

Applied Comparative Genomics

Michael Schatz

August 25, 2025

Lecture I: 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/appliedgenomics2025>

Course Discussions: <https://piazza.com/class/meogfdbmu7x7hf>

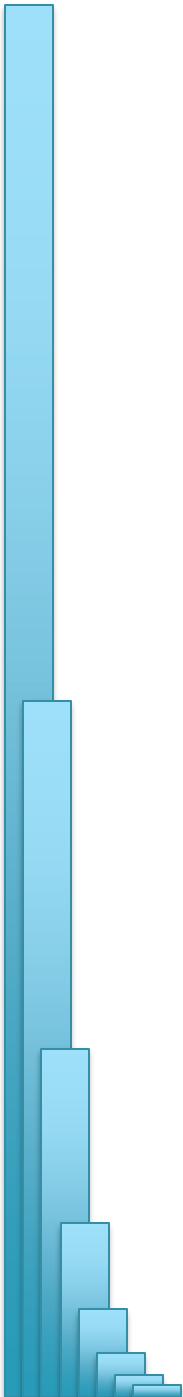
Class Hours: Mon + Wed @ 3:00p – 4:15p, Hodson 316

Schatz Office Hours: TBD and by appointment

TA Office Hours: TBD and by appointment

Please try Piazza first!

TA: Mahler Revsine



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:
<https://www.langmead-lab.org/teaching.html>

Grading Policies

Assessments:

- 5 Assignments: 25% Due at 11:59pm a week later
Practice using the tools and methods we are discussing
- 2 Exams: 40% In Class (Oct 8 & Nov 12)
Assess your performance, focusing on the concepts
- 1 Class Project: 35% Written & Oral presentations
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 (24×4) 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

what's 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

The screenshot shows a GitHub repository page for 'appliedgenomics2025'. The repository is public and contains 1 branch and 0 tags. The main file listed is 'README.md'. The repository description is: 'Materials for EN.601.449/649 Computational Genomics: Applied Comparative Genomics'. It includes links to 'Readme', 'CC0-1.0 license', 'Activity', 'Custom properties', '1 star', '0 watching', and '0 forks'. There are no releases or packages published.

Code | Pull requests | Actions | Security | Insights

main · 1 Branch · 0 Tags

mschatz add syllabus · 16079b8 · 6 minutes ago · 10 Commits

policies · add syllabus · 6 minutes ago

LICENSE · Initial commit · 2 hours ago

README.md · fix text · 44 minutes ago

README · CC0-1.0 license

JHU EN.601.449/EN.601.649: Computational Genomics: Applied Comparative Genomics

Prof: [Michael Schatz](#) (mschatz @ cs.jhu.edu)
TA: [Mahler Revsine](#) (mrevsin1 @ jh.edu)
Class Hours: Monday + Wednesday @ 3:00p - 4:15p Hodson 316
Schatz Office Hours: By appointment
Revsine Office Hours: TBD and by appointment

The primary goal of the course is for students to be grounded in the fundamental theory and applications to 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 genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics. A major focus will be on deep

About

Materials for EN.601.449/649
Computational Genomics: Applied
Comparative Genomics

Readme
CC0-1.0 license
Activity
Custom properties
1 star
0 watching
0 forks
Report repository

Releases

No releases published

Packages

No packages published

<https://github.com/schatzlab/appliedgenomics2025>

Course Webpage

The screenshot shows a GitHub repository page for 'appliedgenomics2025'. The repository is public and contains 1 branch and 0 tags. The main file listed is 'README.md'. The repository description is: 'Materials for EN.601.449/649 Computational Genomics: Applied Comparative Genomics'. It includes links to 'Readme', 'CC0-1.0 license', 'Activity', 'Custom properties', '1 star', '0 watching', and '0 forks'. There are no releases or packages published.

Code

mschatz add syllabus

policies add syllabus

LICENSE Initial commit

README.md fix text

Go to file

Code

About

Materials for EN.601.449/649
Computational Genomics: Applied
Comparative Genomics

Readme

CC0-1.0 license

Activity

Custom properties

1 star

0 watching

0 forks

Report repository

Releases

No releases published

Packages

No packages published

JHU EN.601.449/EN.601.649: Computational Genomics:
Applied Comparative Genomics

Prof: Michael Schatz (mschatz @ cs.jhu.edu)
TA: Mahler Revsine (mrevisn1 @ jh.edu)
Class Hours: Monday + Wednesday @ 3:00p - 4:15p Hodson 316
Schatz Office Hours: By appointment
Revsine Office Hours: TBD and by appointment

The primary goal of the course is for students to be grounded in the fundamental theory and applications to 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 genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics. A major focus will be on deep

<https://github.com/schatzlab/appliedgenomics2025>

Piazza

The screenshot shows a web browser window for the Piazza platform. The URL in the address bar is piazza.com/class/meogfdbmu7x7hf/post/6. The top navigation bar includes links for Q & A, Resources, Statistics, and Manage Class, along with options to Buy a License, Switch to contribution model, and a user profile for Michael Schatz.

The main content area displays a post titled "Welcome to Piazza!" by Michael Schatz. The post was created at 12:12 PM and has 1 view. It contains the following text:

Hi Students, Welcome to Piazza! We'll be conducting all class-related discussion here this term. The quicker you begin as

Below the post, there are interaction icons for Edit, Like (0), Share, and Report, followed by a link to "Followup Discussions".

<https://piazza.com/jhu/fall2025/600449600649>

GradeScope

The screenshot shows the GradeScope course dashboard for EN.601.449/EN.601.649 (Fall 2025). The left sidebar includes links for Dashboard, Assignments, Roster, Extensions, Course Settings, Instructor (Michael Schatz), and Course Actions (Unenroll From Course). The main area displays course details, a "Things To Do" section with links to Roster and Assignments, and a summary of active assignments. A message states "You currently have no assignments." with a "Create Assignment" button. A question mark icon is in the bottom right corner.

www.gradescope.com/courses/1097756

Entry Code: **GVXGV2**

EN.601.449/EN.601.649 | Fall 2025

Course ID: 1097756

Description

EN.601.449/649 Computational Genomics:
Applied Comparative Genomics

Things To Do

- Add students or staff to your course from the [Roster](#) page.
- Create your first assignment from the [Assignments](#) page.

Active Assignments Released Due (EDT) Submissions % Graded Published Regrades

You currently have no assignments.

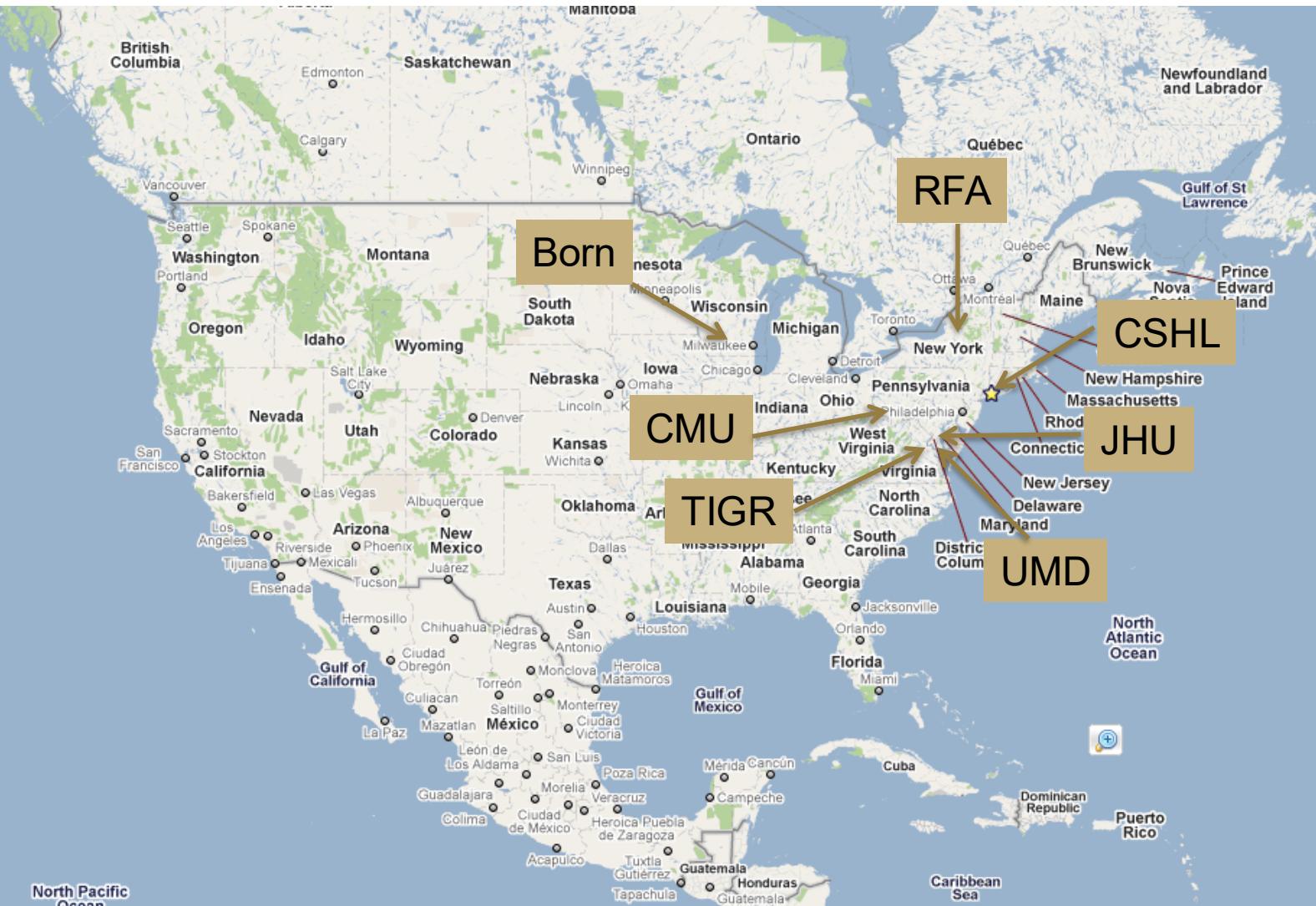
Create an assignment to get started.

