Applied Comparative Genomics

Michael Schatz

August 26, 2024

Lecture 1: Course Overview

Welcome!

The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses.

- We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data.
- The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life.
- The topics will include (pan)-genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

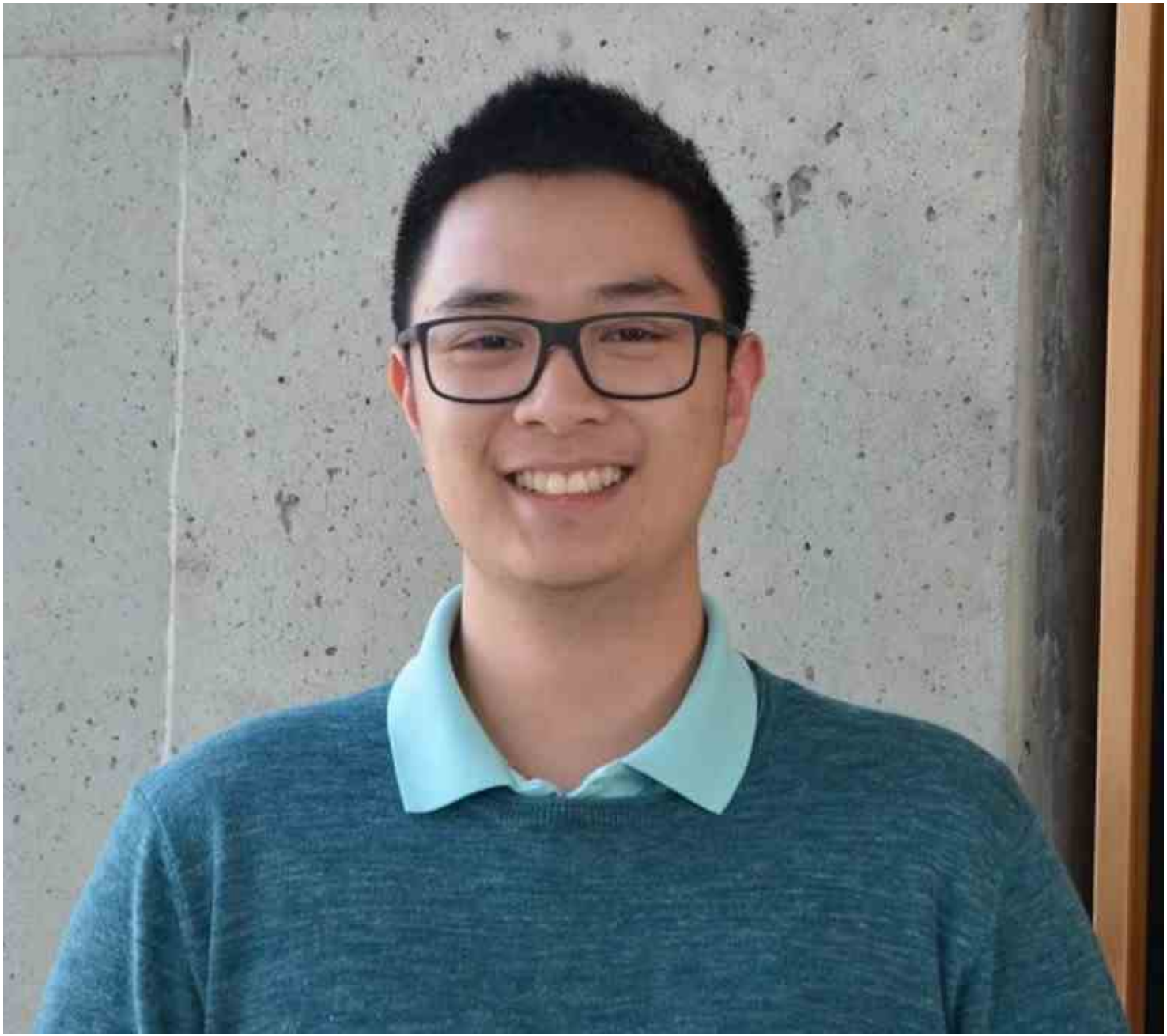
Course Webpage: https://github.com/schatzlab/appliedgenomics2024
Course Discussions: https://piazza.com/jhu/fall2024/600449600649/home

Class Hours: Mon + Wed @ 3:00p – 4:15p, Hodson 311

Schatz Office Hours: TBD and by appointment
Nguyen Office Hours: TBD and by appointment

Please try Piazza first!

TA: Matthew Nguyen



Prerequisites and Resources

Prerequisites

- No formal course requirements
- Access to an Apple or Linux Machine, or Install VirtualBox
- Familiarity with the Unix command line for exercises
  - bash, ls, grep, sed, + install published genomics tools
- Familiarity with a major programming language for project
  - C/C++, Java, R, Perl, Python

Primary Texts

- None! We will be studying primary research papers

Other Resources:

- Google, SEQanswers, Biostars, StackOverflow

- Applied Computational Genomics Course at UU
- https://github.com/quinlan-lab/applied-computational-genomics

- Ben Langmead's teaching materials:
- http://www.langmead-lab.org/teaching-materials/

Grading Policies

Assessments:

- 5 Assignments: 30% Due at 11:59pm a week later
  Practice using the tools and methods we are discussing

- 1 Exam: 30% In Class (Tentatively Nov 4)
  Assess your performance, focusing on the concepts

- 1 Class Project: 40% Presented last week of class
  Significant project developing a novel analysis/method

- In-class Participation: Not graded, but there to help you!

Policies:

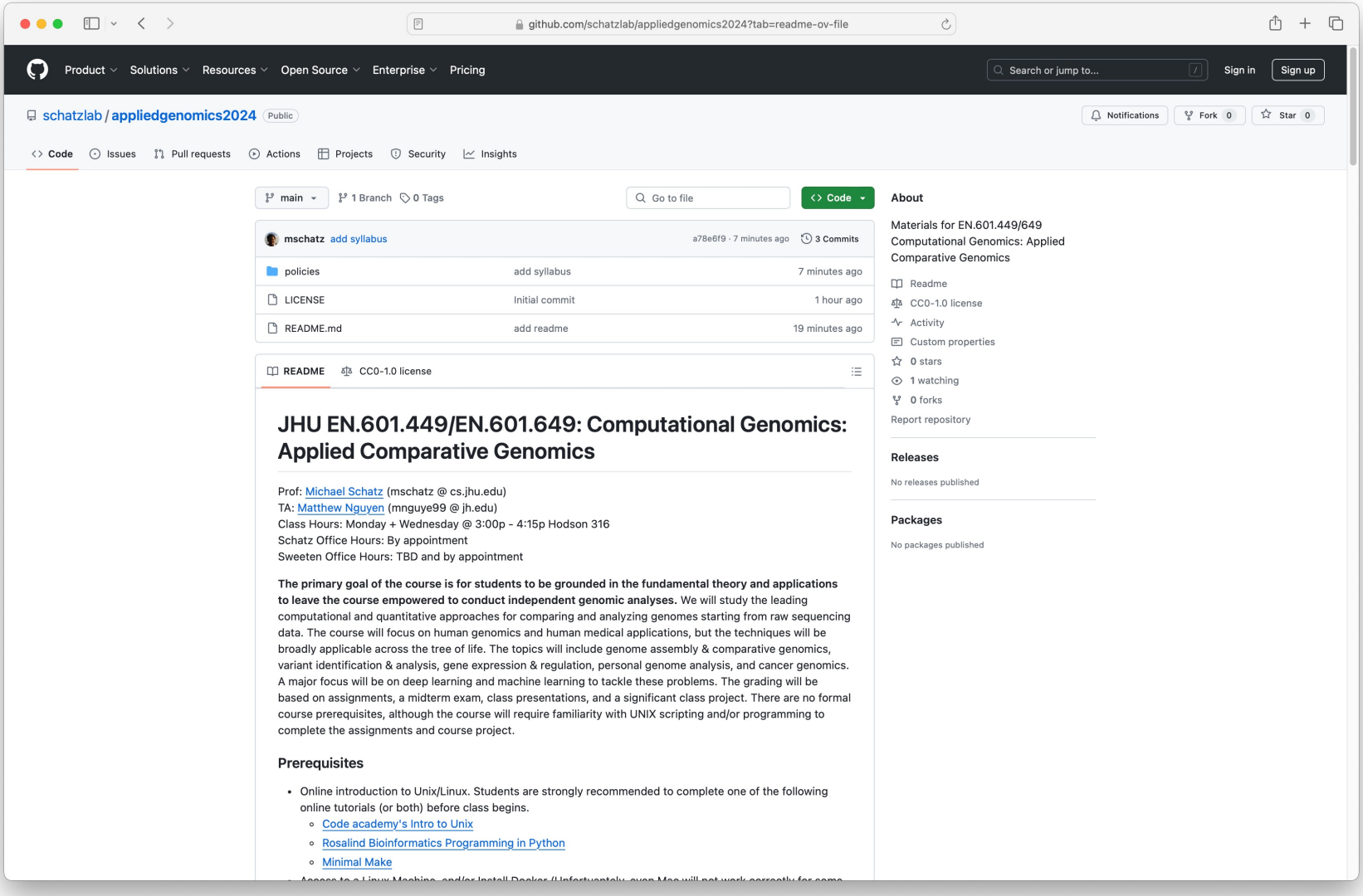
- Scores assigned relative to the highest points awarded
- Automated testing and grading of assignments
- ***Late Days:***
  - A total of 96 hours (24 x 4) can be used to extend the deadline for assignments, but not the class project, without any penalty; after that time assignments will not be accepted

On ChatGPT

whats a good policy on the use of chatgpt for a university class on genomics and data science? the instructor is generally favorable but also wants students to learn and understand the content presented in class. The instructor is also concerned that chatgpt will sometimes make up answers that are completely wrong (hallucinations)

A good policy for the use of ChatGPT in a university class on genomics and data science should emphasize that the tool is a supplement to, not a replacement for, engaging with primary course materials and learning objectives. Students should be encouraged to use ChatGPT to clarify complex concepts, find additional resources, and practice coding, but not to complete assignments or exams directly from AI-generated content without proper understanding and cross-referencing with reputable scientific sources. The policy should require students to disclose when they use ChatGPT and reflect on its contributions and limitations, fostering critical thinking and awareness of potential inaccuracies or "hallucinations." Educating students on the limitations of AI, including how to fact-check responses against trusted academic sources, will enhance their ability to use the tool responsibly. Additionally, promoting ethical use and reinforcing academic integrity are crucial, with the instructor modeling appropriate use and providing clear examples of acceptable and unacceptable practices. This approach will help students develop essential skills in genomics and data science while making the most of AI tools like ChatGPT.

