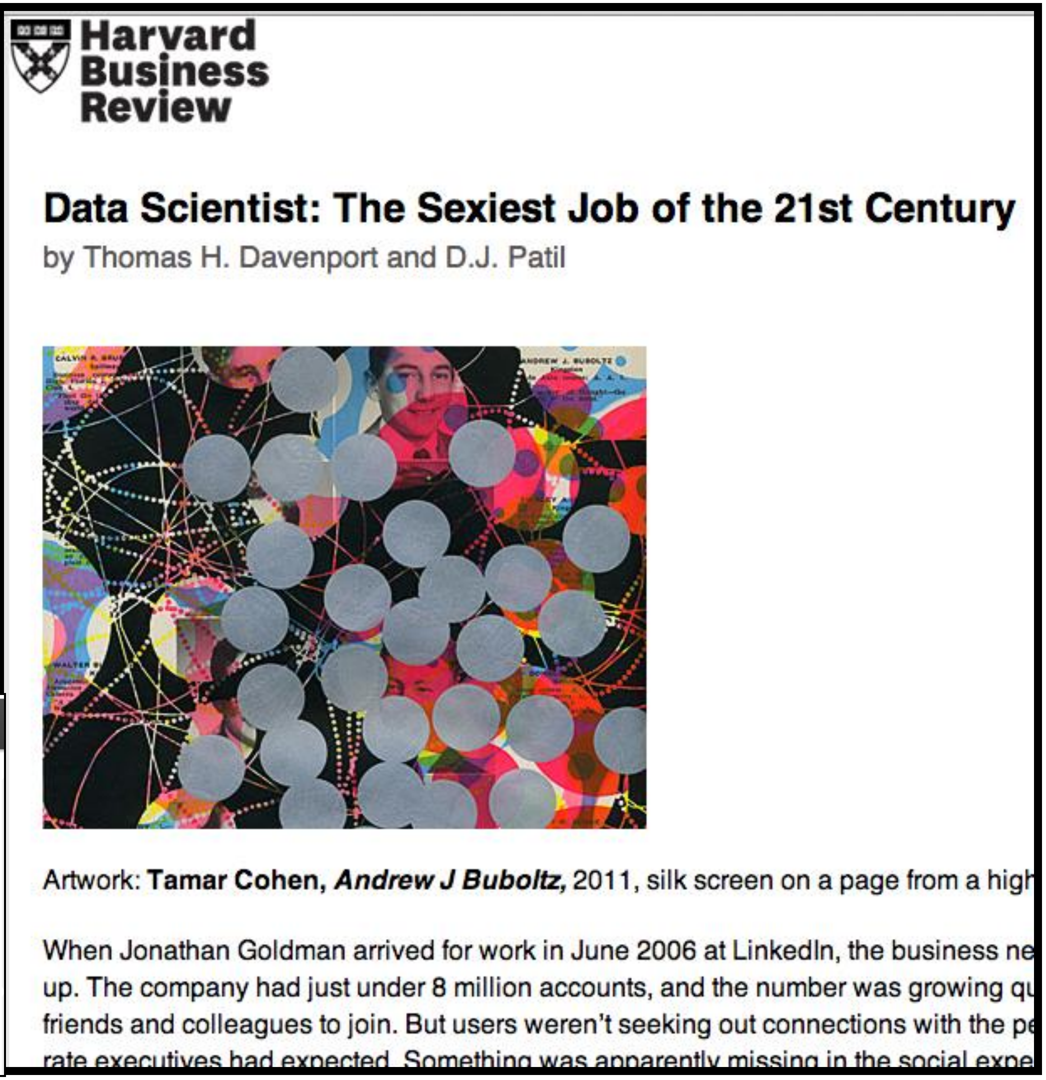
Why Big Data Is All Retailers Want for Christmas

Eric Savitz, Forbes Staff

+ Comment Now + Follow Comments

Guest post written by Quentin Gallivan

Quentin Gallivan is CEO of Pentaho Corp., an Orlando, Florida-based provider of business analytics software.


Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Artwork: **Tamar Cohen, *Andrew J Buboltz*,** 2011, silk screen on a page from a high

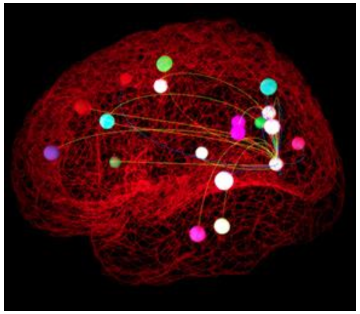
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne up. The company had just under 8 million accounts, and the number was growing qu friends and colleagues to join. But users weren't seeking out connections with the pe rate executives had expected. Something was apparently missing in the social expe

[Oct. '12 issue]

Data Science in Traditional Science

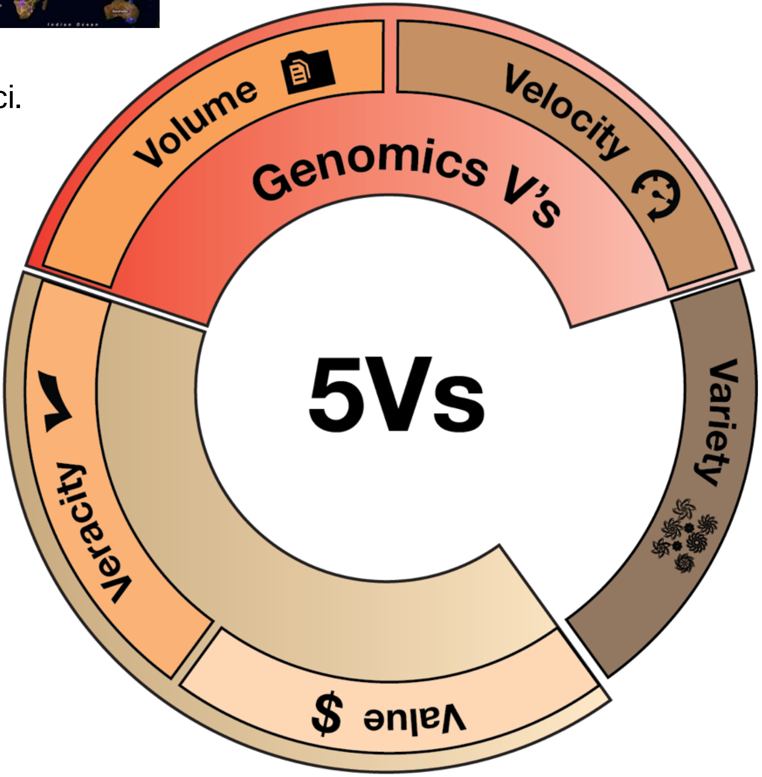
- Pre-dated commercial mining
- Instrument generated
- Large data sets often created by large teams not to answer one Q but to be mined broadly
- Often coupled to a physical/biological model
- Interplay w/ experiments

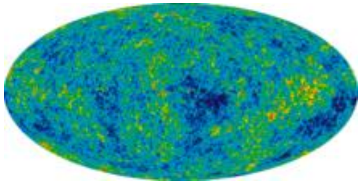
High energy physics - Large Hadron Collider



Neuroscience - The Human Connectome Project

Ecology & Earth Sci. - Fluxnet





Astronomy - Sloan Digital Sky survey

nature

BIG DATA

SCIENCE IN THE PETABYTE ERA

Science

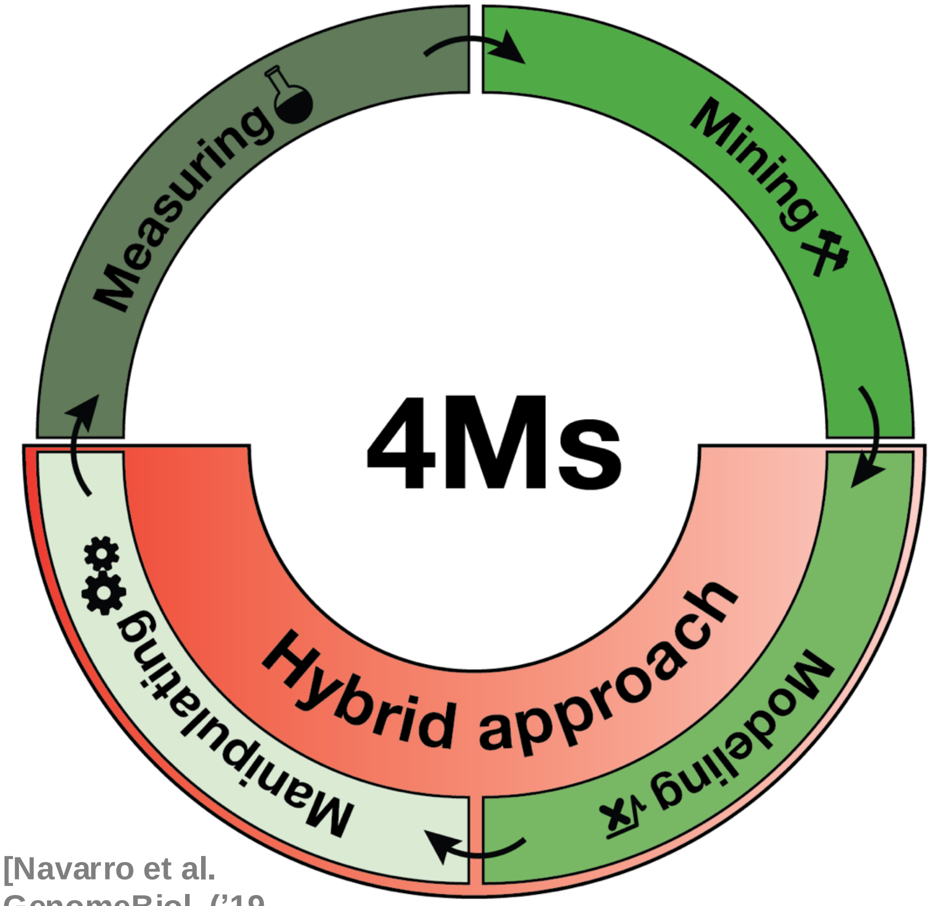
data

Genomics DNA sequencer

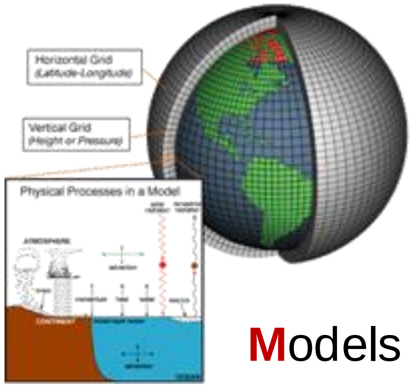
[Navarro et al. GenomeBiol. ('19, in press)]

Coupling of Scientific Data to Models & Experiments

- Scientific data often coupled to a physical/biological model
- Lauffenburger's Sys. Biol. **4Ms:**
  Measurement, Mining, Modeling & Manipulation
  (Ideker et al.'06. Annals of Biomed. Eng.)
- Weather forecasting as an exemplar
  - Physical models & simulation useful but not sufficient ("butterfly" effect)
  - Success via coupling to large-scale sensor data collection



[Navarro et al. GenomeBiol. ('19, in press)]



Horizontal Grid (Latitude-Longitude)
Vertical Grid (Height or Pressure)
Physical Processes in a Model

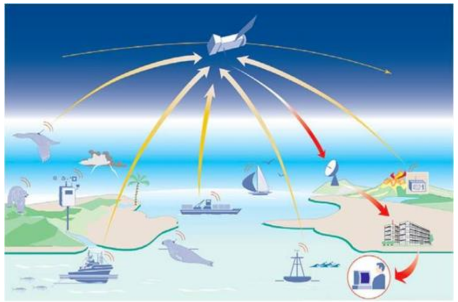
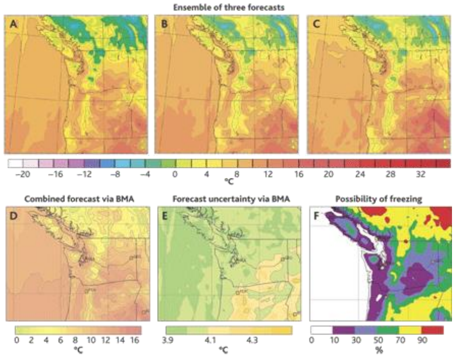


Image from http://web.aibn.uq.edu.au/cssb/ResearchProjects.htm

Models + Data Mining

↓

Forecasts



Ensemble of three forecasts
A
B
C
−20 −16 −12 −8 −4 0 4 8 12 16 20 24 28 32
°C
Combined forecast via BMA
Forecast uncertainty via BMA
Possibility of freezing
D
E
F