Create Assignment

?

<https://www.gradescope.com/>
Entry Code: **GVXGV2**

A Little About Me



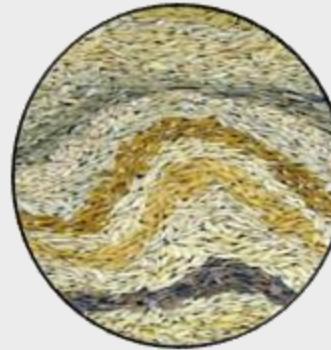
Schatzlab Overview



Human Genetics

Role of mutations
in disease

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



Agricultural Genomics

Genomes &
Transcriptomes

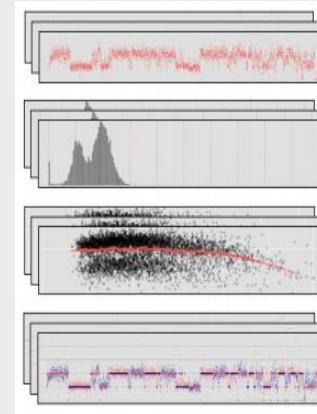
Benoit *et al.* (2025)
Satterlee *et al.* (2024)



Algorithmics & Systems Research

Ultra-large scale
biocomputing

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

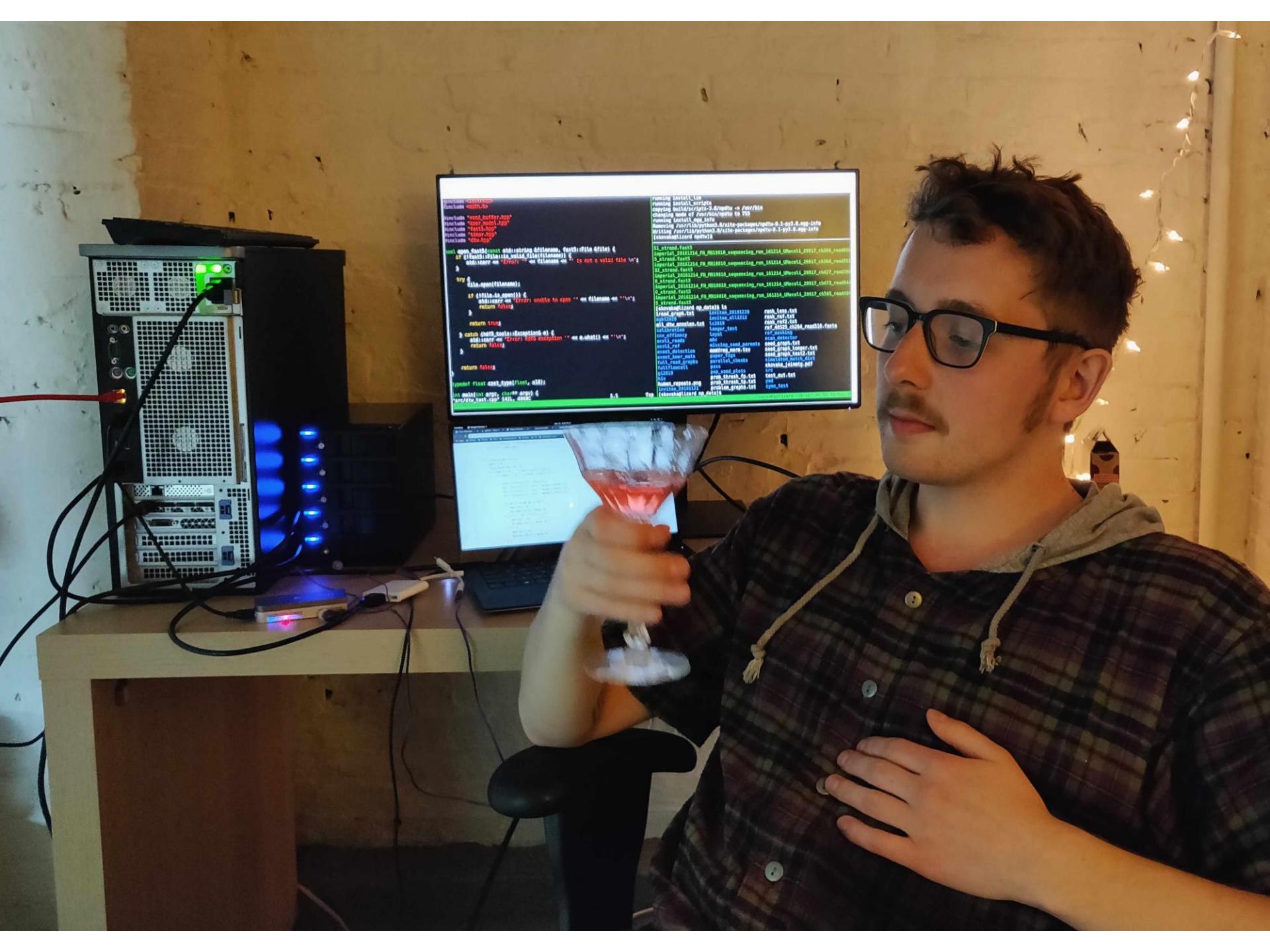


Biotechnology Development

Single Cell + Single
Molecule Sequencing

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

					<hr/>
25	11/19/25	W	In-class presentation		
*	11/24/25	M	Thanksgiving Break		
*	11/26/25	W	Thanksgiving Break		
26	12/1/25	M	In-class presentation		
27	12/3/25	W	In-class presentation		
*	12/10/25	W	Draft Report Due		
*	12/11/25	Th	Final project presentation		
*	12/12/25	F	Final project presentation		
*	12/15/25	M	Final project presentation		
*	12/16/25	Tu	Final Report Due		



Targeted nanopore sequencing × +

nature.com/articles/s41587-020-0731-9

M 24 JHUMail Daily s j P 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 ReadUtil 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?