Course Webpage

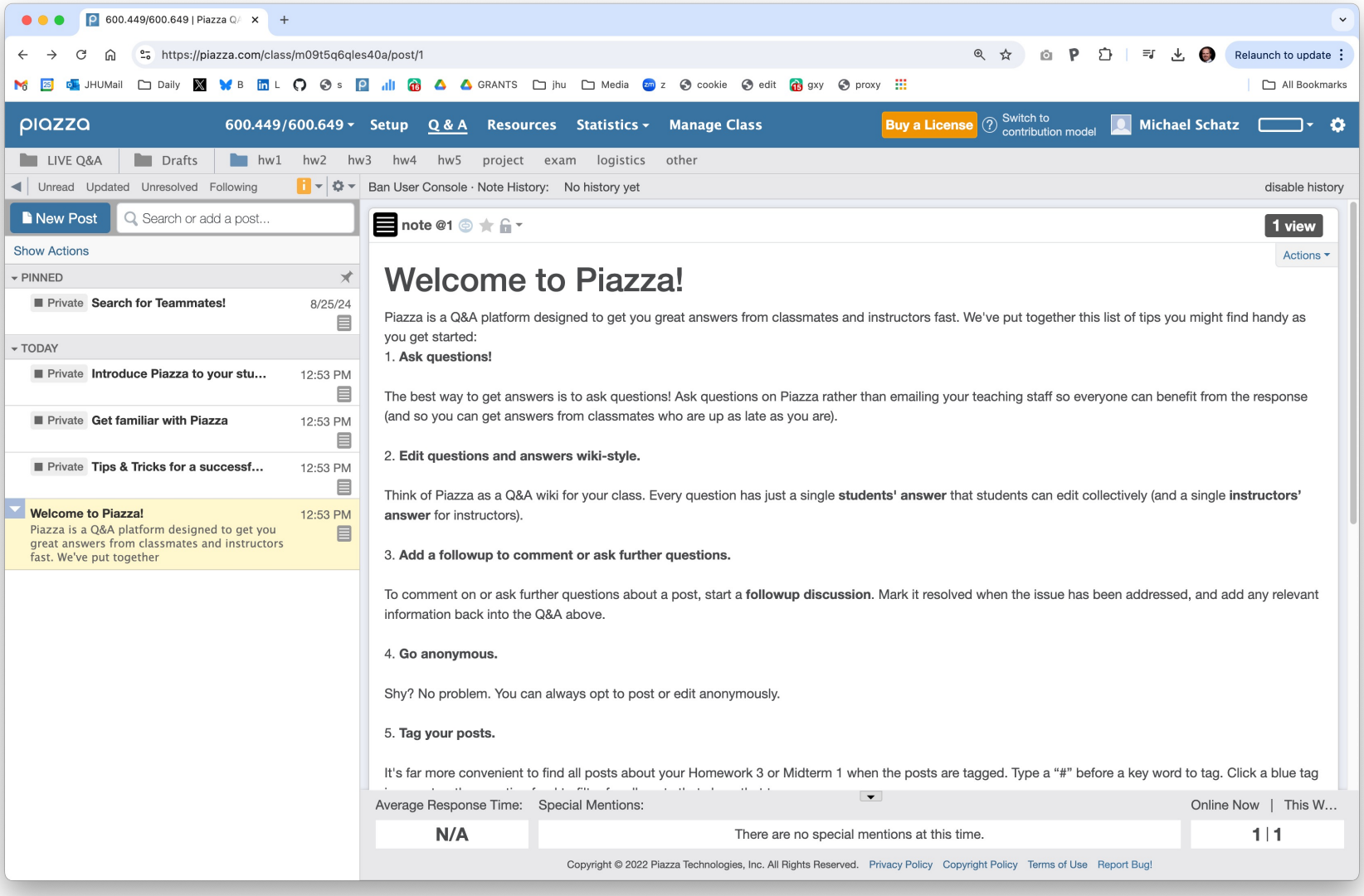


https://github.com/schatzlab/appliedgenomics2024

No class on November 13

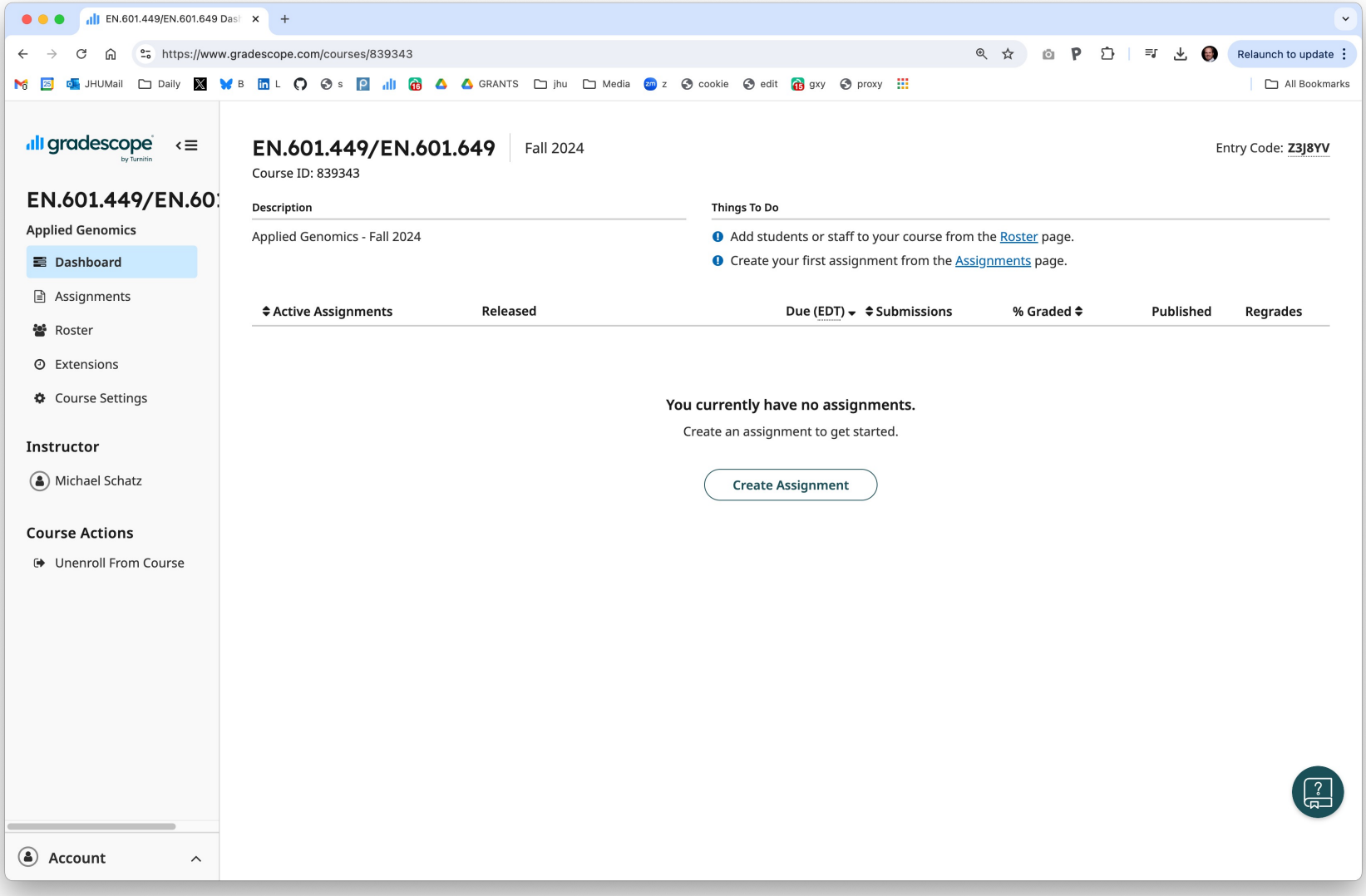
https://meetings.cshl.edu/meetings.aspx?meet=data&year=24

Piazza



https://piazza.com/class/m09t5q6qles40a

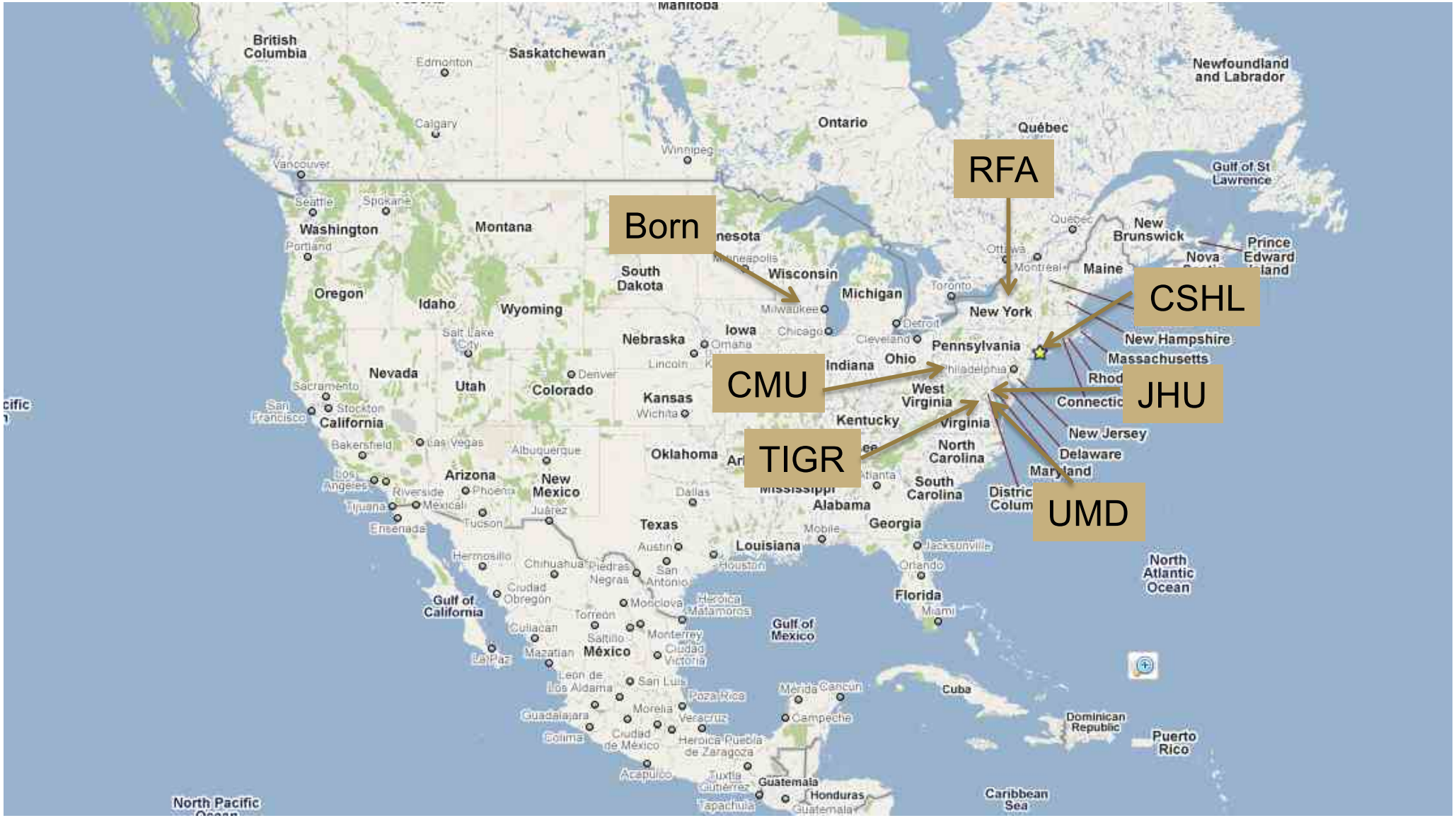
GradeScope



https://www.gradescope.com/

Entry Code: **Z3J8YV**

A Little About Me



Schatzlab Overview

Human Genetics

Role of mutations in disease

Nurk *et al.* (2022)
Aganezov *et al.* (2020)

Agricultural Genomics

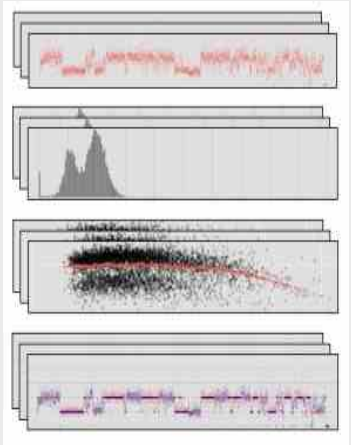
Genomes & Transcriptomes

Satterlee *et al.* (2024)
Naish *et al.* (2021)

Algorithmics & Systems Research

Ultra-large scale biocomputing

Kirsche *et al.* (2023)
Schatz *et al.* (2022)

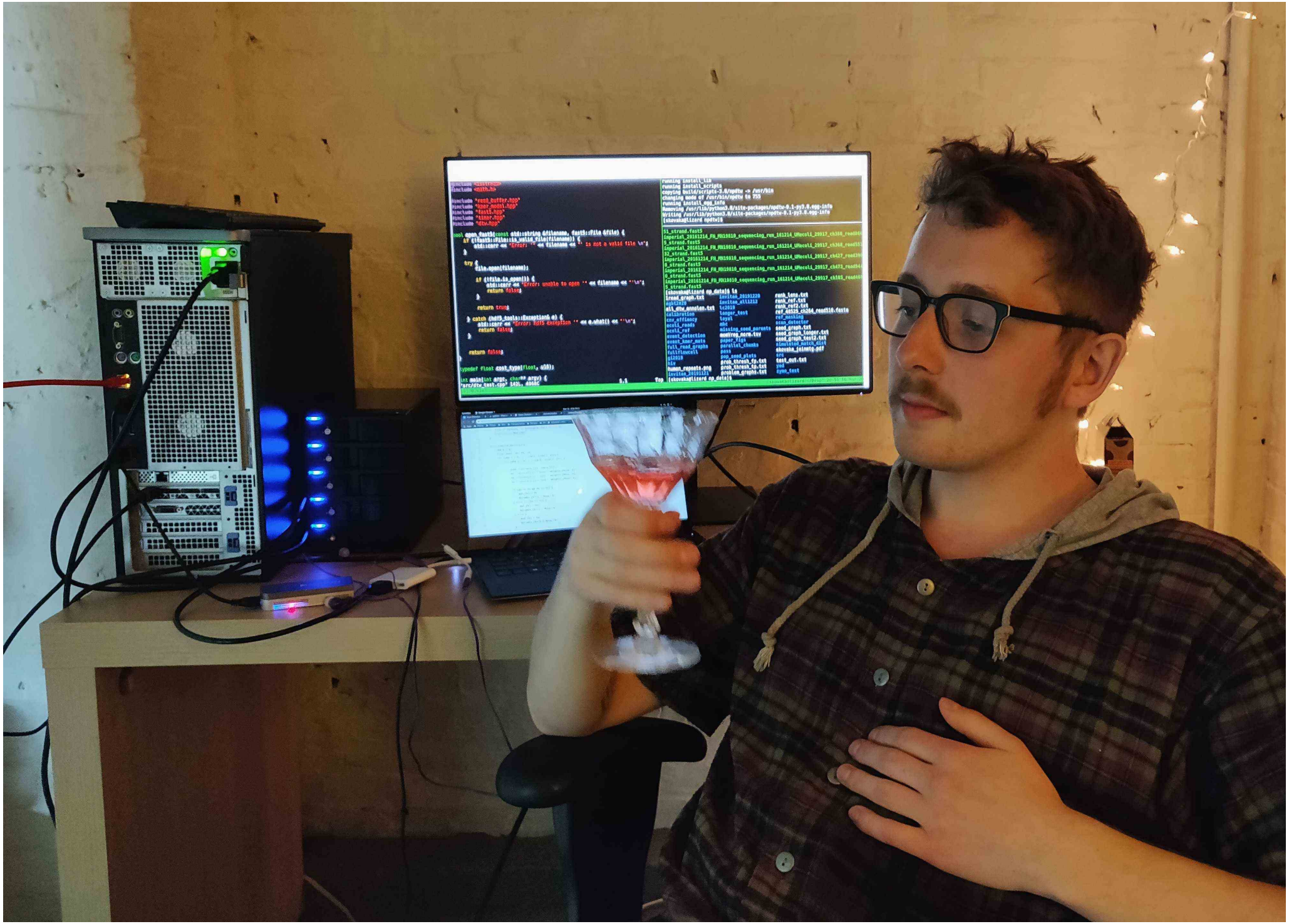


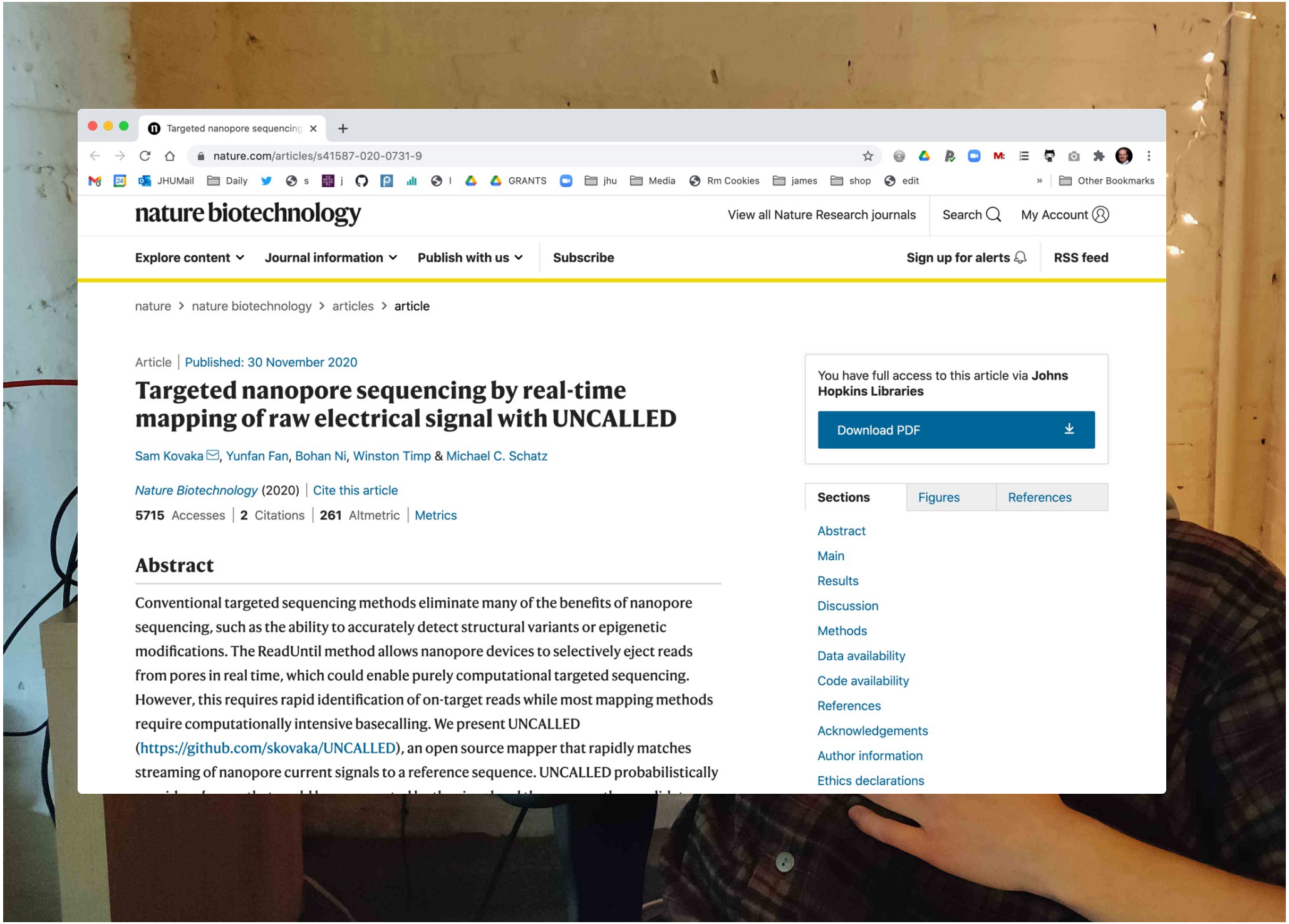
Biotechnology Development

Single Cell + Single Molecule Sequencing

Kovaka *et al.* (2024)
Rozowsky *et al.* (2023)

