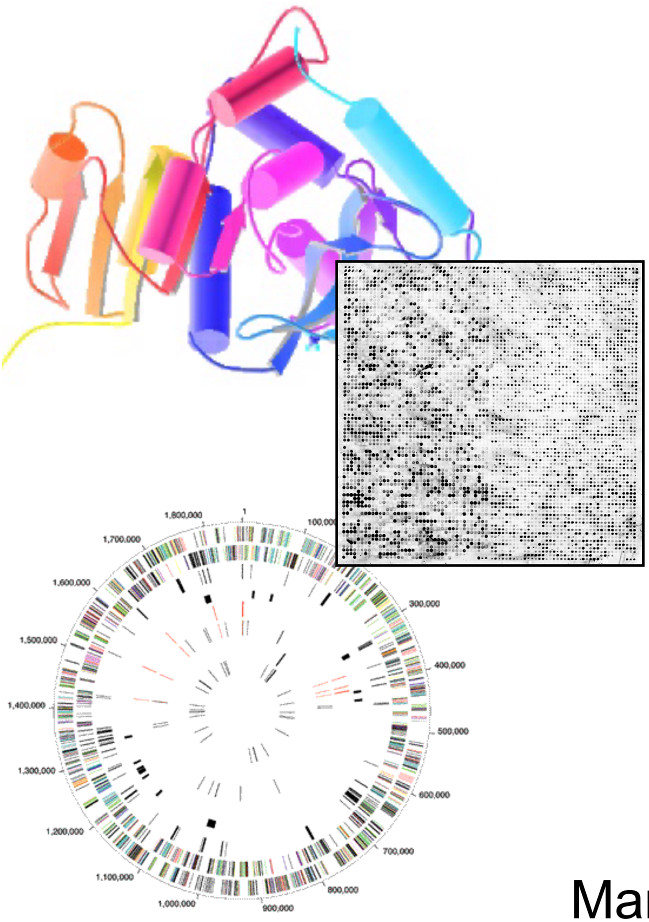
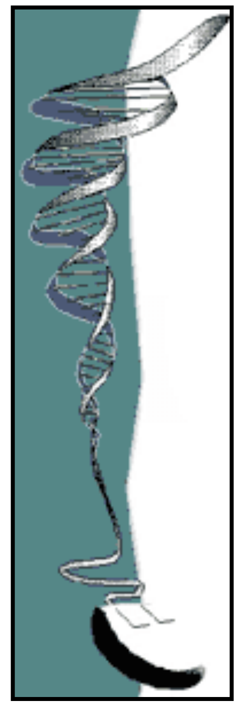
Biomedical Data Science: Introduction

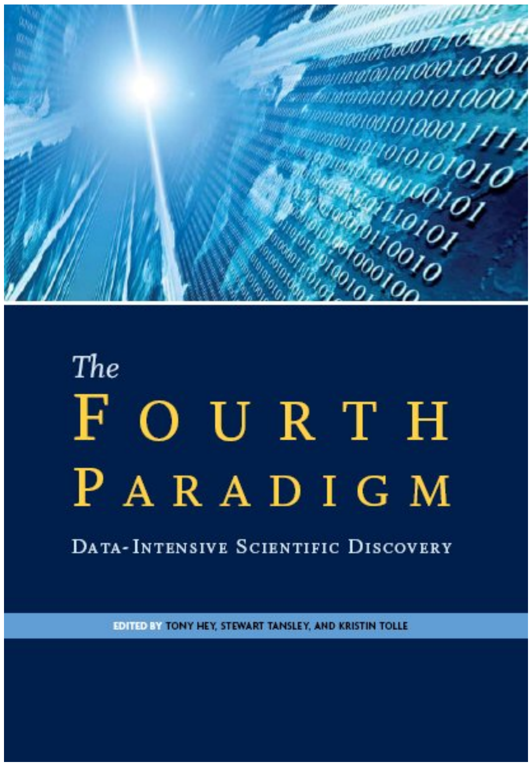




Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring ’18)

Overview:
what is
Biomed. Data science?
(Placing it into context)

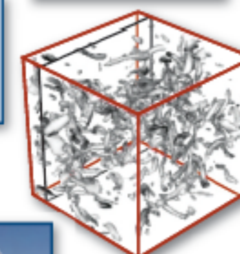
Jim Gray's 4th Paradigm



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

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



- 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

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

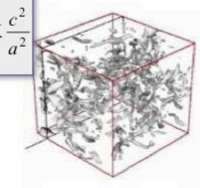
Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena

- Last few hundred years:
  theoretical branch
  using models, generalizations

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

- 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

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…

Forbes · New Posts +36 posts this hour · Most Popular 16-Year-Old Innovator · Lists Promising Companies

108 Share · 349 Tweet · 193 Share · 353 Submit · 12 +1

CIO Network
INSIGHTS AND IDEAS FOR TECHNOLOGY LEADERS.
+ Follow (469)
Cognizant

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

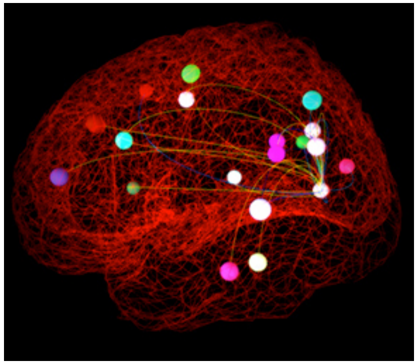
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne up. The company had just under 8 million accounts, and the number was growing qu friends and colleagues to join. But users weren't seeking out connections with the pe rate executives had expected. Something was apparently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

[Oct. '12 issue]

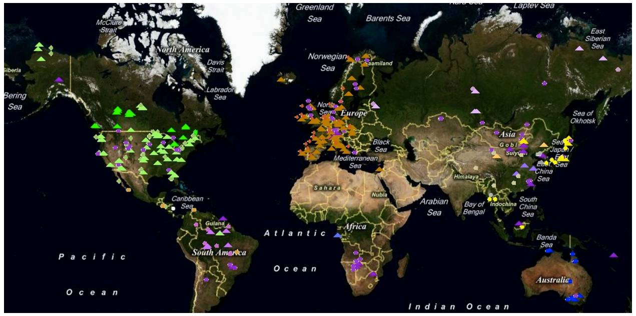
Data Science in Traditional Science

- 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

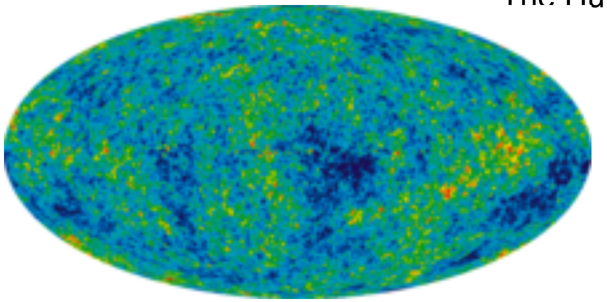


Neuroscience - The Human Connectome Project



Ecology & Earth Sci. - Fluxnet

Genomics DNA sequencer



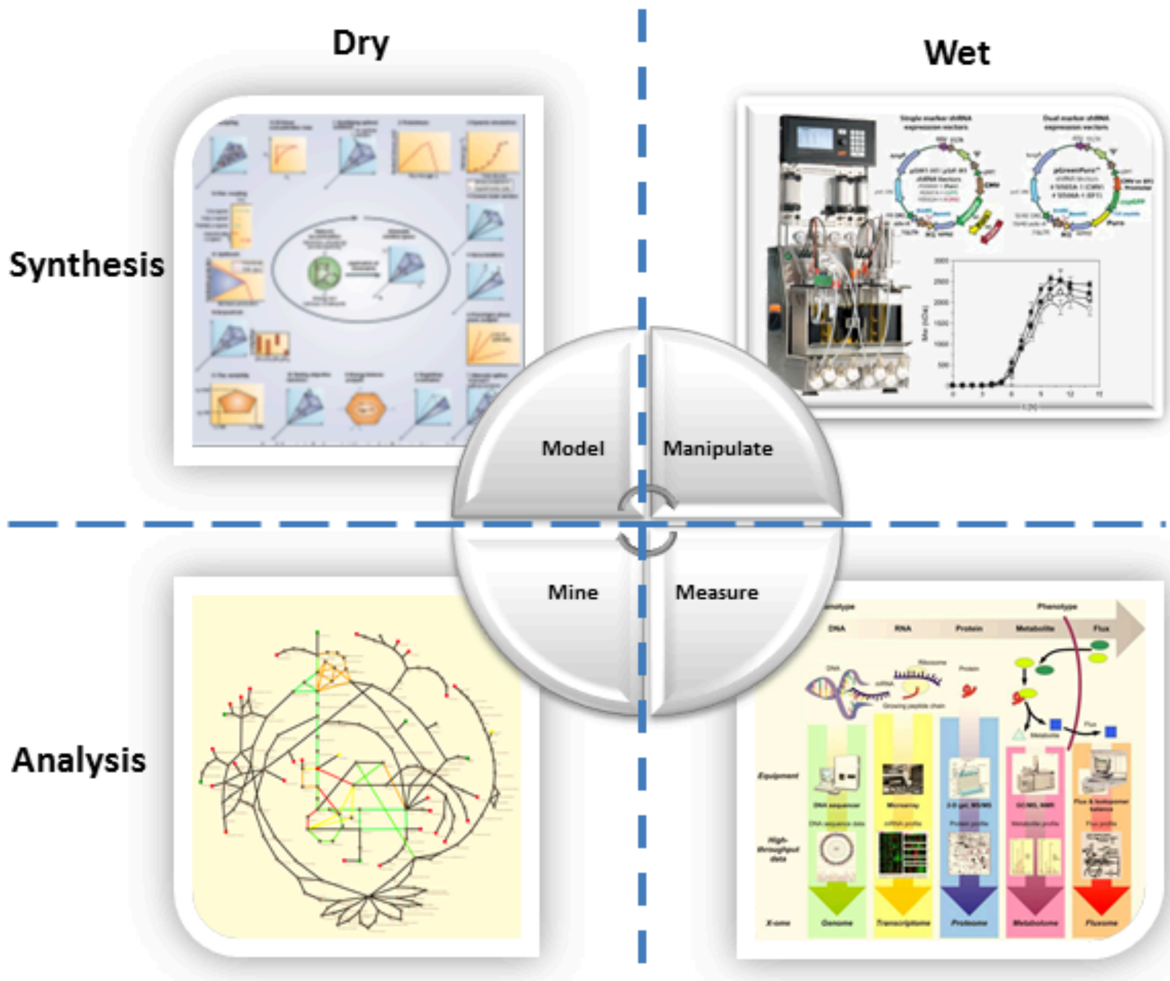
Astronomy - Sloan Digital Sky survey

4Ms:

Measurement,
Mining,
Modeling
& Manipulation

TREY IDEKER, L. RAIMOND WINSLOW & A. DOUGLAS LAUFFENBURGER ('06). "Bioengineering and Systems Biology," Annals of Biomedical Engineering DOI: 10.1007/s10439-005-9047-7

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



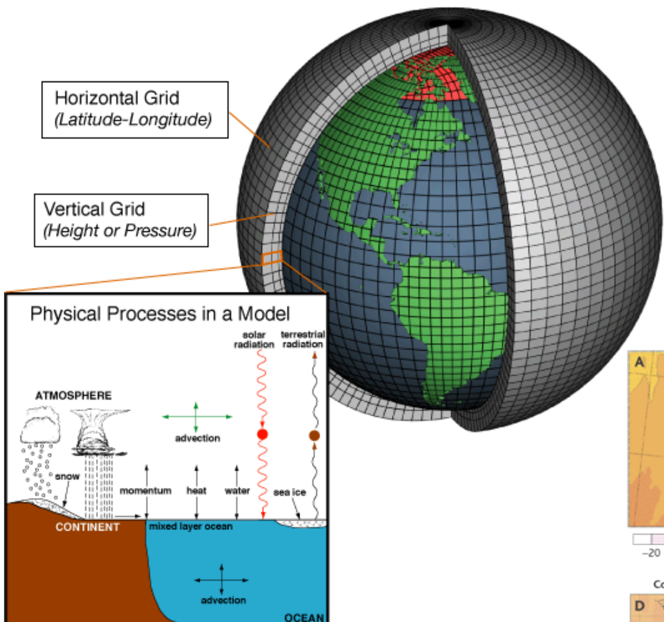
Weather forecasting

Lampooned but actually very successful

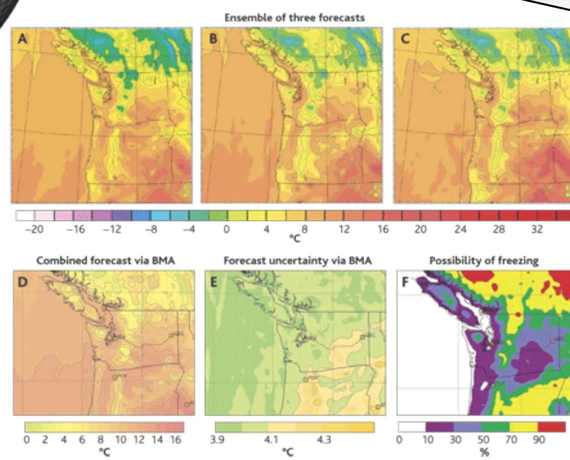
- No ability to predict a century ago;
- now forecasts checked by billions every day
- Interpretable & useful statistical predictions,
- informing everything from clothing choices to commerce

How do they do it?

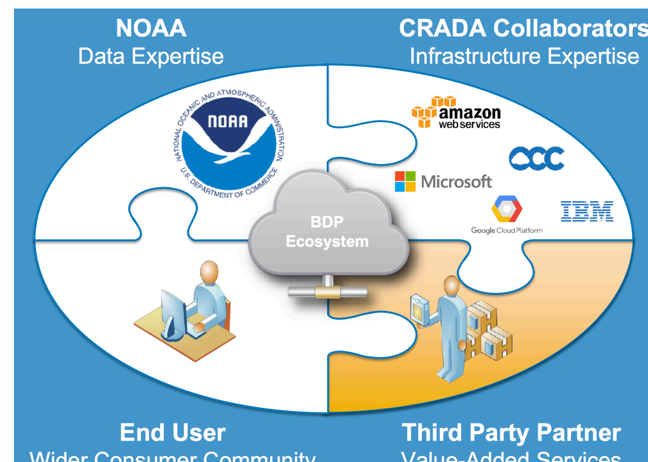
- Physical models & massive sim. useful
- (but not sufficient - think “butterfly” effect.)
- Large-scale data collection via sensors

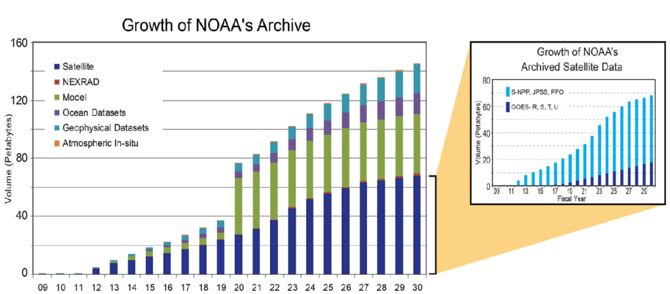


1964, first climate model



90s, ensemble methods





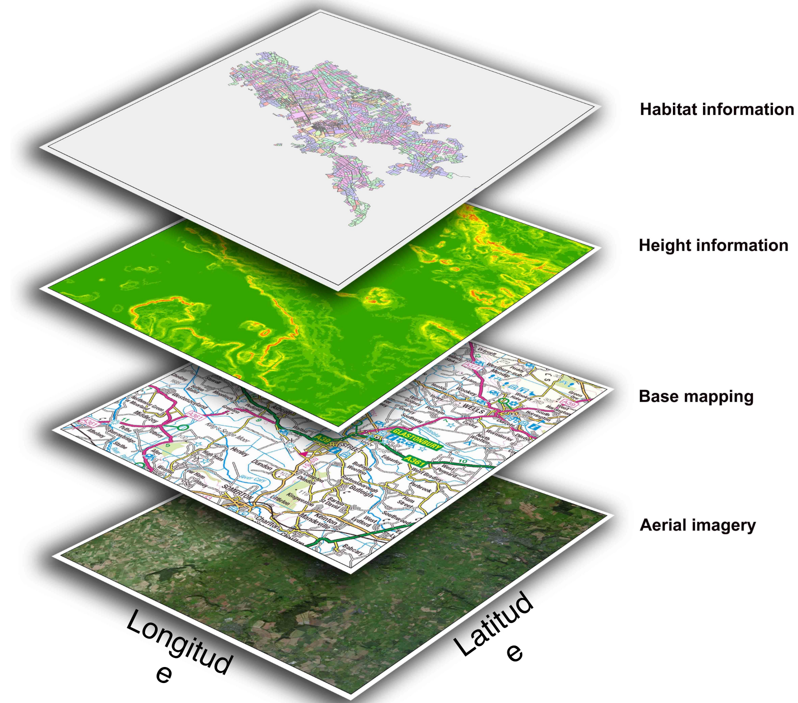
2010s, big data project

Biomedical Data Science

- The ambition of **data map** & eventually model of the genome, connectome, organs…
- The recent success of genomics (to highlight) but maybe a **changing landscape**
- How **scaling** is integral to the changing landscape
- Using large-scale data as an **anchor** for heterogeneous phenotype/medical data

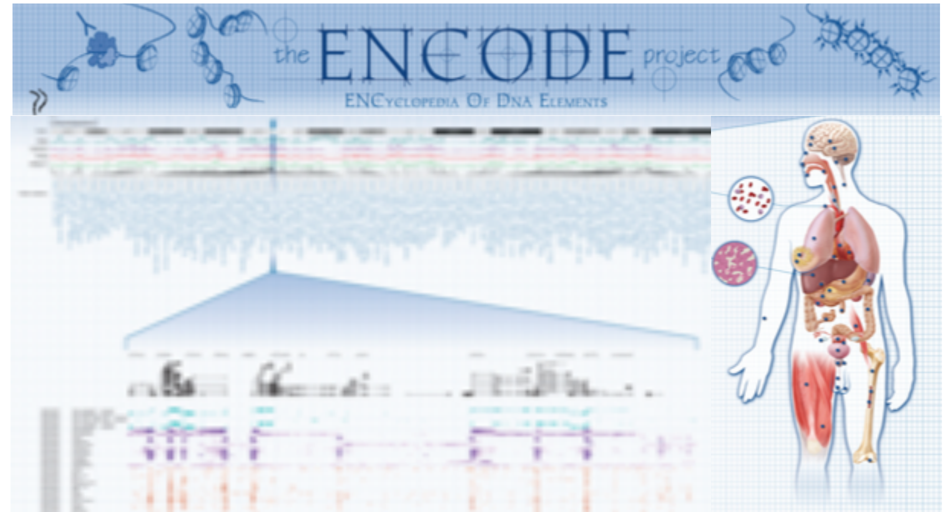
Human genome annotation — a non-intuitive map