Unsolved Questions in Biology

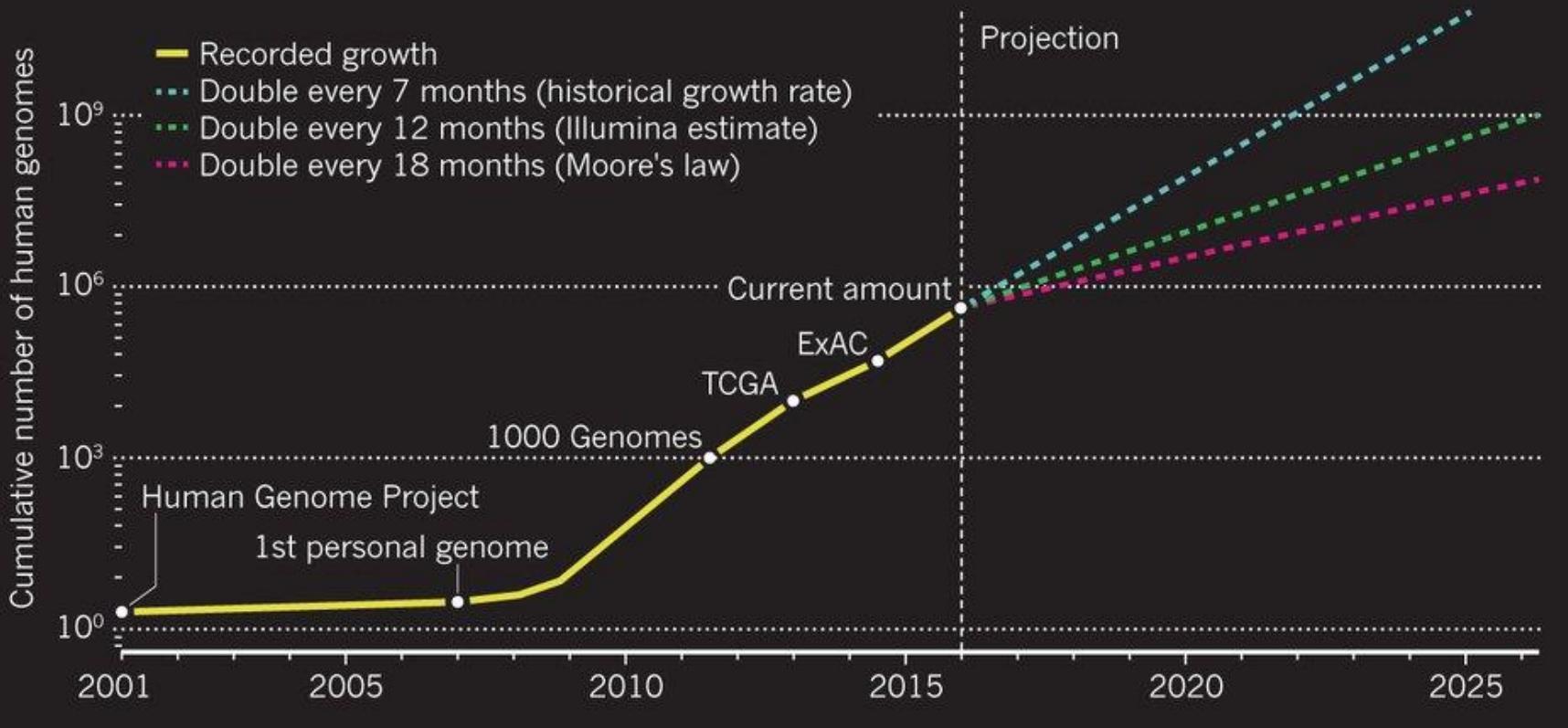
- What is your genome sequence?
- How does your genome compare to my genome?
- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?
- How does methylation change during development?
- How does chromatin change during development?
- How does 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

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TCGA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



Big Data: Astronomical or Genomical?

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

Sequencing Capacity

DNA SEQUENCING SOARS

Human
aggreg
the Ex
three p

The instruments provide the data, but
none of the answers to any of these
questions.

Cumulative number of human genomes

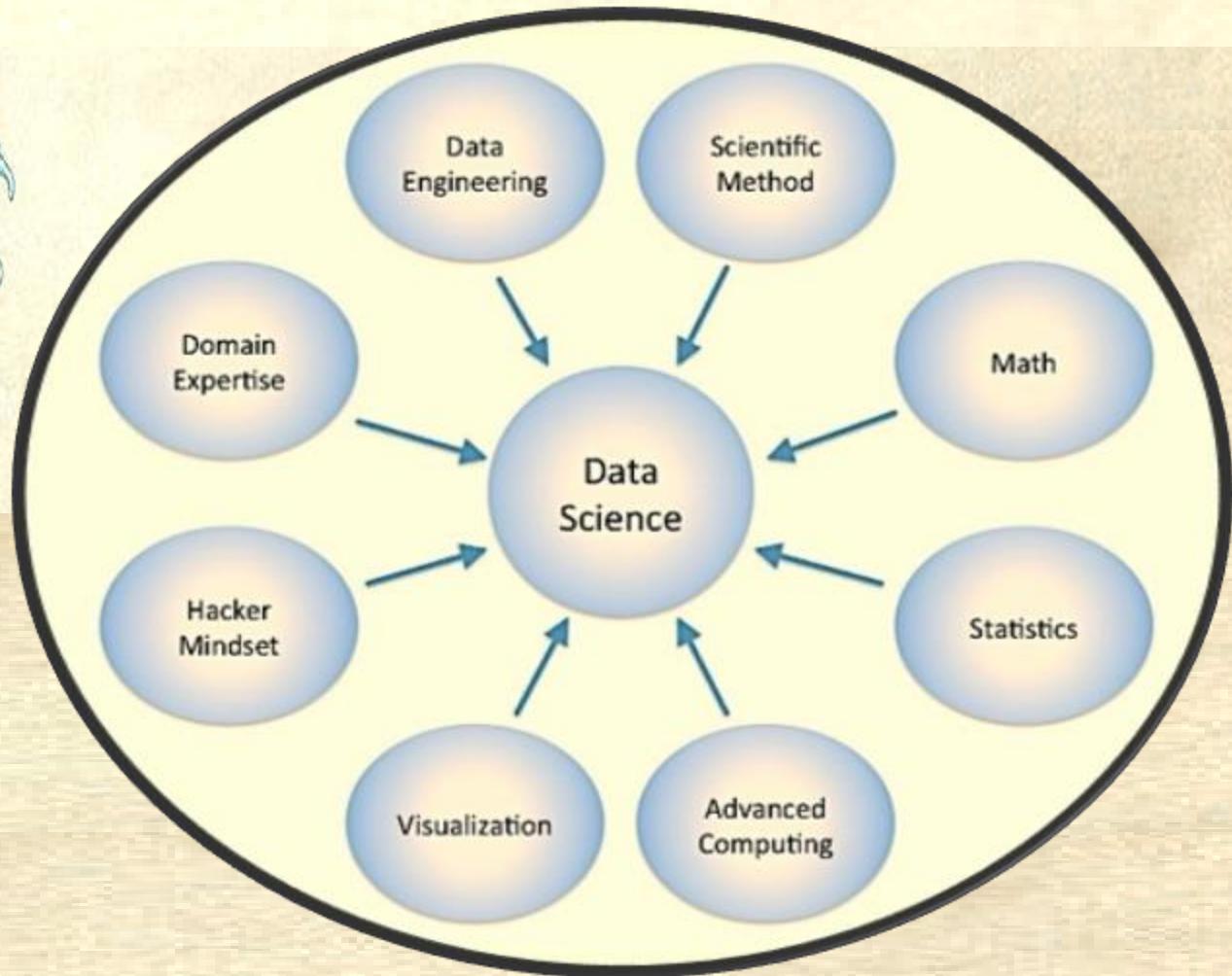
Year	Approximate Cumulative Number of Genomes (10^9)
2001	~10^0
2005	~10^1
2010	~10^2
2015	~10^3
2020	~10^4
2025	~10^5

What software and systems will?

And who will create them?