20	4-Nov	Mon	Midterm * in class exam *		
21	6-Nov	Wed	Human Genetic Diseases		* Genome-Wide Association Studies (Bush & Moore, 2012, PLOS Comp Bio) * The contribution of de novo coding mutations to autism spectrum disorder (Iossifov et al, 2014, Nature)
22	11-Nov	Mon	Metagenomics	Prelim Report Due; Final Report Assigned	* Kraken: ultrafast metagenomic sequence classification using exact alignments (Wood and Salzberg, 2014, Genome Biology) * Chapter 12: Human Microbiome Analysis (Morgan and Huttenhower)
23	13-Nov	Wed	No class -		
24	18-Nov	Mon	Cancer Genomics		* The Hallmarks of Cancer (Hanahan & Weinberg, 2000, Cell) * Evolution of Cancer Genomes (Yates & Campbell, 2012, Nature Reviews Genetics) * Comprehensive molecular portraits of human breast tumours (TCGA, 2012, Nature)
25	20-Nov	Wed	In class project presentation		
*	25-Nov	Mon	Thanksgiving Break		
*	27-Nov	Wed	Thanksgiving Break		
26	2-Dec	Mon	In class project presentation		
27	4-Dec	Wed	In class project presentation		
*	16-Dec	Mon	Final Report Due	Final Report Due	



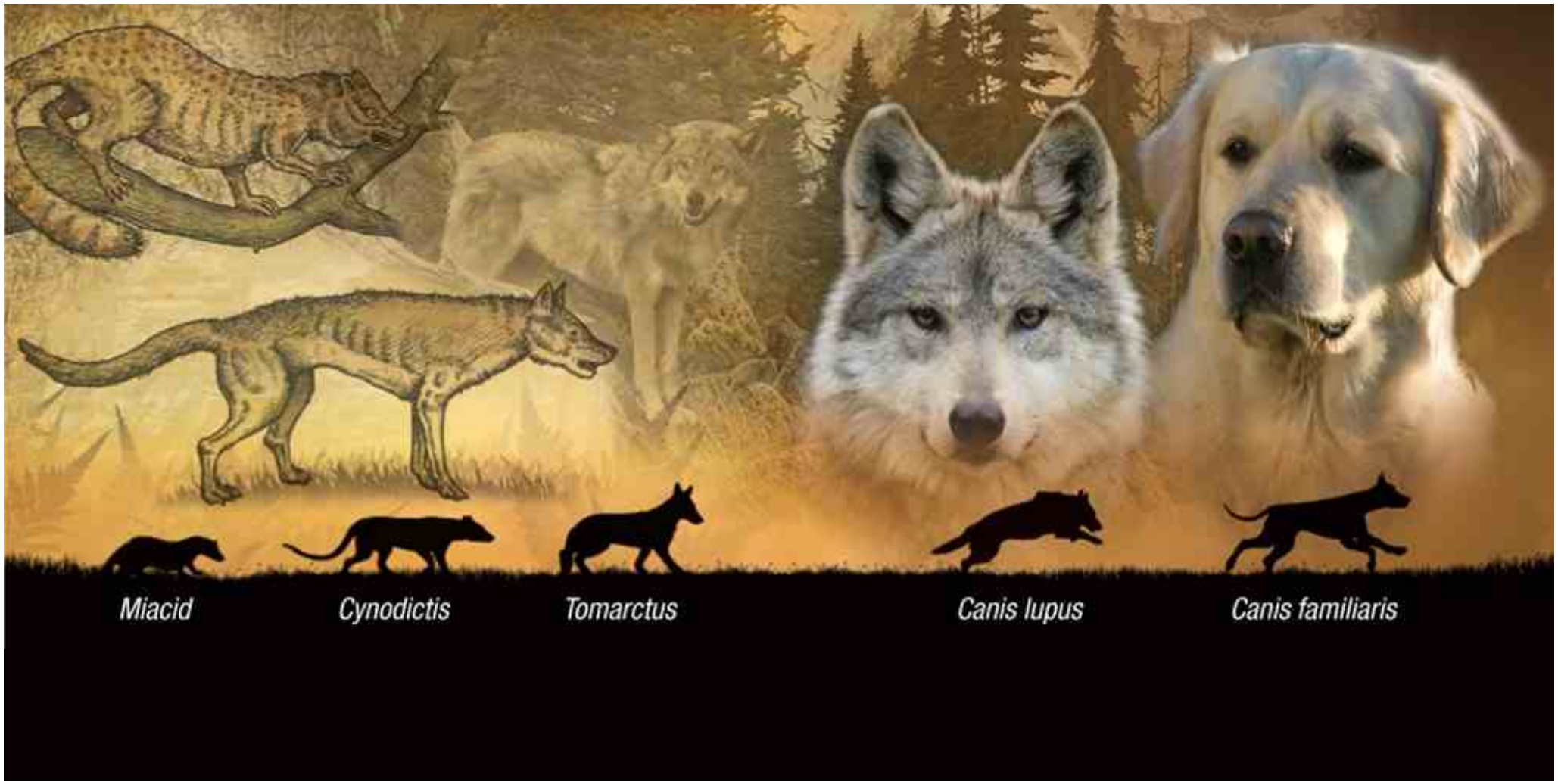

Targeted nanopore sequencing
nature.com/articles/s41587-020-0731-9
JHUMail
Daily
GRANTS
jhu
Media
Rm Cookies
james
shop
edit
Other Bookmarks
nature biotechnology
View all Nature Research journals
Search
My Account
Explore content
Journal information
Publish with us
Subscribe
Sign up for alerts
RSS feed
nature > nature biotechnology > articles > article
Article | Published: 30 November 2020
Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED
Sam Kovaka, Yunfan Fan, Bohan Ni, Winston Timp & Michael C. Schatz
Nature Biotechnology (2020) | Cite this article
5715 Accesses | 2 Citations | 261 Altmetric | Metrics
Abstract
Conventional targeted sequencing methods eliminate many of the benefits of nanopore sequencing, such as the ability to accurately detect structural variants or epigenetic modifications. The ReadUntil method allows nanopore devices to selectively eject reads from pores in real time, which could enable purely computational targeted sequencing. However, this requires rapid identification of on-target reads while most mapping methods require computationally intensive basecalling. We present UNCALLED (https://github.com/skovaka/UNCALLED), an open source mapper that rapidly matches streaming of nanopore current signals to a reference sequence. UNCALLED probabilistically
You have full access to this article via Johns Hopkins Libraries
Download PDF
Sections
Figures
References
Abstract
Main
Results
Discussion
Methods
Data availability
Code availability
References
Acknowledgements
Author information
Ethics declarations

Why Genomics?

Earliest Genomics

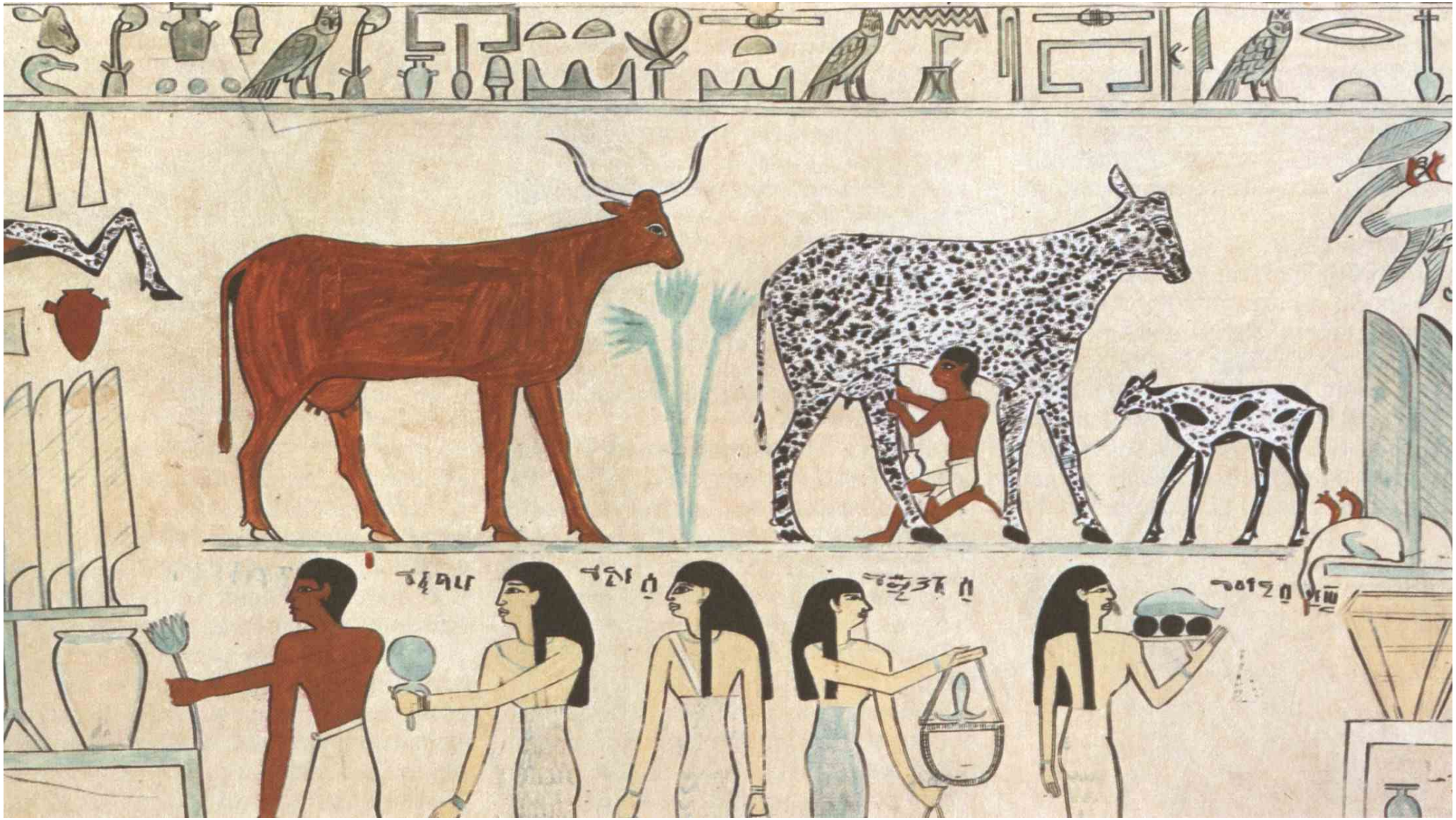
Any Guesses?