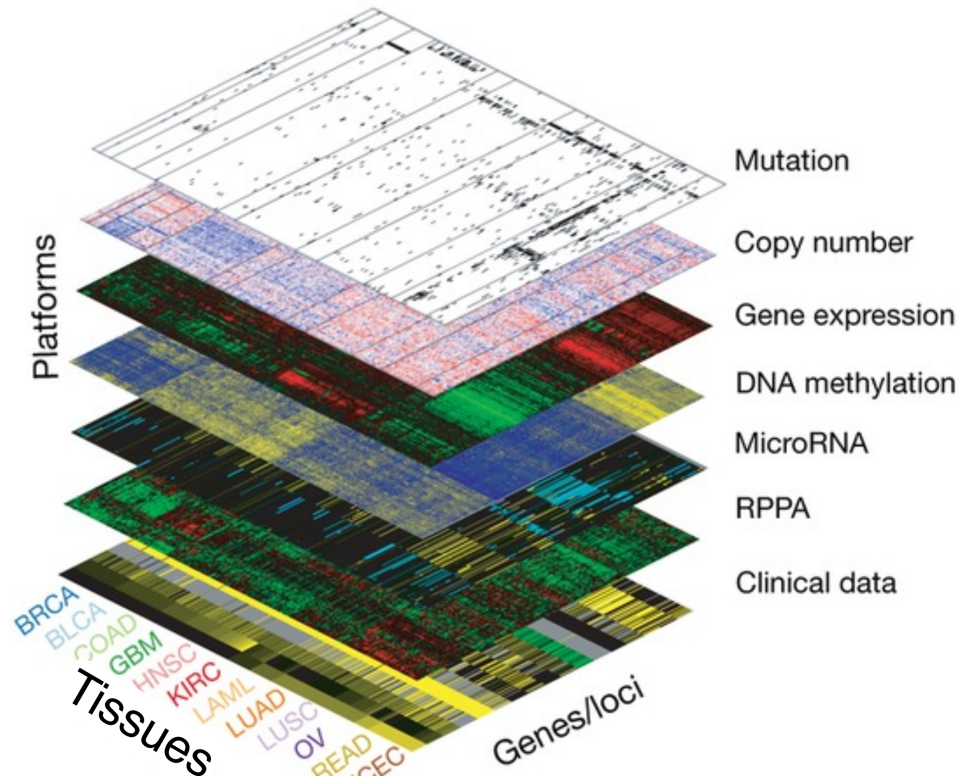
geographical information



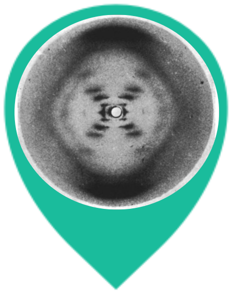
- Large-scale organisation providing an overview of the genome
- Integration of heterogeneous data

genomic information





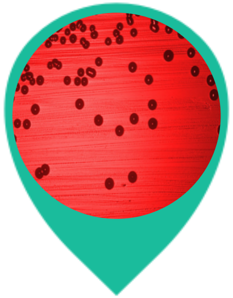
Biomed. Data Sci. via Example: Huge Success in Amassing Genotype-Phenotype Relationships



1953

Double Helix

The discovery of double helix by James Watson and Francis Crick



1995

Sequenced Genome

Haemophilus influenzae as the first organism's genome completely sequenced

2008

Thousand Genomes

By far the most detailed catalogue of human genetic variation

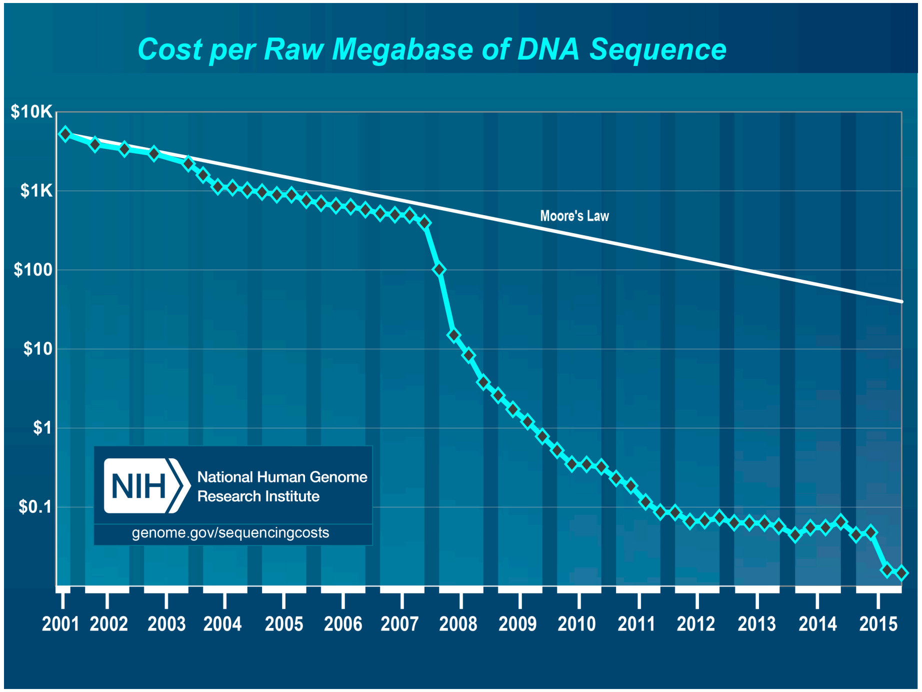
2015

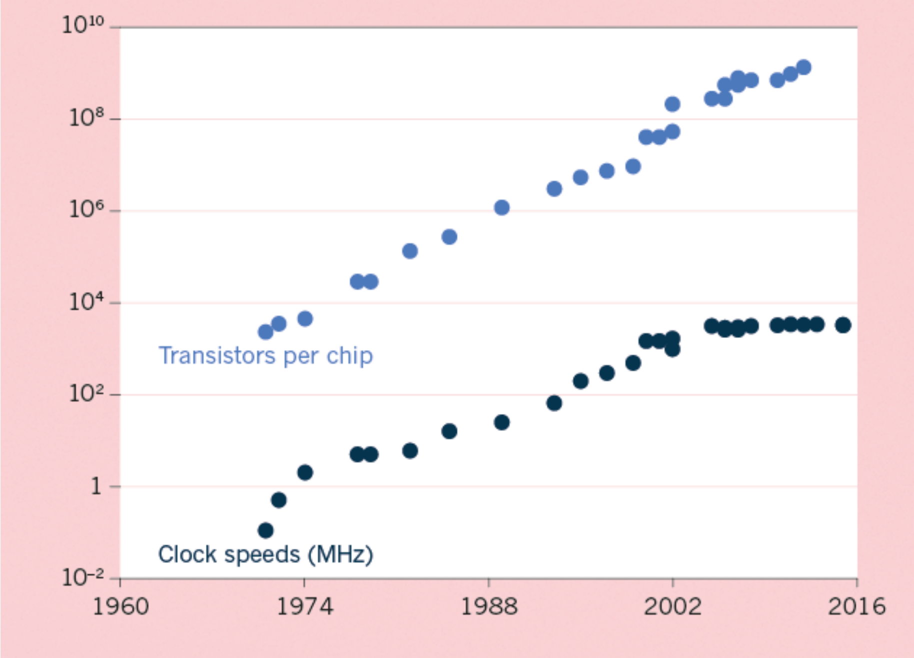
Integrated health data

Study with over 0.5M participants collecting integrated data from genotypes to phenotypic details and clinical information

Sequencing Data Explosion:

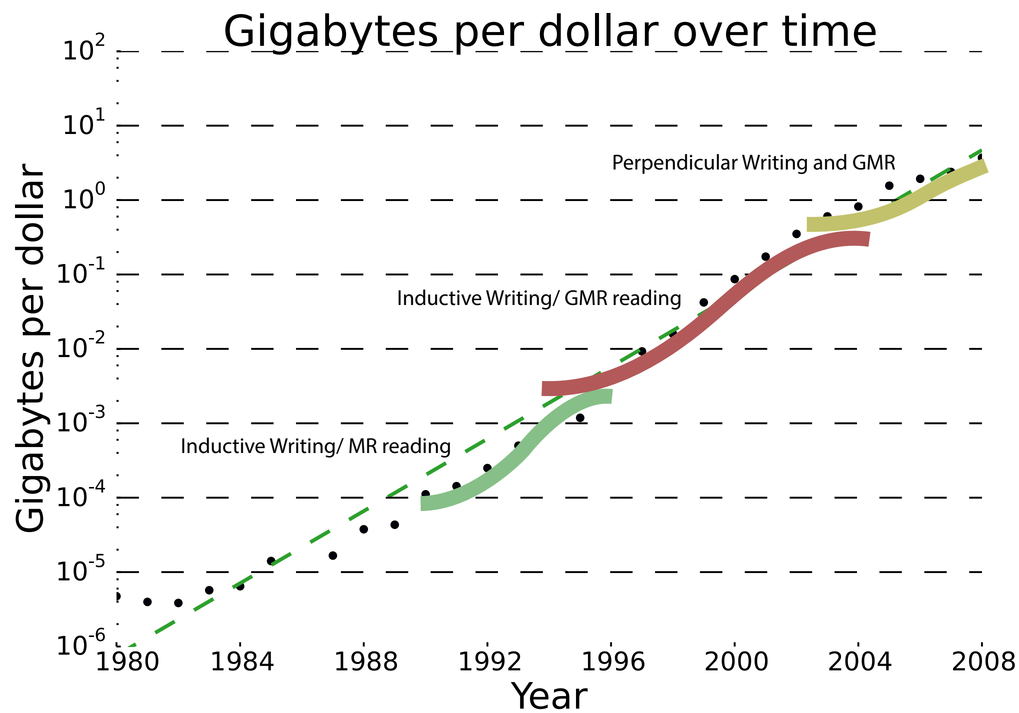
Powered by hyper-exponential incr. in data & exponential increase in computing (Moore's Law)

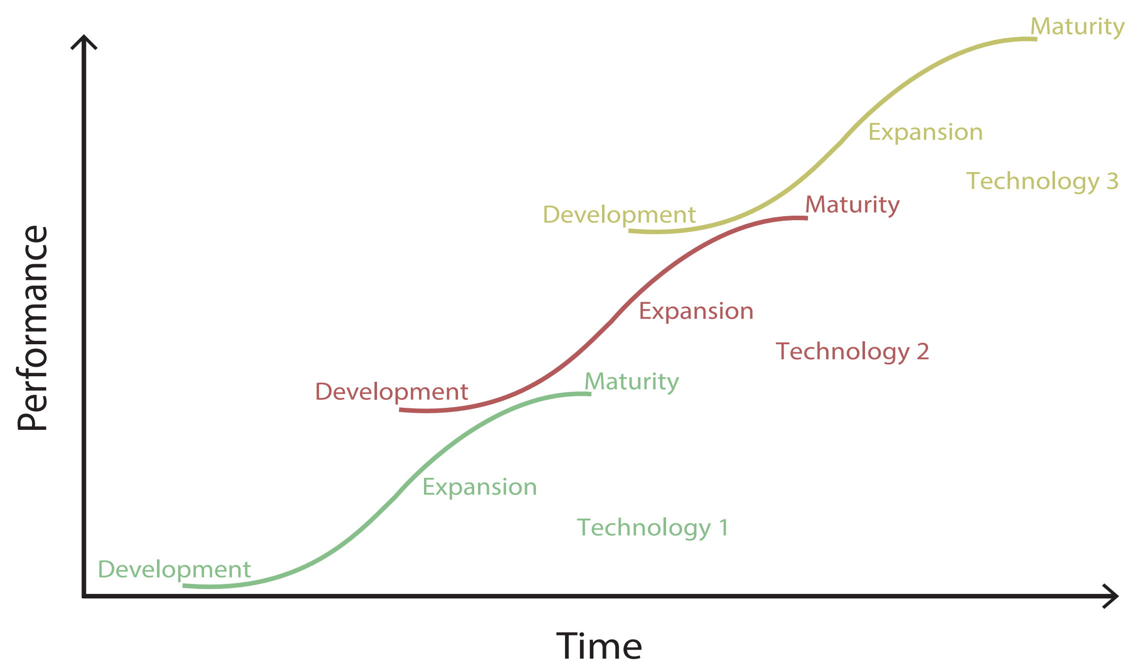




Kryder's Law and S-curves underlying exponential growth

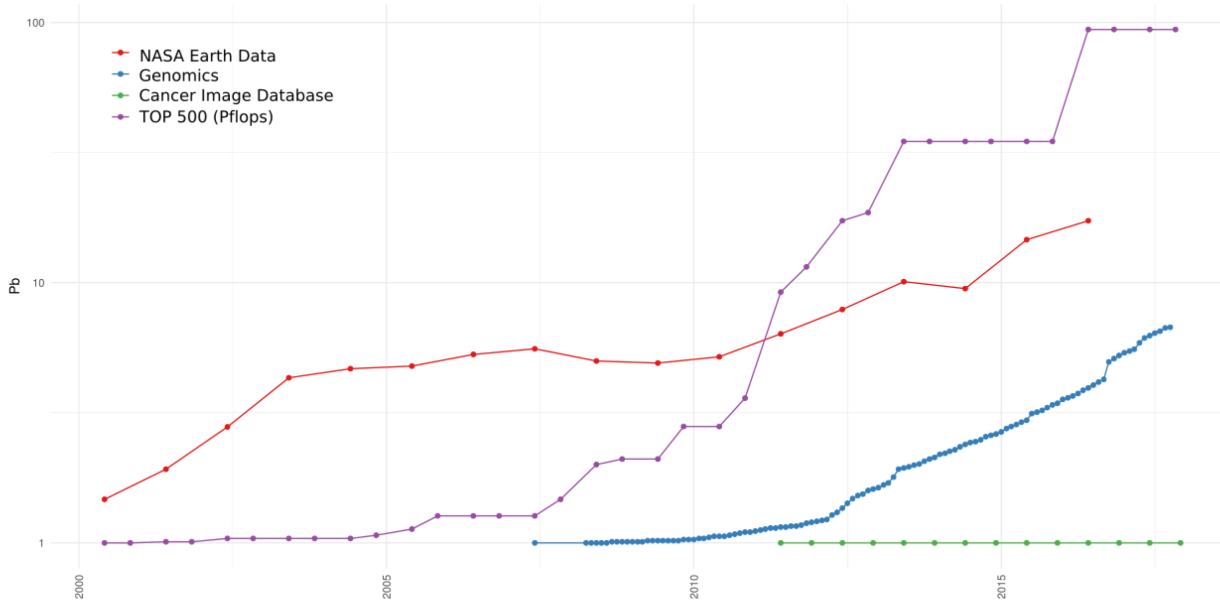
- Moore's & Kryder's Laws
  - As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
- Exponential increase seen in Kryder's law is a superposition of S-curves for different technologies

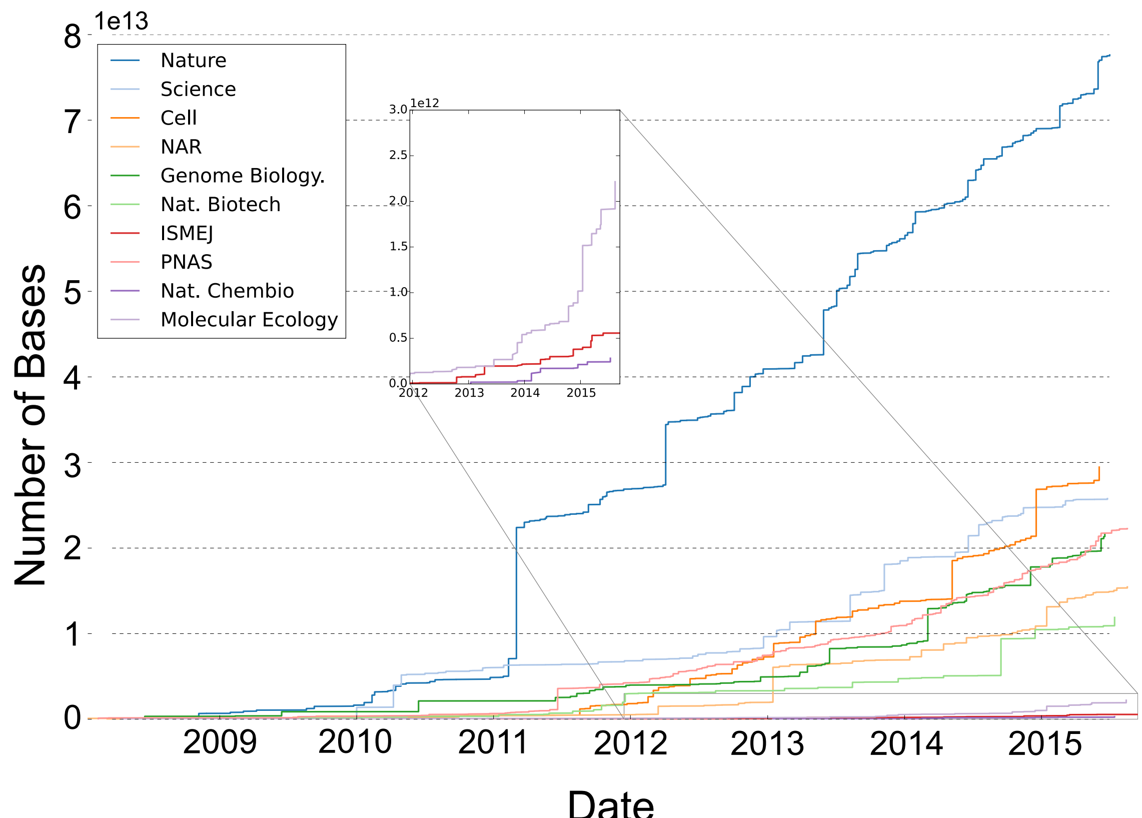




[Muir et al. ('15) GenomeBiol.]

Sequencing cost reductions have resulted in an explosion of data

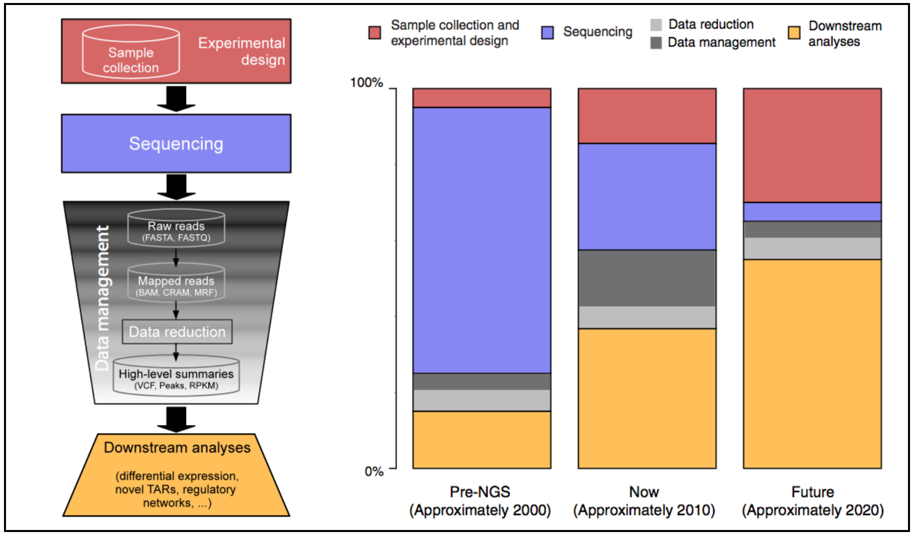




- The type of sequence data deposited has changed as well.

[Muir et al. (‘15) GenomeBiol.]