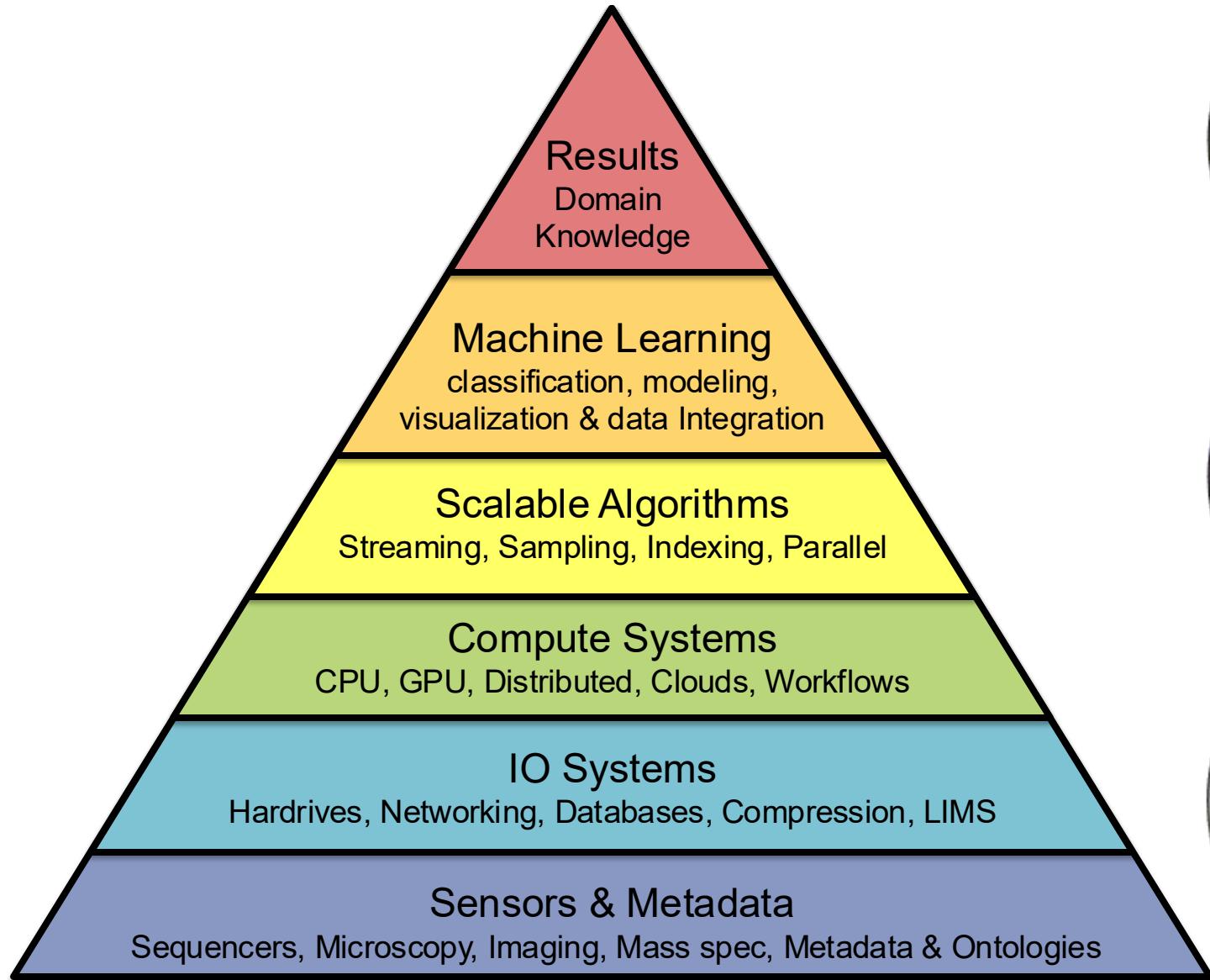


Who is a Data Scientist?



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

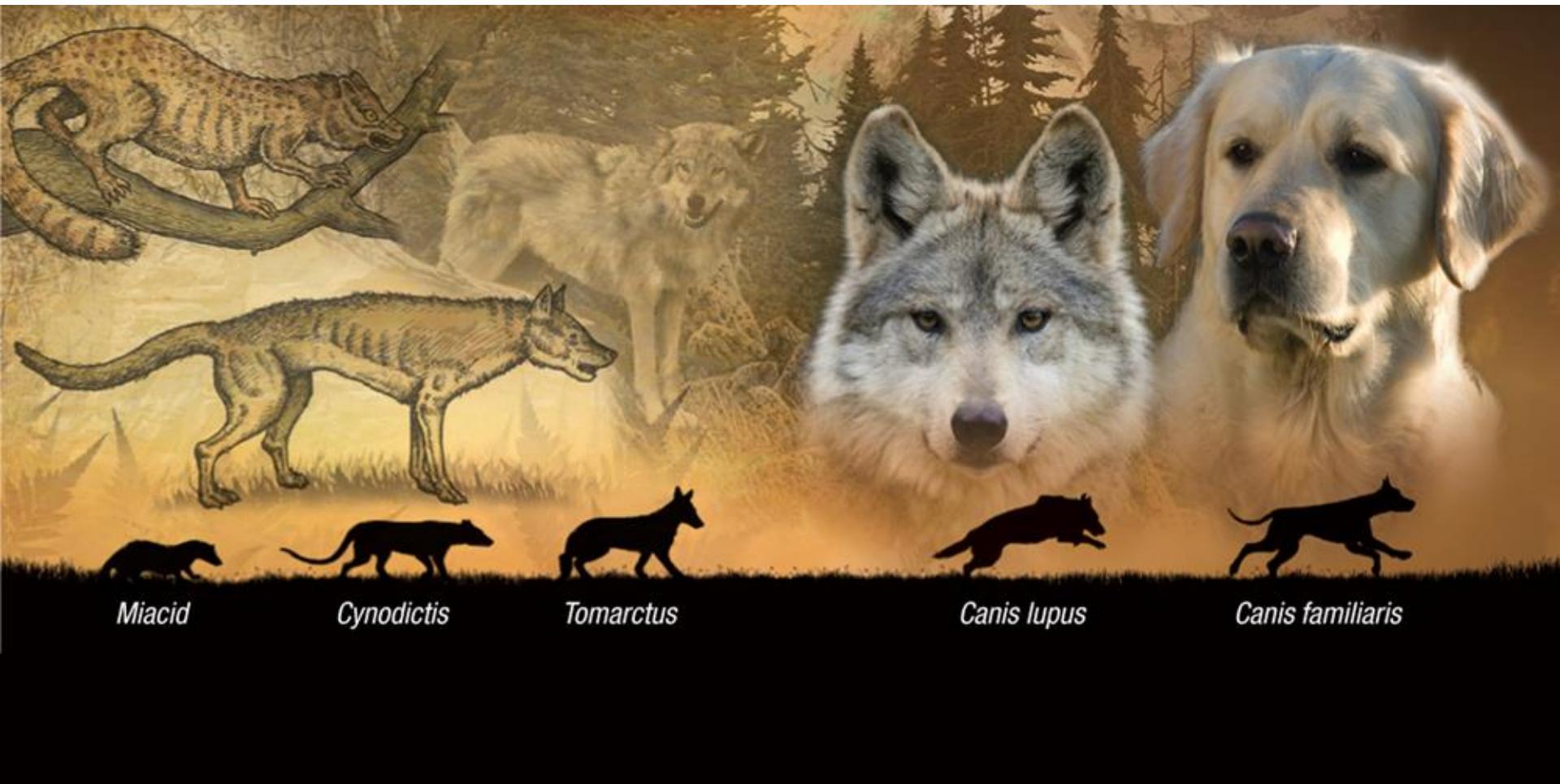
Applied Genomics



Earliest Genomics

Any Guesses?