Earliest Genomics



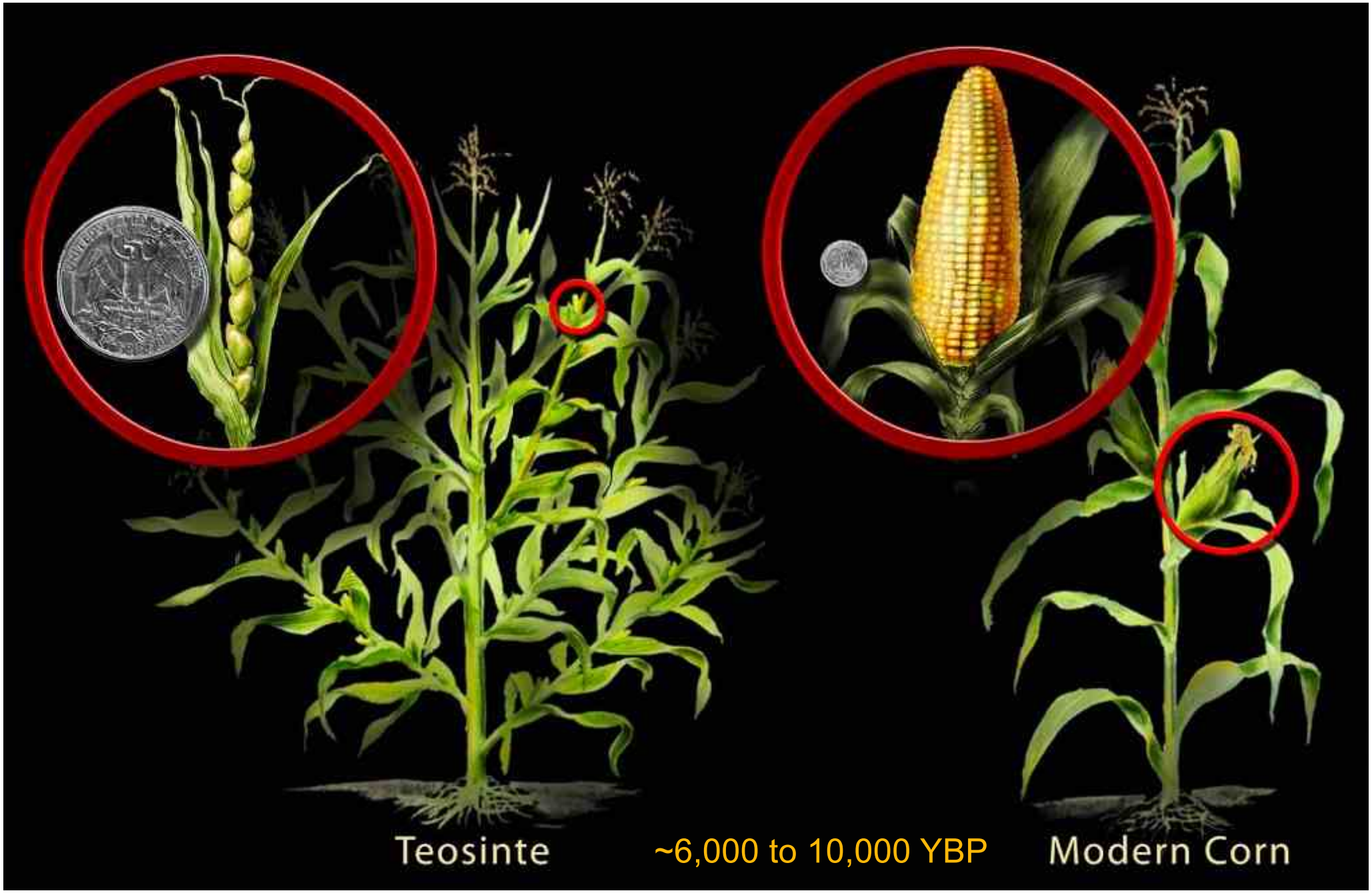
15,000 to 35,000 YBP

Earliest Genomics



~1,000 to 10,000 YBP

Earliest Genomics



Discovery of Chromosomes

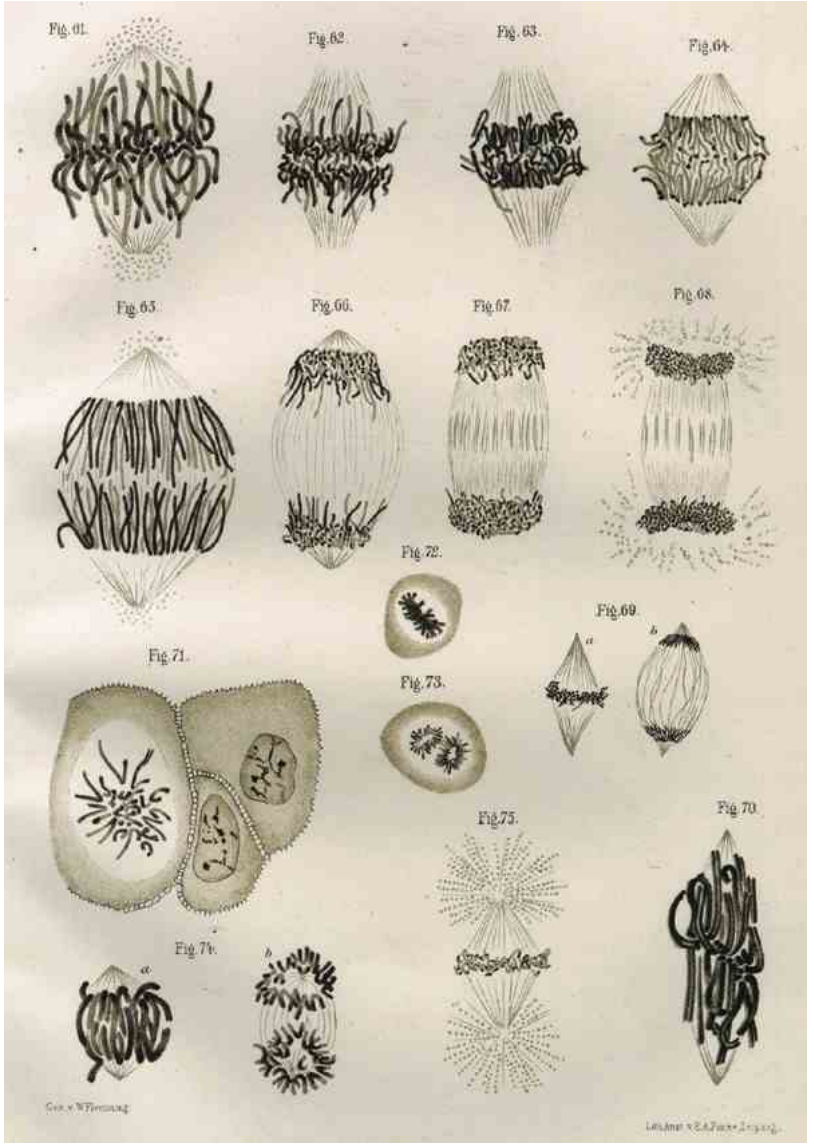
By the mid-1800s, microscopes were powerful enough to observe the presence of unusual structures called “chromosomes” that seemed to play an important role during cell division.

It was only possible to see the chromosomes unless appropriate stains were used

“Chromosome” comes from the Greek words meaning “color body”

Today, we have much higher resolution microscopes, and a much richer varieties of dies and dying techniques so that we can visualize particular sequence elements.

When you see something unexpected that you think might be interesting, give it a name

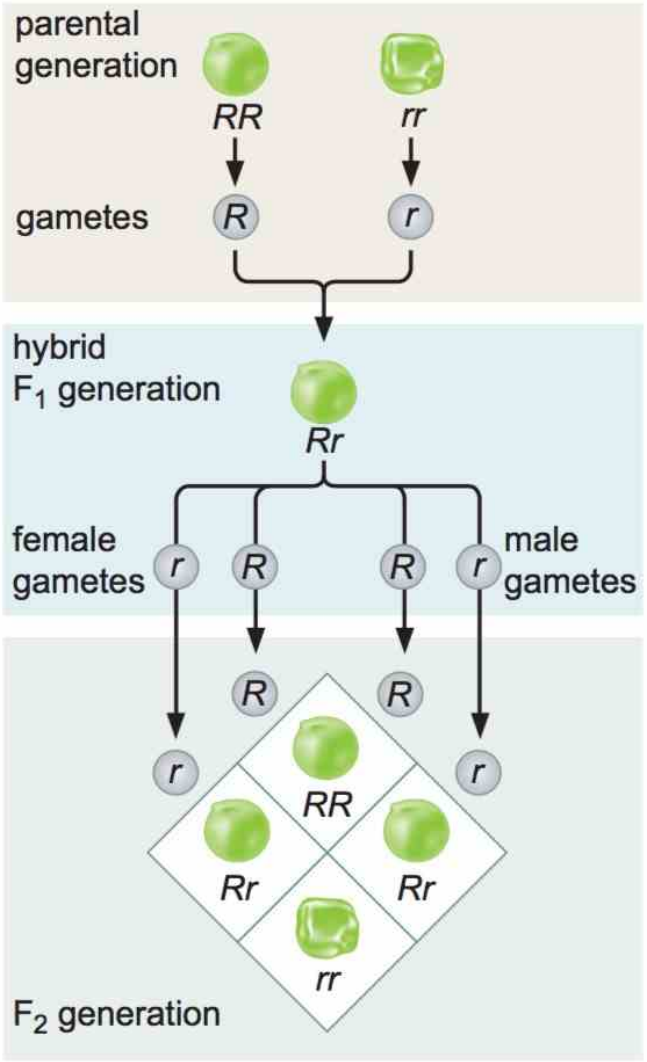


Drawing of mitosis by Walther Flemming.
Flemming, W. Zellsubstanz, Kern und Zelltheilung (F. C. W. Vogel, Leipzig, 1882).

The “first” quantitative biologist

Any Guesses?

Laws of Inheritance



Seed		Flower	Pod		Stem	
Form	Cotyledons	Color	Form	Color	Place	Size
Grey & Round	Yellow	White	Full	Yellow	Axial pods, Flowers along	Long (6-7ft)
White & Wrinkled	Green	Violet	Constricted	Green	Terminal pods, Flowers top	Short (½-1ft)
1	2	3	4	5	6	7

http://en.wikipedia.org/wiki/Experiments_on_Plant_Hybridization

Observations of 29,000 pea plants and 7 traits

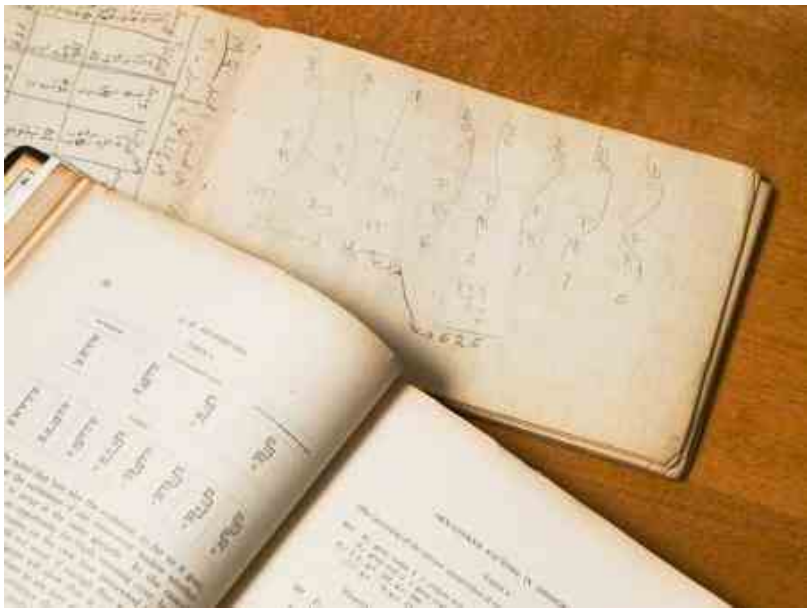
Generation	A	Aa	a	in Verhältniss gestellt: A : Aa : a
1	1	2	1	1 : 2 : 1
2	6	4	6	3 : 2 : 3
3	28	8	28	7 : 2 : 7
4	120	16	120	15 : 2 : 15
5	496	32	496	31 : 2 : 31
n				2^n-1 : 2 : 2^n-1

Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

The first genetic map

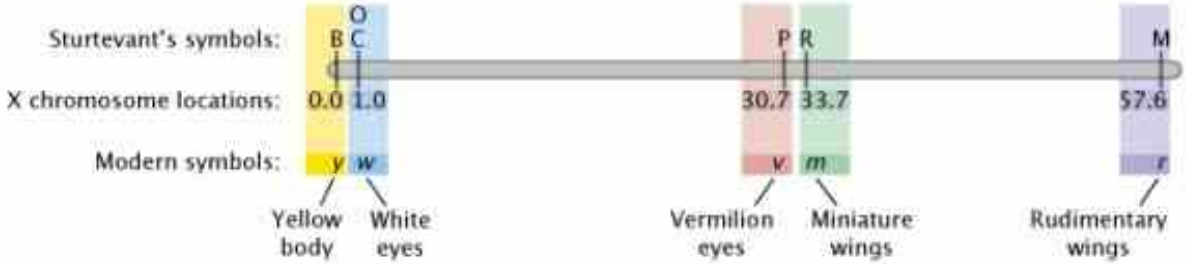
Mendel's Second Law (The Law of Independent Assortment) states alleles of one gene sort into gametes independently of the alleles of another gene: ***Pr(smooth/wrinkle) is independent of Pr(yellow/green)***



http://www.caltech.edu/news/first-genetic-linkage-map-38798

Morgan and Sturtevant noticed that the probability of having one trait given another was **not** always 50/50– those traits are ***genetically linked***

Sturtevant realized the probabilities of co-occurrences could be explained if those alleles were arranged on a linear fashion: traits that are most commonly observed together must be locates closest together

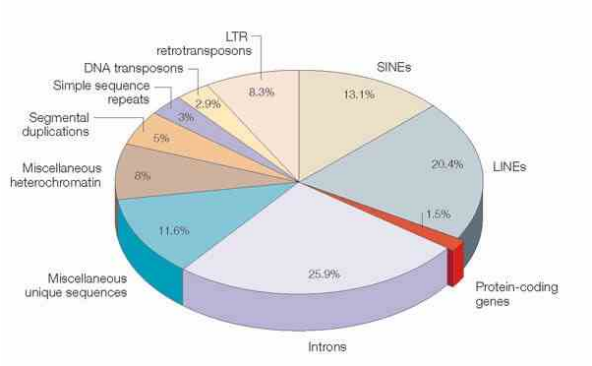


The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) *Journal of Experimental Zoology*, 14: 43-59

Jumping Genes

Previously, genes were considered to be stable entities arranged in an orderly linear pattern on chromosomes, like beads on a string

Careful breeding and cytogenetics revealed that some elements can move (cut-and-paste, DNA transposons) or copy itself (copy-and-paste, retrotransposons)



(Gregory, 2005, Nature Reviews Genetics)

(Much) later analysis revealed that nearly 50% of the human genome is composed of transposable elements, including LINE and SINE elements (long/short interspersed nuclear elements) which can occur in 100k to 1M copies

"The genome is a graveyard of ancient transposons"

The origin and behavior of mutable loci in maize.
McClintock, B. (1950) *PNAS*. 36(6):344–355.
Nobel Prize in Physiology or Medicine in 1983

Discovery of the Double Helix

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).

[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).

[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

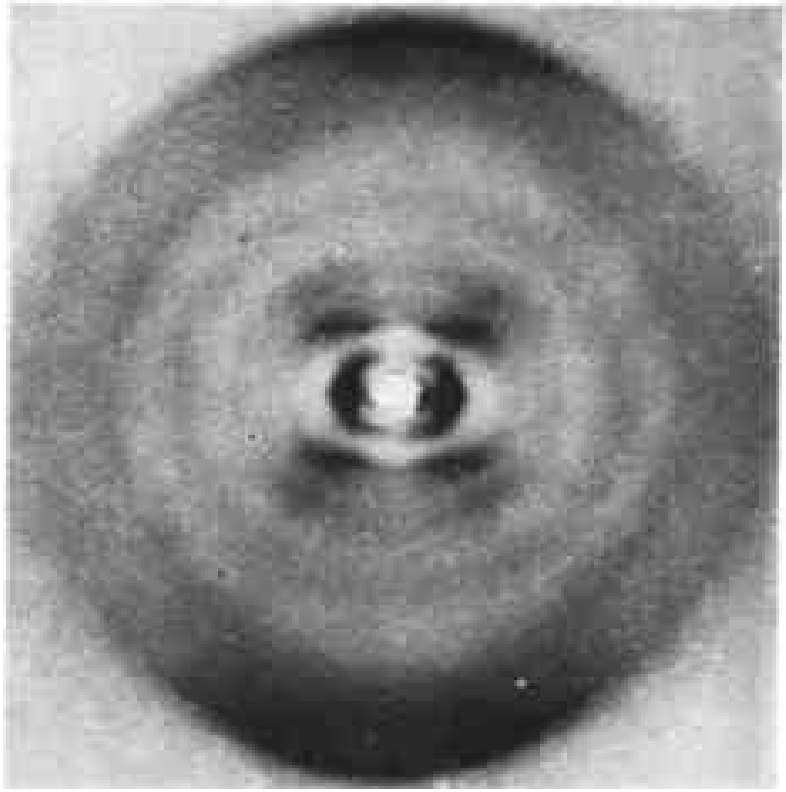
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at



It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the con-

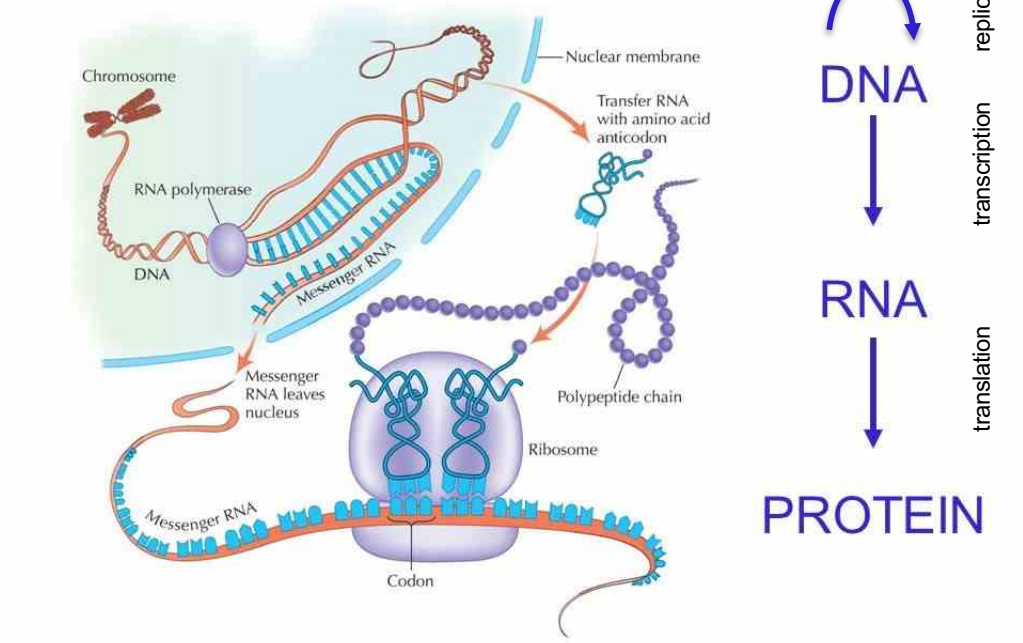
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

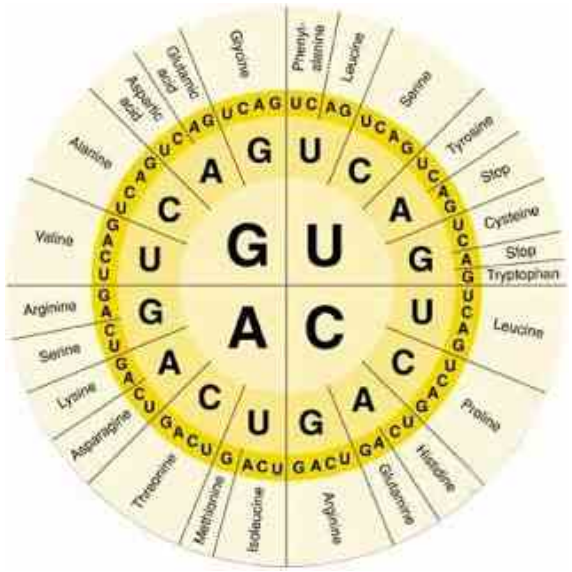
Watson JD, Crick FH (1953). *Nature* 171: 737–738.