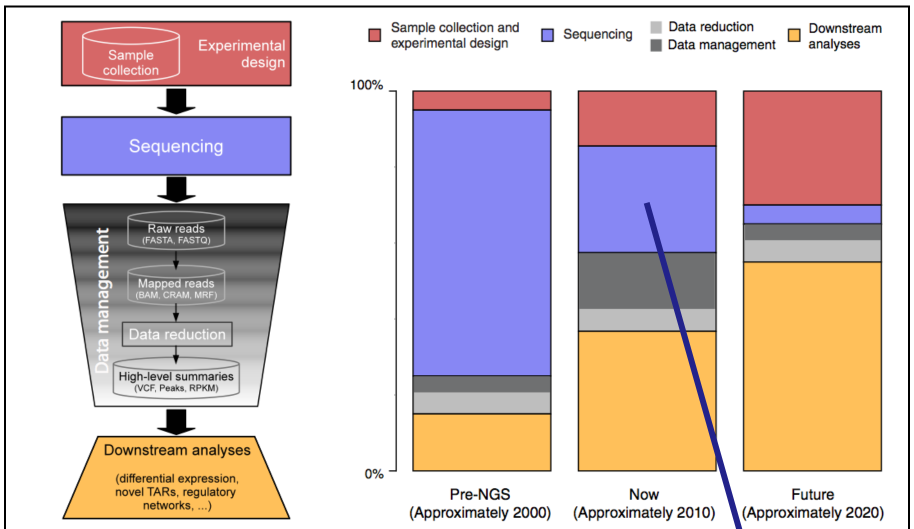
The changing costs of a sequencing pipeline



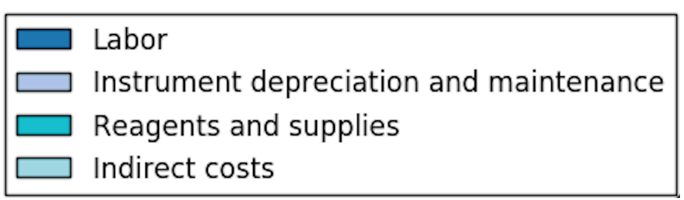
From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

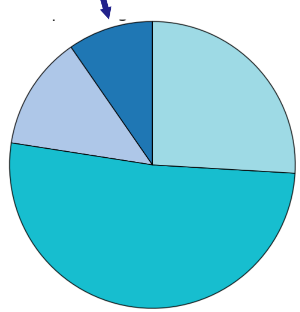
[Sboner et al. (‘11), Muir et al. (‘15) Genome Biology]

The changing costs of a sequencing pipeline



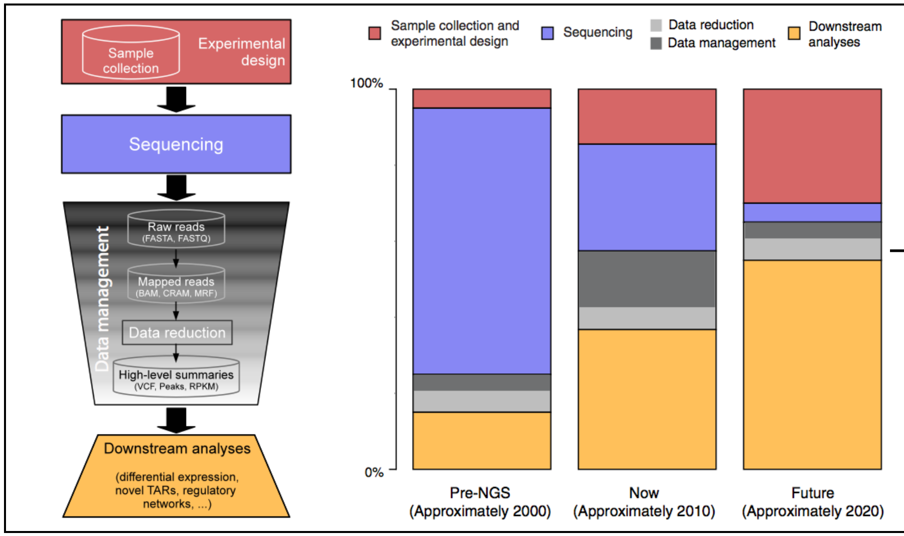
From '00 to ~' 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

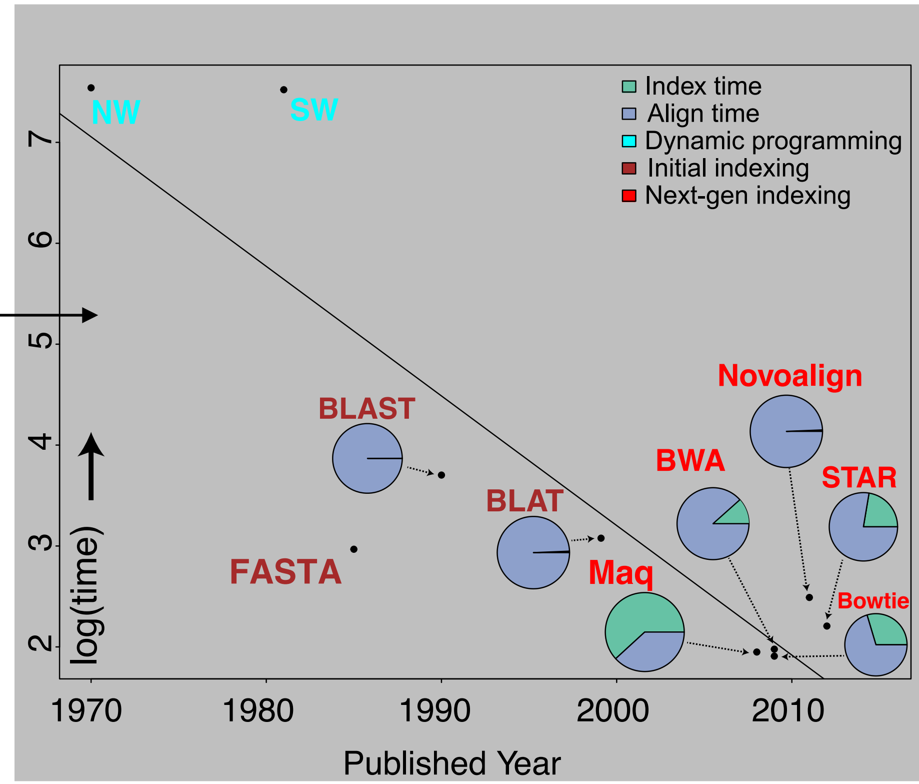




[Sboner et al. ('11), Muir et al. ('15) Genome Biology]

The changing costs of a sequencing pipeline



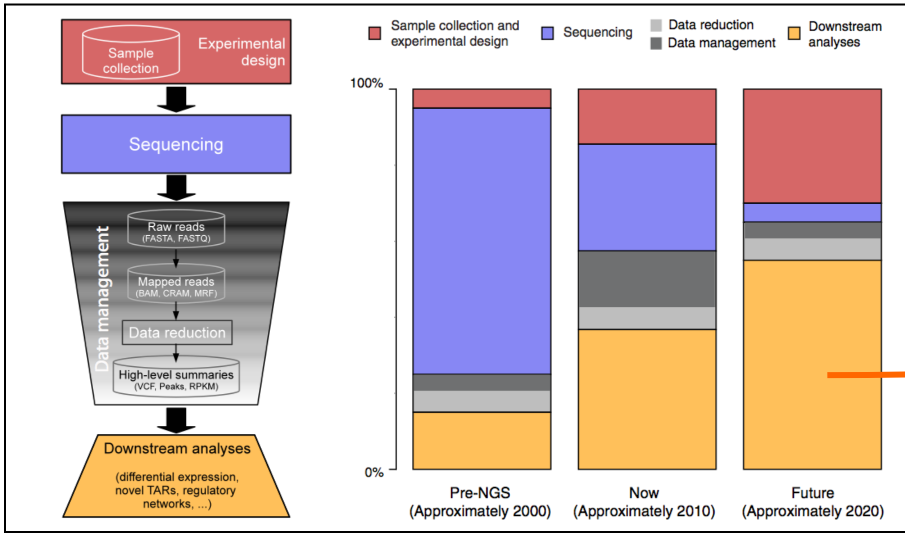


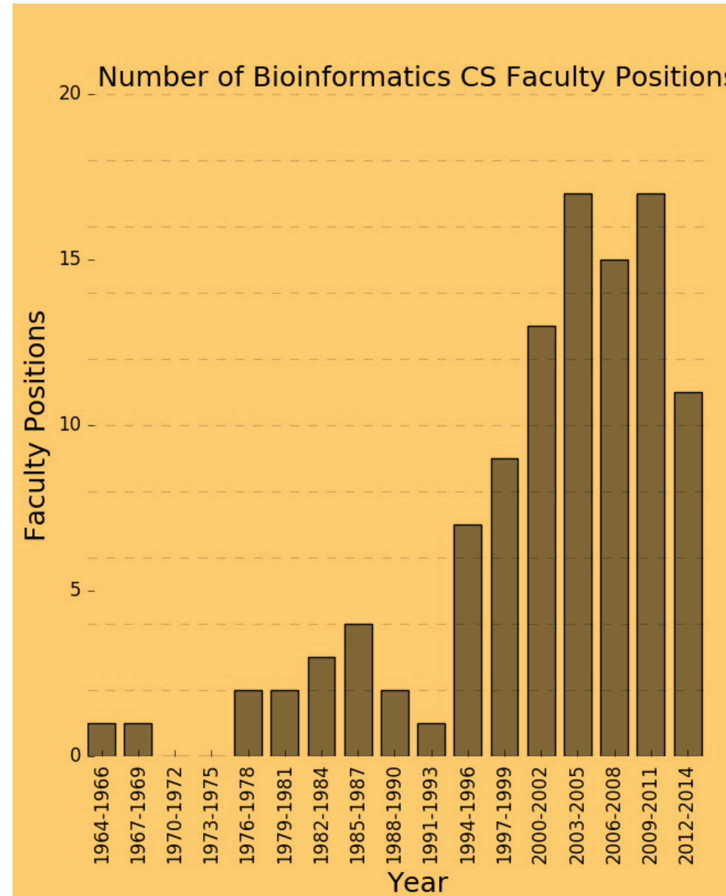
From '00 to ~'20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

Alignment algorithms scaling to keep pace with data generation

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

The changing costs of a sequencing pipeline



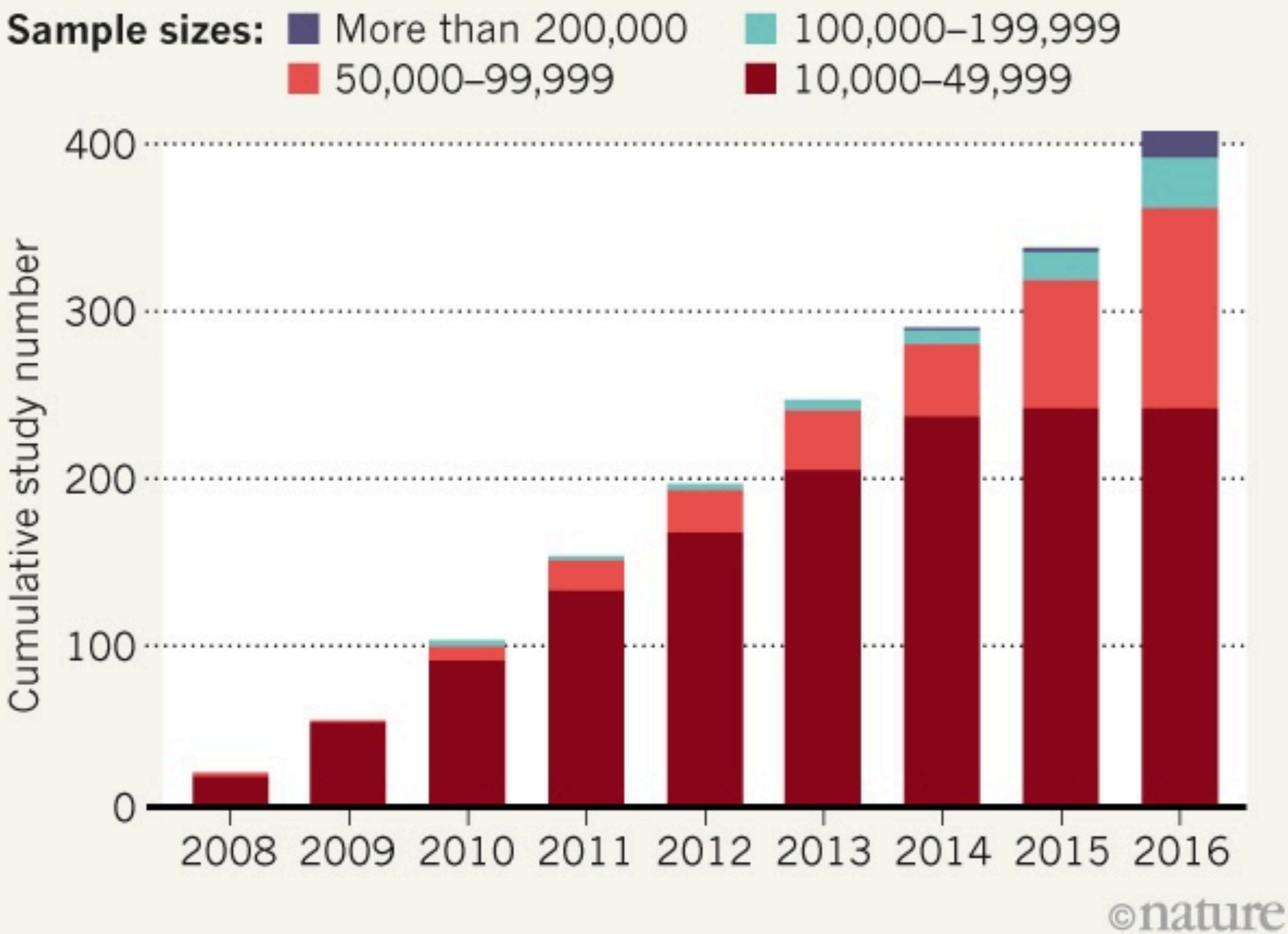


From ‘00 to ~’ 20,
cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

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

THE 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.



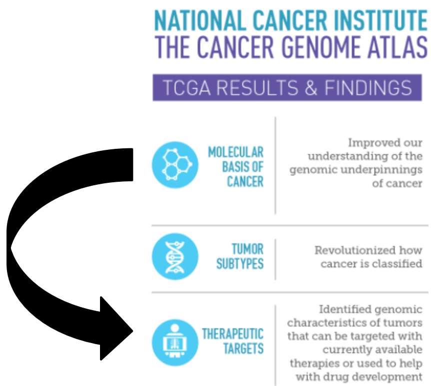
Basic Science to Medical Practice

Research Initiatives and Biomedical Startups

Large-scale genomics data as a anchor to organize large amounts of phenotype data – EMRs, wearables…

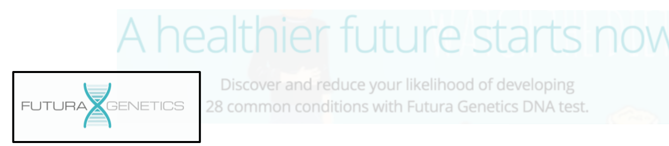
1. Genomics of disease-focused cohorts; GWAS [2002-present], TCGA, PCAWG [2006-present]
2. Integration of genomic data with rich clinical phenotypes; UKBiobank, All of Us [2016-present]
3. Integration of genomic data in EMRs for clinical decision support & wearables; [Near future]
4. Home-based routine sequencing of DNA and RNA in blood as part of preventive care [Speculative future]

INITIATIVES



PCAWG
PanCancer Analysis of WHOLE GENOMES

STARTUPS



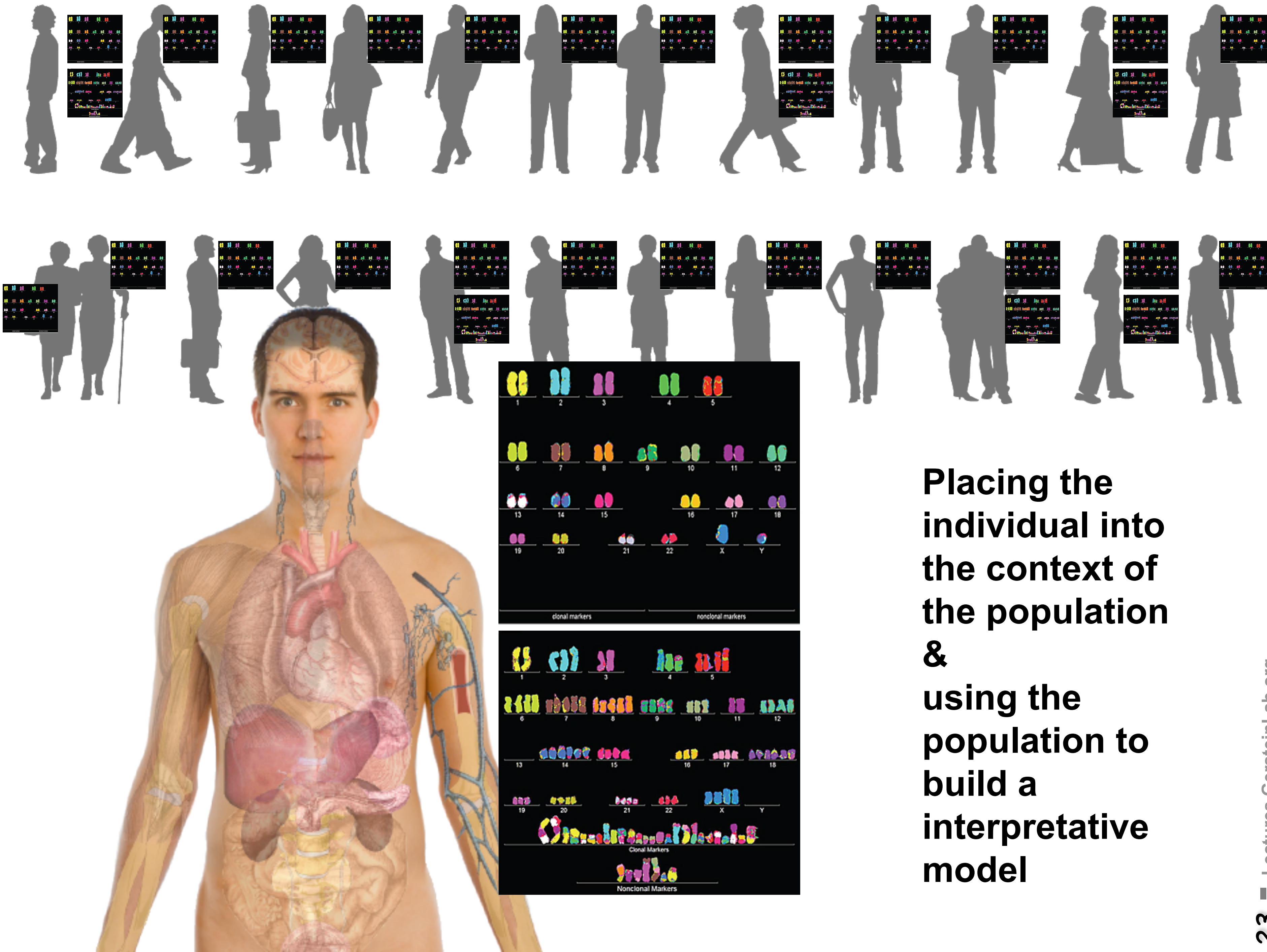
Medical Big Data: Promise and Challenges (Lee and Yoon , *Kidney Res. Clin. Pract.*, 2017)

EX of ’omics research on focused patient cohorts: Many Yale Researchers Involved in Neurogenomics

- Involved national initiatives:
  psychENCODE, CMG,
  BrainSpan, BSMN,
  NIDA Neuroproteomics

- Yale investigators:
  M Gunel, N Sestan, F Vaccarino, J Noonan,
  J Gelernter, A Nairn