Earliest Genomics



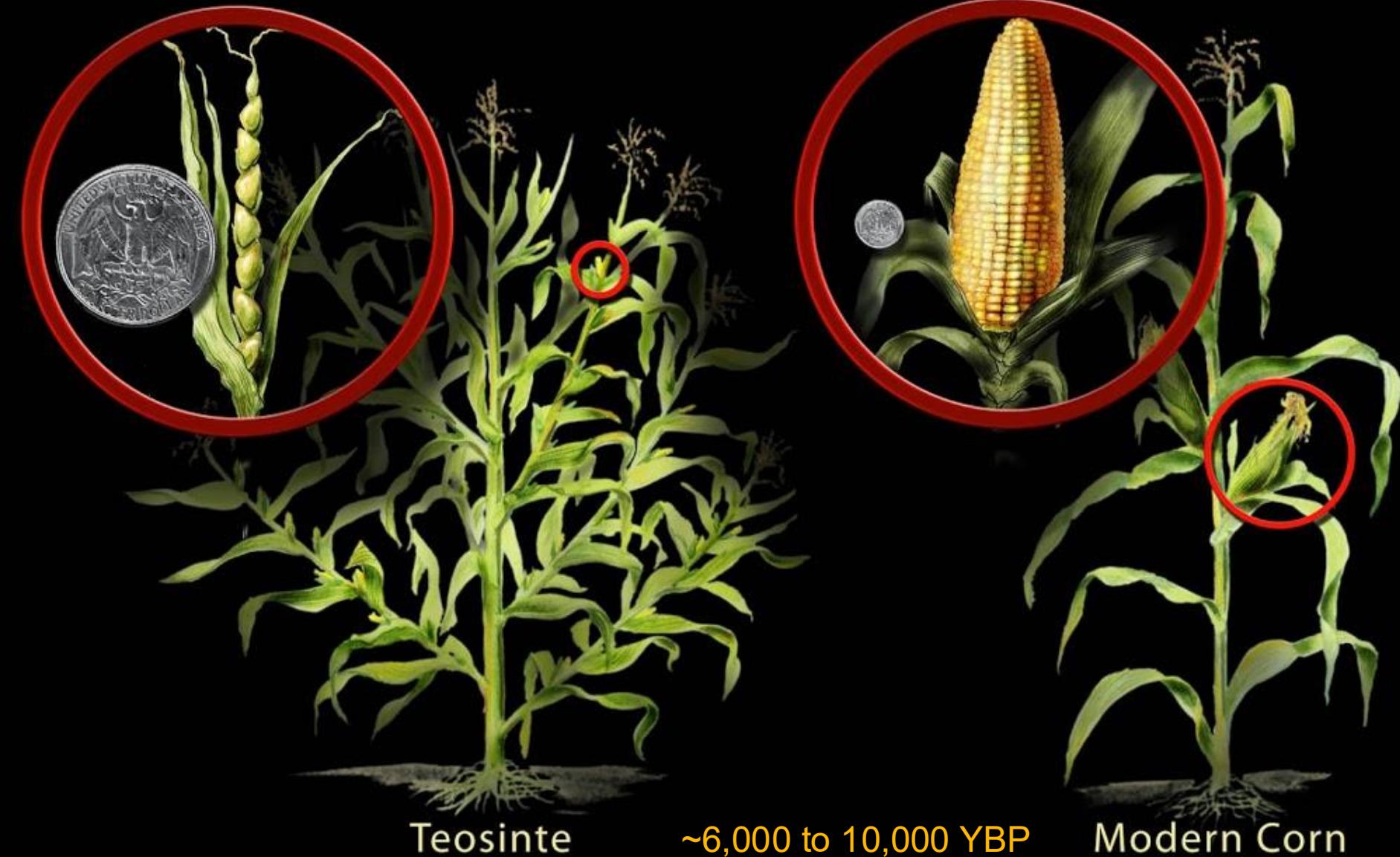
15,000 to 35,000 YBP

Earliest Genomics



~1,000 to 10,000 YBP

Earliest Genomics



Discovery of Chromosomes

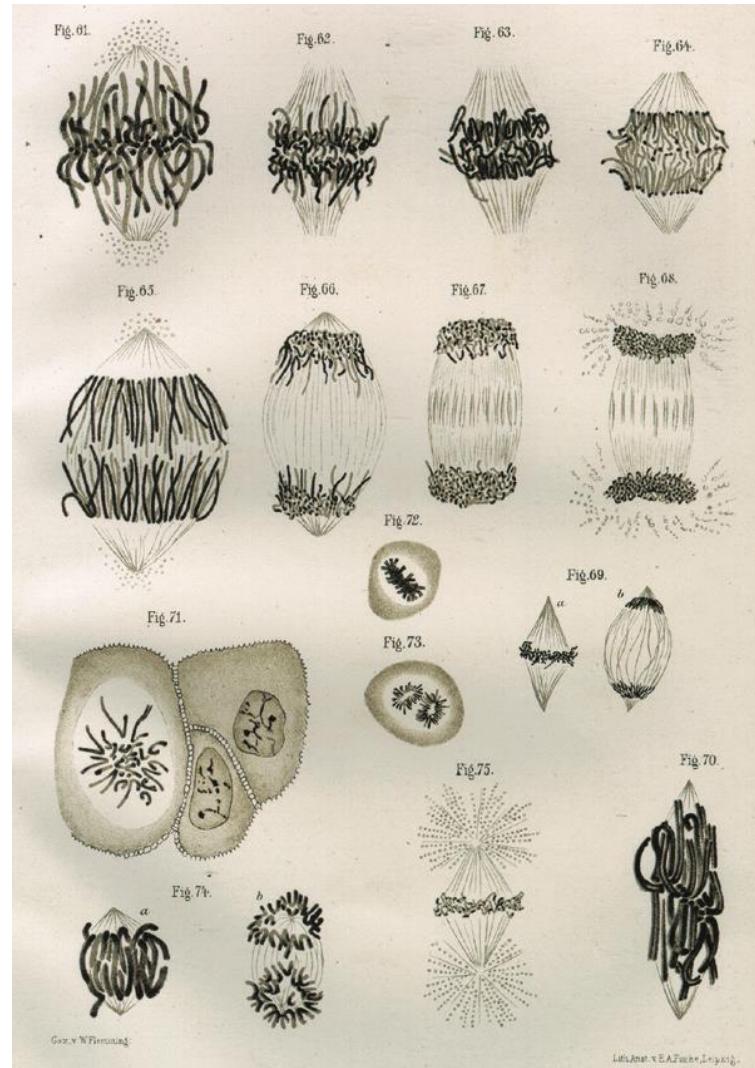
By the mid-1800s, microscopes were powerful enough to observe the presence of unusual structures called “chromosomes” that seemed to play an important role during cell division.

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



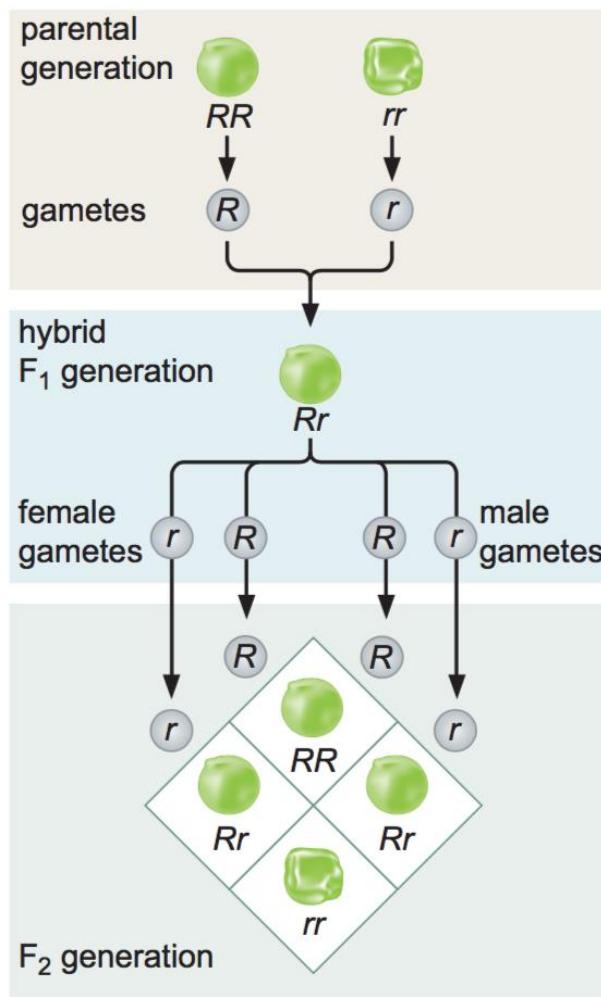
Drawing of mitosis by Walther Flemming.

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 (1-1ft)
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	in Verhältniss gestellt:			
	A	Aa	a	$A : Aa : a$
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).



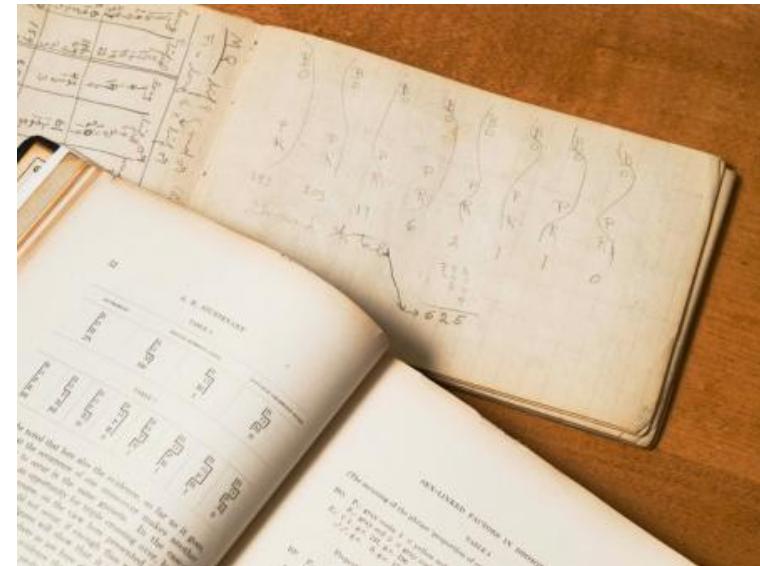
Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

The first genetic map

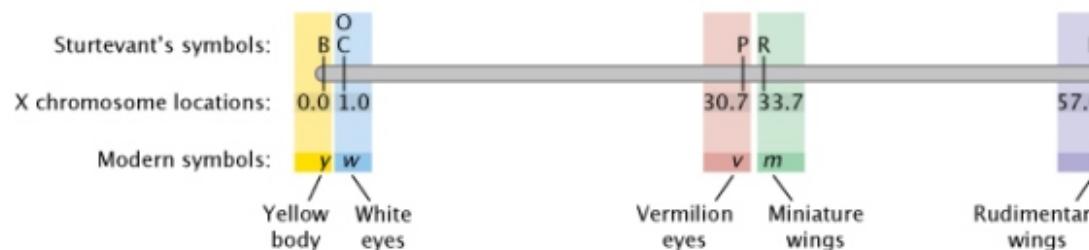
Mendel's Second Law (The Law of Independent Assortment) states alleles of one gene sort into gametes independently of the alleles of another gene: ***Pr(smooth/wrinkle) is independent of Pr(yellow/green)***

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 located closest together



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



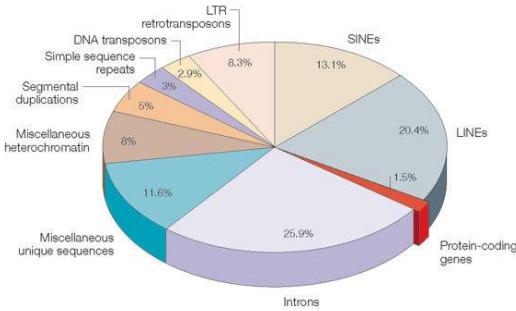
The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) *Journal of Experimental Zoology*, 14: 43-59

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

NO. 4356 April 25, 1953

NATURE

737

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., Gerard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1921).