Nobel Prize in Physiology or Medicine in 1962

Central Dogma of Molecular Biology

"Once 'information' has passed into protein it cannot get out again. In more detail, the transfer of information ***from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible***, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein"



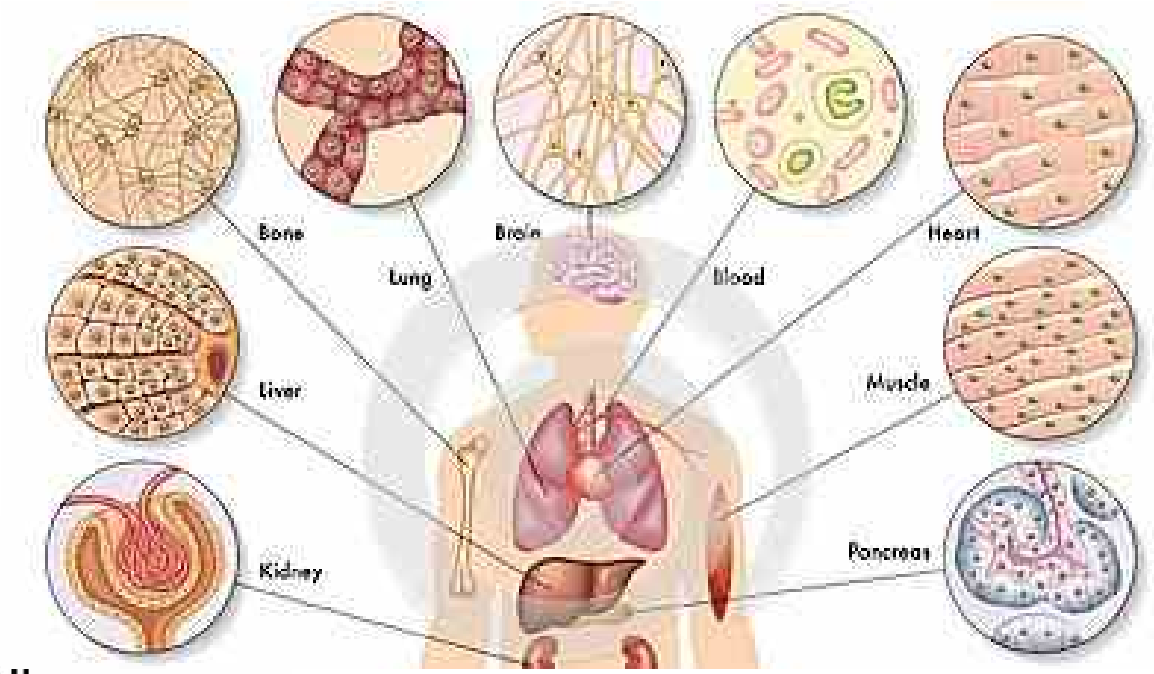


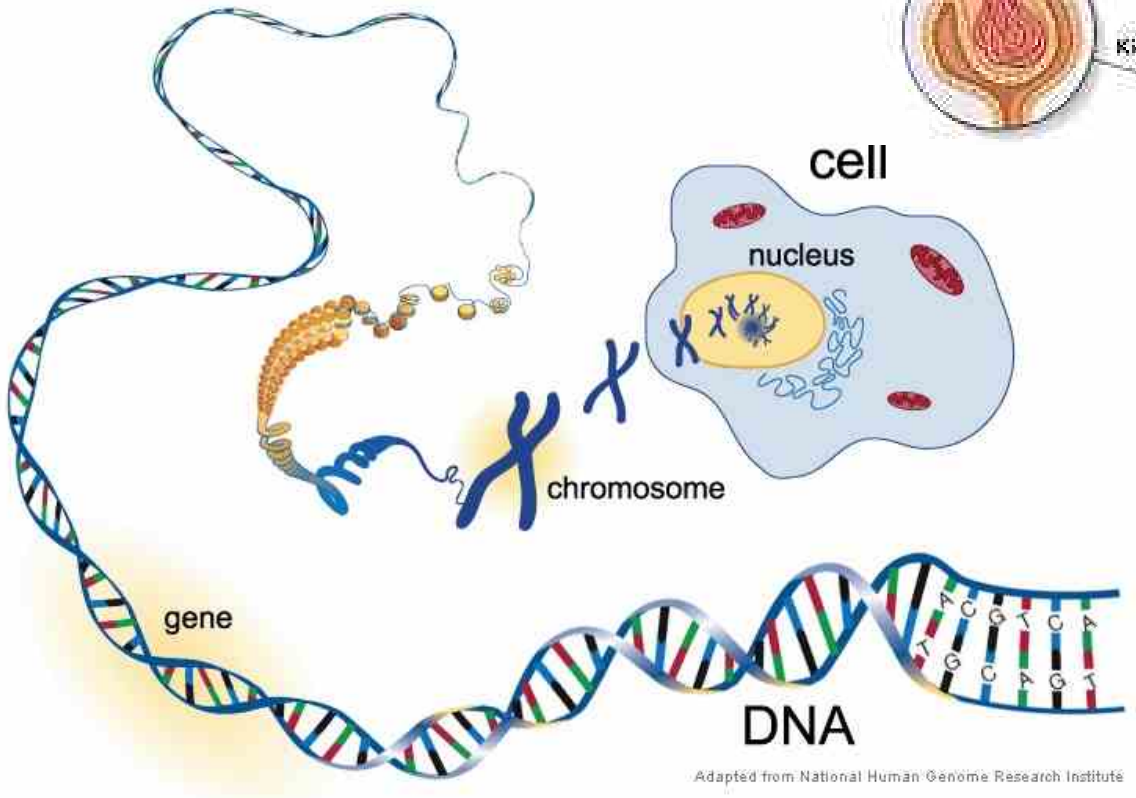
On Protein Synthesis

Crick, F.H.C. (1958). Symposia of the Society for Experimental Biology pp. 138–163.

One Genome, Many Cell Types

Each cell of your body contains an exact copy of your 3 billion base pair genome.





Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Unsolved Questions in Biology

- What is your genome sequence?
- How does your genome compare to my genome?

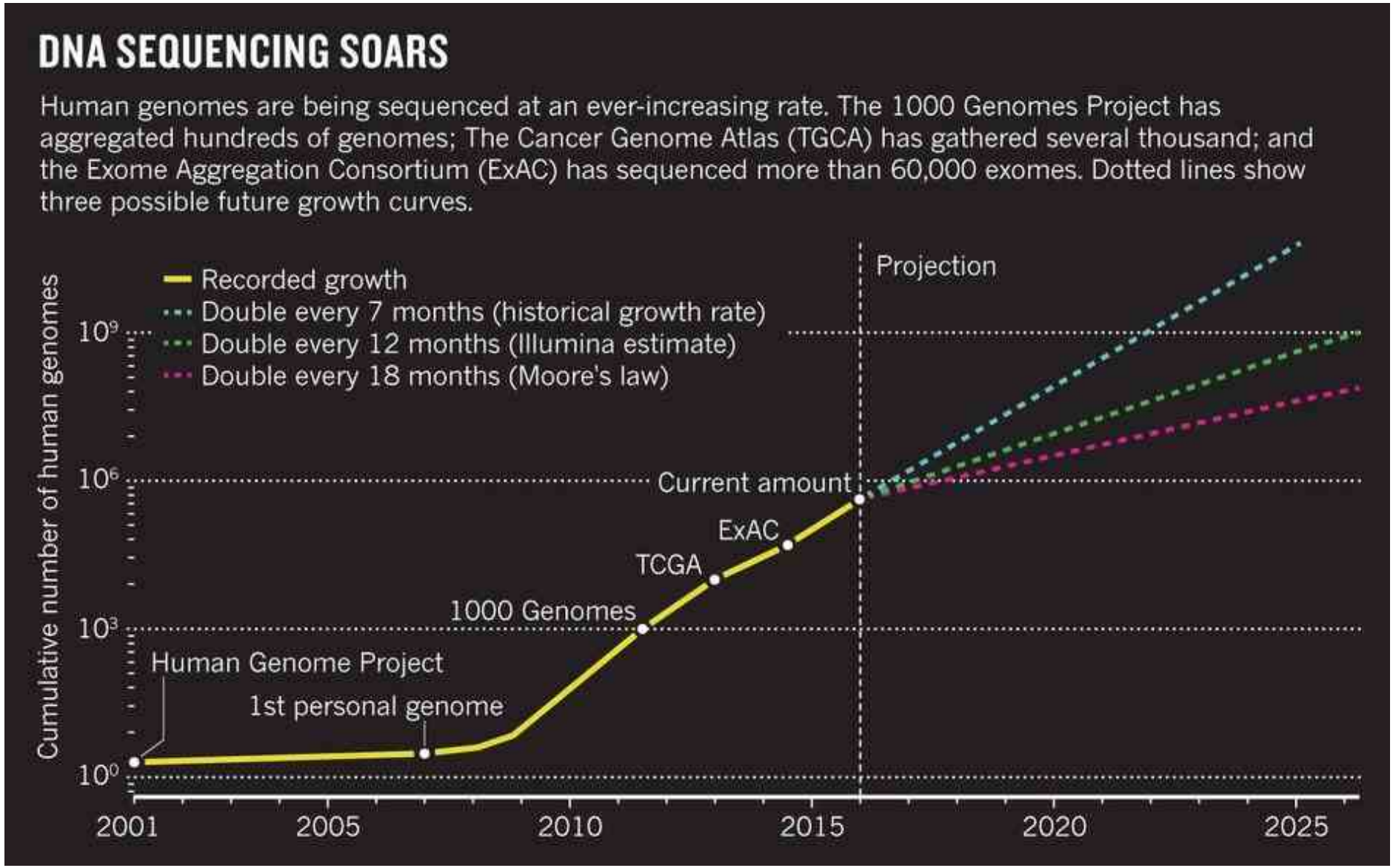
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs and treatments should we give you?

- ***Plus thousands and thousands more***

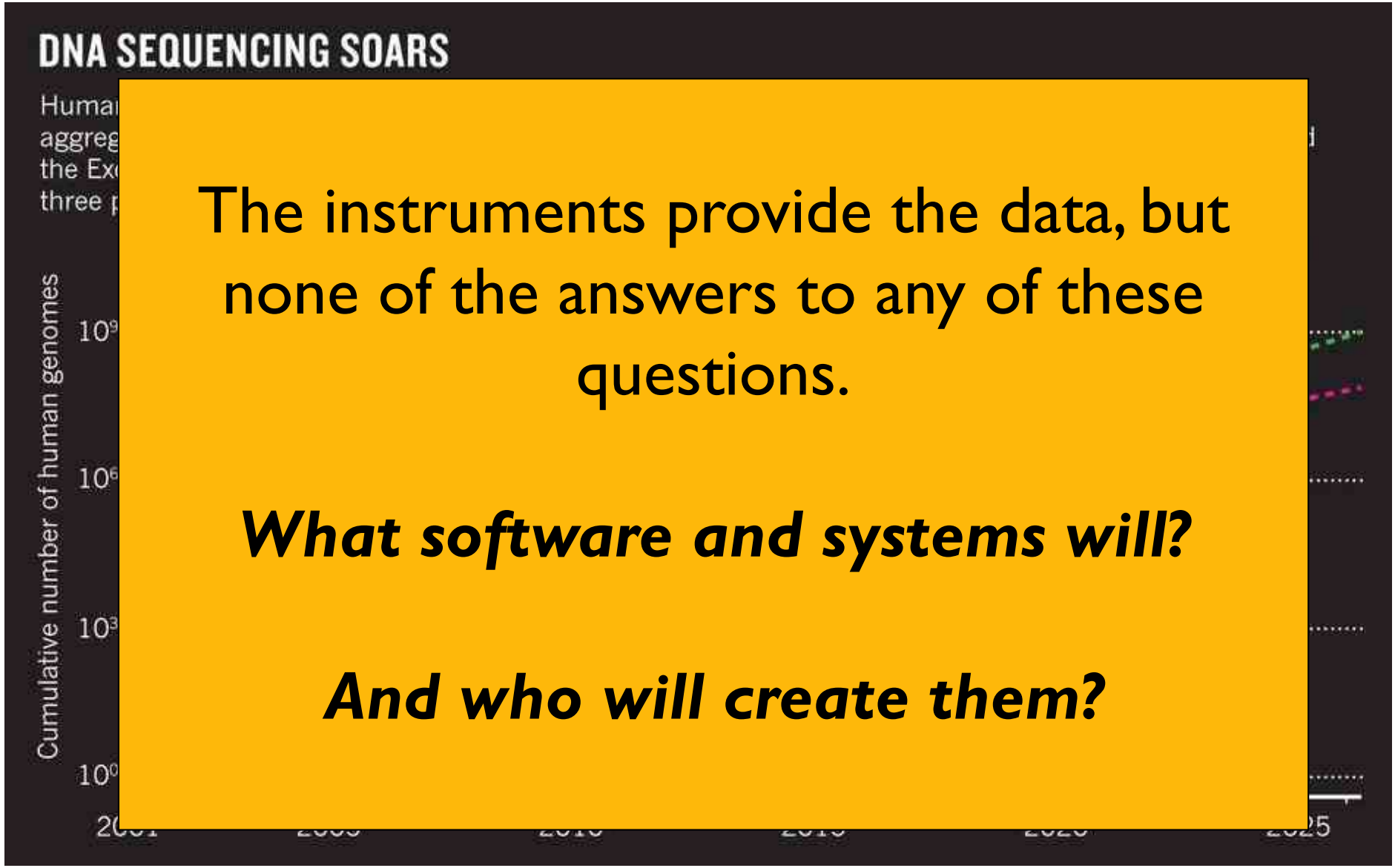
Sequencing Capacity



Big Data: Astronomical or Genomical?