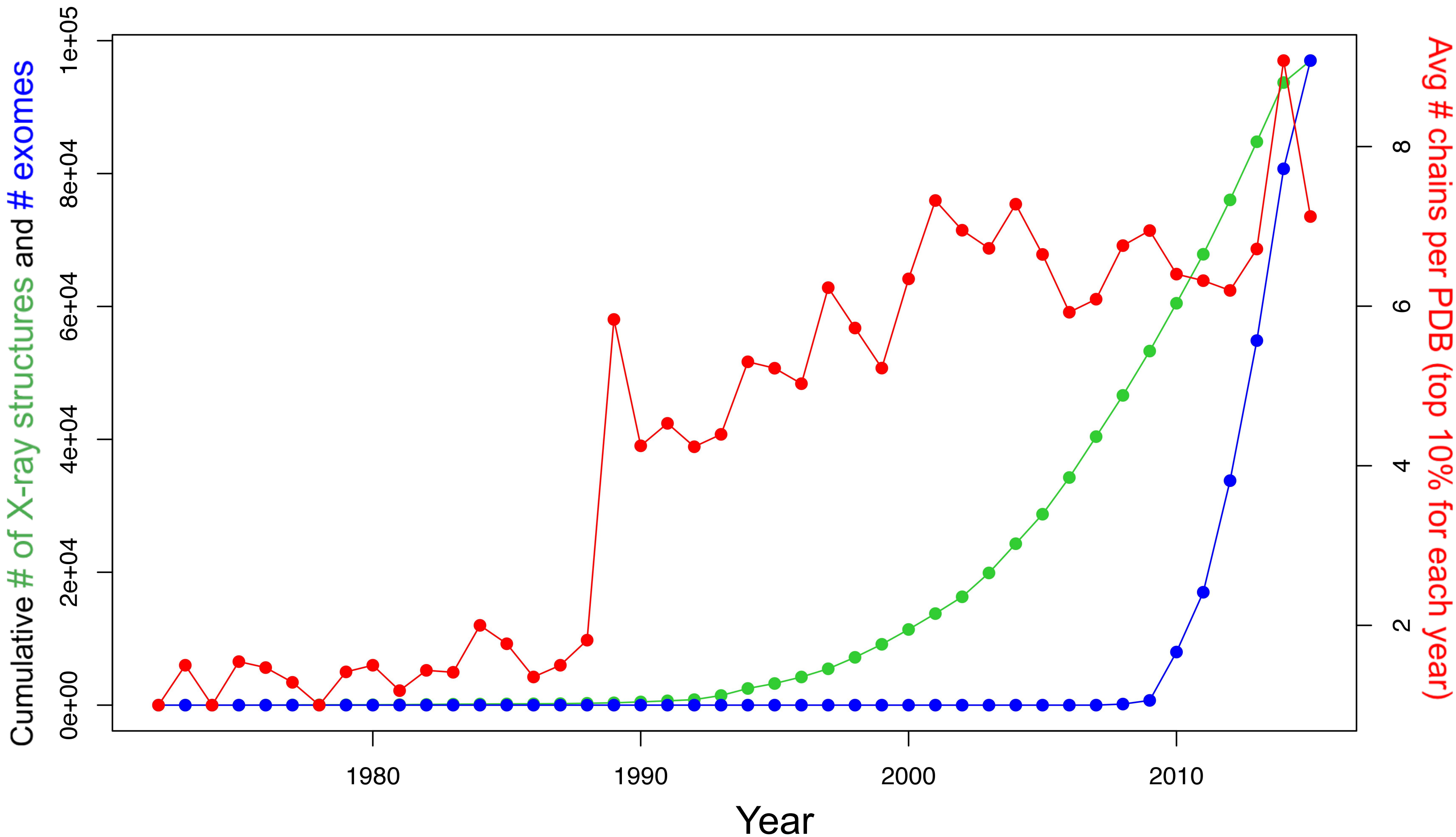
- DNA variants, altered protein & RNA levels in brains in development & various diseases (eg ASD, SCZ)



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

Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)

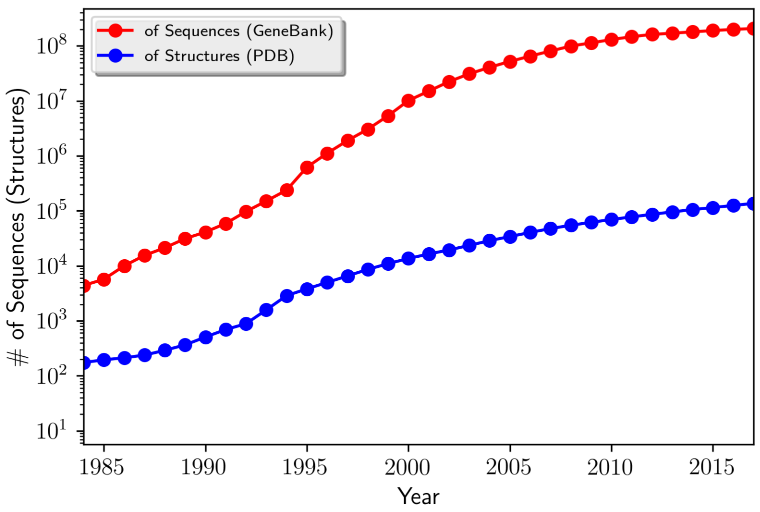
The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.

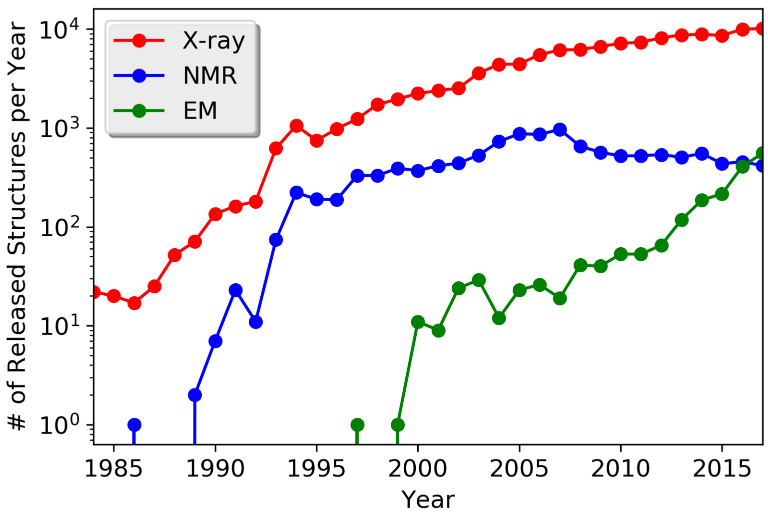


Exome data hosted on NCBI Sequence Read Archive (SRA)

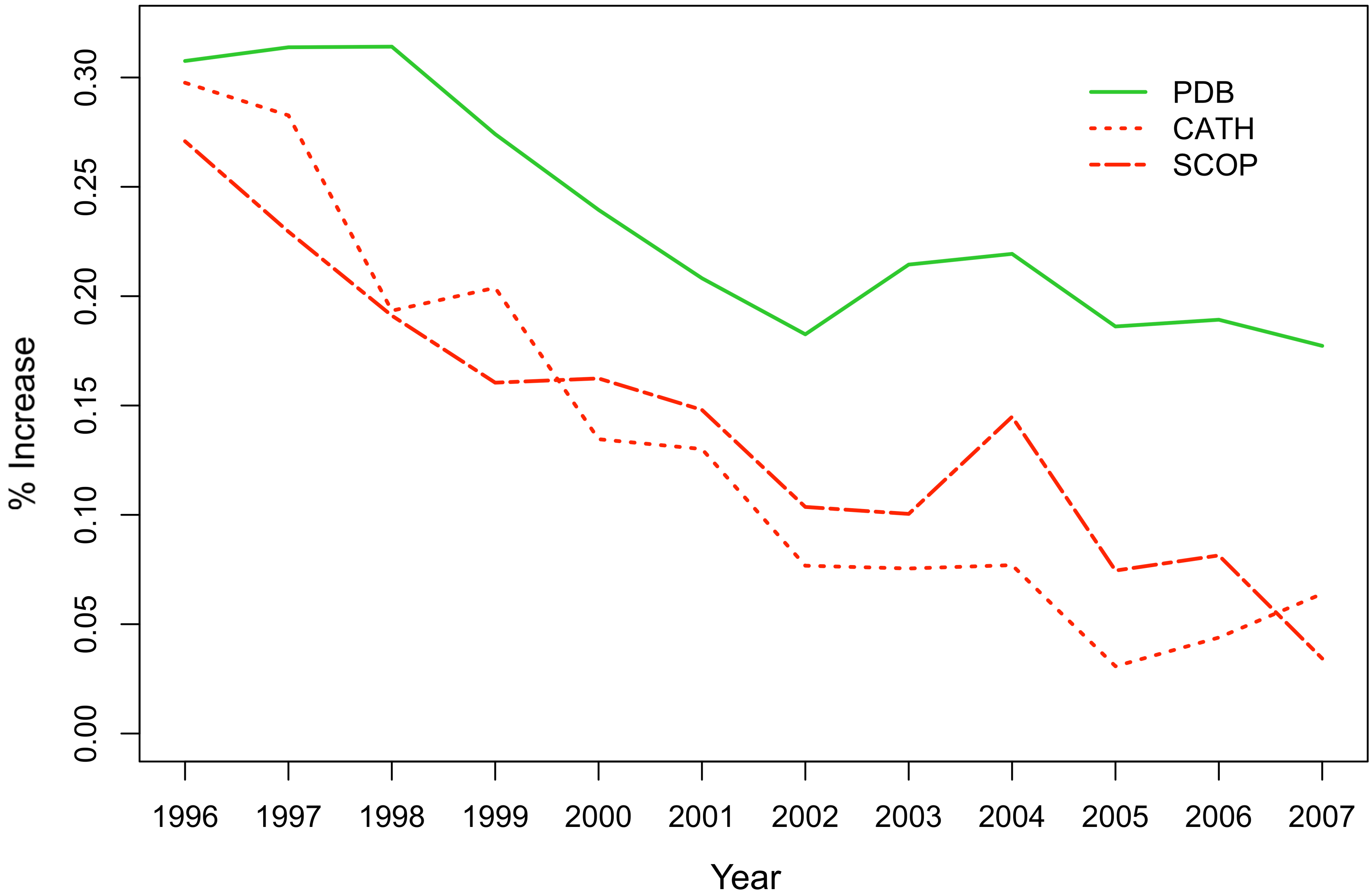
[Sethi et al. COSB ('15)]

Experimental determination of 3D structures can not keep up with the explosive growth of sequence information
The Electron Microscopy (EM) has emerged as a powerful tool in determining 3D structures





Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes – Gene & Struc. Families as main organizing principle

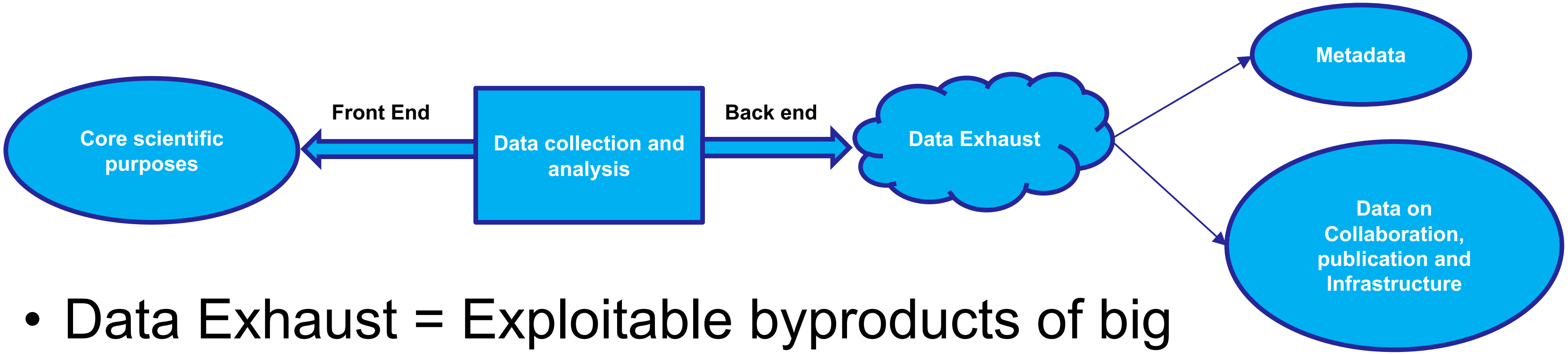


PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

[Sethi et al. COSB ('15)]

Core Qs v Creative Use of the Data

Data Exhaust

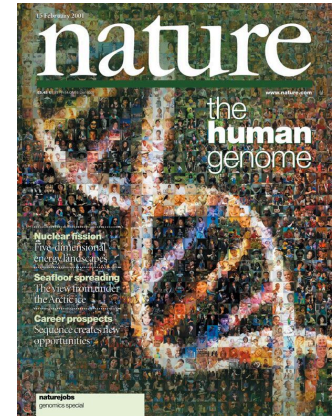


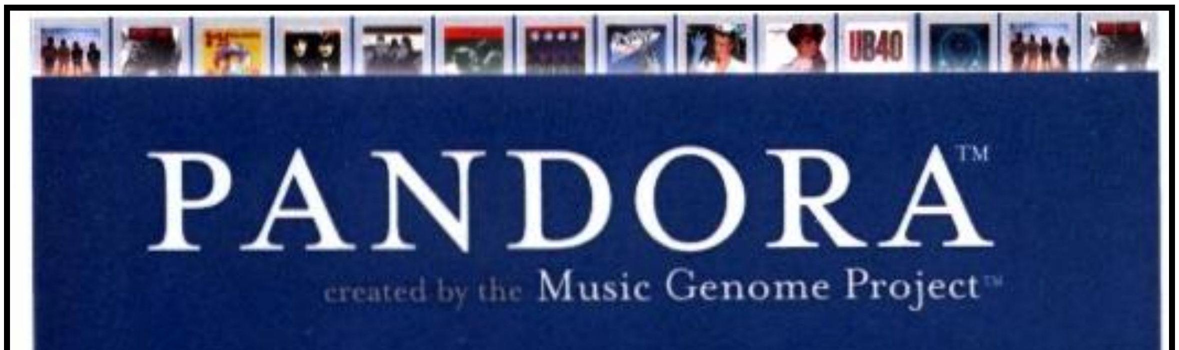
- Data Exhaust = Exploitable byproducts of big data collection and analysis
- Creative use of Data is key to Data Science !
- Aspects of Privacy but also Science of Science

[PHOTO: RELAXNEWS; from http://www.lapresse.ca]

Genomics: as Data Science sub-discipline

- Developing ways of organizing & mining categorizing information on a large scale
  - Very fundamental & early form of "Big Data", feeding into other enterprises (classification approach, R)
  - Also importing tech. developed in other big data disciplines (Hadoop)