² Longstaff-Higgin, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Suppl.* **S**, 285 (1949).

³ Von Arx, W. S., Woods Hole Papers in Phys. Oceanogr. Meteor., **11** (1956).

⁴ Ekman, V. W., *Actie. 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 Corey¹. 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 diagram 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.

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 this proposal on several 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 model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The sequence of 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 Å. 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 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. 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-coordinates. 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 I 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 cytosine (pyrimidine), and guanine

(purine) with thymine (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^{2,3} 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.

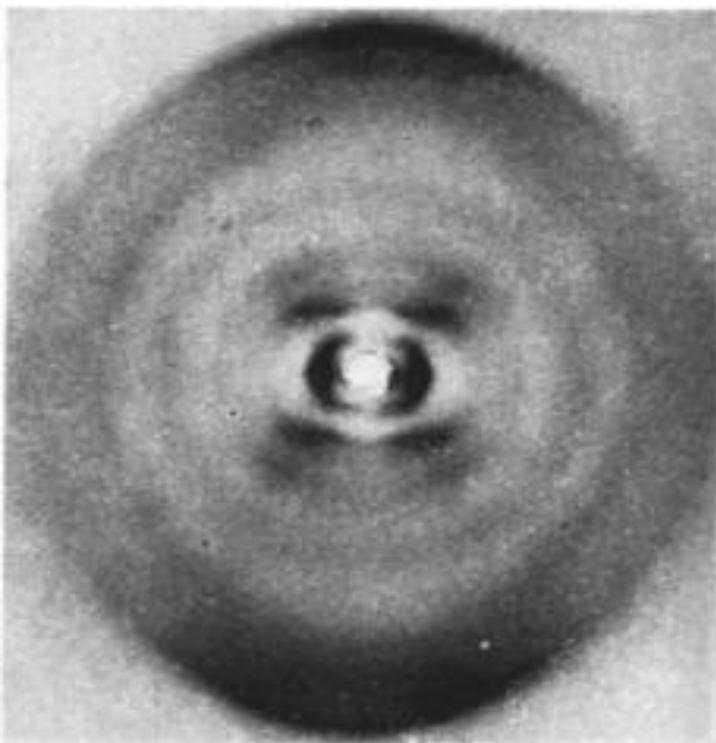
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data^{4,5} on deoxyribose nucleic acids 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 con-

This figure is purely diagrammatic. The two vertical lines symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases which link the chains together. The vertical line marks the fibre axis



STRUCTURAL ORGANIZATION.

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-

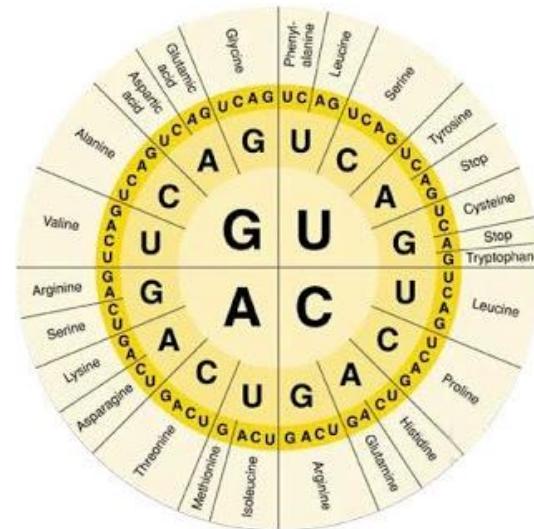
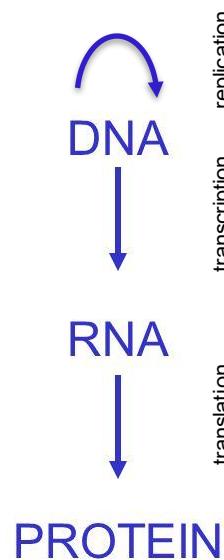
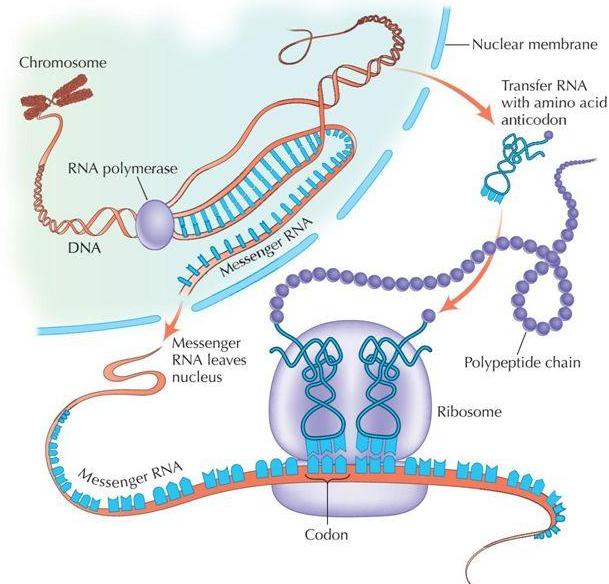
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

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”