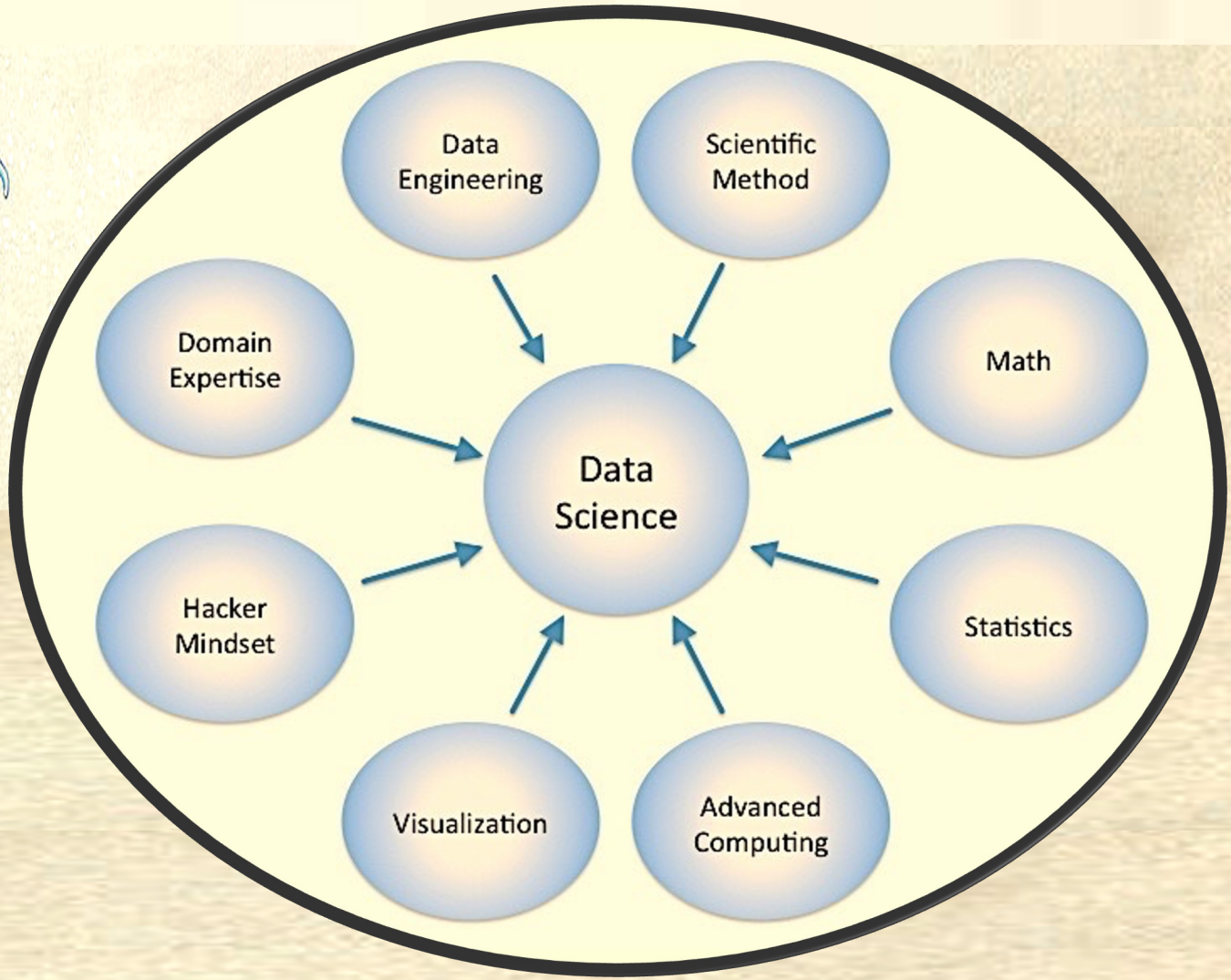
Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

Sequencing Capacity



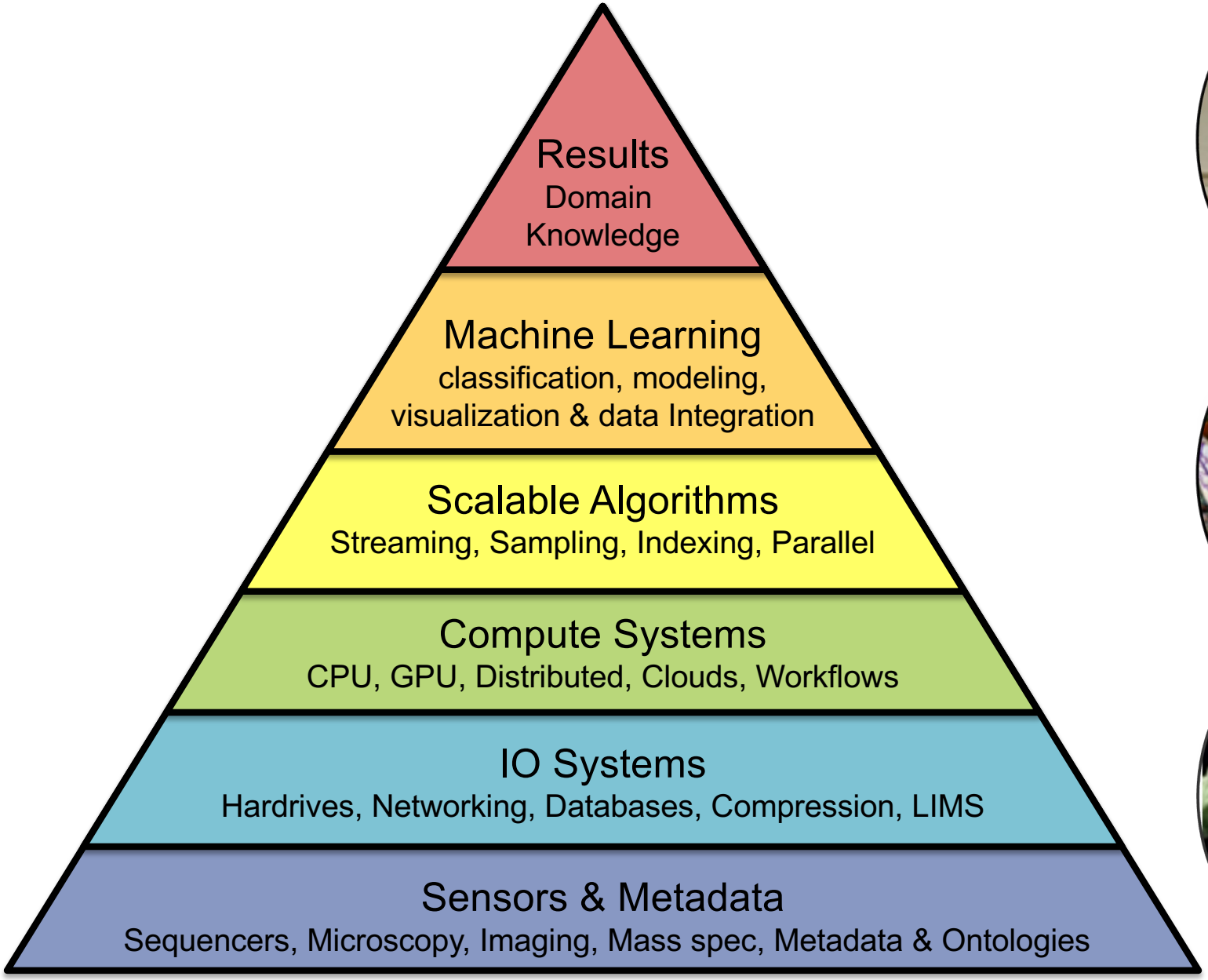
Big Data: Astronomical or Genomical?
Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

Who is a Data Scientist?