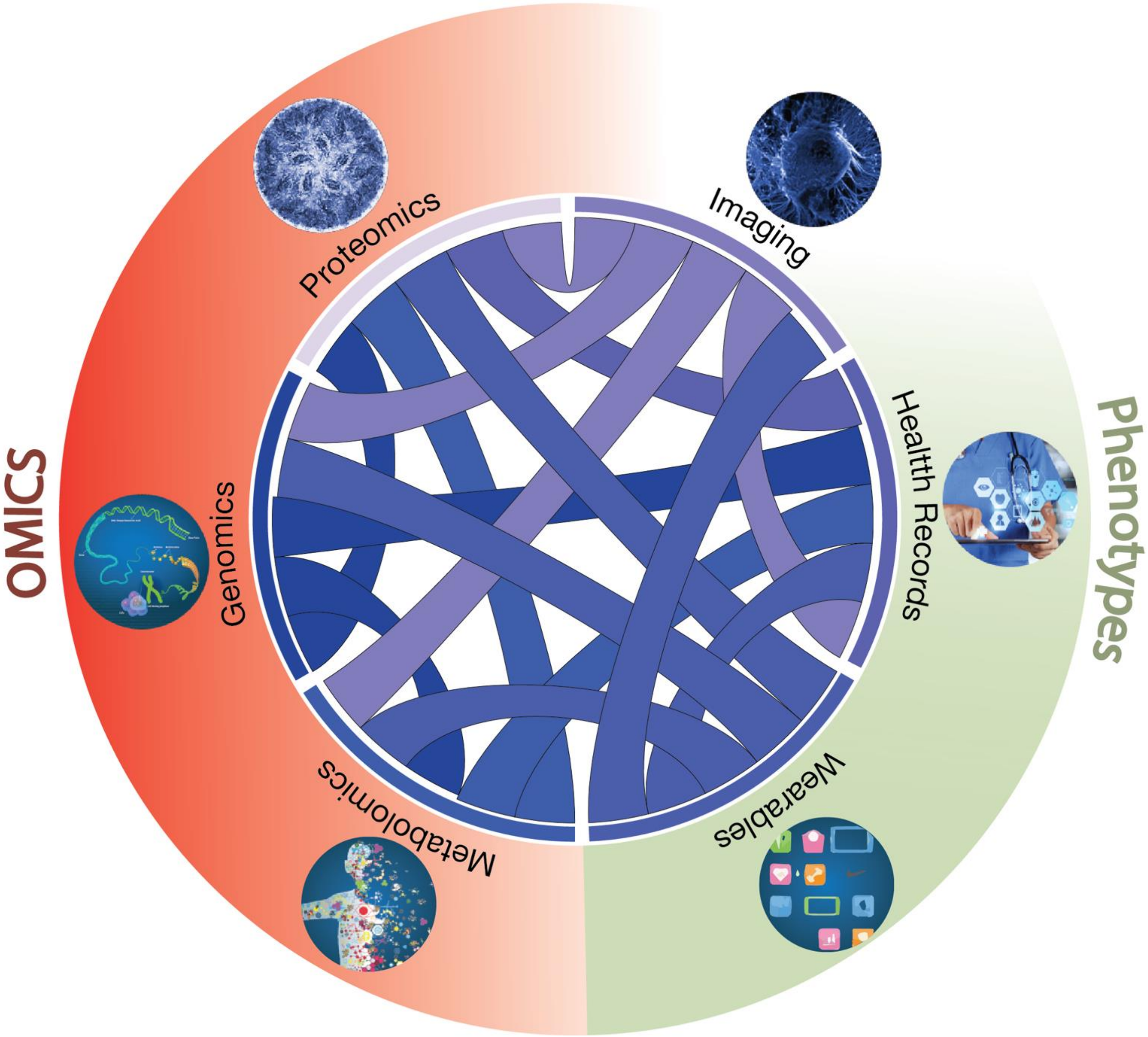
NOAA

Biomed. Data science:

Scaling & Integration

Drivers of Biomedical Data Science

- **Integration** across data types
- **Scaling** of individual data types



[Navarro et al. GenomeBiol. ('19, in press)]

Case Study: Amazing Progress in Scaling & Integration with Genotype-Phenotype Relationships

1953

Double Helix

Watson & Crick

1995

Sequenced Genome

H. influenzae

2008

1000 Genomes

Catalogue of human variation

2015

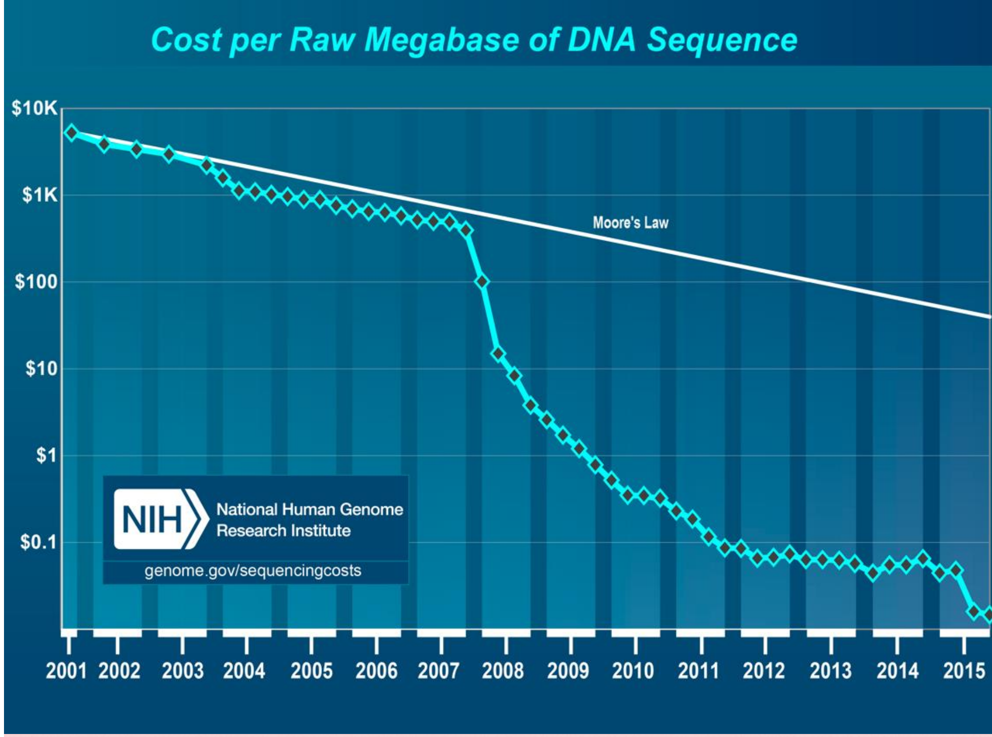
Integrated health data

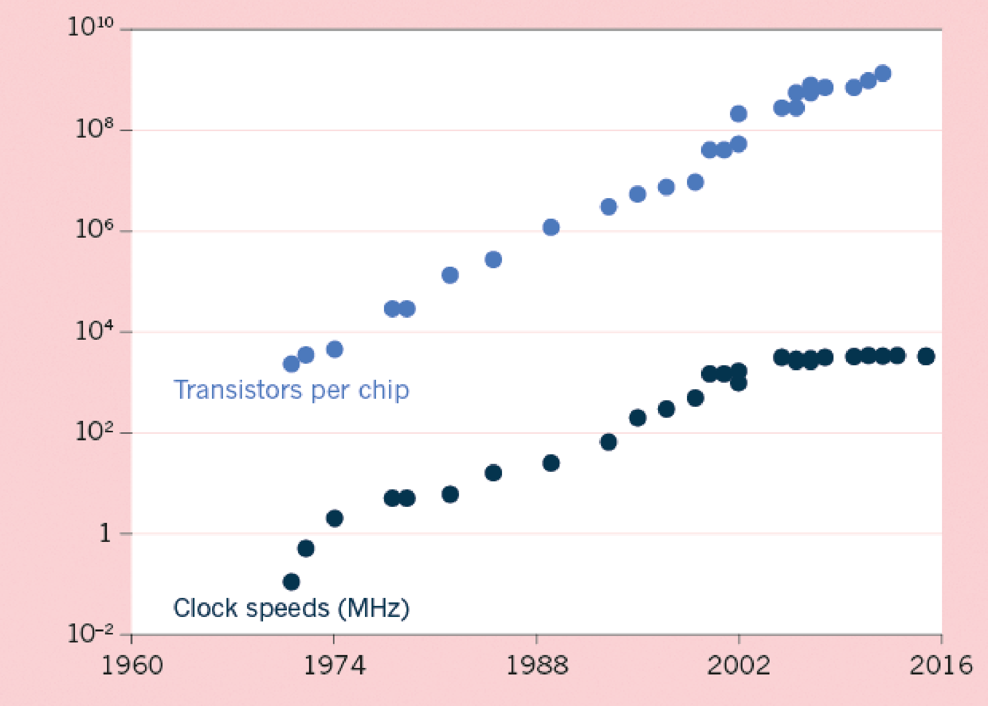
UKBB study with over 500K participants, genotypes to phenotypic details & clinical information

The Scaling of Genomic Data Science:

Powered by exponential increases in data & computing

(Moore’s Law)

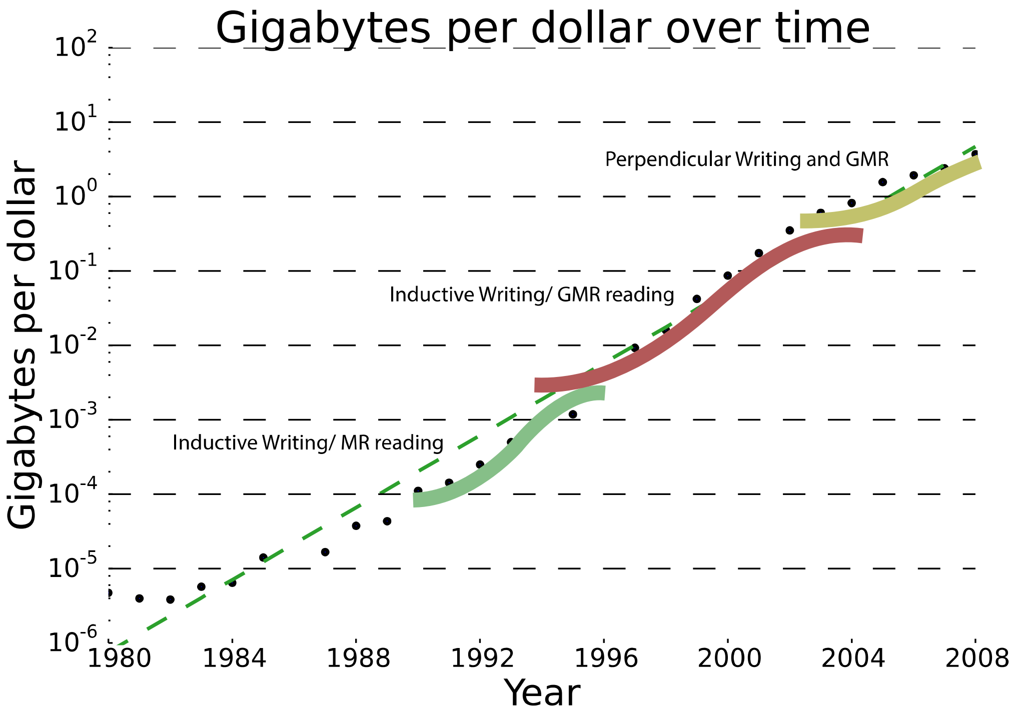


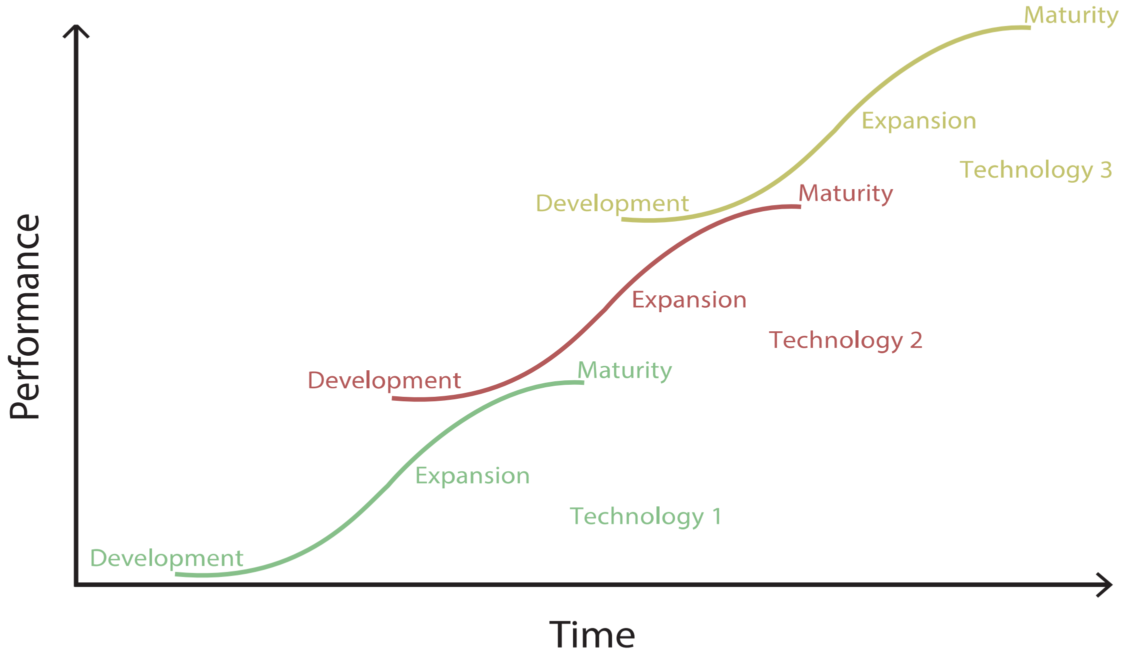


[NHGRI website + Waldrop (‘15) Nature]

Kryder's Law and S-curves underlying exponential growth

- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves (sigmoids) for different technologies