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

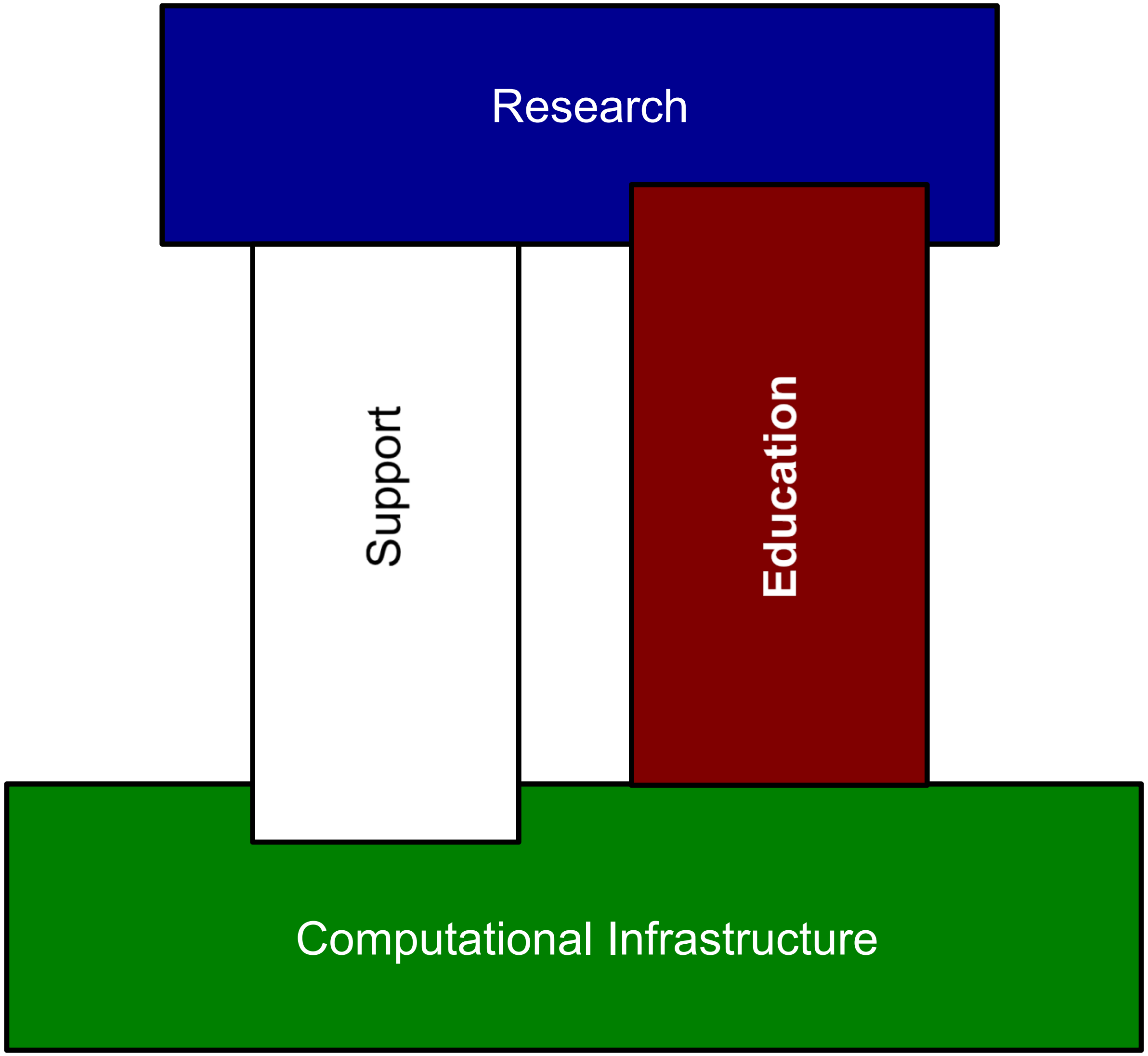
THE ART GENOME PROJECT

BY MATTHEW ISRAEL, JESSICA BACKUS AND OLIVIA JENE FAGON

FEB 9TH, 2016 5:00 AM

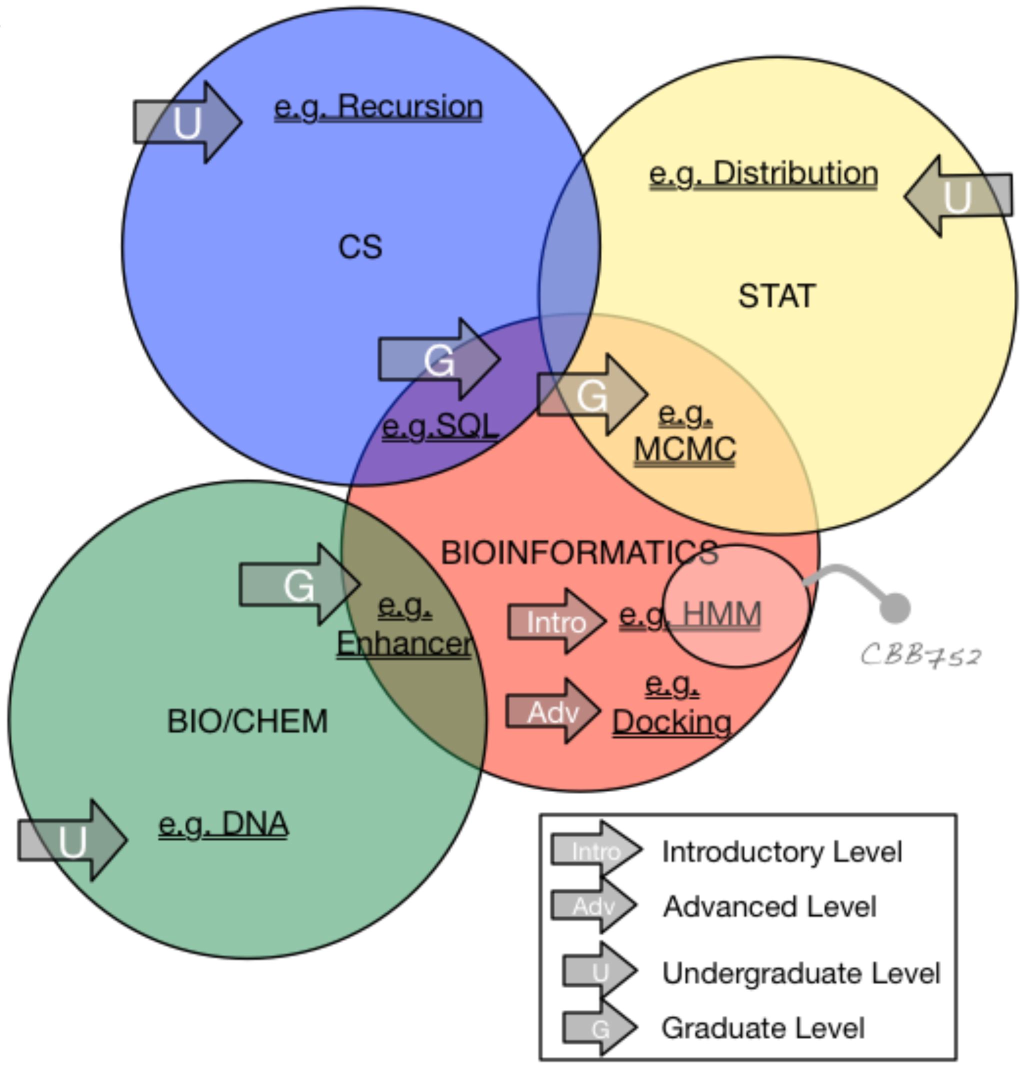
General Thoughts on the Course

Elements of Bioinformatics as a discipline

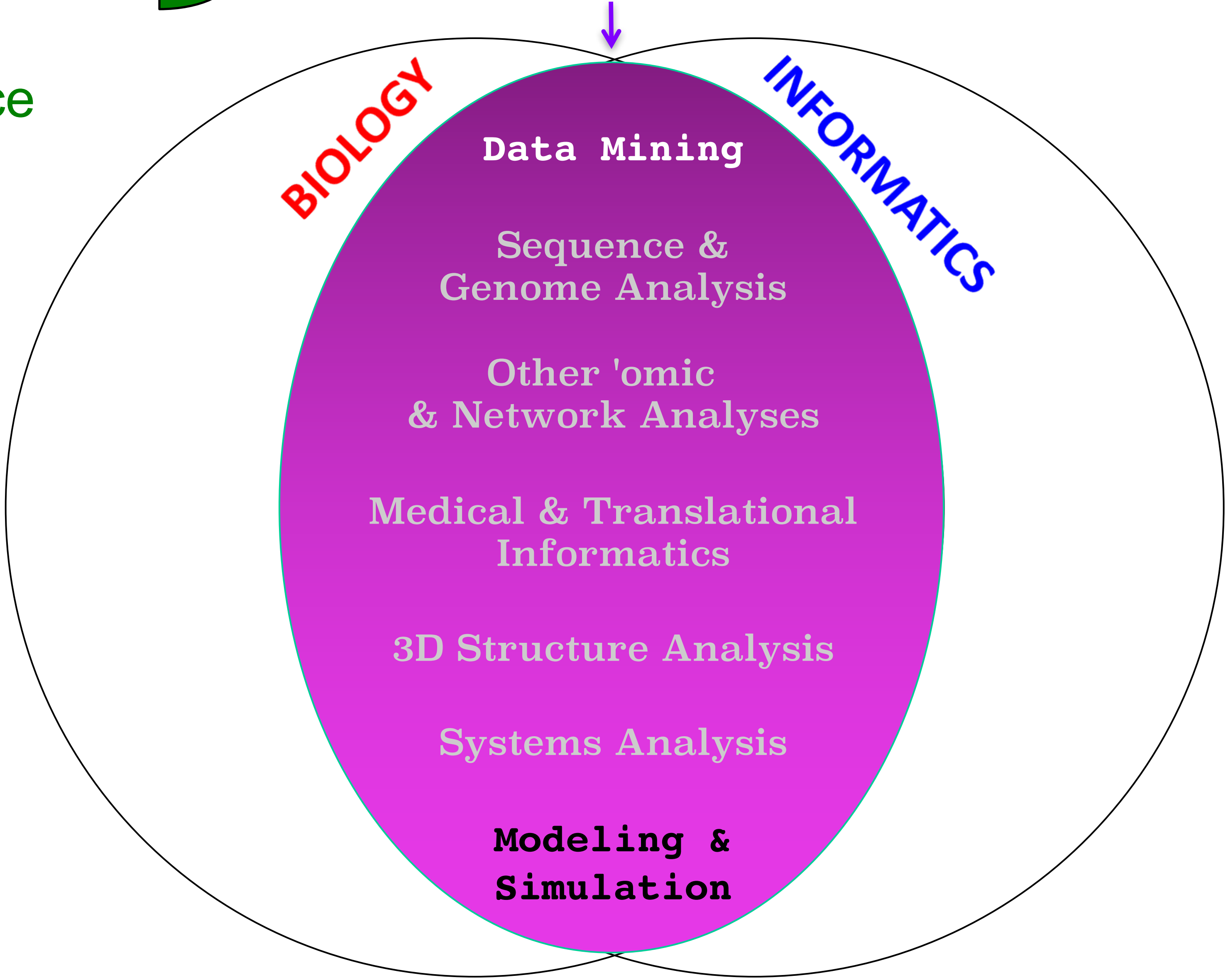


Defining Bioinformatics – by crowd-sourced judgement

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



Biomedical Data Science ⊃ (Molecular) BIOINFORMATICS



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

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]

Class Web Page

GersteinLab.org/courses/452

Assignment #0 Page

goo.gl/**BfSpQV**

Office Hours

Right after class & tomorrow at 10 am (in Bass 432)

More details on Bioinformatics as a subdiscipline of Biomedical Data Science

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]

What Information to Organize?

- **Sequences** (DNA & Protein)
- 3D Structures
- Network & Pathway Connectivity
- Phylogenetic tree relationships
- Large-scale gene expression & functional genomics data
- Phenotypic data & medical records....

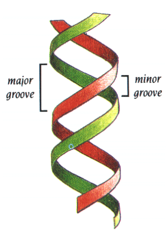
What is the Information?
Molecular Biology as an Information Science

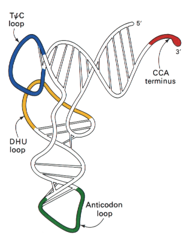
- Central Dogma
  of Molecular Biology

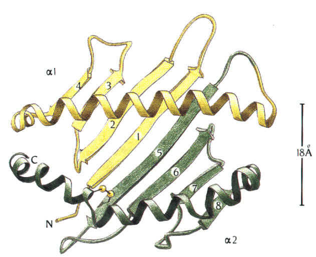
```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

- Central Paradigm
  for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Biological Function
     -> Organismal Phenotype
```







•Genetic material

•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity

Molecular Biology Information - DNA

- Raw DNA Sequence
  - 4 bases: AGCT
  - ~1 K in a gene, ~2 M in genome
  - ~3 Gb Human

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .
```

```
. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccatttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Protein Sequence