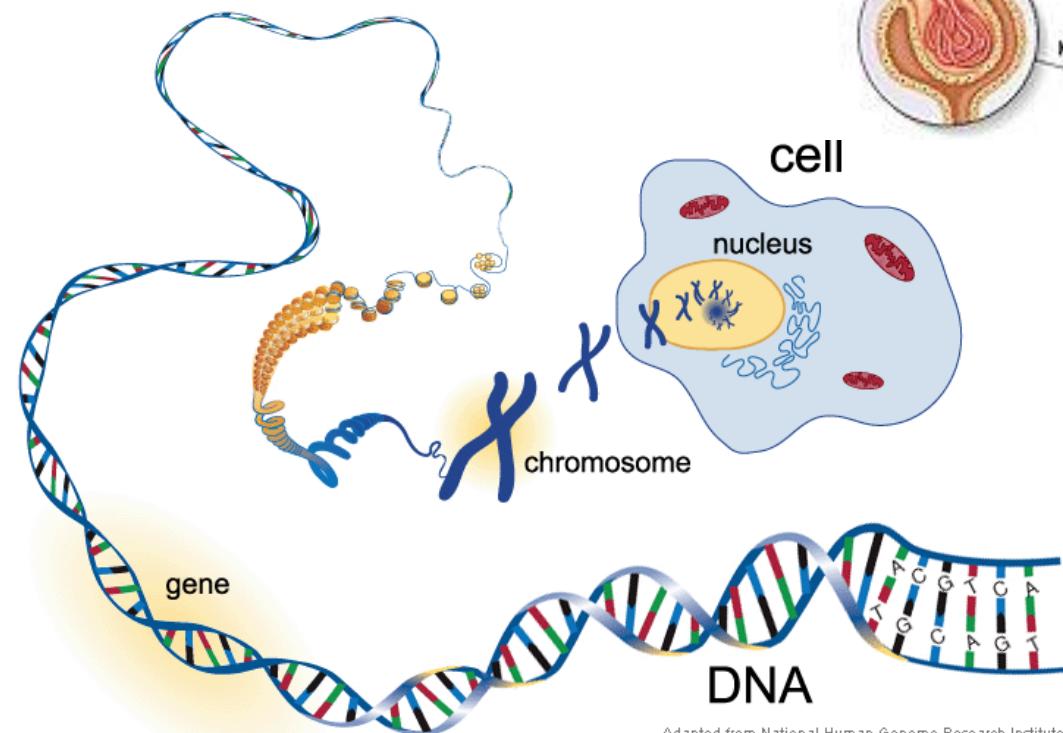


On Protein Synthesis

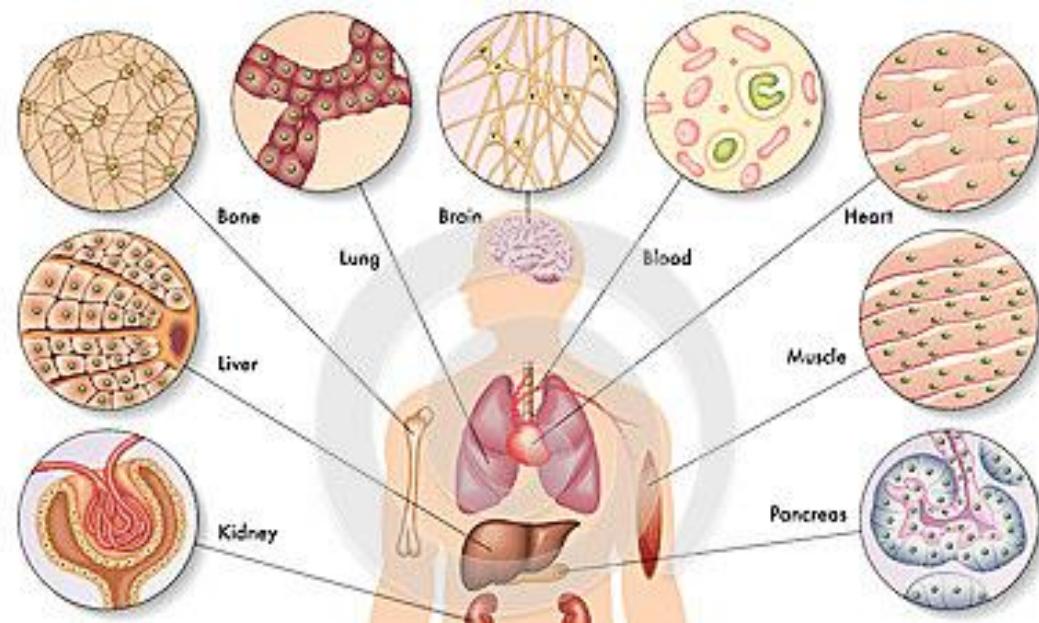
Crick, F.H.C. (1958). *Symposia of the Society for Experimental Biology* pp. 138–163.