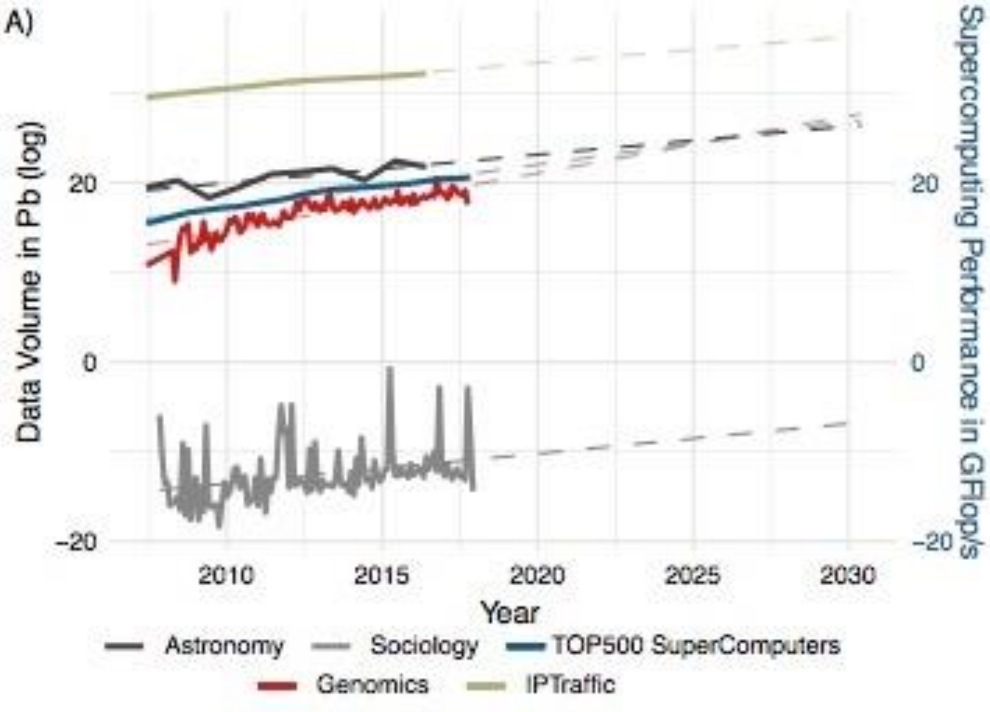


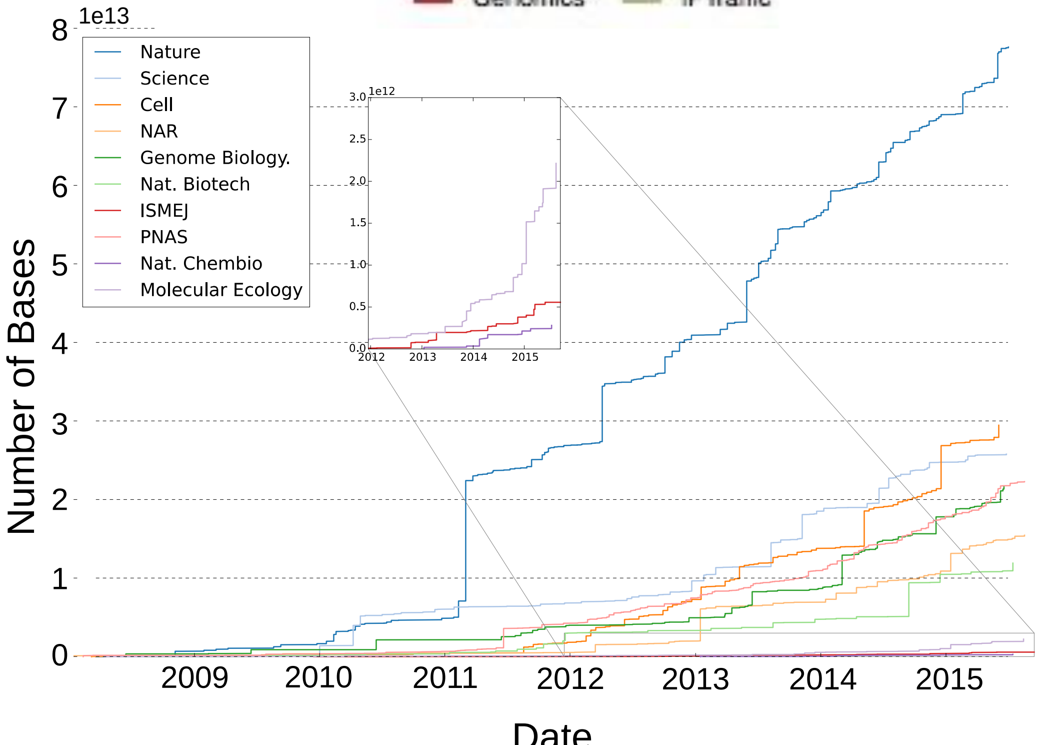


[Muir et al. ('15) GenomeBiol.]

Sequencing cost reductions have resulted in an explosion of data

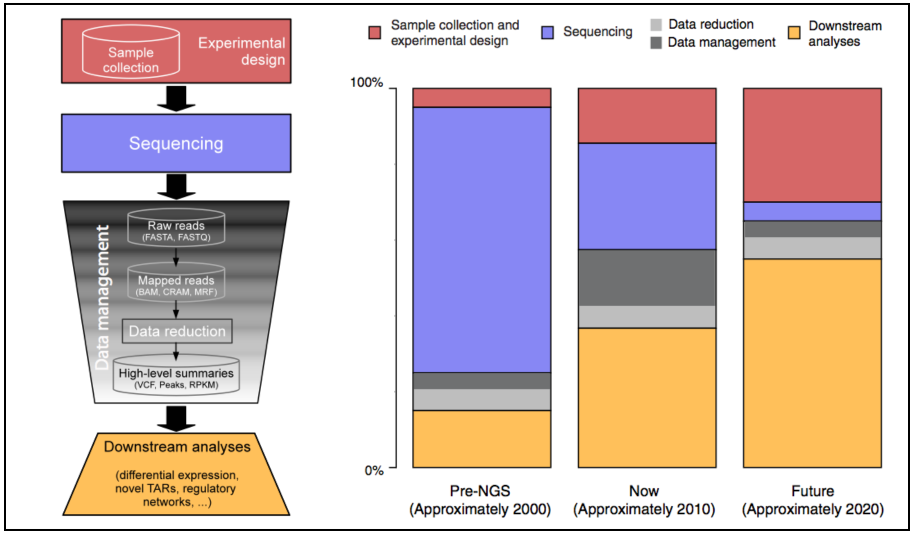
- The type of sequence data deposited has changed as well.





[Muir et al. ('15) GenomeBiol.]

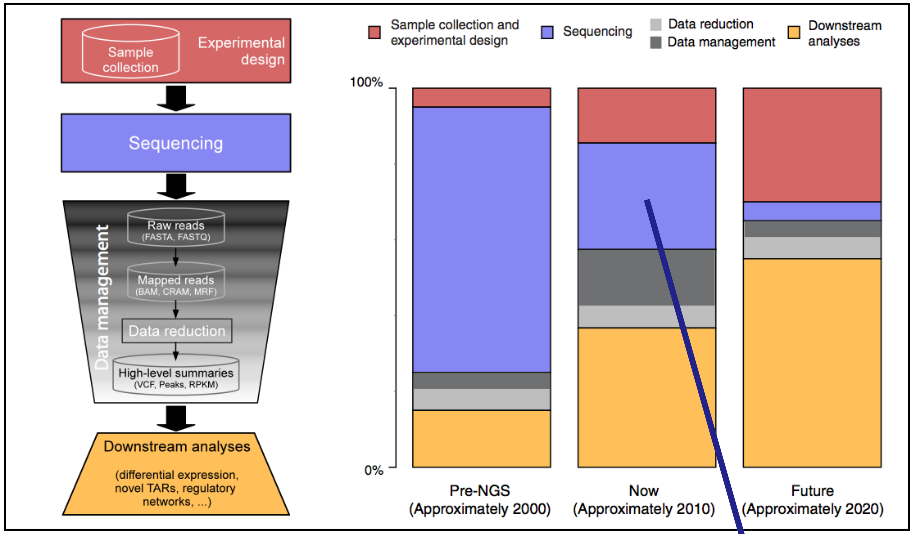
The changing costs of a sequencing pipeline



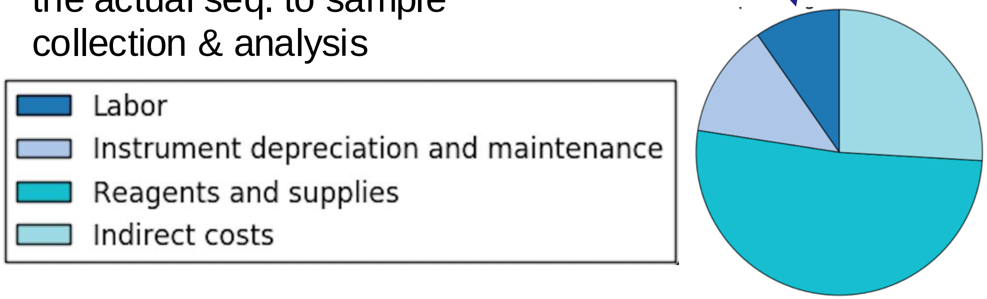
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from
the actual seq. to sample
collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

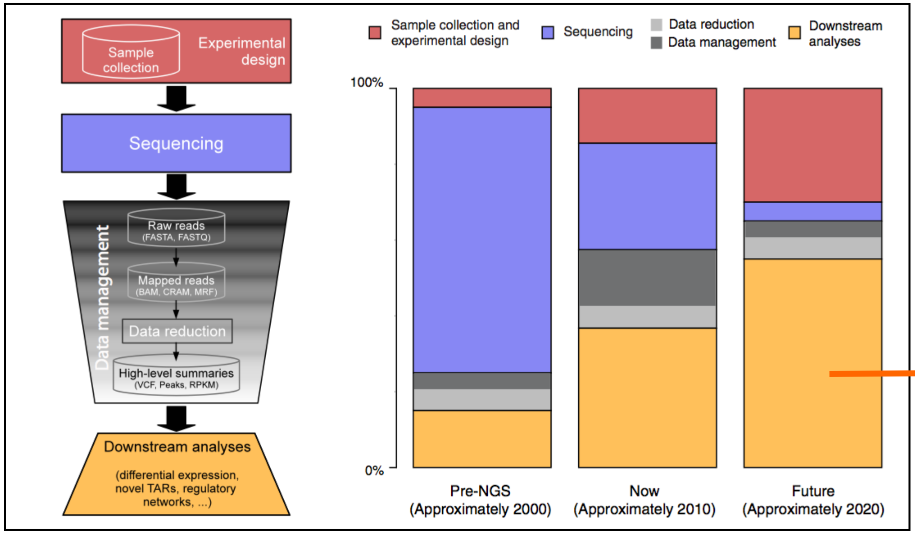


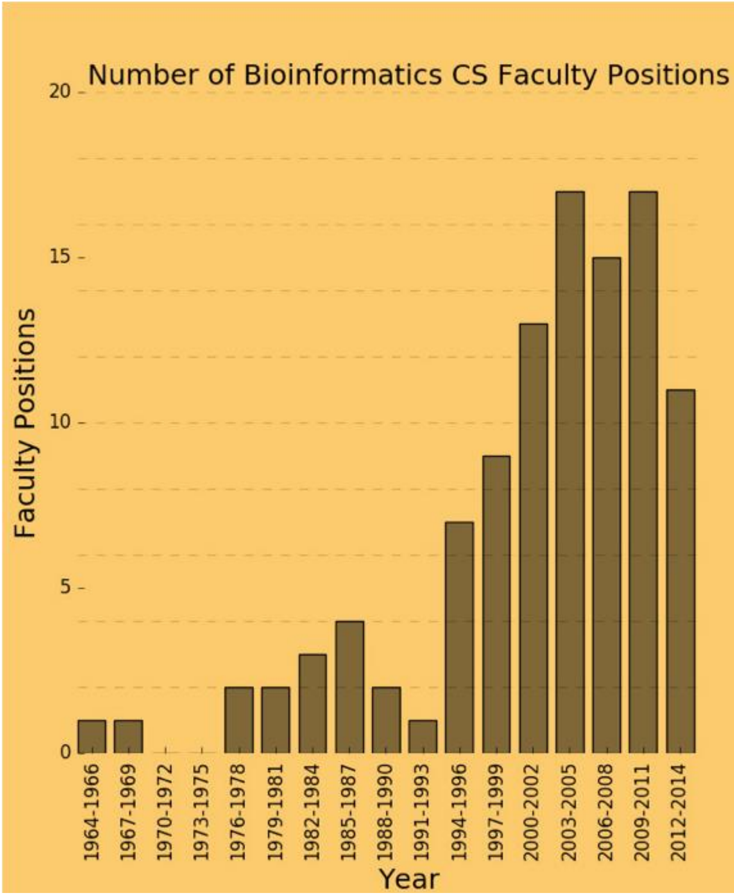
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis



[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline

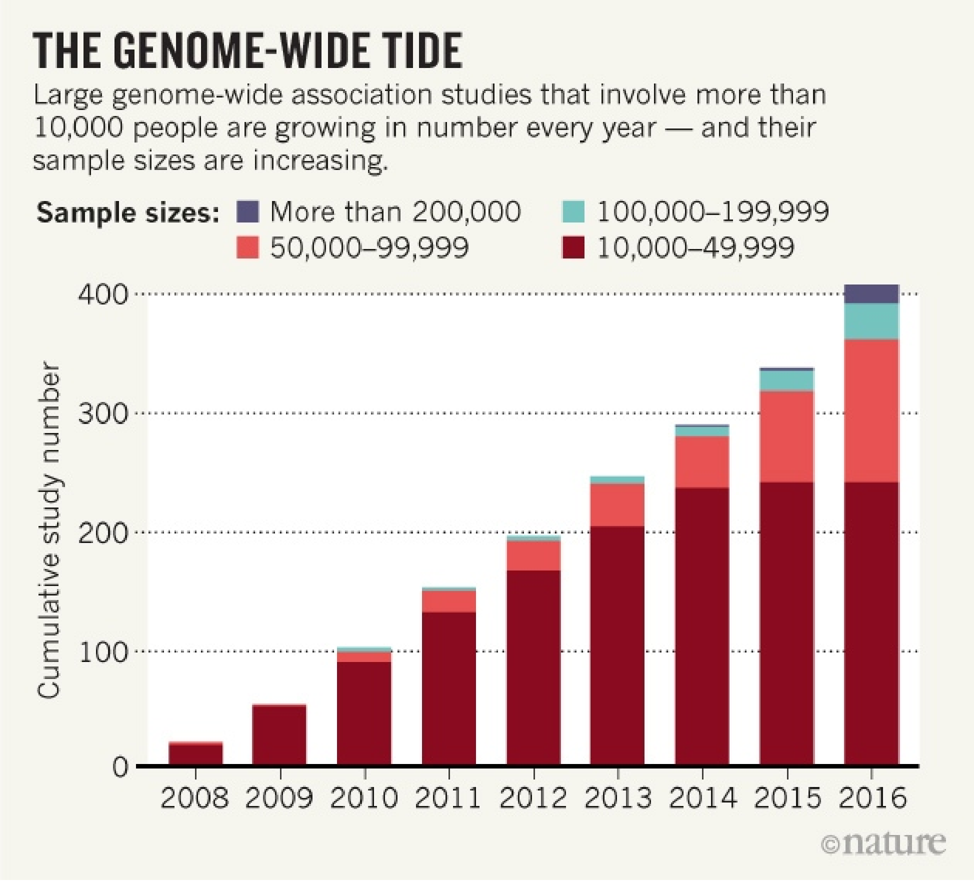




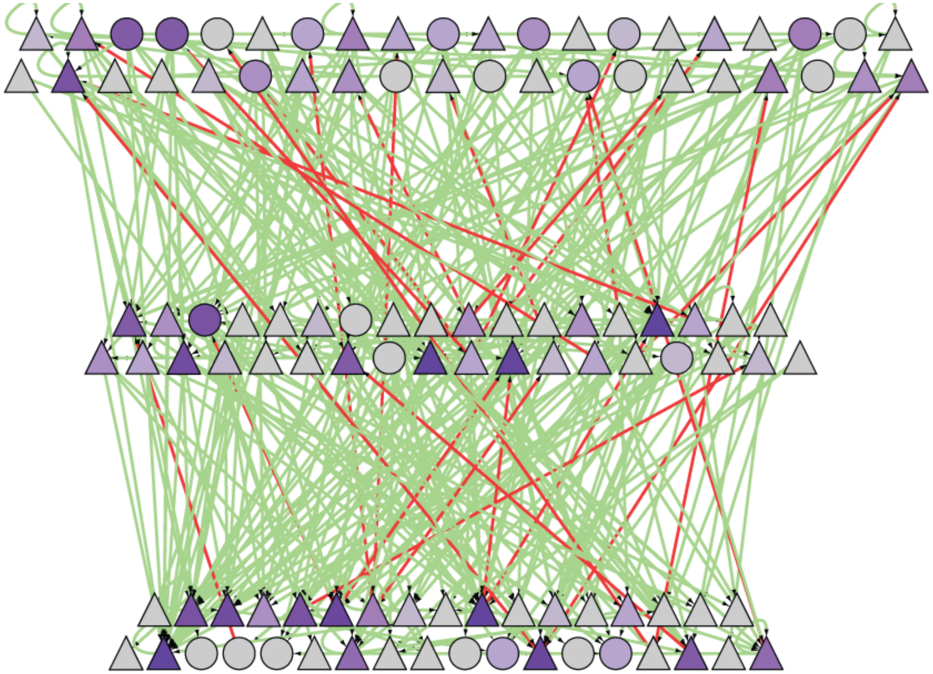
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