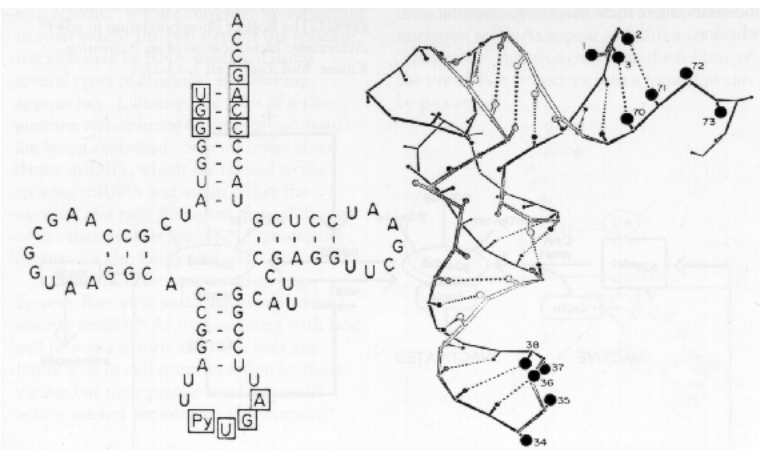
- 20 letter alphabet
  - ACDEFGHIKLMNPQRSTVWY   but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- >12 M known protein sequences (uniprot, http://www.ebi.ac.uk/uniprot/TrEMBLstats/, 2011)

```
d1dhfa_  LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__  LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_  ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__  TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_  LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__  LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_  ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__  TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_  VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__  VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_  ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__  ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_  -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__  -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_  -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE----- IMVIGGGRVYEQFLPKA
d3dfr__  -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

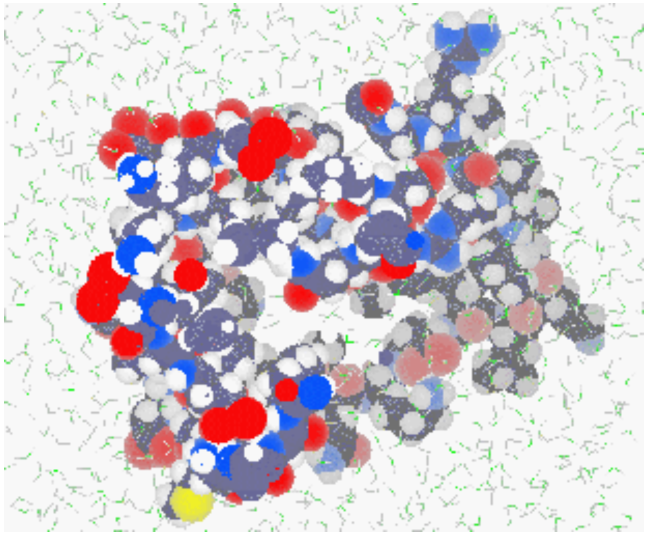
Molecular Biology Information: Macromolecular Structure

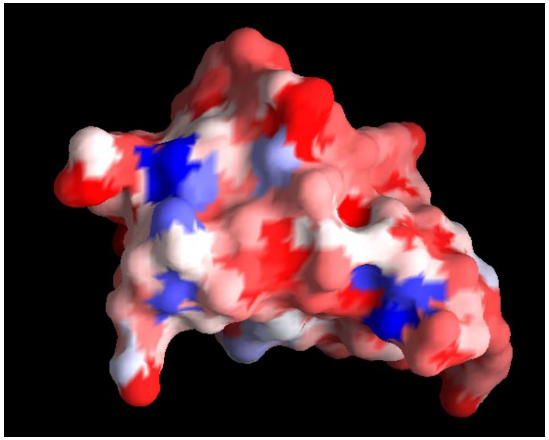
- DNA/RNA/Protein
  - Mostly protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)



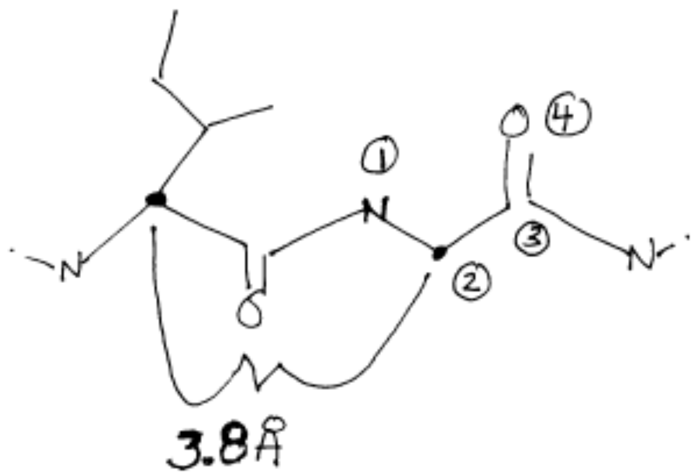
'Identity elements' in *Escherichia coli* glutamine tRNA.





Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - 200 residues/domain => 200 CA atoms, separated by 3.8 A
  - Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    => ~1500 xyz triplets (=8x200) per p
  - >100K Domains, ~1200 folds (scop 1



ATOM	1	C	ACE	0	9.401	30.166	60.595	1.00	49.88	1GKY 67
ATOM	2	O	ACE	0	10.432	30.832	60.722	1.00	50.35	1GKY 68
ATOM	3	CH3	ACE	0	8.876	29.767	59.226	1.00	50.04	1GKY 69
ATOM	4	N	SER	1	8.753	29.755	61.685	1.00	49.13	1GKY 70
ATOM	5	CA	SER	1	9.242	30.200	62.974	1.00	46.62	1GKY 71
ATOM	6	C	SER	1	10.453	29.500	63.579	1.00	41.99	1GKY 72
ATOM	7	O	SER	1	10.593	29.607	64.814	1.00	43.24	1GKY 73
ATOM	8	CB	SER	1	8.052	30.189	63.974	1.00	53.00	1GKY 74
ATOM	9	OG	SER	1	7.294	31.409	63.930	1.00	57.79	1GKY 75
ATOM	10	N	ARG	2	11.360	28.819	62.827	1.00	36.48	1GKY 76
ATOM	11	CA	ARG	2	12.548	28.316	63.532	1.00	30.20	1GKY 77
ATOM	12	C	ARG	2	13.502	29.501	63.500	1.00	25.54	1GKY 78
. . .										
ATOM	1444	CB	LYS	186	13.836	22.263	57.567	1.00	55.06	1GKY1510
ATOM	1445	CG	LYS	186	12.422	22.452	58.180	1.00	53.45	1GKY1511
ATOM	1446	CD	LYS	186	11.531	21.198	58.185	1.00	49.88	1GKY1512
ATOM	1447	CE	LYS	186	11.452	20.402	56.860	1.00	48.15	1GKY1513
ATOM	1448	NZ	LYS	186	10.735	21.104	55.811	1.00	48.41	1GKY1514
ATOM	1449	OXT	LYS	186	16.887	23.841	56.647	1.00	62.94	1GKY1515
TER	1450		LYS	186						1GKY1516

Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae rd." *Science* 269: 496-512.

  (Picture adapted from TIGR website, http://www.tigr.org)

- Timeline

  1995, HI (bacteria): 1.6 Mb & 1600 genes done

  1997, yeast: 13 Mb & ~6000 genes for yeast

  1998, worm: ~100Mb with 19 K genes

  1999: >30 completed genomes!

  2000, draft human

  2003, human: 3 Gb & 100 K genes...

  2010, 1000 human genomes!

  2017, 13K human genomes

1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* **269**: 496]

1997

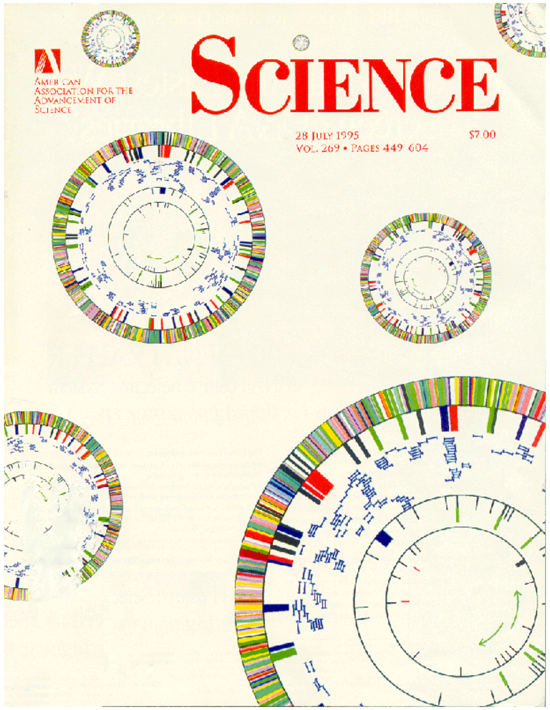
Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]

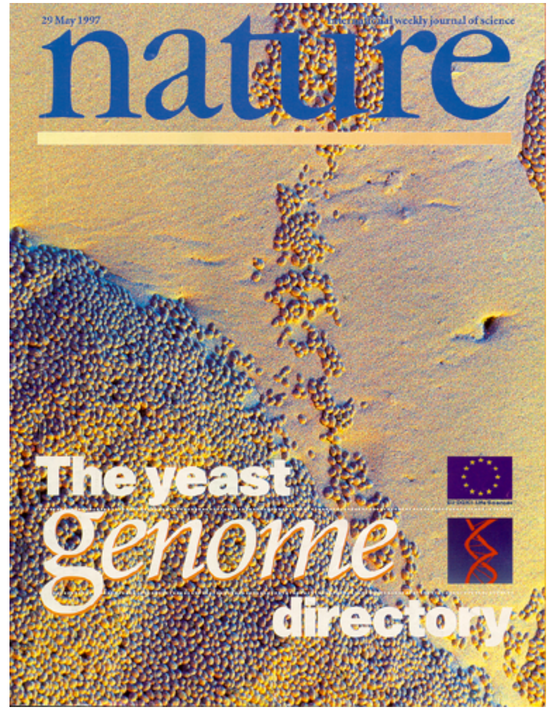
1998

Animal,
~100 Mb,
~20K genes
[*Science* **282**:
1945]

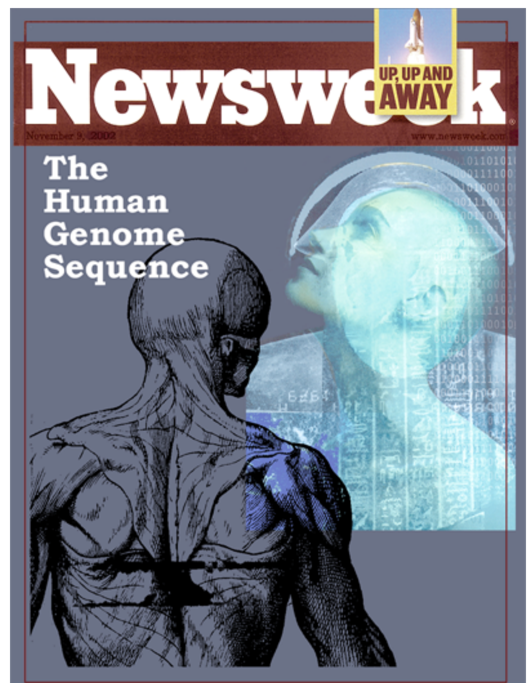
2000?

Human,
~3 Gb,
~20K genes





A Bioinformatics prediction that came true!

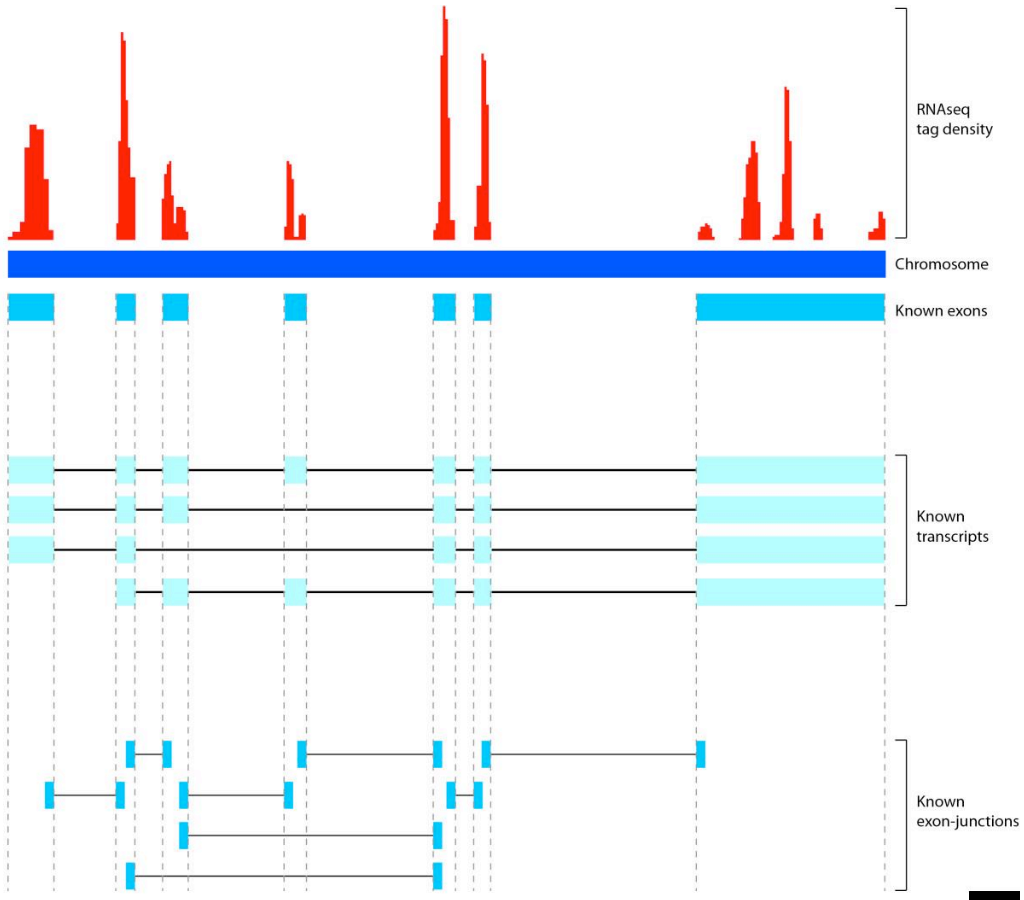


real thing, Apr '00

↔

'98 spoof

Gene Expression Data: On & Off



- Early experiments yeast
  - Complexity at 10 time points, 6000 x 10 = 60K floats
- Then tiling array technology
  - 50 M data points to tile the human genome at ~50 bp res.
- Now Next-Gen Sequencing (RNAseq)
  - 10M+ reads on the human genome, counts
- Can only sequence genome once but can do an infinite variety of expression experiments

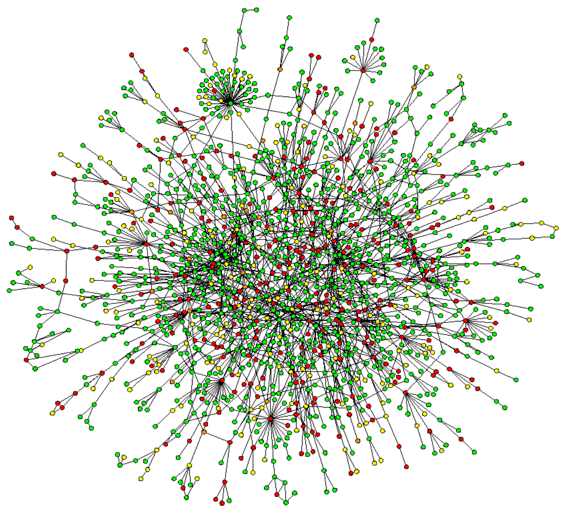
Molecular Networks: Connectivity

Regulatory Networks

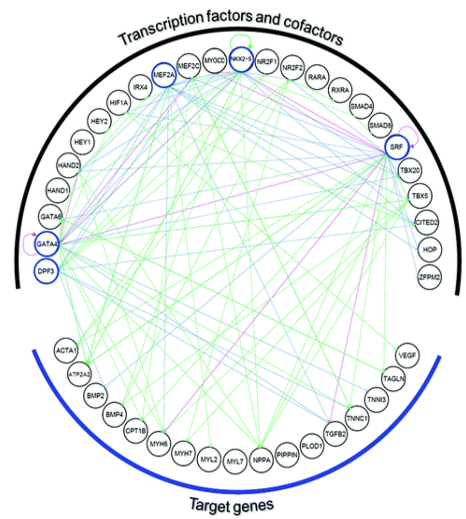
- Get readouts of where proteins bind to DNA : Chip-chip then chip-seq

Protein Interaction Networks

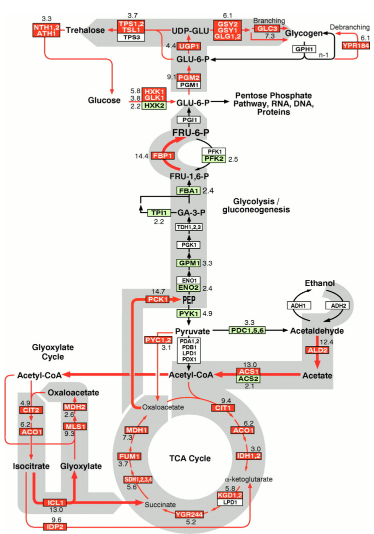