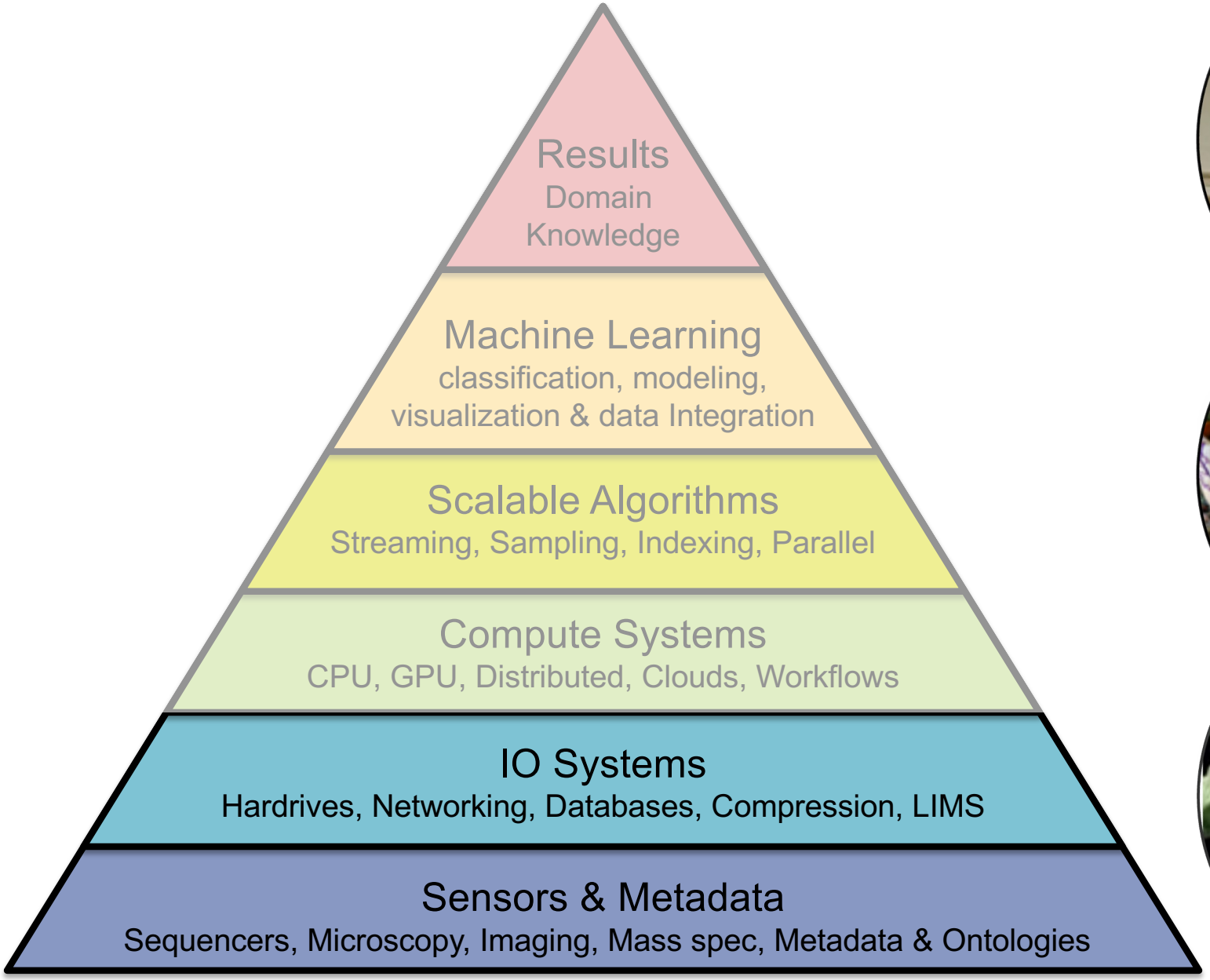


http://en.wikipedia.org/wiki/Data_science

Applied Genomics



Applied Genomics

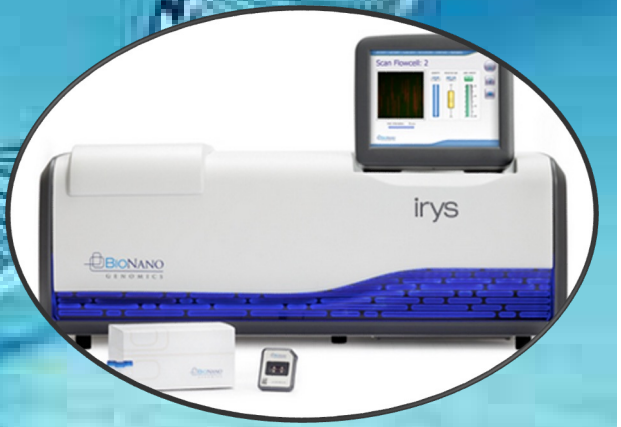


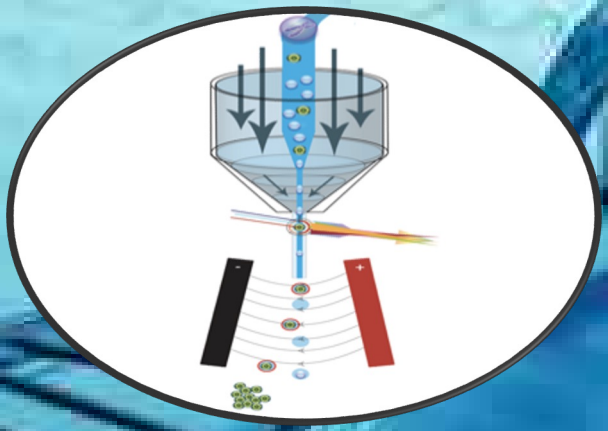
Genomics Arsenal in the year 2024

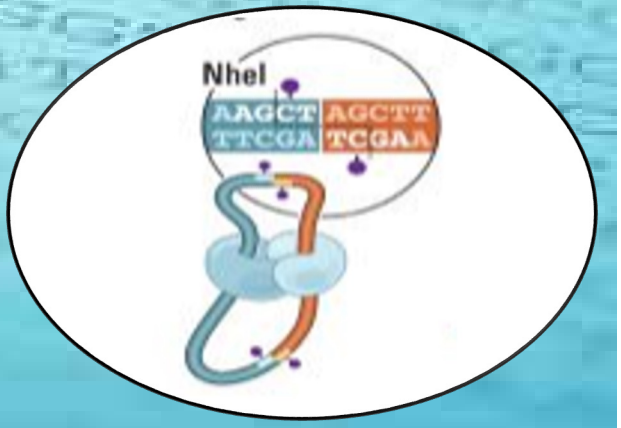
Sample Preparation

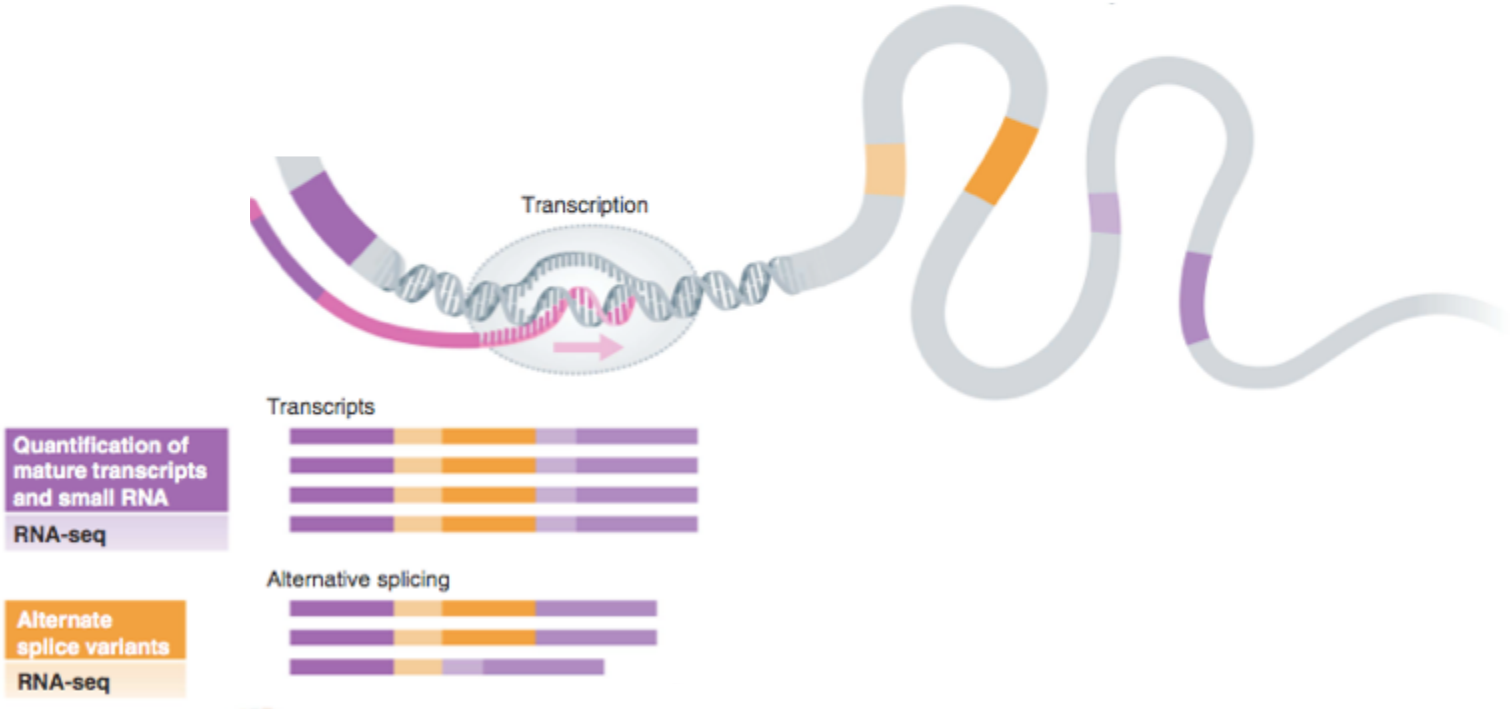
Sequencing

Chromosome Mapping

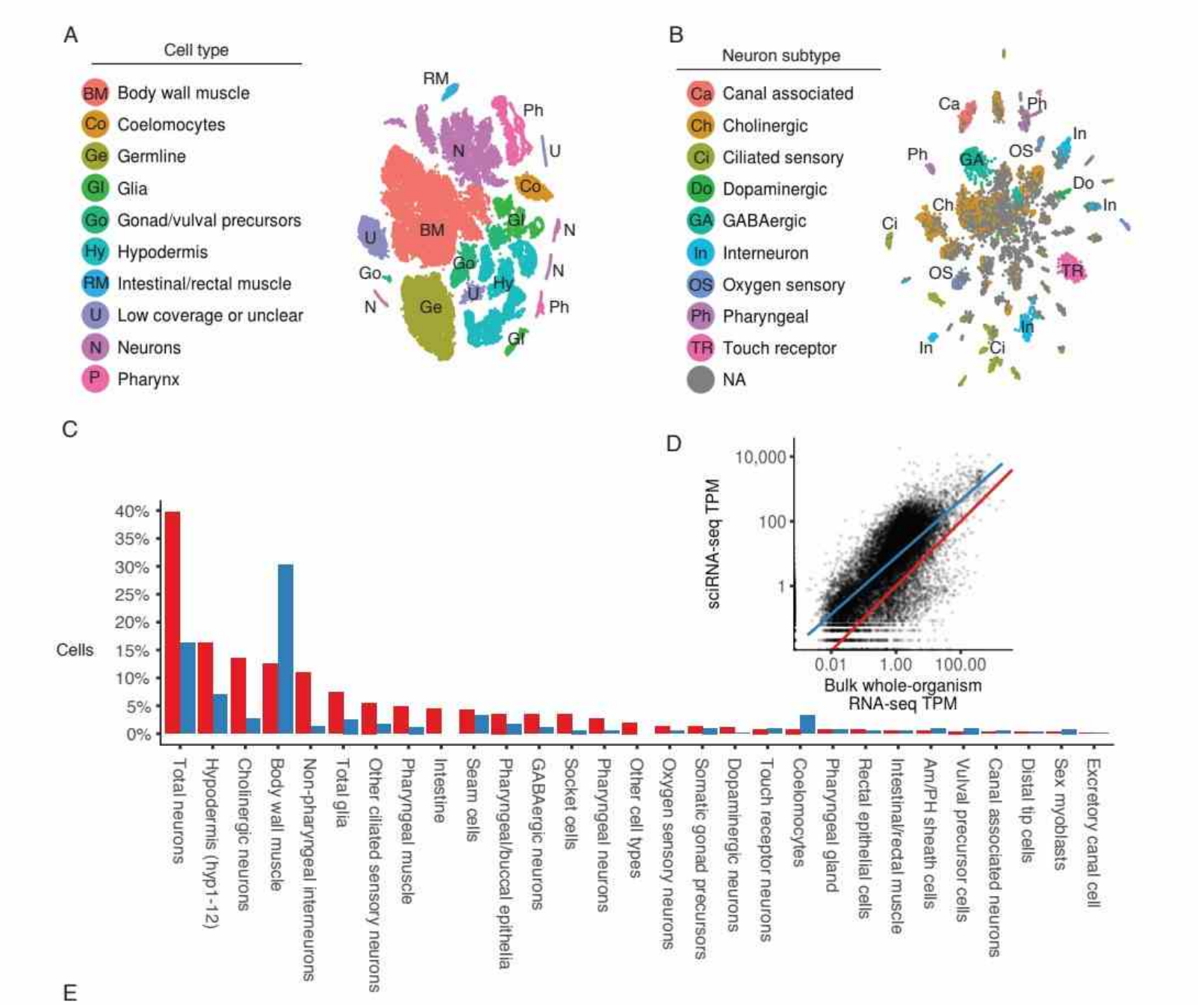






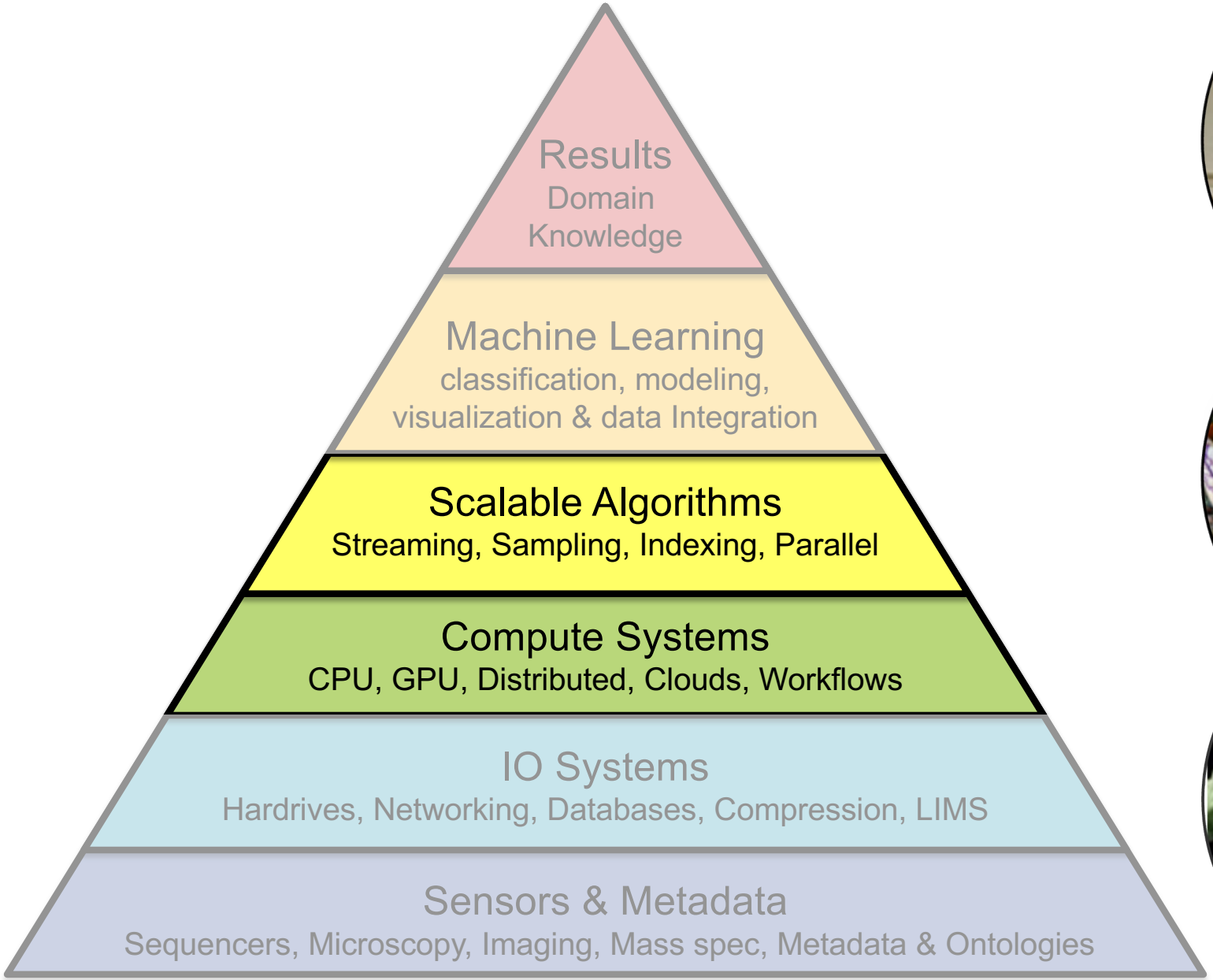


Soon et al., Molecular Systems Biology, 2013



Comprehensive single-cell transcriptional profiling of a multicellular organism
Cao, *et al. (2017) Science. doi: 10.1126/science.aam8940*

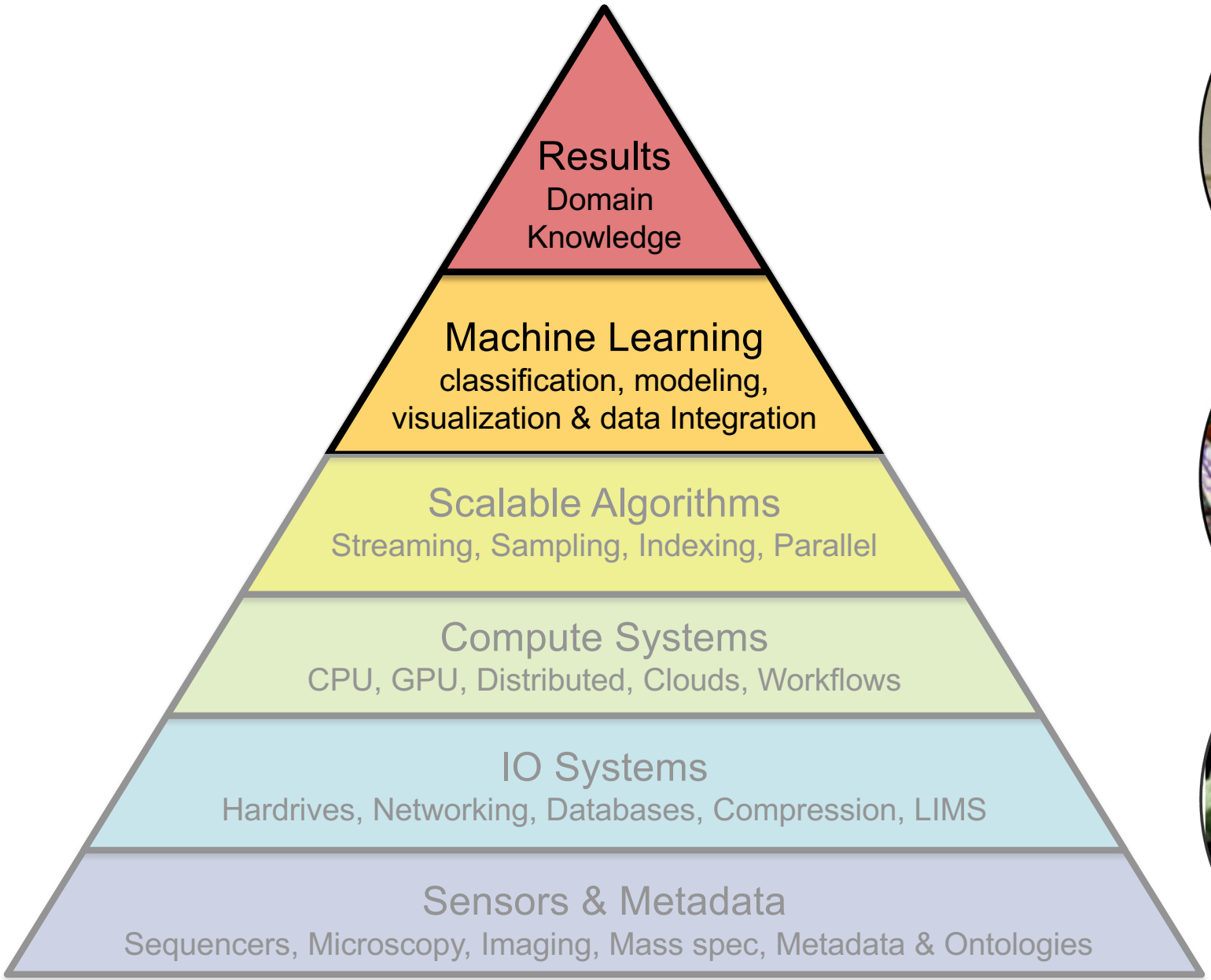
Applied Genomics



Potential Topics

- Genome assembly, whole genome alignment
- Full text indexing: Suffix Trees, Suffix Arrays, FM-index
- Dynamic Programming: Edit Distance, sequence similarity
- Read mapping & Variant identification
- Gene Finding: HMMs, Plane-sweep algorithms
- RNA-seq: mapping, assembly, quantification
- ChIP-seq: Peak finding, motif finding
- Methylation-seq: Mapping, CpG island detection
- HiC: Domain identification, scaffolding
- Chromatin state analysis: ChromHMM
- Scalable genomics: Cloud computing, scalable data structures
- Population & single cell analysis: clustering, pseudotime
- Disease analysis, cancer genomics, Metagenomics
- Deep learning in genomics

Applied Genomics



Genetic Basis of Autism Spectrum Disorders

Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?