A Success of Scale & Integration: Many GWAS variants found, most not in genes, but affecting regulatory network


THE GENOME-WIDE TIDE

Large genome-wide association studies that involve more than 10,000 people are growing in number every year — and their sample sizes are increasing.



- A 1st GWAS done at Yale, for AMD: (Klein et al. 05, Science)
- Many since then
- Most SNVs fall into non-coding regulatory regions
  (major contributions by Yale groups to this ENCODE annotation effort)

[*Nature* 489: 91]

Basic Science to Medicine

INITIATIVES

AGTC

MVP

biobank uk

TCGA

Genomics england

PCAWG PanCancer Analysis of WHOLE GENOMES

THE PRECISION MEDICINE INITIATIVE

STARTUPS

color

23andMe

Veritas The Genome Company

FOUNDATION MEDICINE®

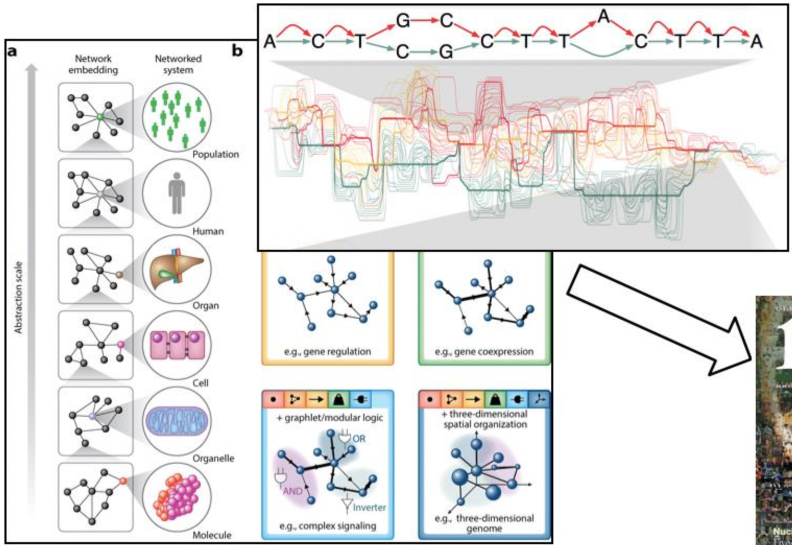
- Large-scale ‘omics data as an anchor to organize phenotypic data – EMRs, wearables…
- 1st [‘05-]: Exomes & chips of disease-focused cohorts – init. GWAS, TCGA, PGC
- 2nd [‘15-]: Integration of full WGS with rich & diverse phenotypes - UKBiobank, TopMed, Genomics England, PCAWG, All of Us

Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

Examples of Imports & Exports to/from Genomics & Other Data Science Application Areas

Technical Imports

Networks and graphs

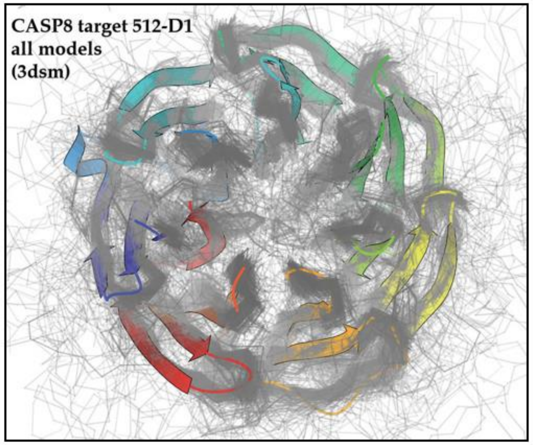


hadoop

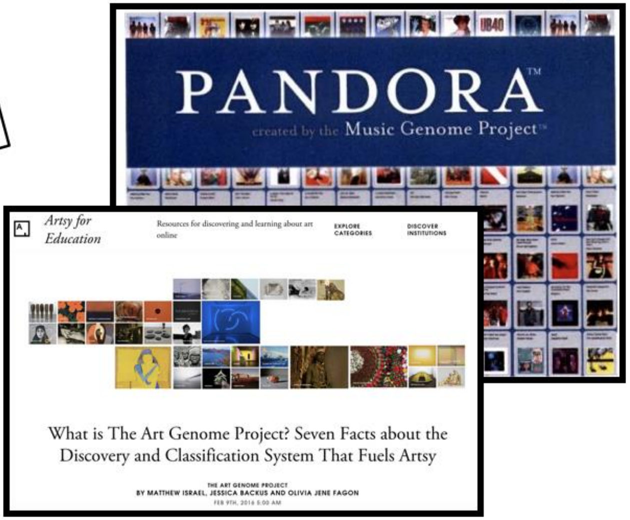
Importing tech. developed in other big data disciplines

Cultural Imports

CASP



Cultural Exports



Artsy for Education

Resources for discovering and learning about art online

EXPLORE CATEGORIES

DISCOVER INSTITUTIONS

What is The Art Genome Project? Seven Facts about the Discovery and Classification System That Fuels Artsy

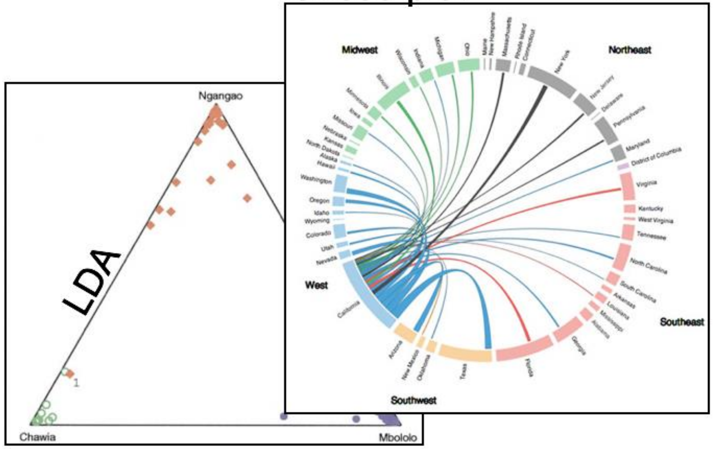
THE ART GENOME PROJECT

BY MATTHEW ISRAEL, JESSICA BACKUS AND OLIVIA JENE FAGON

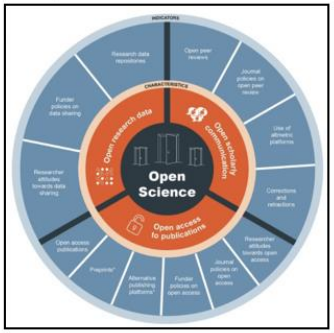
FEB 9TH, 2016 5:00 AM

Technical Exports

Circos plot

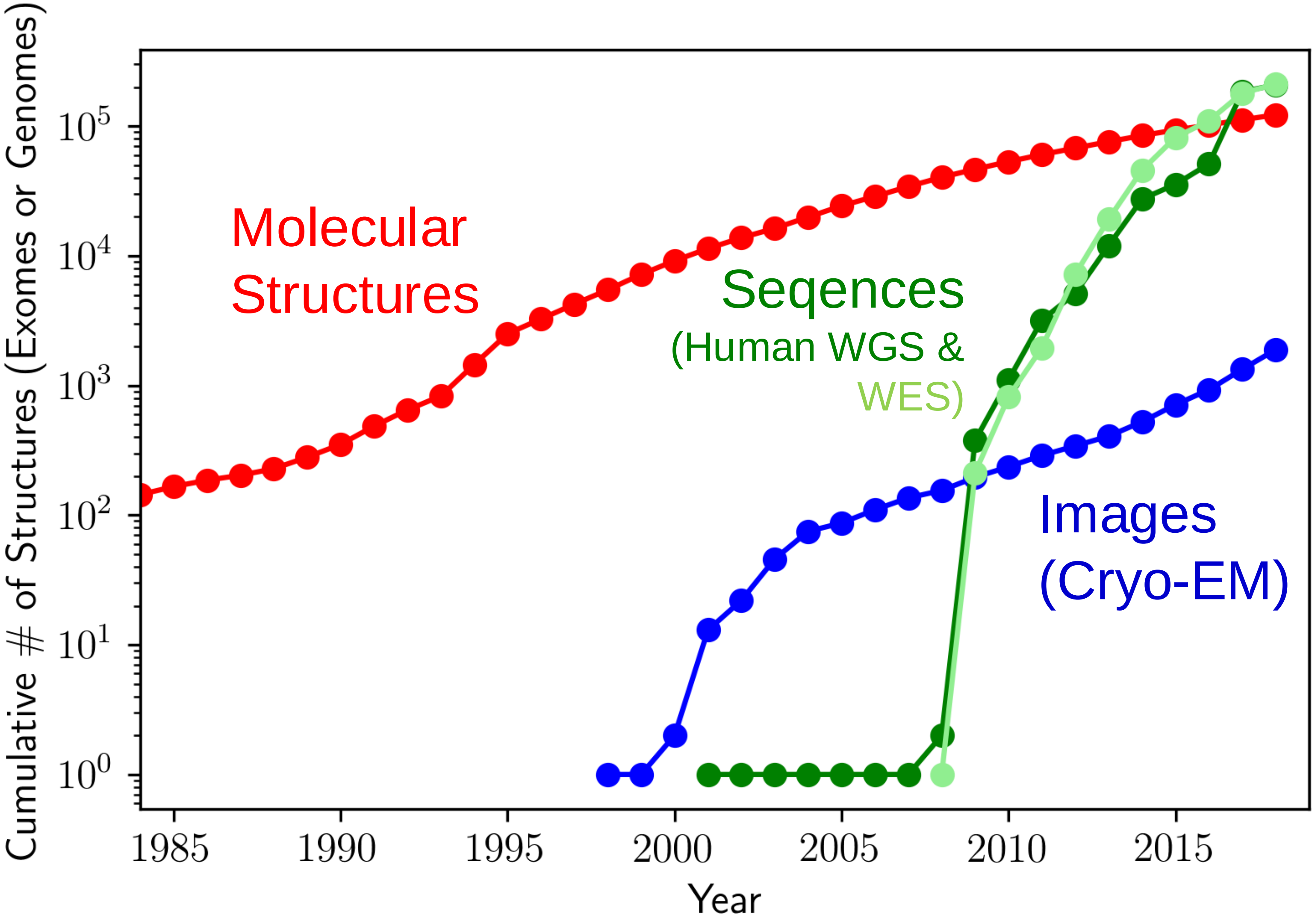


Open Science


INDICATORS
CHARACTERISTICS
Open Science
Open research data
Open scholarly communication
Open access to publications
Research data repositories
Open peer reviews
Journal policies on open peer review
Use of altmetric platforms
Corrections and retractions
Researcher attitudes towards open access
Journal policies on open access
Funder policies on open access
Alternative publishing platforms*
Preprints*
Open access publications
Researcher attitudes towards data sharing
Funder policies on data sharing

[Navarro et al. GenomeBiol. ('19, in press)]

How will the Data Scaling Continue? The Past, Present & Future Ecosystem of Large-scale Biomolecular Data