- For yeast: 6000 x 6000 / 2 ~ 18M possible interactions (maybe ~30K real)



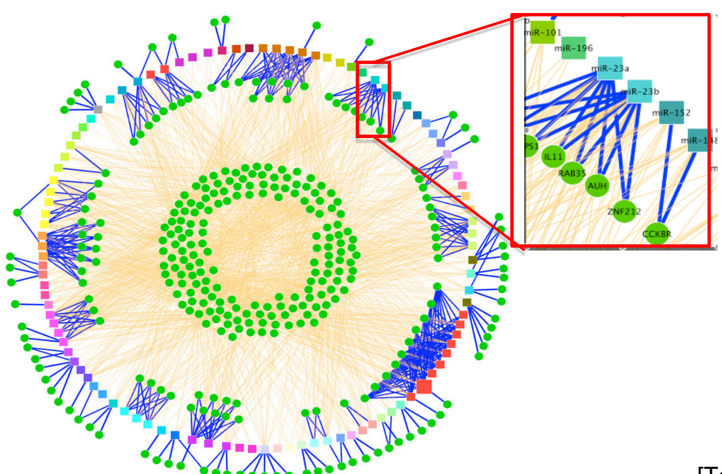
Protein-protein Interaction networks



TF-target-gene Regulatory networks



Metabolic pathway networks

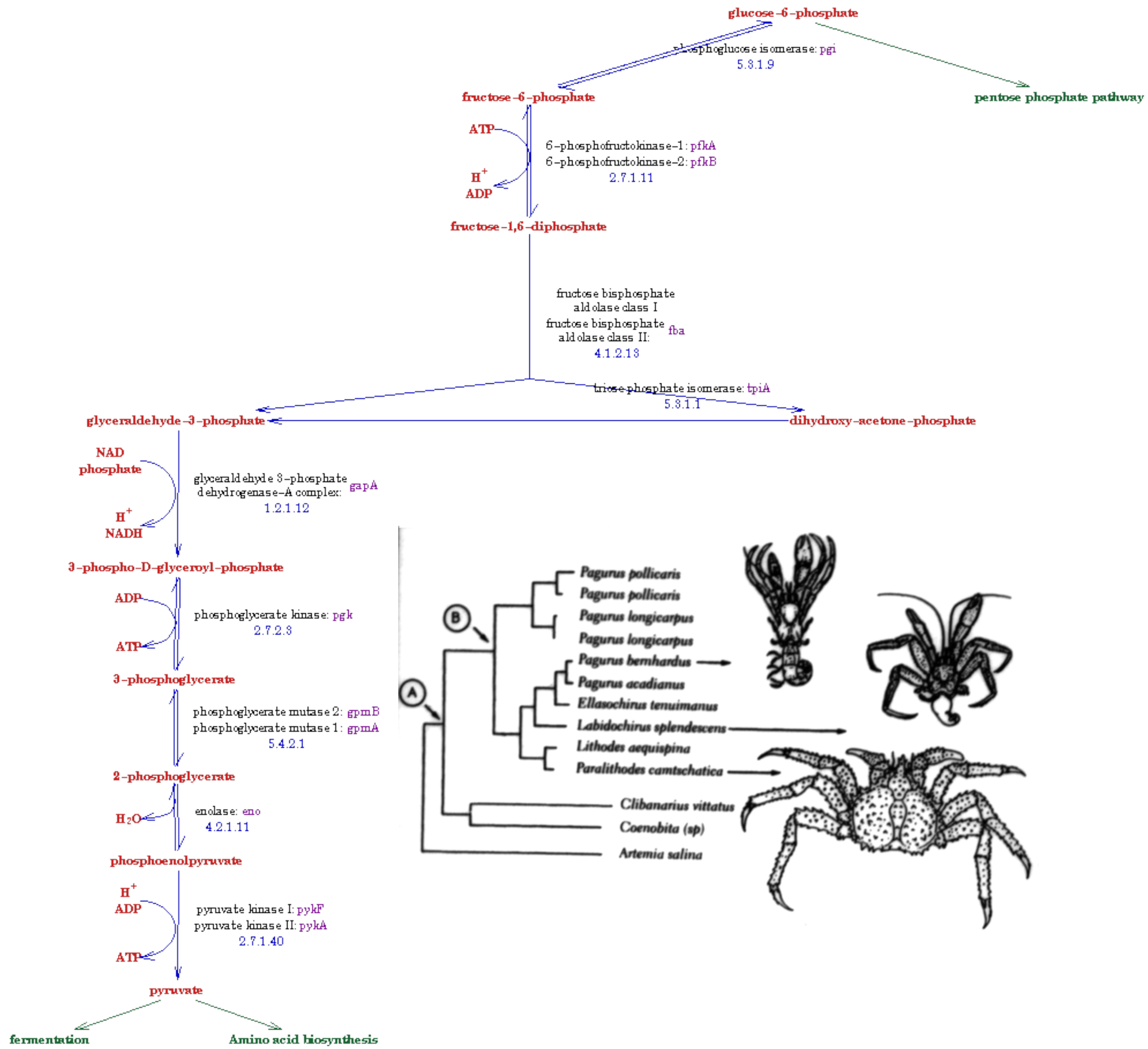


miRNA-target networks

[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686]

Molecular Biology Information: Other Integrative Data

- Information to understand genomes
  - Whole Organisms Phylogeny, traditional zoology
  - Environments, Habitats, ecology
  - Phenotype Experiments (large-scale KOs, transposons)
  - The Literature (MEDLINE)
- The Future....



(Pathway drawing from P Karp's EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)

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**.

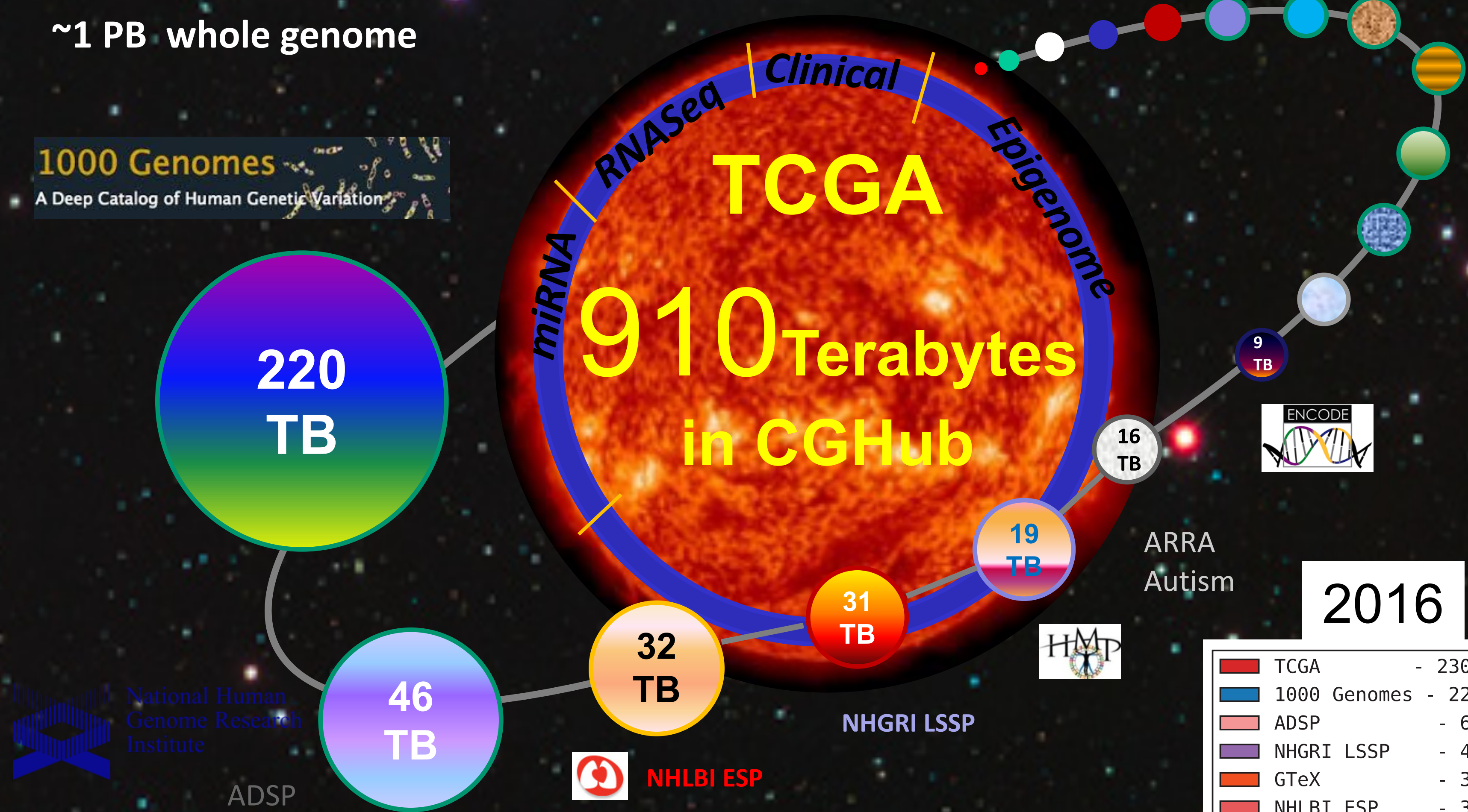
[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Seq Universe

[from Heidi Sofia, NHGRI]

SRA >1 petabyte

TCGA endpoint: ~2.5 Petabytes
- ~1.5 PB exome
- ~1 PB whole genome



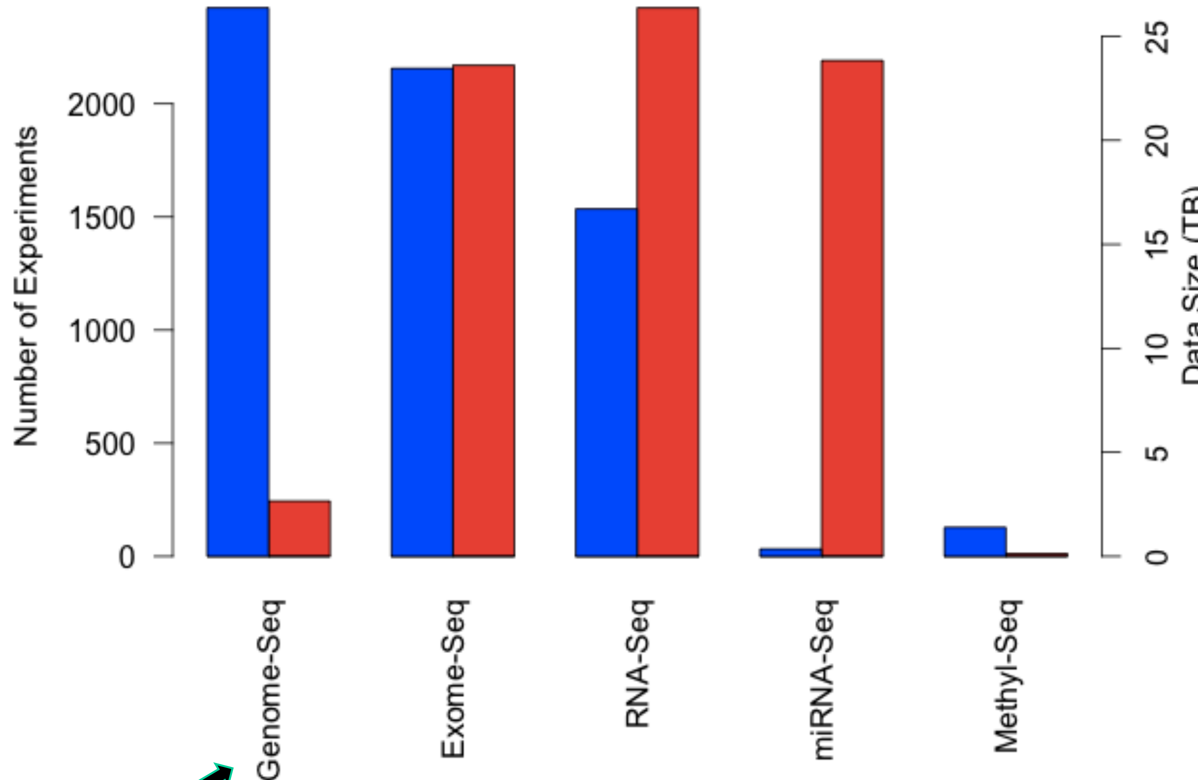
2016

TCGA	- 2300 TB
1000 Genomes	- 222 TB
ADSP	- 68 TB
NHGRI LSSP	- 40 TB
GTeX	- 34 TB
NHLBI ESP	- 32 TB
HMP	- 29 TB
ARRA Autism	- 24 TB
ENCODE	- 9 TB

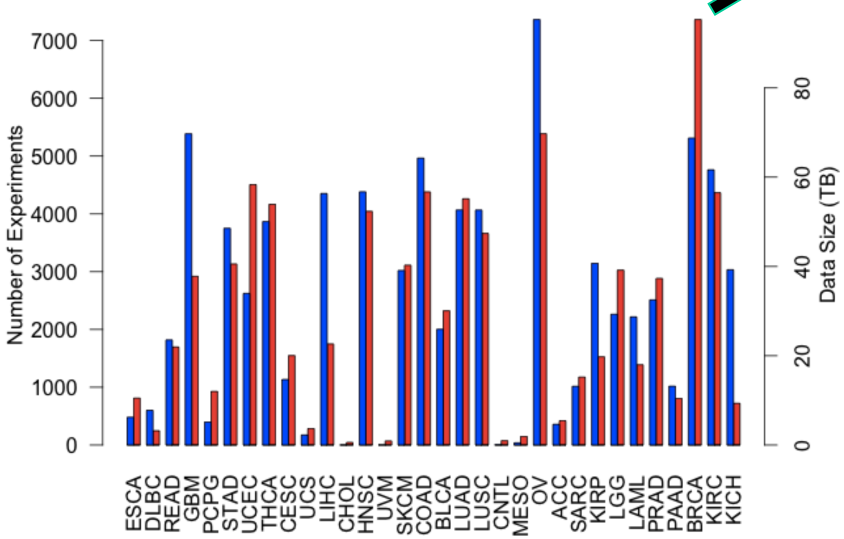
Sofia, 2-28-14

TCGA: What's in a petabyte?

- >73,000 Expt
- 34 Cancer Types
- ~5,000 Patients



TCGA Cancer Types



Breast Cancer Expt. Types

What is Bioinformatics?

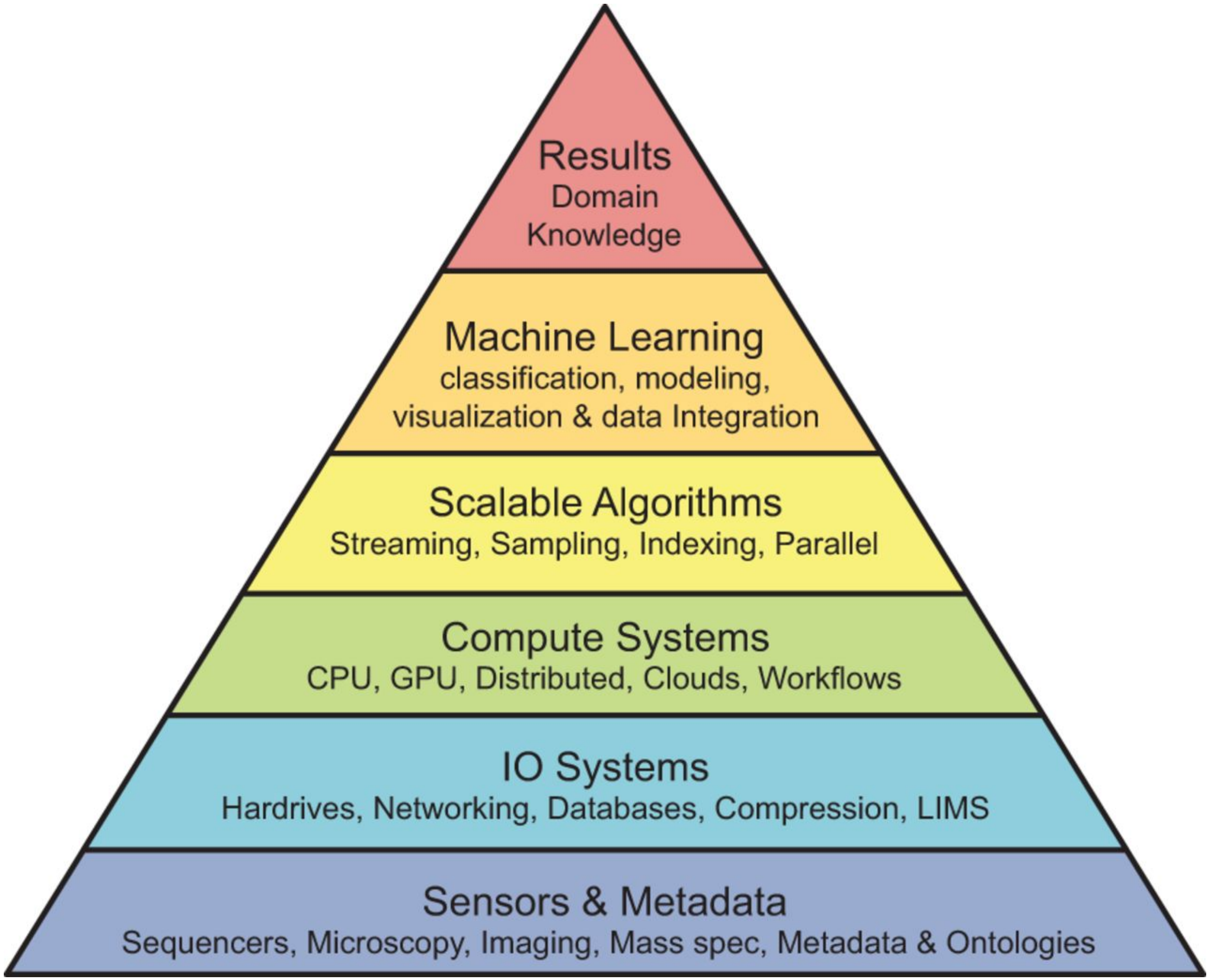
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, 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]

General Types of "Informatics" techniques in Computational Biology – a mix between mining & modeling

- **Databases**
  - Building, Querying
  - Representing Complex data
- **Data mining**
  - Machine Learning techniques
  - Clustering & Tree construction
  - Rapid Text String Comparison & textmining
  - Detailed statistics of significance & association
- **Network Analysis**
  - Analysis of Topology (eg Hubs)
  - Predicting Connectivity
- **Structure Analysis & Geometry**
  - Graphics (Surfaces, Volumes)
  - Comparison & 3D Matching (Vision, recognition, docking)
- **Physical Modeling**
  - Newtonian Mechanics
  - Minimization & Simulation
  - Modeling Chemical Reactions & Cellular Processes

Data science analysis stack.



Michael C. Schatz Genome Res. 2015;25:1417-1422

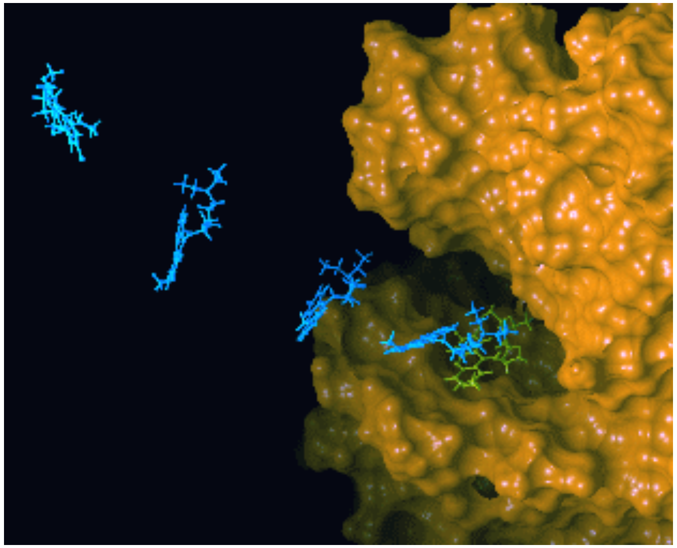
What is Bioinformatics?

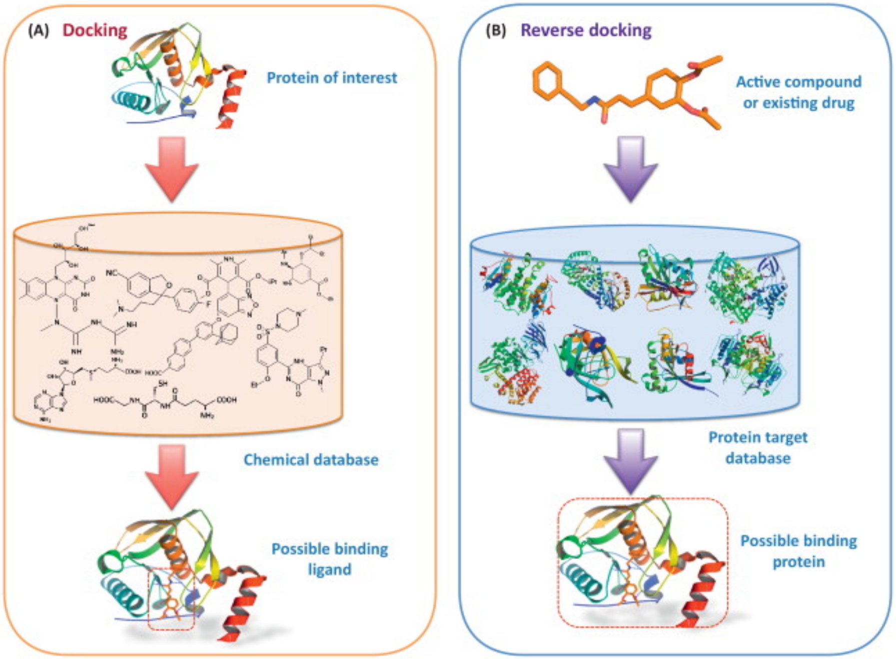
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, 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]

Major Application I: Designing Drugs

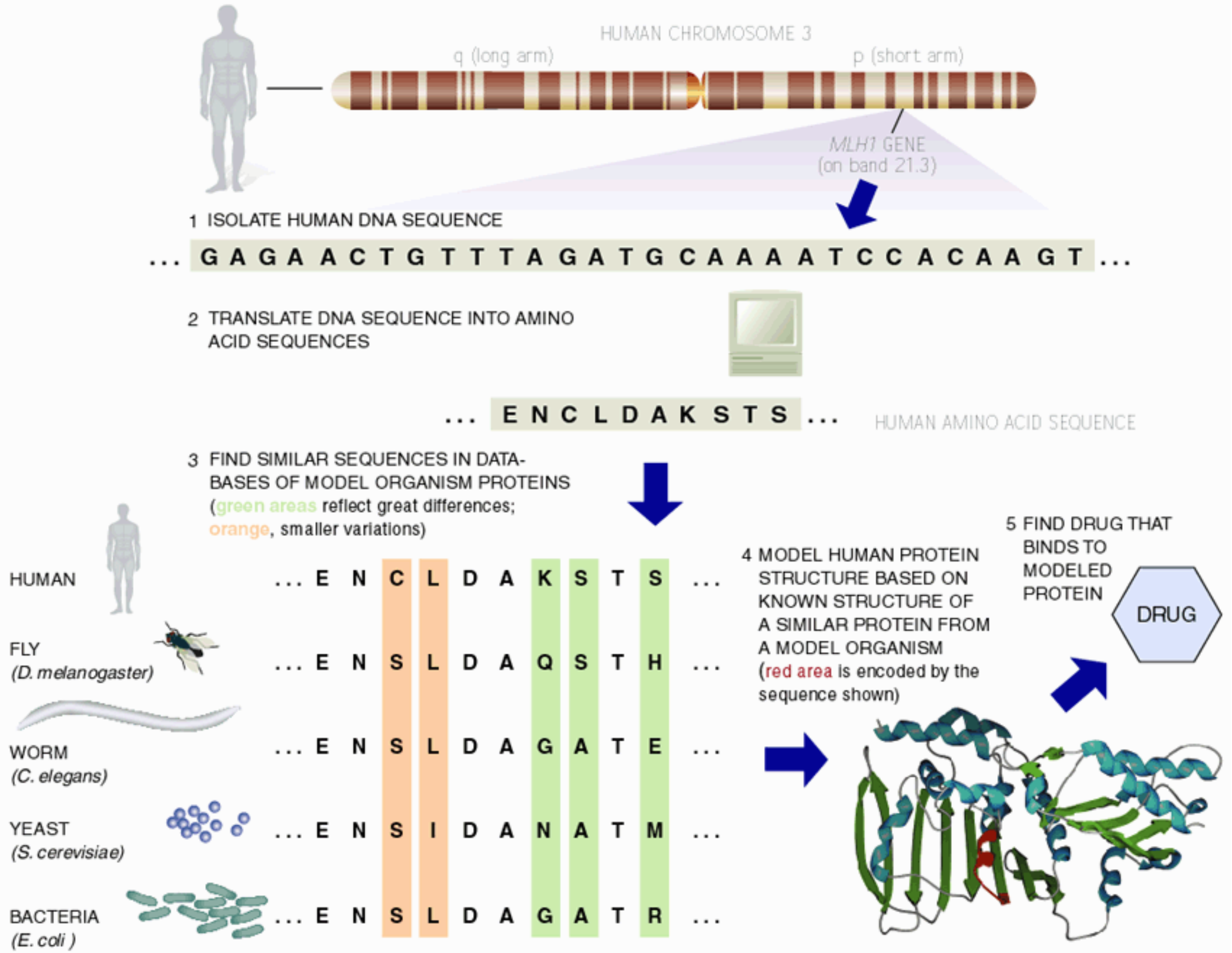
- Understanding how structures bind other molecules
- Designing inhibitors using docking, structure modeling
- *In silico* screens of chemical and protein databases





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

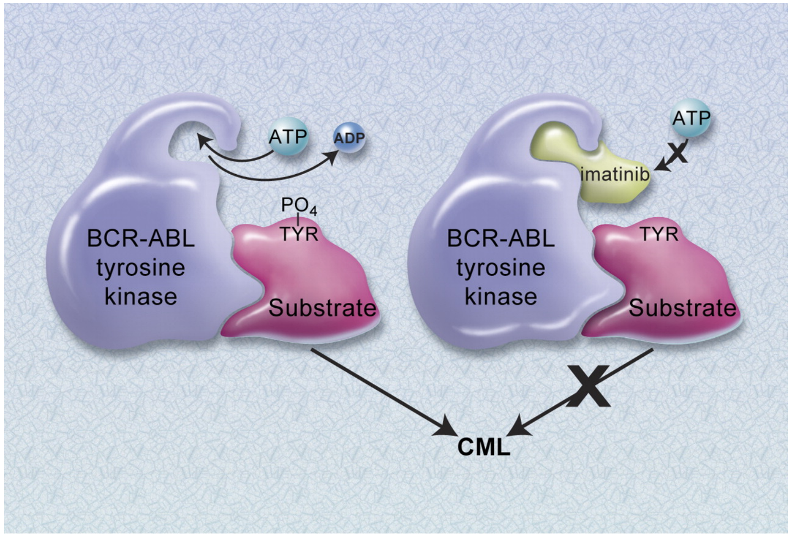
Major Application II: Finding Homologs

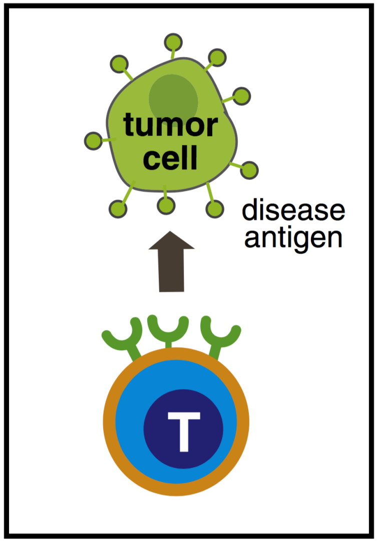


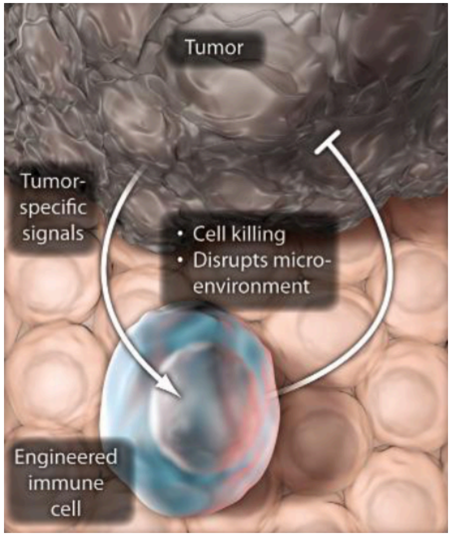
[Adapted from Sci. Am.]

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



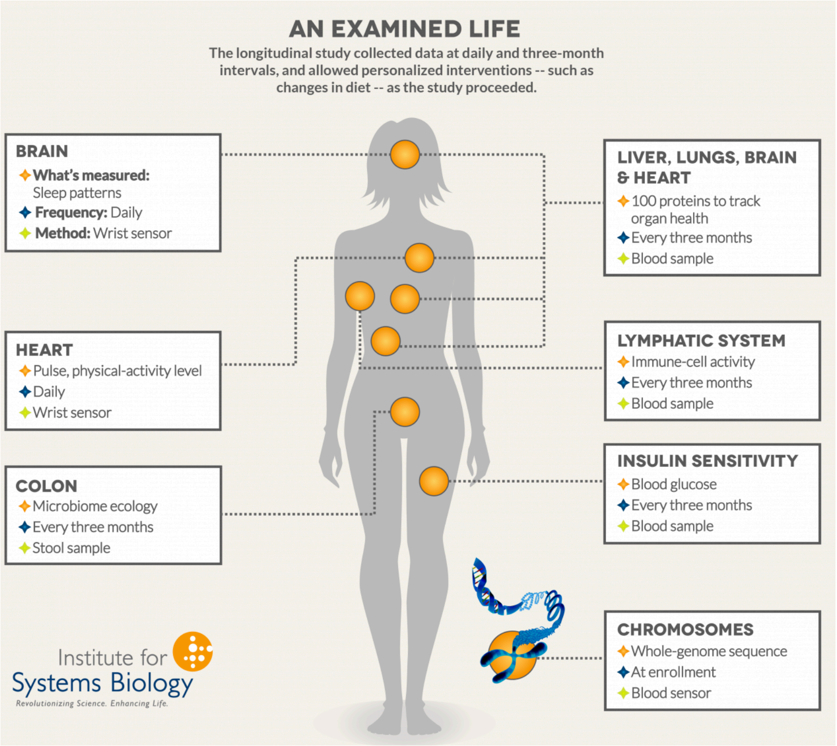




(From left to right, figures adapted from Druker BJ. Blood 2008 and the Lim Lab at UCSF)

Major Application IV: Personal Genome Characterization

- Identify mutations in personal genomes.
  - SNPs, structural variants
- Estimate phenotypic (deleterious or protective) impact of variants.
- Compare one person to wider population.
- Track changes over time.
  - Transcriptome studies
  - Longitudinal health studies (e.g. 100K wellness project, Framingham Heart Study)



(Figure from Institute for Systems Biology)

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]

Personal Genomics

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.

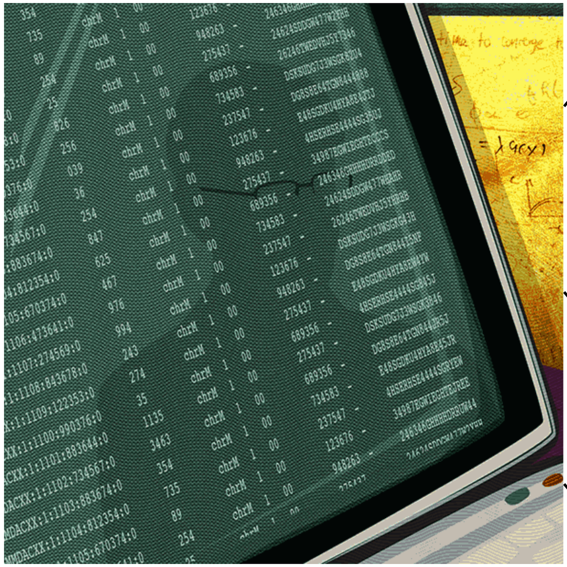
- Personal genome annotation combined with multi-omic and longitudinal health data can inform new links between genotype and phenotype relevant to an individual and the larger population.

- 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.

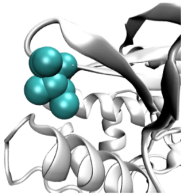
- We will also use analysis techniques including systems and network modeling as well as structural modeling to contextualize and interpret the mechanisms through which these variants impact health.

Analyzing Carl Zimmer's genome



SNV AAGCT → ACGCT

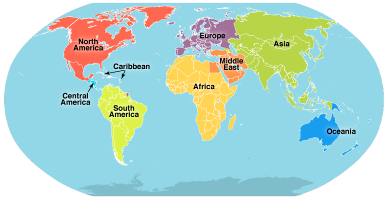
Protein Structure