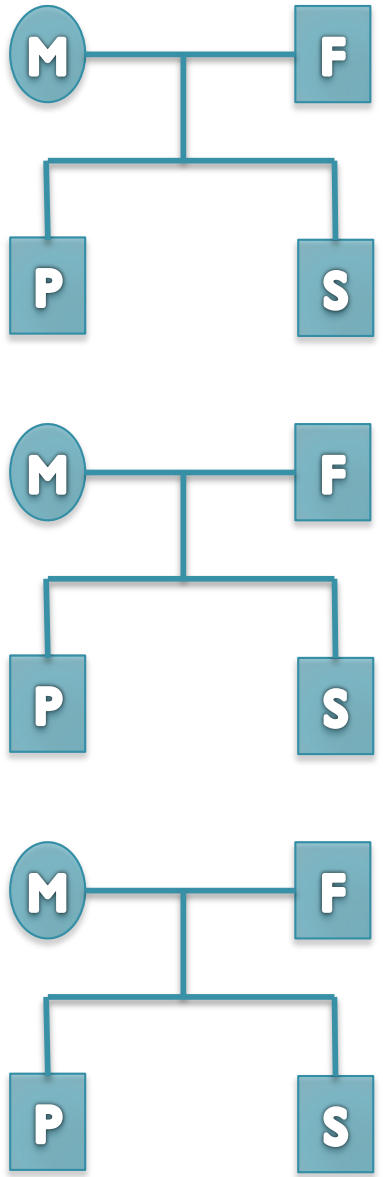
http://www.autismspeaks.org/what-autism

Searching for the genetic risk factors

Search Strategy

- Thousands of families identified from a dozen hospitals around the United States
- Large scale genome sequencing of "simplex" families: mother, father, affected child, unaffected sibling
- Unaffected siblings provide a natural control for environmental factors

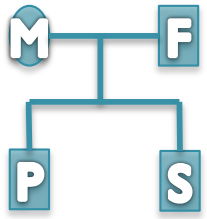
Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?



SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

De novo mutation discovery and validation

De novo mutations:
Sequences not inherited from your parents.




```
Reference:   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Father(2):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Mother(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother(2):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Sibling(1):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling(2):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Proband(1):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):  ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

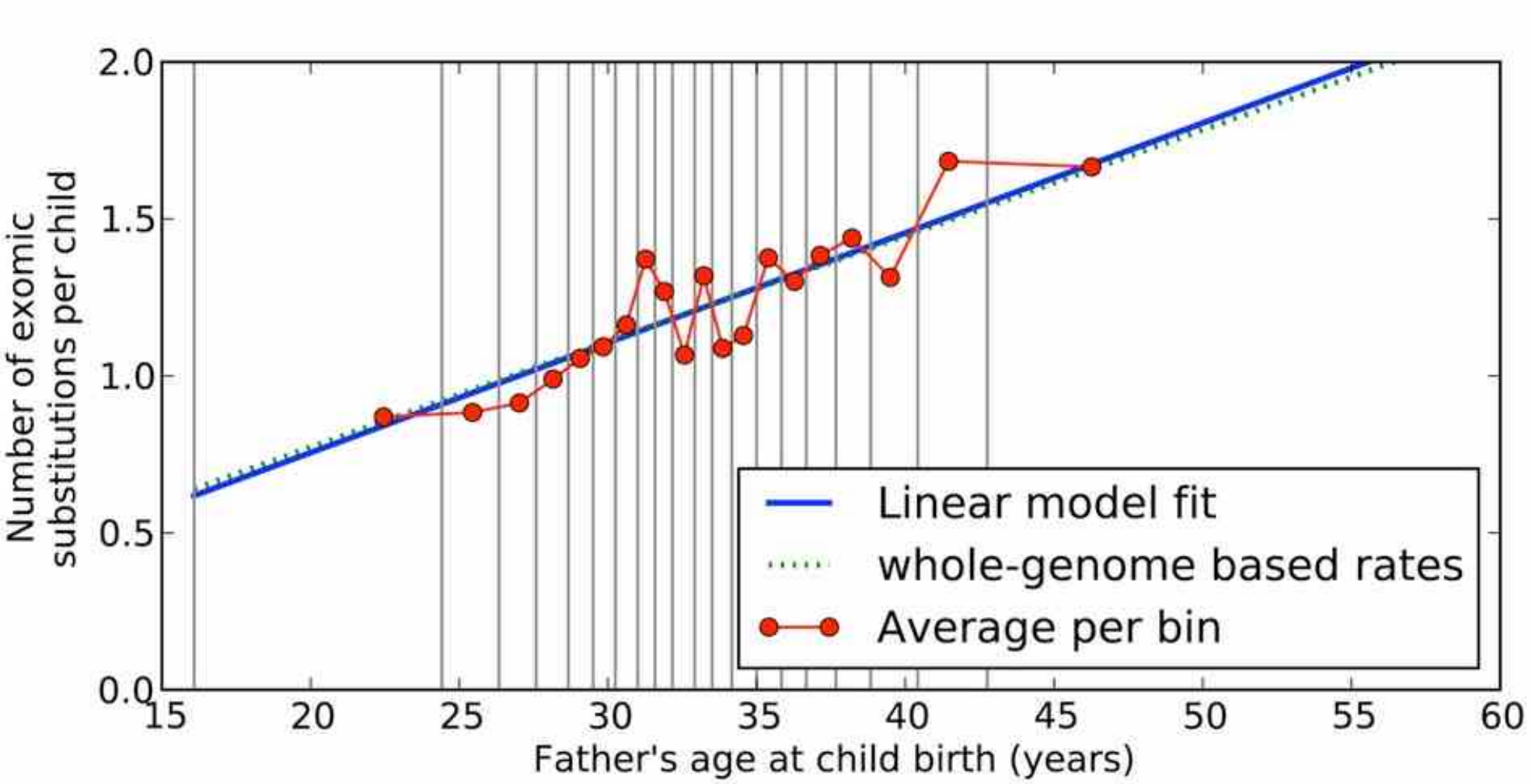
De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

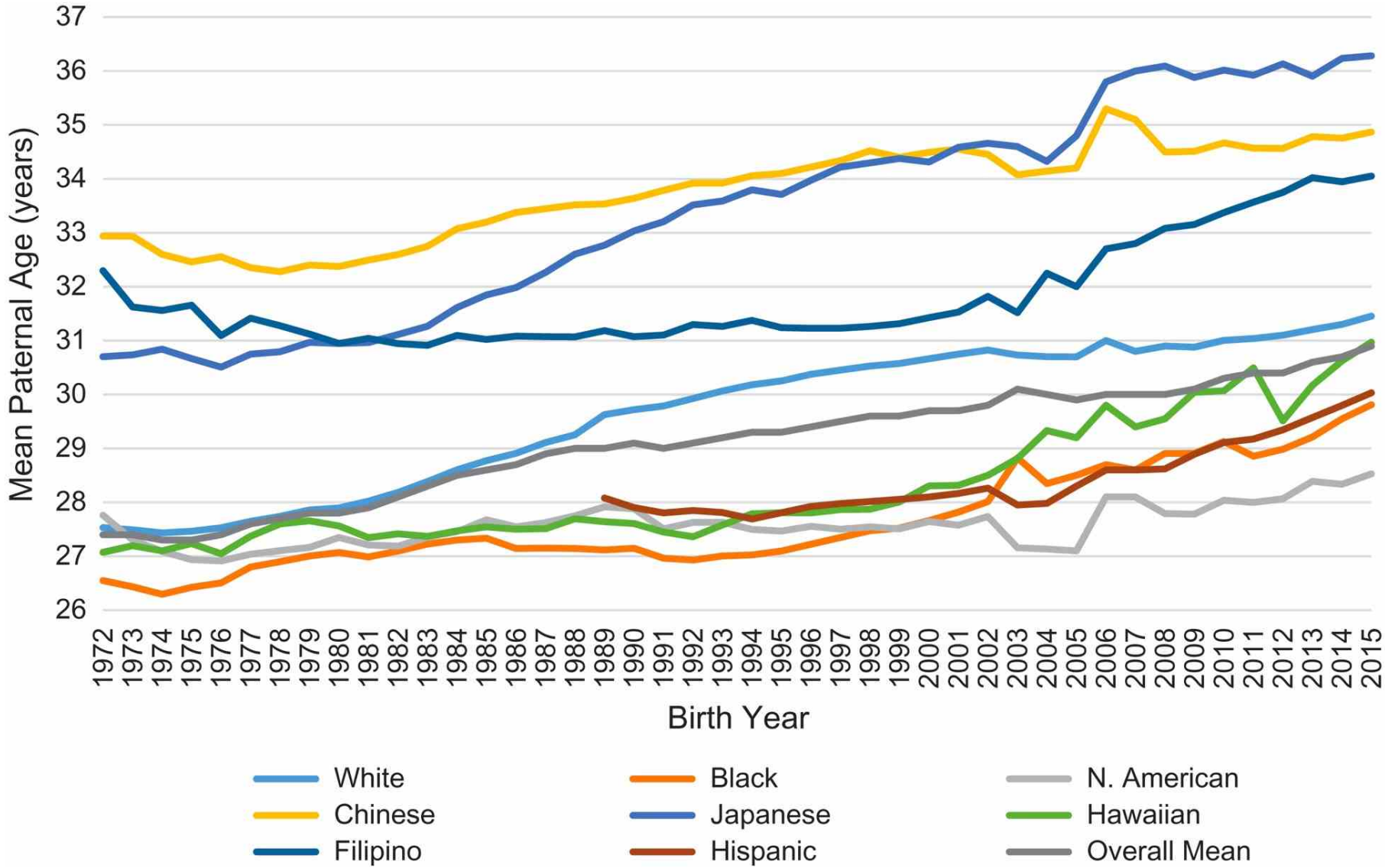
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

De novo Mutations in Men



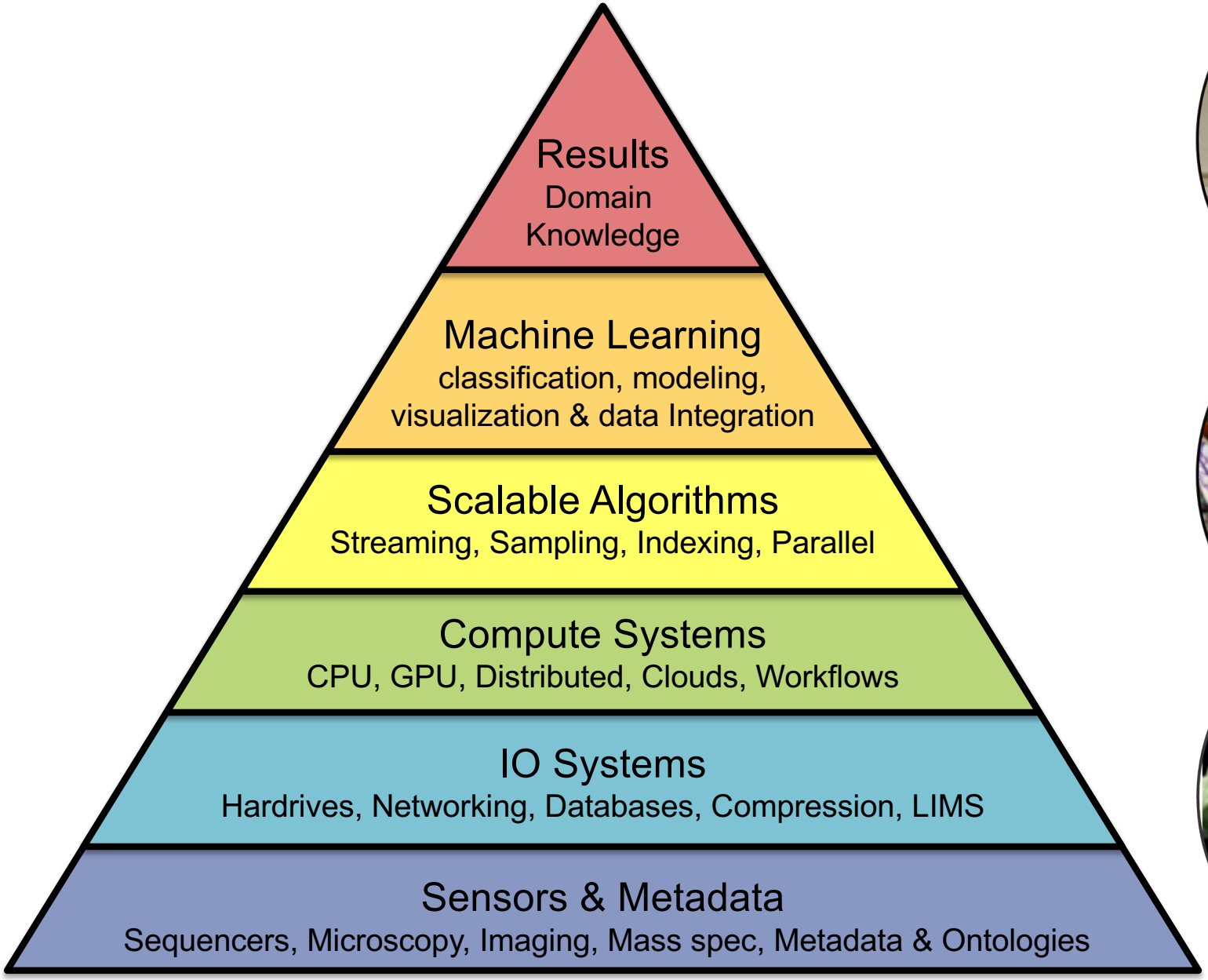
The contribution of de novo coding mutations to autism spectrum disorder
Iossifov *et al* (2014) *Nature*. doi:10.1038/nature13908

Age of Fatherhood



The age of fathers in the USA is rising: an analysis of 168 867 480 births from 1972 to 2015
Khandwala *et al* (2017) *Human Reproduction. https://doi.org/10.1093/humrep/dex267*

Applied Genomics



Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Register on Piazza
4. Get Ready for assignment 1
    1. Set up conda, set up Docker
    2. Set up Dropbox for yourself!
    3. Get comfortable on the command line