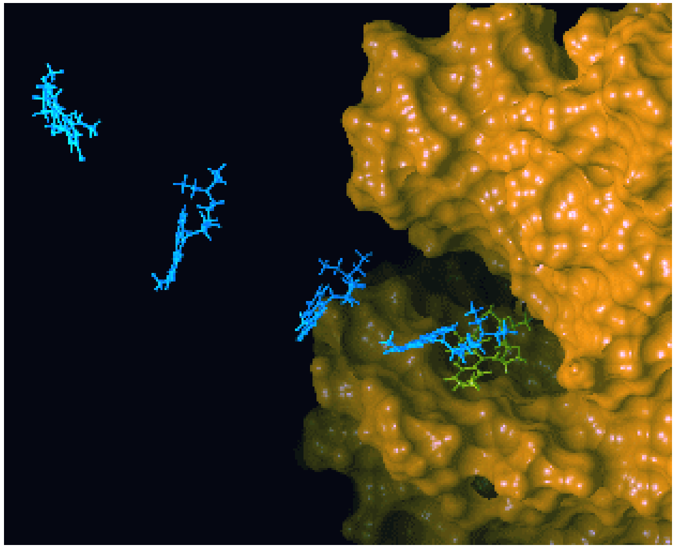


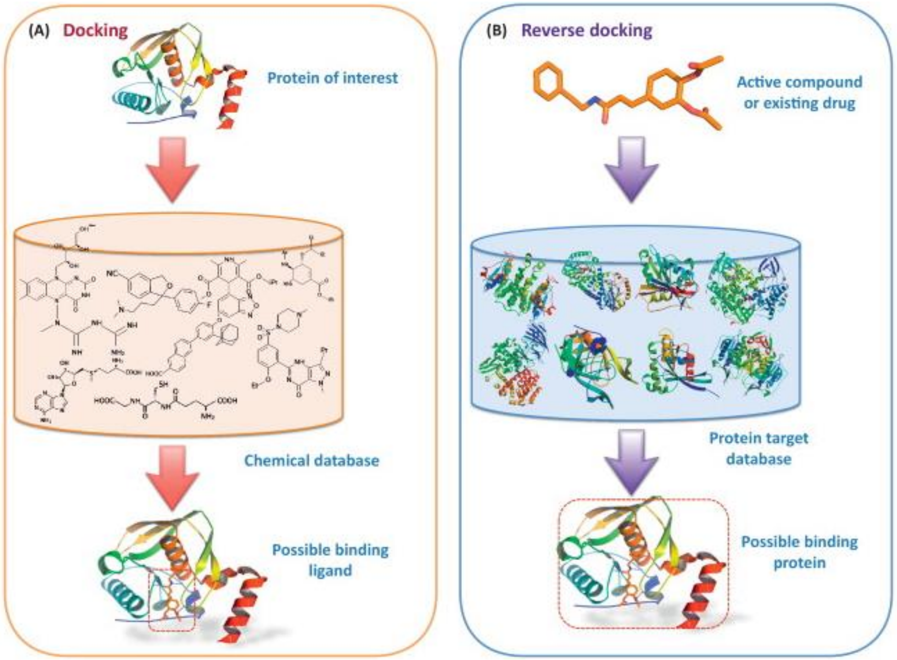
Biomed. Data science:

Applications

Major Application I:
Designing Drugs from Structural Targets

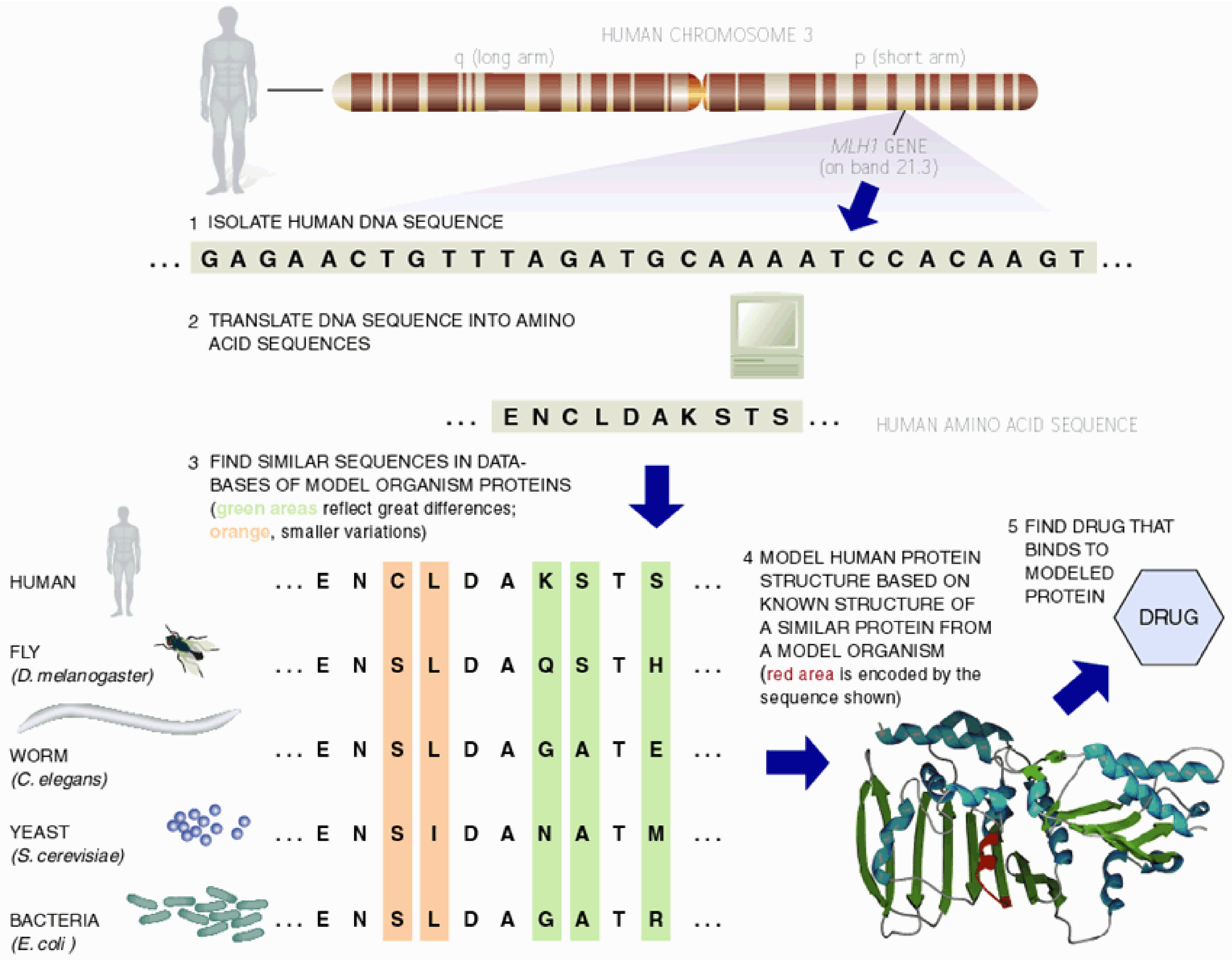
- Understanding how structures bind other molecules
- Designing inhibitors using docking, structure modeling
- *In silico* screens of chemical and protein databases





(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

... GAGAACTGTTTAGATGCAAAATCCACAAGT ...

2 TRANSLATE DNA SEQUENCE INTO AMINO ACID SEQUENCES

... ENCLDAKSTS ... HUMAN AMINO ACID SEQUENCE

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	...

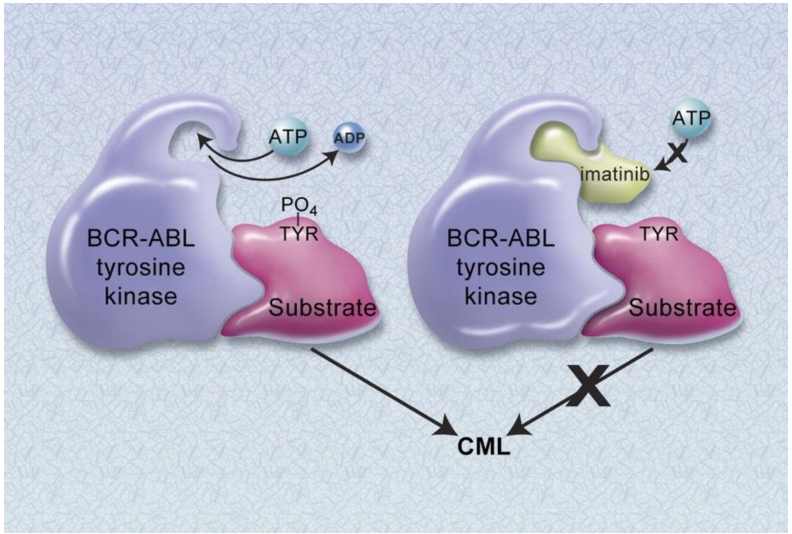
4 MODEL HUMAN PROTEIN STRUCTURE BASED ON KNOWN STRUCTURE OF A SIMILAR PROTEIN FROM A MODEL ORGANISM (red area is encoded by the sequence shown)

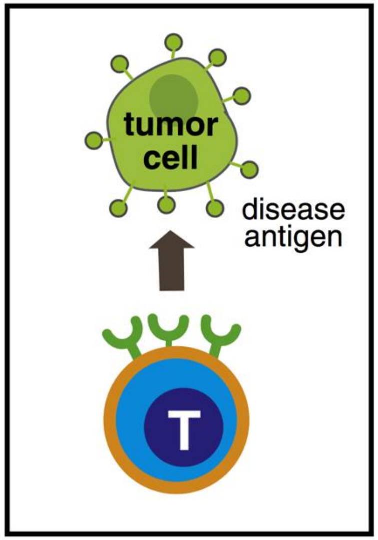
5 FIND DRUG THAT BINDS TO MODELED PROTEIN

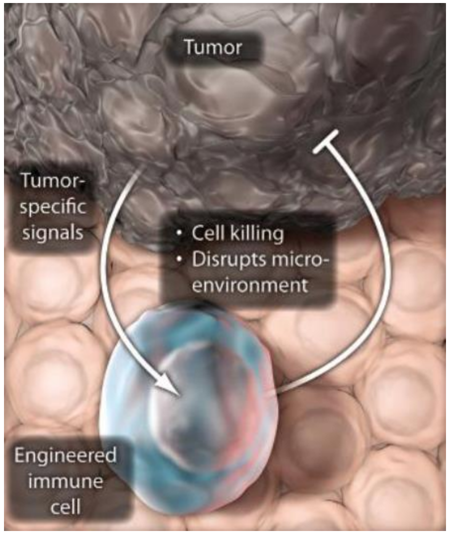
[Adapted from Sci. Am.]

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





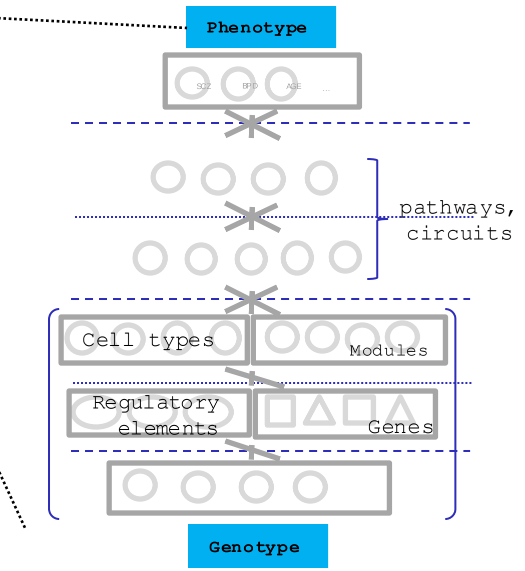


(From left to right, figures adapted from Druker BJ. Blood 2008 and the Lim Lab at UCSF)

Lectures.GersteinLab.org

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%

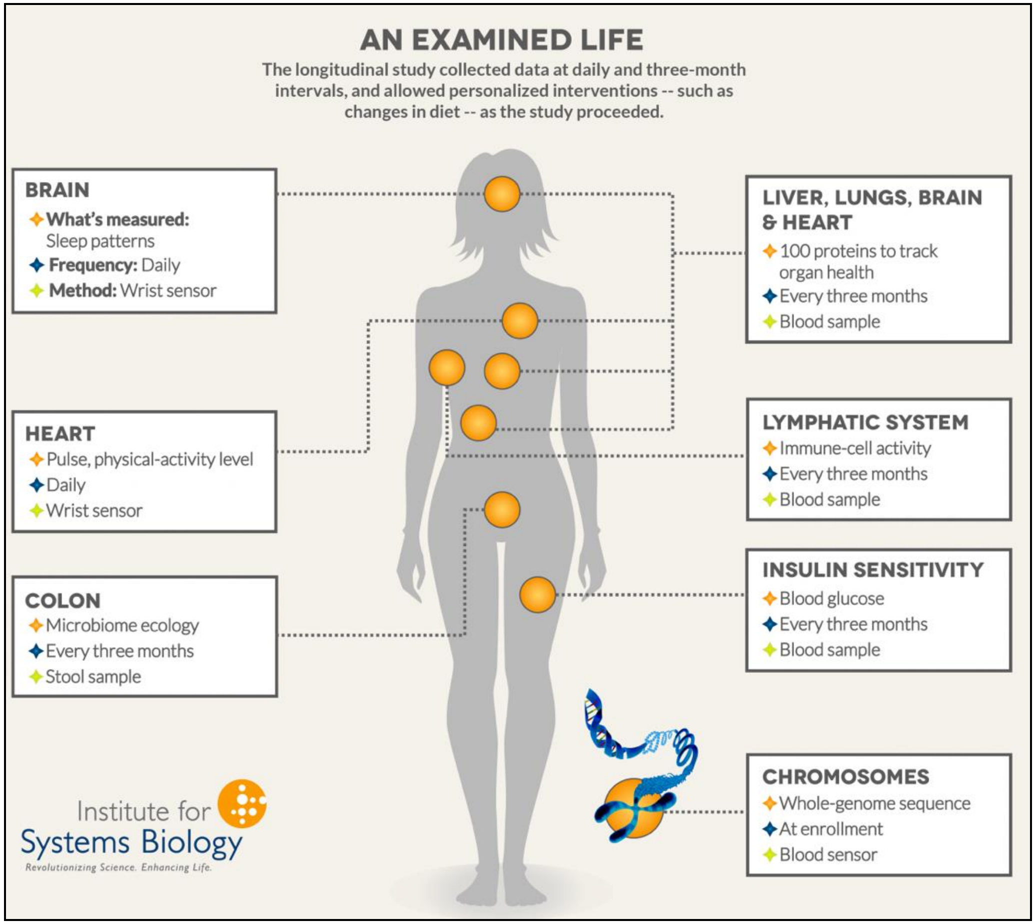
Thus, interested in developing predictive models of psychiatric traits which:

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

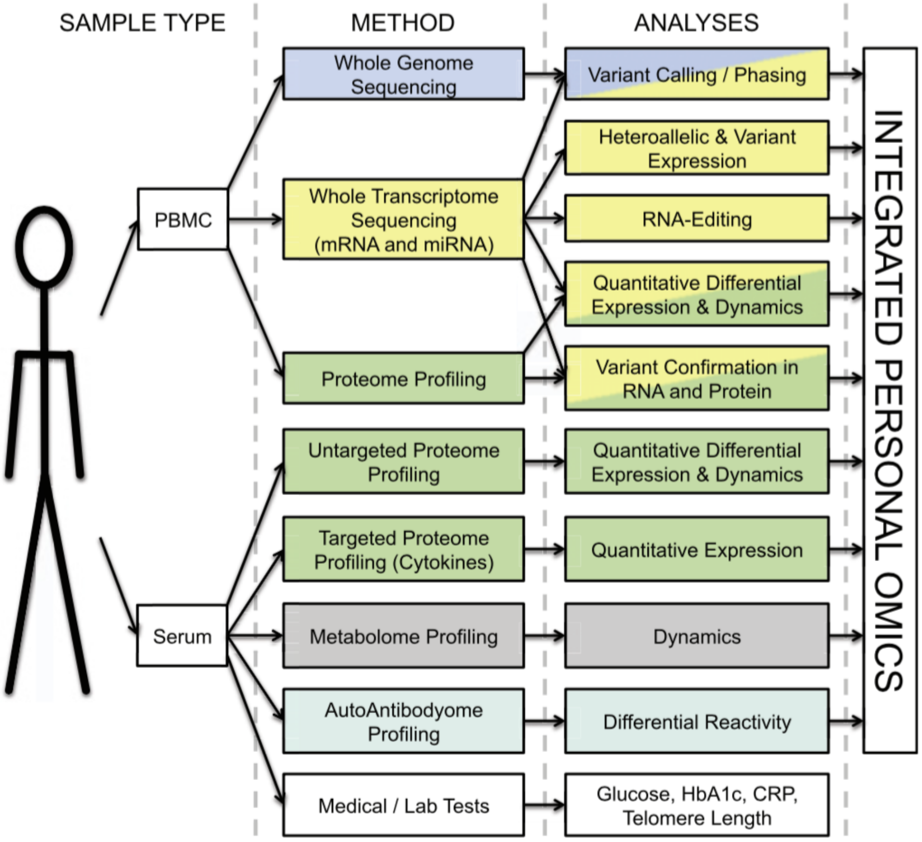

AN EXAMINED LIFE
The longitudinal study collected data at daily and three-month intervals, and allowed personalized interventions -- such as changes in diet -- as the study proceeded.
BRAIN
What's measured: Sleep patterns
Frequency: Daily
Method: Wrist sensor
HEART
Pulse, physical-activity level
Daily
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)

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. - https://www.cell.com/cell/fulltext/S0092-8674(12)00166-3

Expanding personalized medicine beyond the genome.

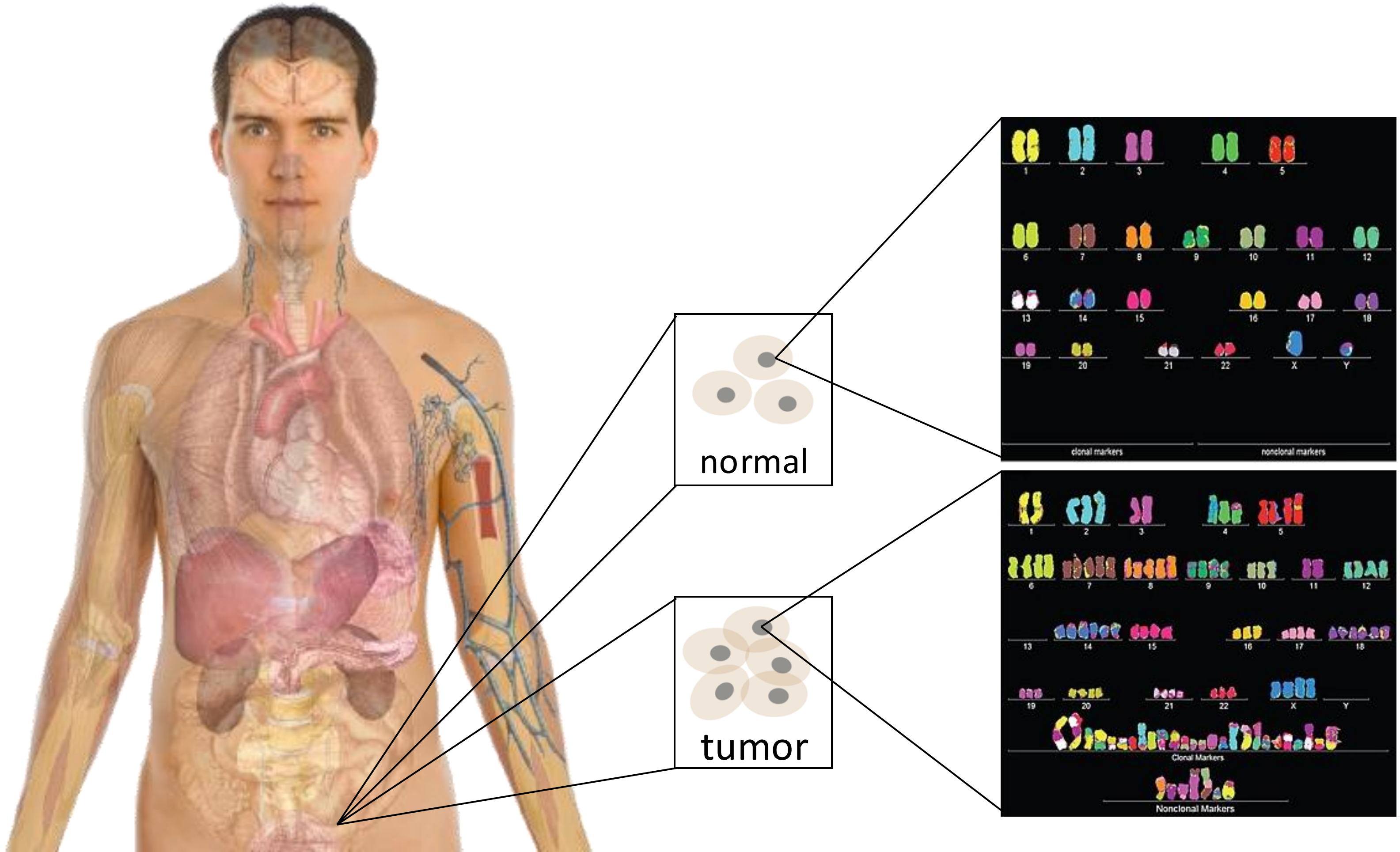
- An integrated personal omics profile (iPOP) is an example of a more comprehensive version of personalized medicine.

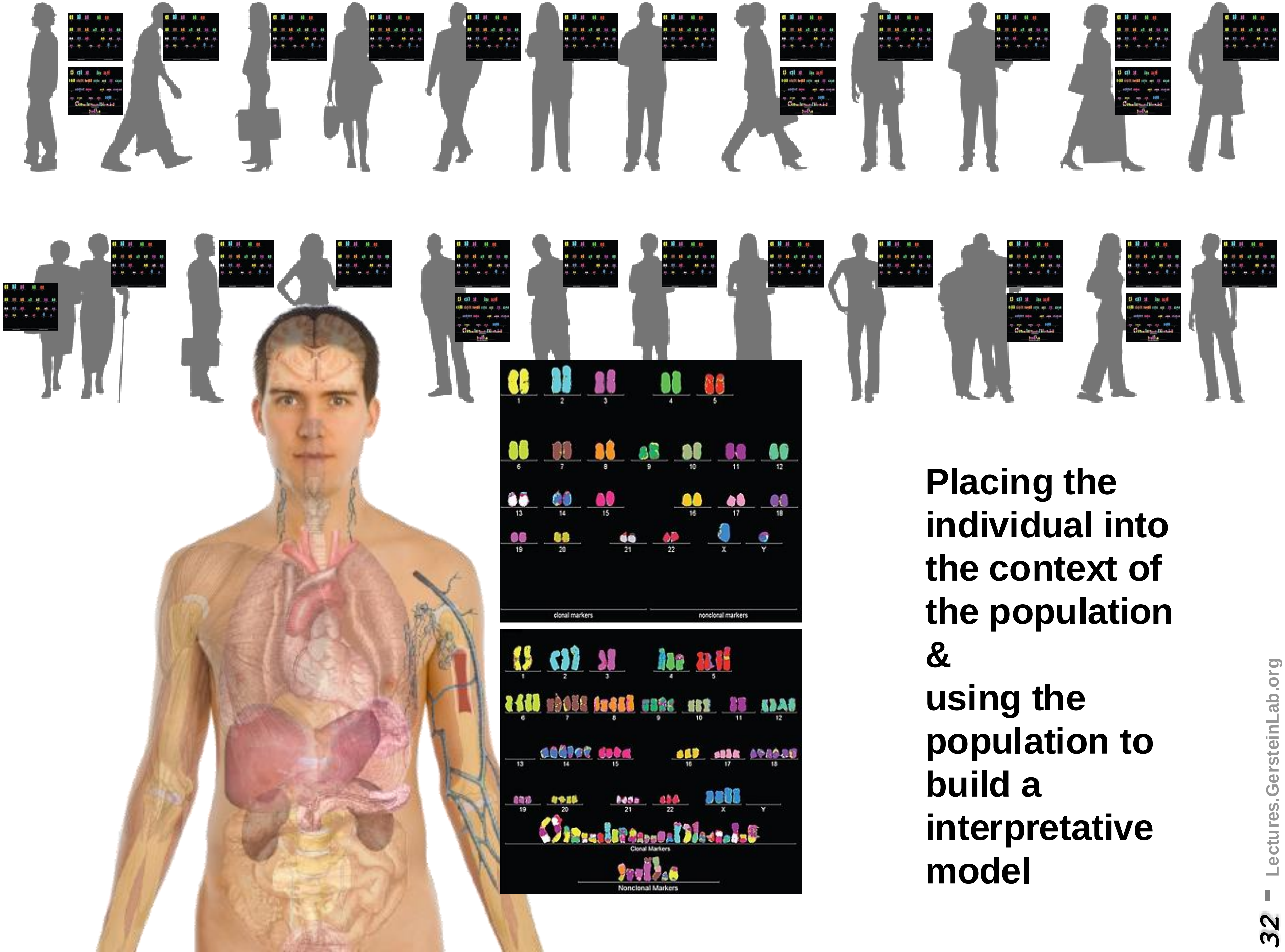
- Michael Snyder had his genome sequenced and collected many other large scale datasets over an extended period of time.

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

Our field as future Gateway – Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.





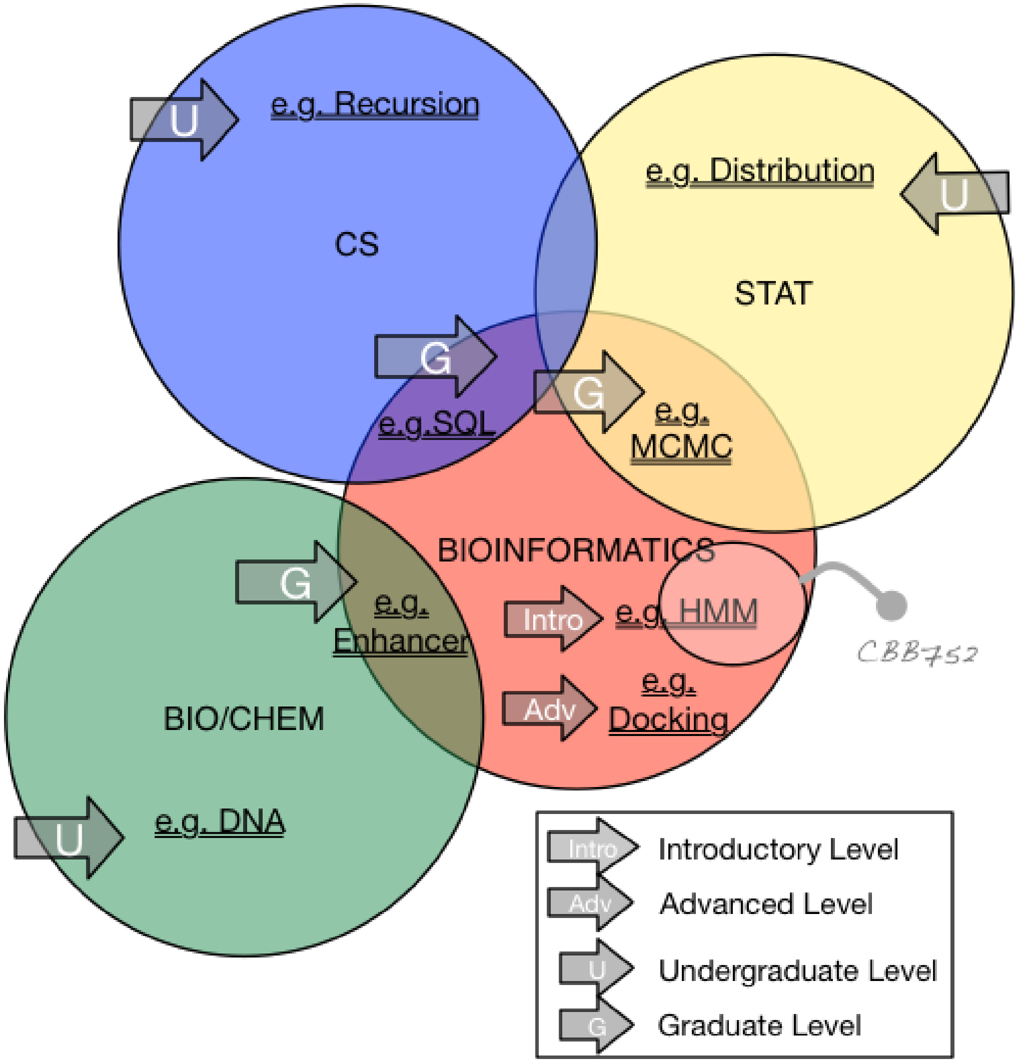
Placing the individual into the context of the population
&
using the population to build a interpretative model

Biomed. Data science:

The Course

Defining the field – by crowd-sourced judgement

- Bioinformatics
  - Related terms
    - Biological Data Science
    - Bioinformatics & / or / vs Computational Biology
    - Bio-computing
    - Systems Biology
    - “Qbio”
- What are its boundaries
  - Determining the "Support Vectors”



Overview of Topics Surveyed

Introduction & Overview of the Data

- Genomics & Sequencing
- Proteomics & Structure
- Databases

Data Mining & Machine Learning

- ”Classic” Supervised & Unsupervised Approaches
  - Decision Tree & SVMs
  - Clustering & SVD
- Application to ’Omics Data
  - Comparing sequences
  - Processing single cell & epigenomic data

Network Analysis

- Topology & Connectivity
- Gene Networks

Deep Learning

- Basic Theory & Applications

Physical Modeling

- Macromolecular Simulation
- Markov Models
- Molecular Packing

Additional Topics

- Privacy
- Personal Genome Analysis
- Image Analysis

What is Bioinformatics?