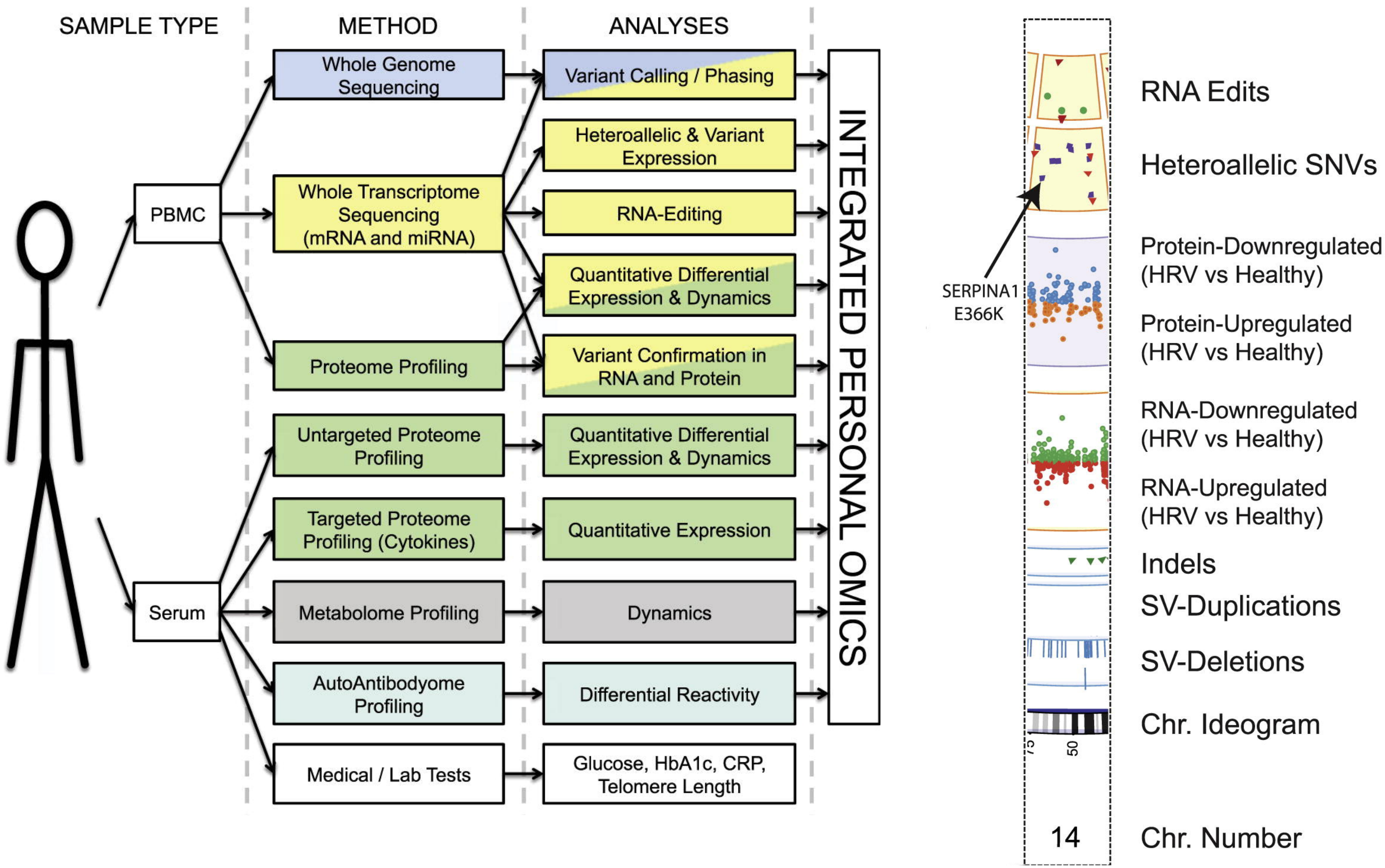
Wild-type

Mutated

Ancestry



Personal Omics Profiling



(Figure from Chen et al. Cell 2012)

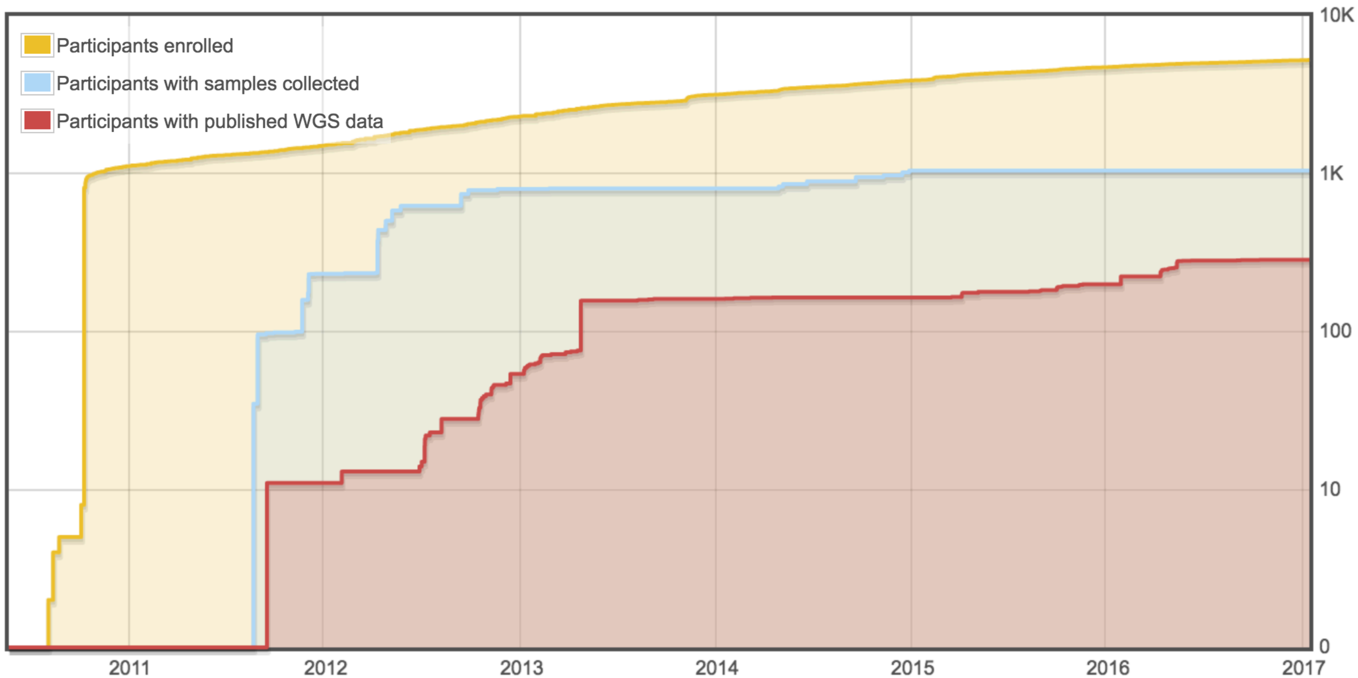
Personal Genome Project

Sharing Personal Genomes

The Personal Genome Project was founded in 2005 and is dedicated to creating public genome, health, and trait data. Sharing data is critical to scientific progress, but has been hampered by traditional research practices—our approach is to invite willing participants to publicly share their personal data for the greater good.

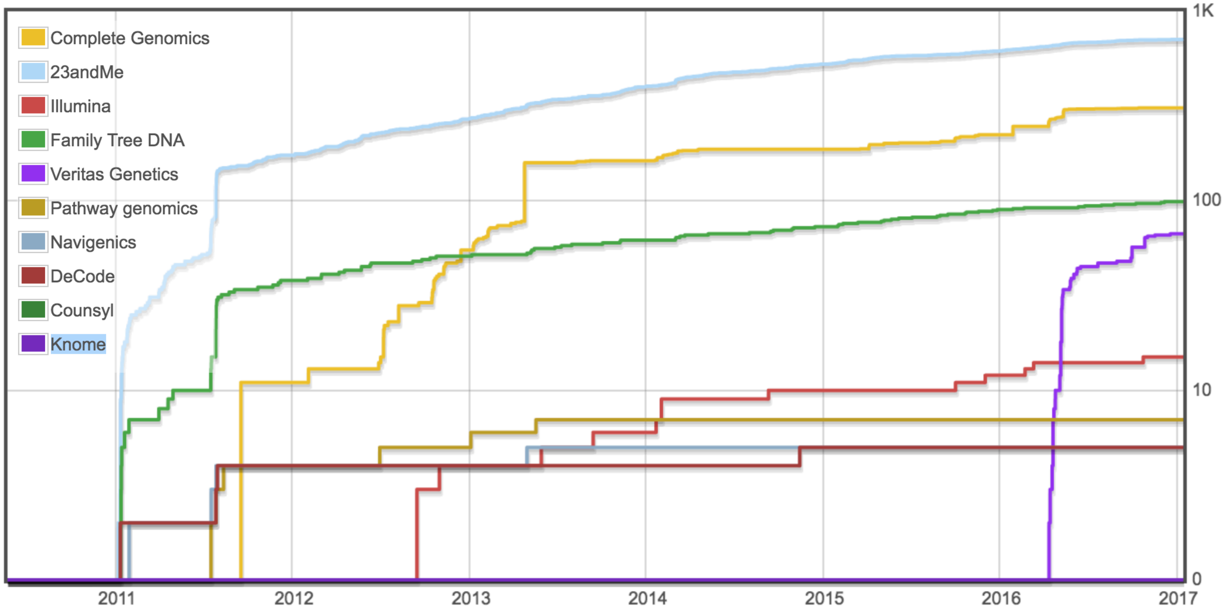
Learn more ›

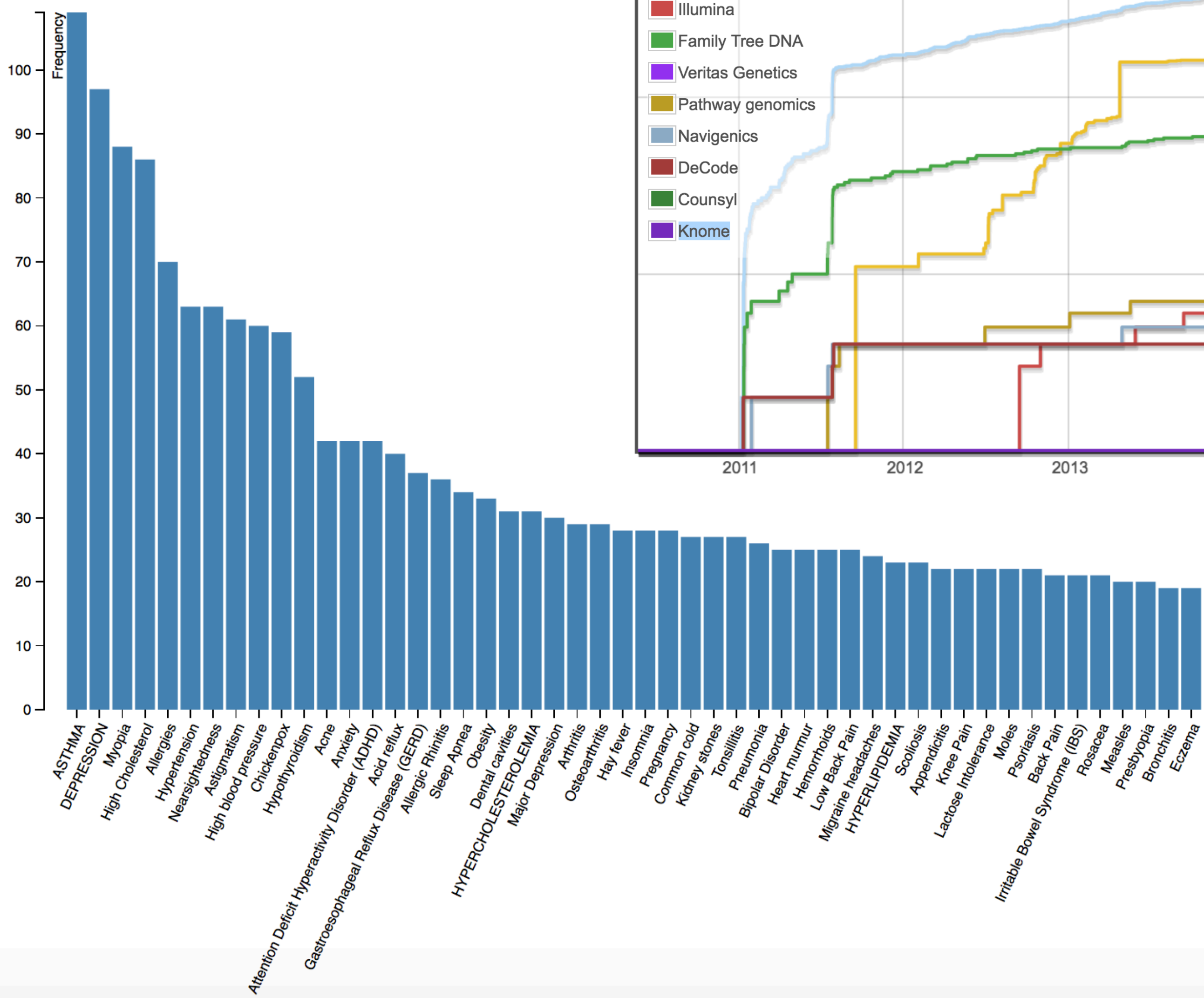
Pipeline: enrolled → samples collected → WGS data published



Data Types in the Personal Genome Project

Number of participants per data type

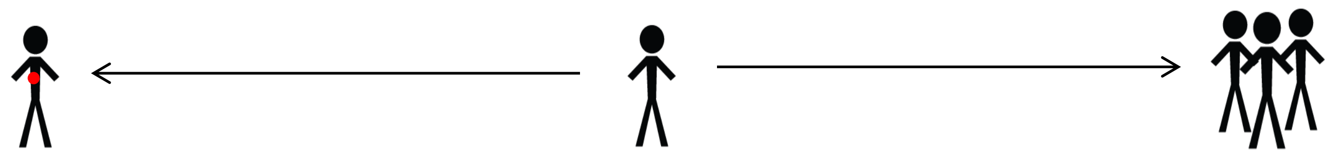




Conditions

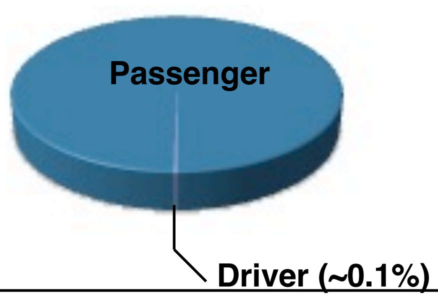
Human Genetic Variation

A Cancer Genome | **A Typical Genome** | **Population of 2,504 peoples**



Origin of Variants

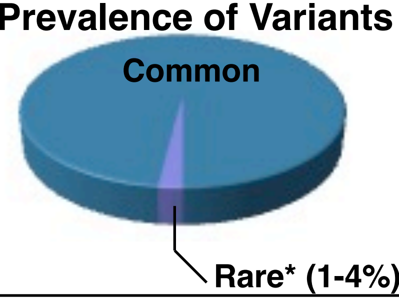
	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K



Class of Variants

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

Prevalence of Variants

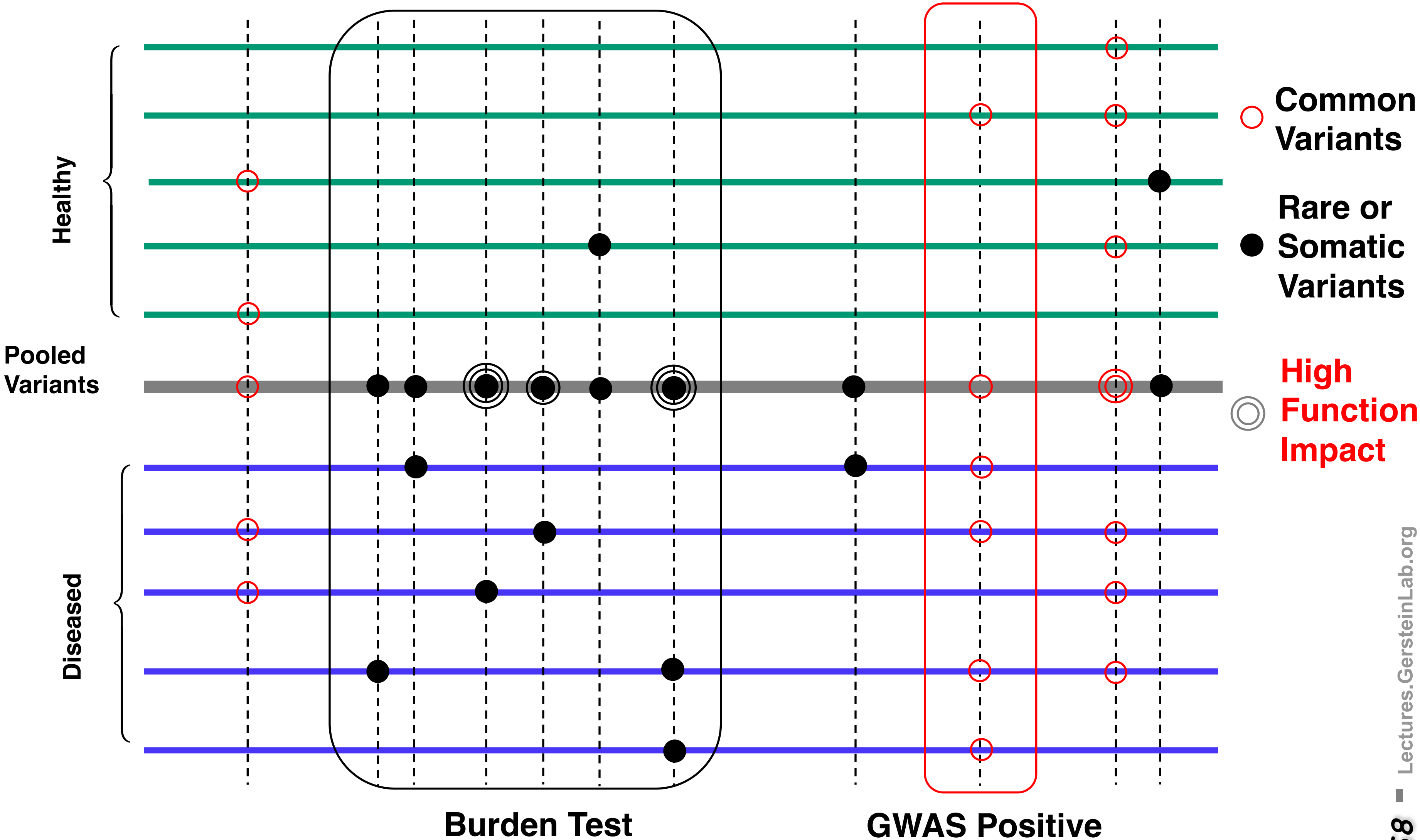


SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M

* 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



Extra stuff related to 1st Assignment

Class Web Page

GersteinLab.org/courses/452

Assignment #0 Page

goo.gl/**BfSpQV**

Are They or Aren't They Comp. Bio.? (#1)

- (Digital Libraries & Medical Record Analysis
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature
- (Motif Discovery Using Gibb's Sampling
- (Methods for Structure Determination
  - Computational Crystallography
    - Refinement
  - NMR Structure Determination
    - (Distance Geometry
- (Metabolic Pathway Simulation
- (The DNA Computer

Are They or Aren't They Comp. Bio.? (#1, Answers)

- (YES?) Digital Libraries & Medical Record Analysis
  - Automated Bibliographic Search and Textual Comparison
  - Knowledge bases for biological literature
- (YES) Motif Discovery Using Gibb's Sampling
- (NO?) Methods for Structure Determination
  - Computational Crystallography
    - Refinement
  - NMR Structure Determination
    - (YES) Distance Geometry
- (YES) Metabolic Pathway Simulation
- (NO) The DNA Computer

Are They or Aren't They Comp. Bio.? (#2)

- (Gene identification by sequence characteristics
  - Prediction of splice sites
- (DNA methods in forensics
- (Modeling of Populations of Organisms
  - Ecological Modeling (predator & prey)
- (Modeling the nervous system
  - Computational neuroscience
  - Understanding how brains think & using this to make a better computer
- (Molecular phenotype discovery – looking for gene expression signatures of cancer
  - What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#2, `Answers`)

- **(YES)** Gene identification by sequence characteristics
  - Prediction of splice sites
- **(YES)** DNA methods in forensics
- **(NO)** Modeling of Populations of Organisms
  - Ecological Modeling (predator & prey)
- **(NO?)** Modeling the nervous system
  - Computational neuroscience
  - Understanding how brains think & using this to make a better computer
- **(YES)** Molecular phenotype discovery – looking for gene expression signatures of cancer
  - What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#3)

- (RNA structure prediction
- (Radiological Image Processing
  - Computational Representations for Human Anatomy (visible human)
- (Artificial Life Simulations
  - Artificial Immunology / Computer Security
  - (Genetic Algorithms in molecular biology
- (Homology Modeling & Drug Docking
- (Char. drugs & other small molecules (QSAR)
- (Computerized Diagnosis based on Pedigrees
- (Processing of NextGen sequencing image files
- (Module finding in protein networks

Are They or Aren't They Comp. Bio.? (#3, Answers)

- (YES) RNA structure prediction
- (NO) Radiological Image Processing
  - Computational Representations for Human Anatomy (visible human)
- (NO) Artificial Life Simulations
  - Artificial Immunology / Computer Security
  - (NO?) Genetic Algorithms in molecular biology
- (YES) Homology Modeling & Drug Docking
- (YES) Char. drugs & other small molecules (QSAR)
- (NO) Computerized Diagnosis based on Pedigrees
- (NO) Processing of NextGen sequencing image files
- (YES) Module finding in protein networks