- *(Molecular)* **Bio** - **informatics**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Thoughts on the Class

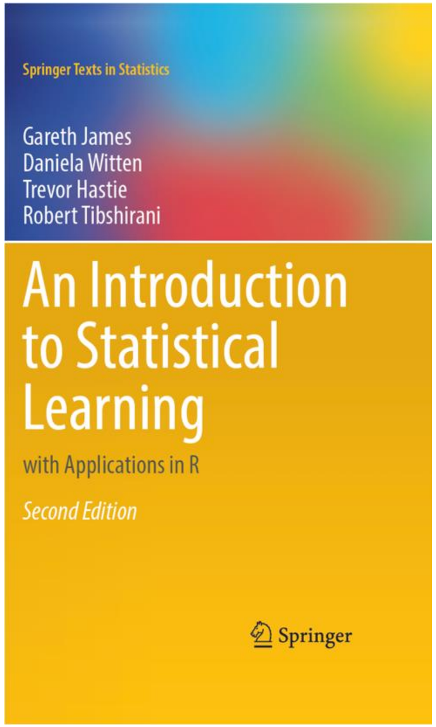
GersteinLab.org/courses/452
(Class Web Page)

- Broad overview with a few deep dives
  - Fundamentally interdisciplinary field
  - Here, focusing on molecular bioinformatics
  - Some deep dives into sequence comparison, Bayesian approaches, low-dimensional representations
  - Steering away from material in related Yale classes
- Goal is good intuition on approaches & the application area
  - Apply to related problems
- Lectures provide structure of knowledge to be assimilated
  - Varied backgrounds
  - Variety of learning approaches
- Sections for interaction & more hands-on treatment
- Quizzes & homework for individual command of basic knowledge
- Final Project for teamwork
- cbb752@gersteinlab.org for issues

Lectures (& Readings)

- Lectures form the backbone of what you need to know
  - We will post final pptx & pdf AFTER the lecture
  - Also, will have current, lecture-hall recorded videos put up quickly on canvas
  - Class-produced lecture summaries about a week after each lecture (see course website)
- No book
  - Key readings for each lecture listed in the slides
  - ISLR as close as we can get to a text
  - Section papers

- Past Year's As a Guide
  - Convention for numbering lectures: **YYMN** = (**Y**)ear, (**M**)odule, (**N**)umber e.g. 23m3, 22m3, (21) M3
  - If you want to look ahead, we will mostly follow the flow in 2021-2023 (See the notations at the top of each slide pack for key differences.)
  - Mostly 2021 has well-produced videos, with a few from following years



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

Springer

Key References for i1+i2a

(ranked from most #1 to least important)

1. Navarro et al (2019) Genome Biology
   https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1724-1
   (Read Abstract & Introduction)

2. Muir et al (2016) Genome Biology
   https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0917-0
   (Read first 3 sections)

3. Zimmer (2023). STAT
   https://www.statnews.com/feature/game-of-genomes/season-one
   (Just look at the first season & read more if you want.)

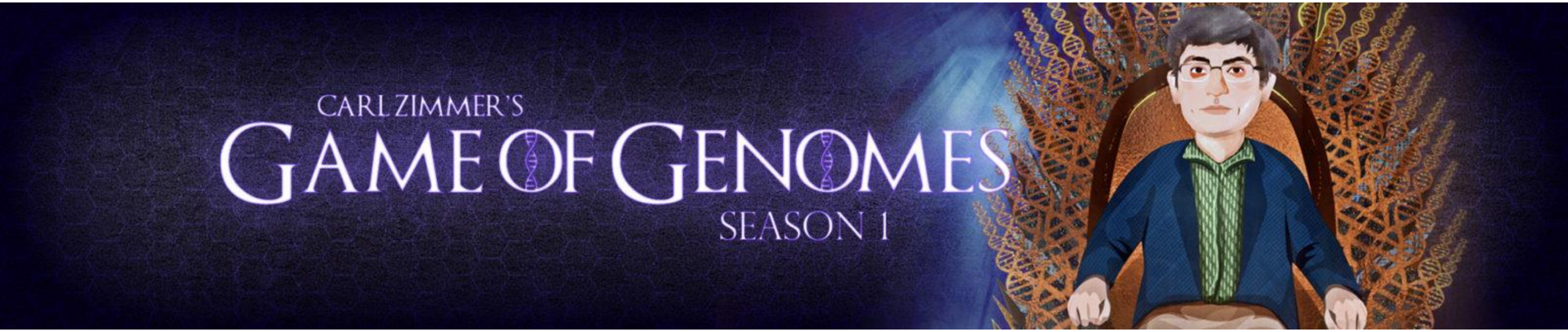
4. Luscombe et al (2001) Methods Inform Med
   http://archive.gersteinlab.org/papers/e-print/whatis-mim/text.pdf
   (Read Introduction)

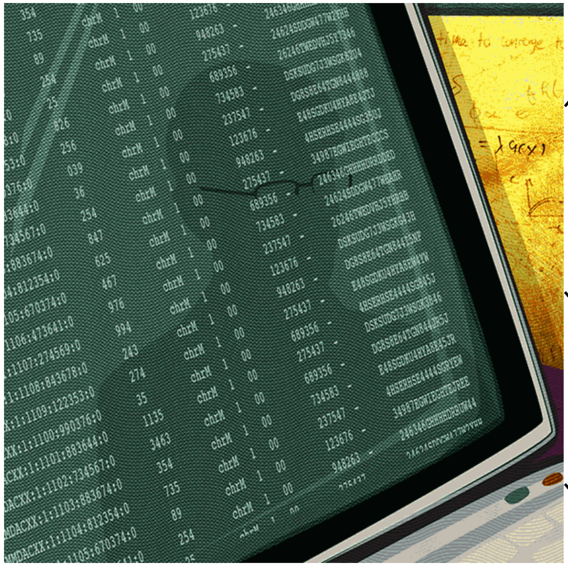
5. Babu et al. (2023). Annual Review of Medicine
   https://doi.org/10.1146/annurev-med-052422-020437
   (Just Skim)

Biomed. Data science:

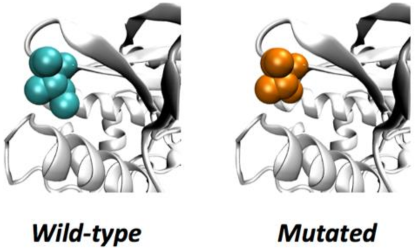
The Final Project

Analyzing Carl Zimmer's genome





SNV AAGCT → ACGCT



Protein Structure

Ancestry

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

2017

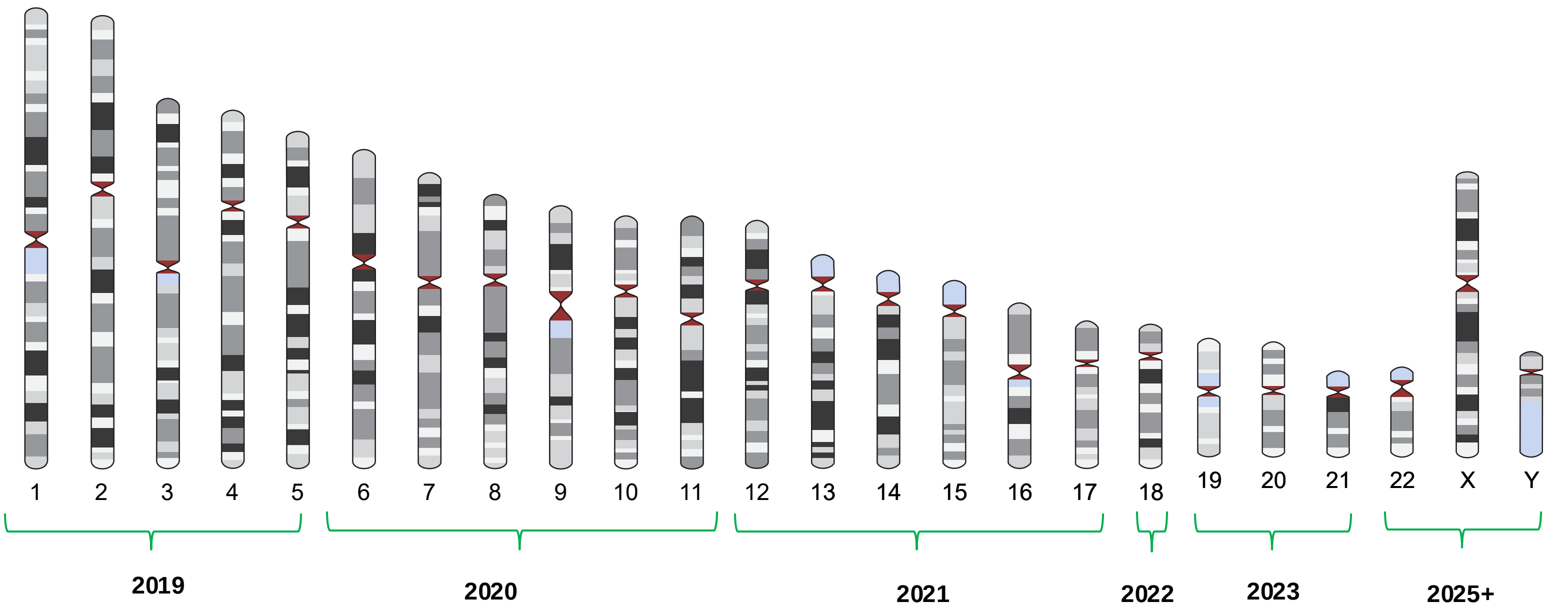
- Each group created a GitHub page detailing the work of each team
  - Additionally, each group has a power point presentation:
- Topics of projects include:
  - Comparative analysis of personal genomes
  - Personal genomes and personalized medicine (CRISPR)
  - Network analysis of personal genomes
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

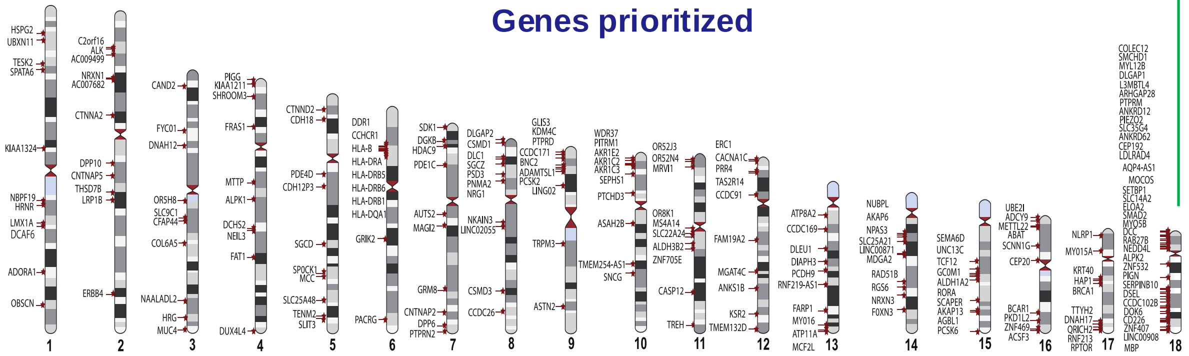
2019 – 2023

- 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

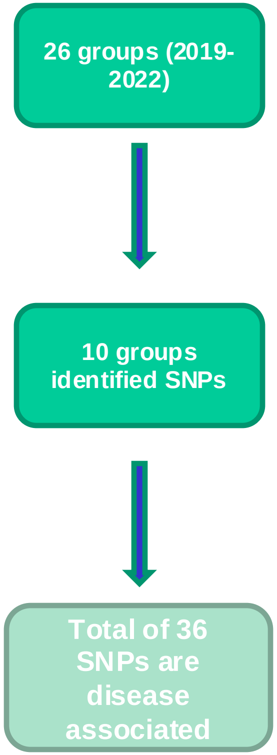


History of Analyzing the “Zimmerome” in Class

Genes prioritized

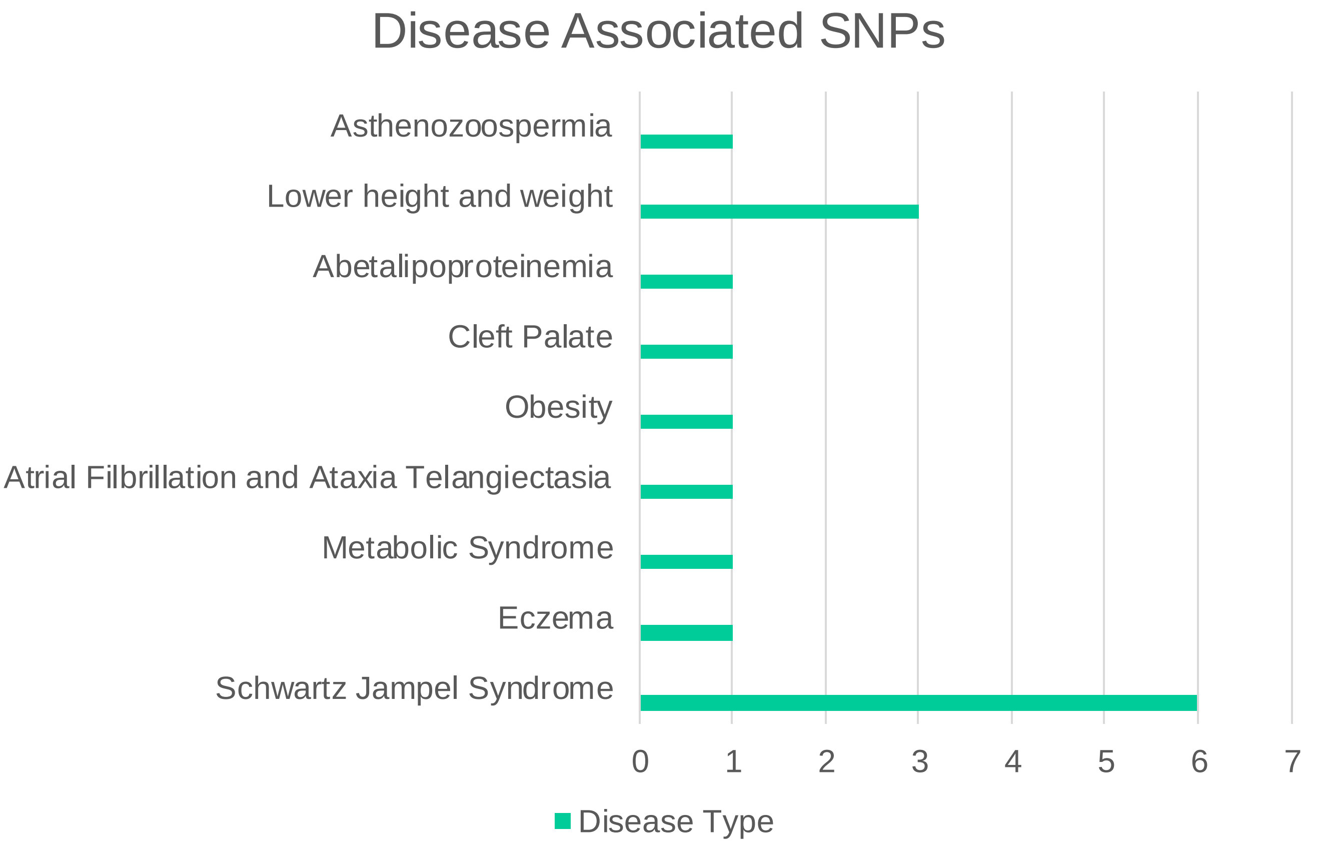


History of the Analysis of the “Zimmerome” in the Class

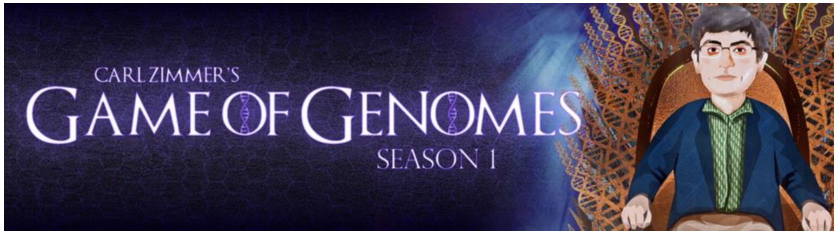

26 groups (2019-2022)

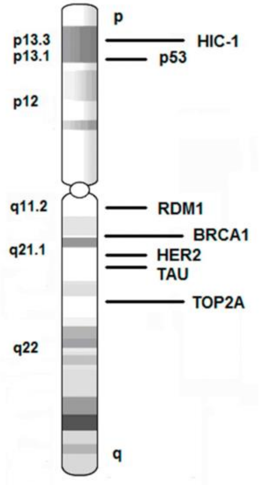
10 groups identified SNPs

Total of 36 SNPs are disease associated



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





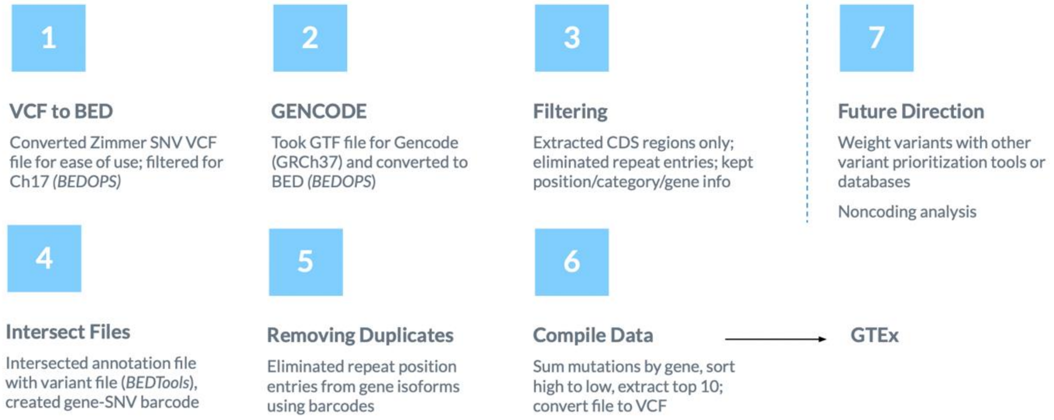
Team based approach
- Assigned Teams (4-5 people in your section, assigned by TFs)
- Each team focuses on a single chromosome
- Cross-disciplinary

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 → **GTEx**

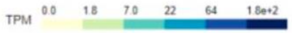
Sum mutations by gene, sort high to low, extract top 10; convert file to VCF

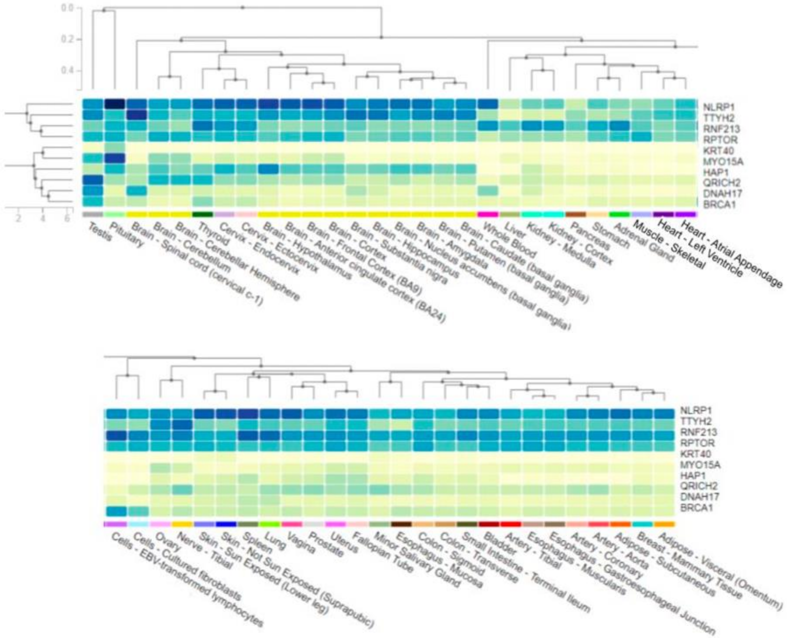
Computational Pipeline

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

2. Biological/Biomedical Interpretation

Tissue Specific Expression Extracted from GTEx

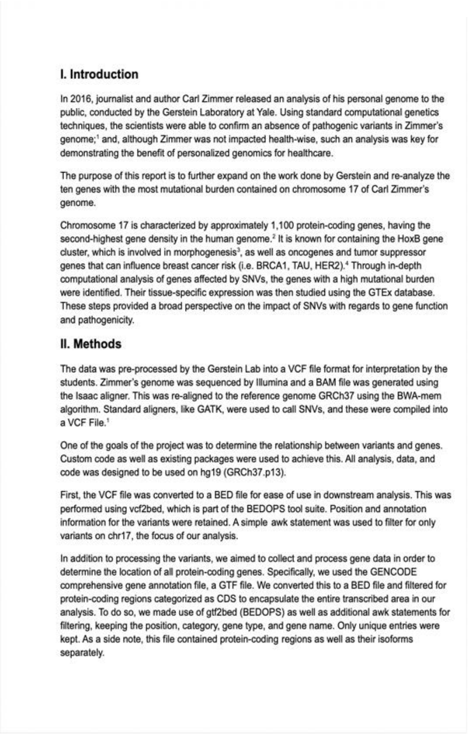




Interpretation of Results

- 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.

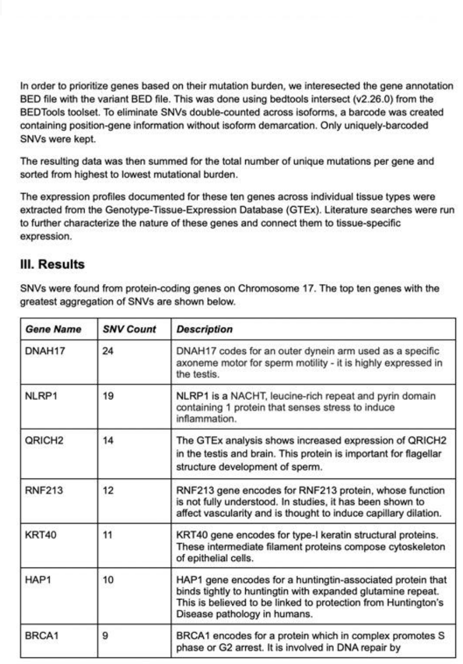
II. Methods

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 23
  - 2 to 3 minute mp4 recording per group (nominate 1 person to make the recording)
  - We will play these recordings in class on 4/23
- in your Discussion Section of Week April 23
  - Approximately 10 minute presentation by other members of the group

Written Report

- Due: May 5, 2025
- At least 1000 words

Summary Slide

- 1 summary slide giving an overview of your project

Summary Metadata File

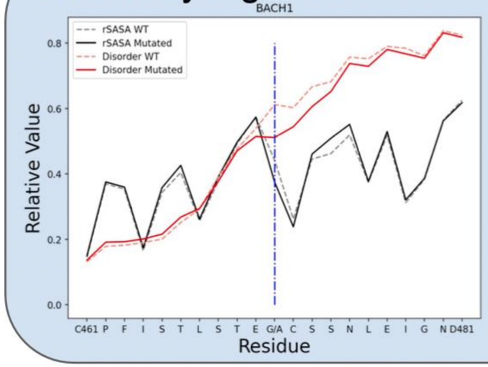
- A single text file containing relevant information
- More description in assignment file

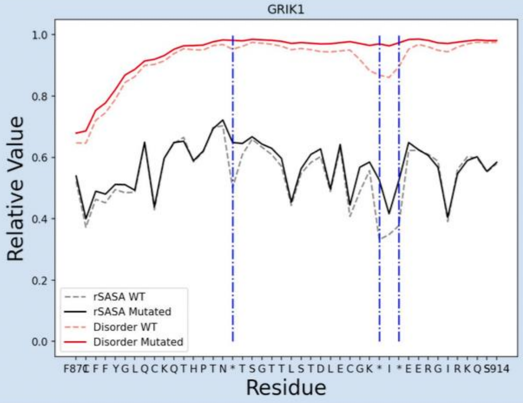
2023 group 3 (chr 21)

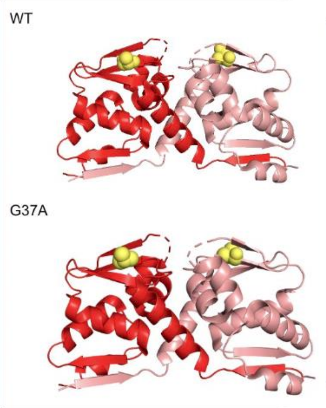
Top 10 Prioritized Genes

1. DSCAM
2. RUNX1
3. NCAM2
4. KCNJ6
5. TSPEAR
6. GRIK1
7. ERG
8. APP
9. CHODL
10. BACH1

Summary Figures:







Summary:

1. Prioritization approach: protein coding genes with highest mutational burden
2. Downstream analysis: protein structure (rSASA) analysis with NetSurfP
3. Findings:
   a. APP causes protein aggregates that are well-linked to Alzheimer's: found no protein coding variants (good news!)
   b. GRIK1 L902S has a correlation with ADHD
      - it is a cationic channel in the Cerebellum and hypothalamus; binds to excitatory neurotransmitter L-glutamate

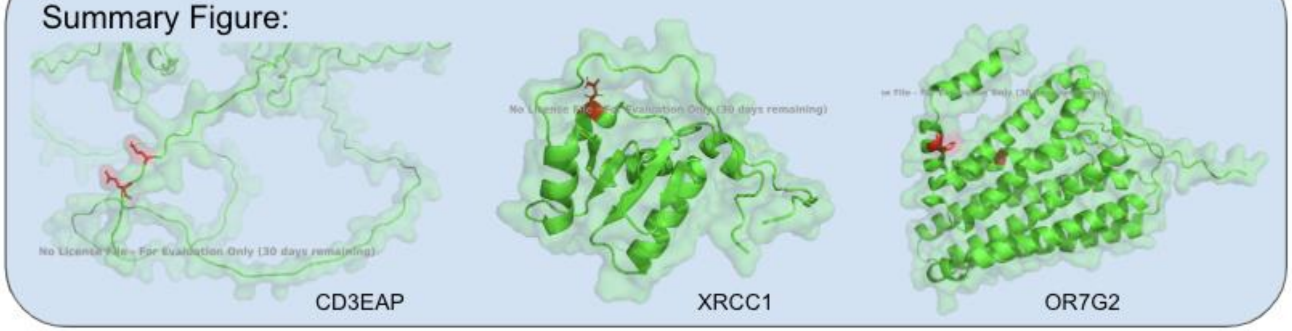
Lectures.GersteinLab.org

2023 Group 1
Section 3 (chr 19)

Top 10 Prioritized Genes

1. CD3EAP
2. XRCC1
3. OR7G2
4. OR10H5
5. LILRB4
6. KIR2DL3
7. KIR2DL1
8. KIR3DL2
9. ZFP28
10. GP6

Summary Figure:



Summary:

1. Resistance to Cisplatin-based cancer treatment
2. Dysregulation of certain immune cell subtypes ability to distinguish host and tumor cells
3. Dysregulation of collagen-based platelet adhesion

Notes on the Course

- Surveys: Please make sure these are done quickly
  - will count in overall grade
  - Available from GersteinLab.org/courses/452

- Sections:
  - Discussion section assignments will be sent out shortly
  - Will start next week
  - Lecture summaries will start on Wed.

- For issues, please discuss with TFs right after class or email cbb752@gersteinlab.org
  - e.g. for students registered as “guest student” on canvas