One Genome, Many Cell Types

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

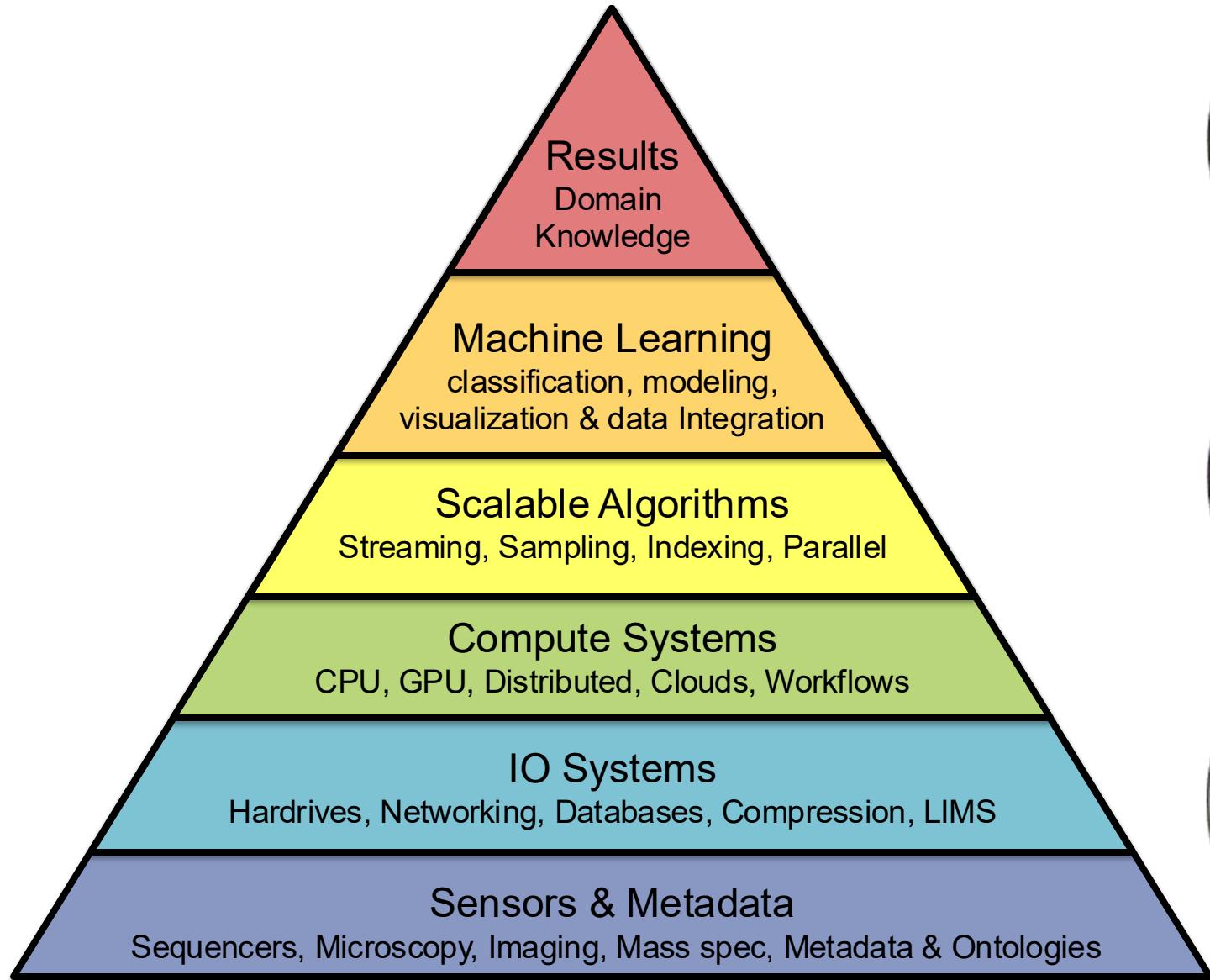


Adapted from National Human Genome Research Institute

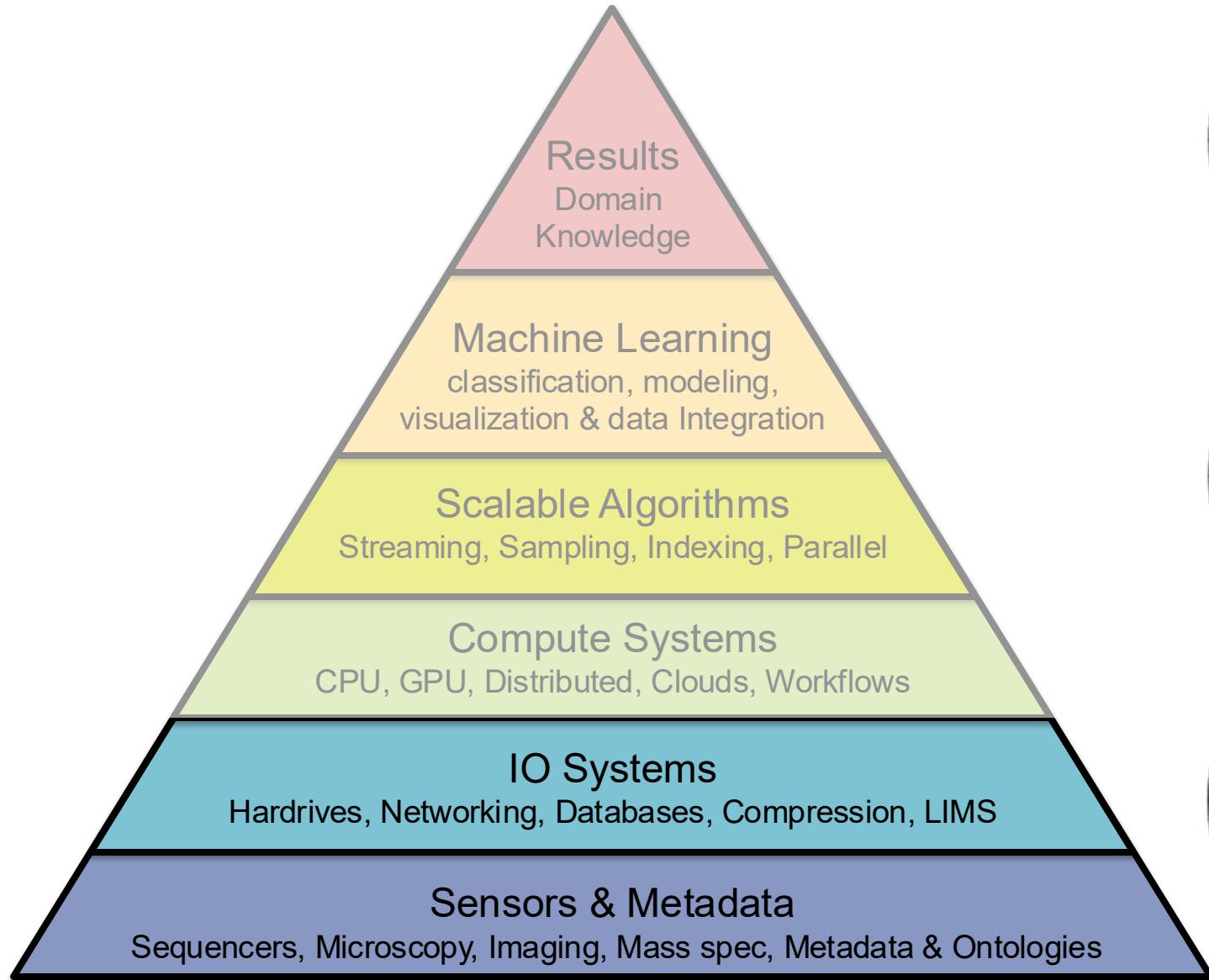


Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Applied Genomics



Applied Genomics



Genomics Arsenal in the year 2025

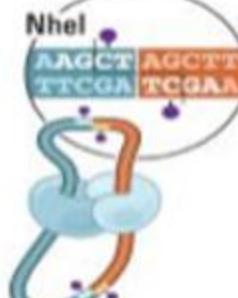
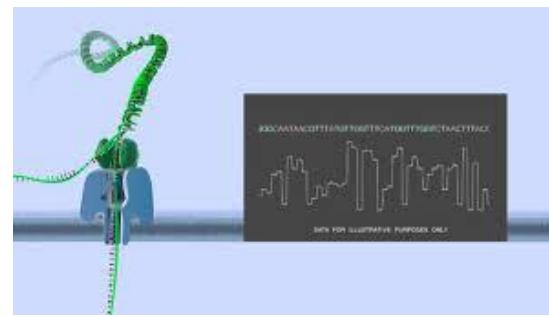
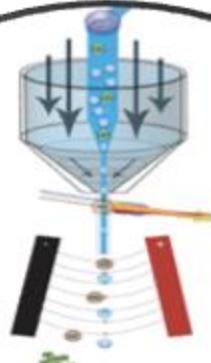
Sample Preparation

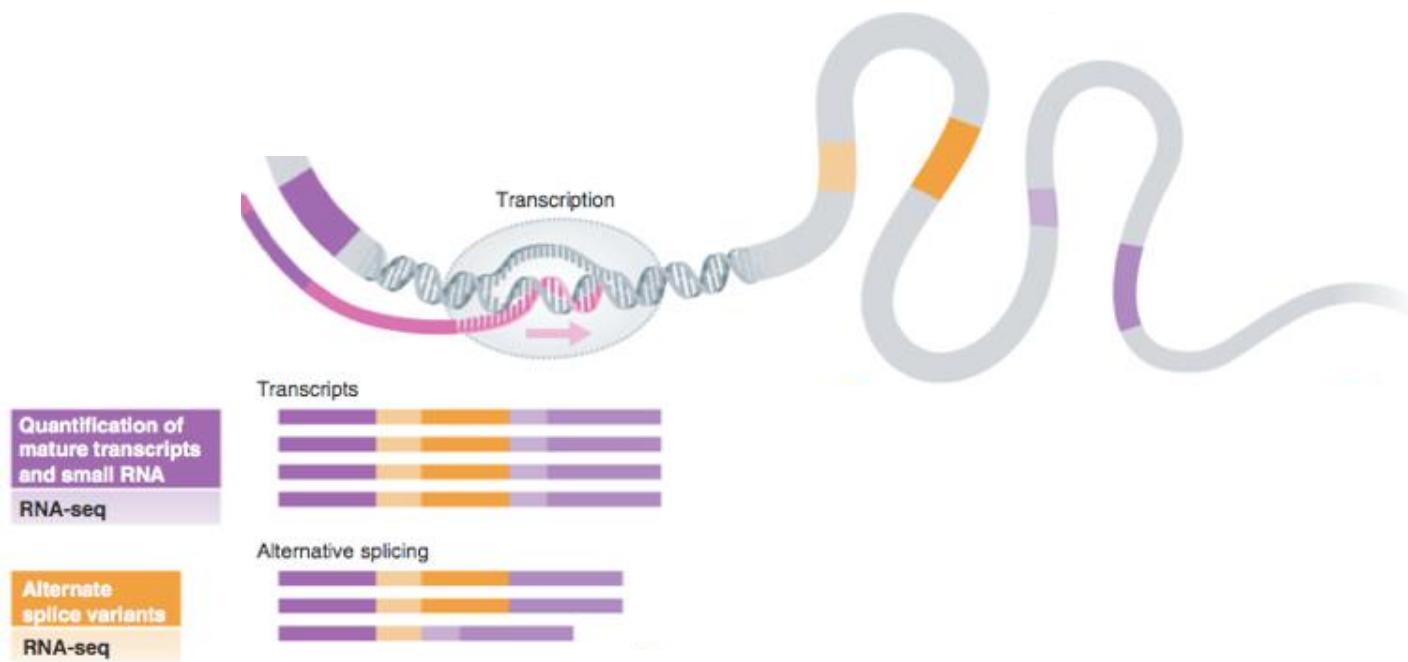


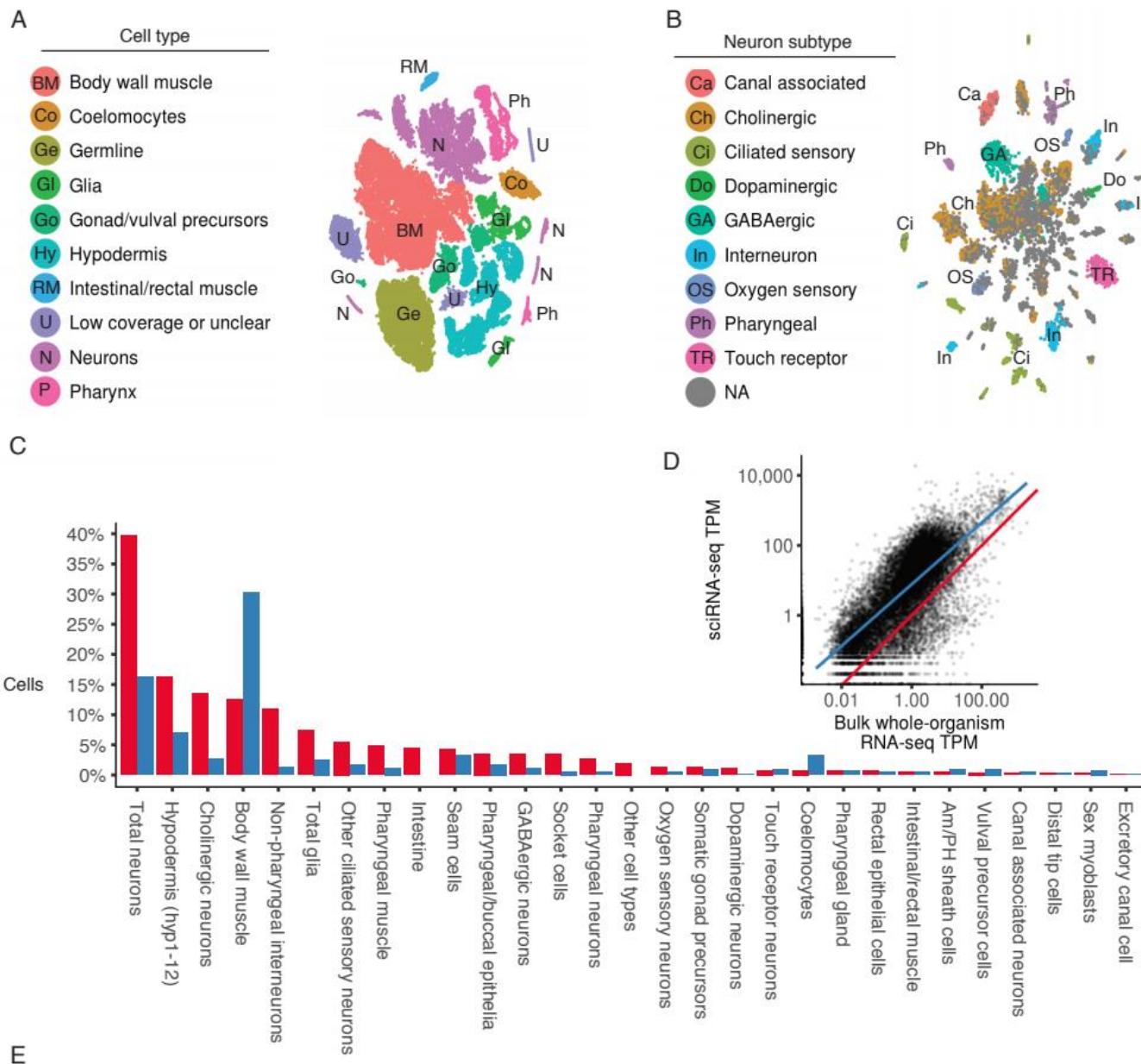
Sequencing



Chromosome Mapping

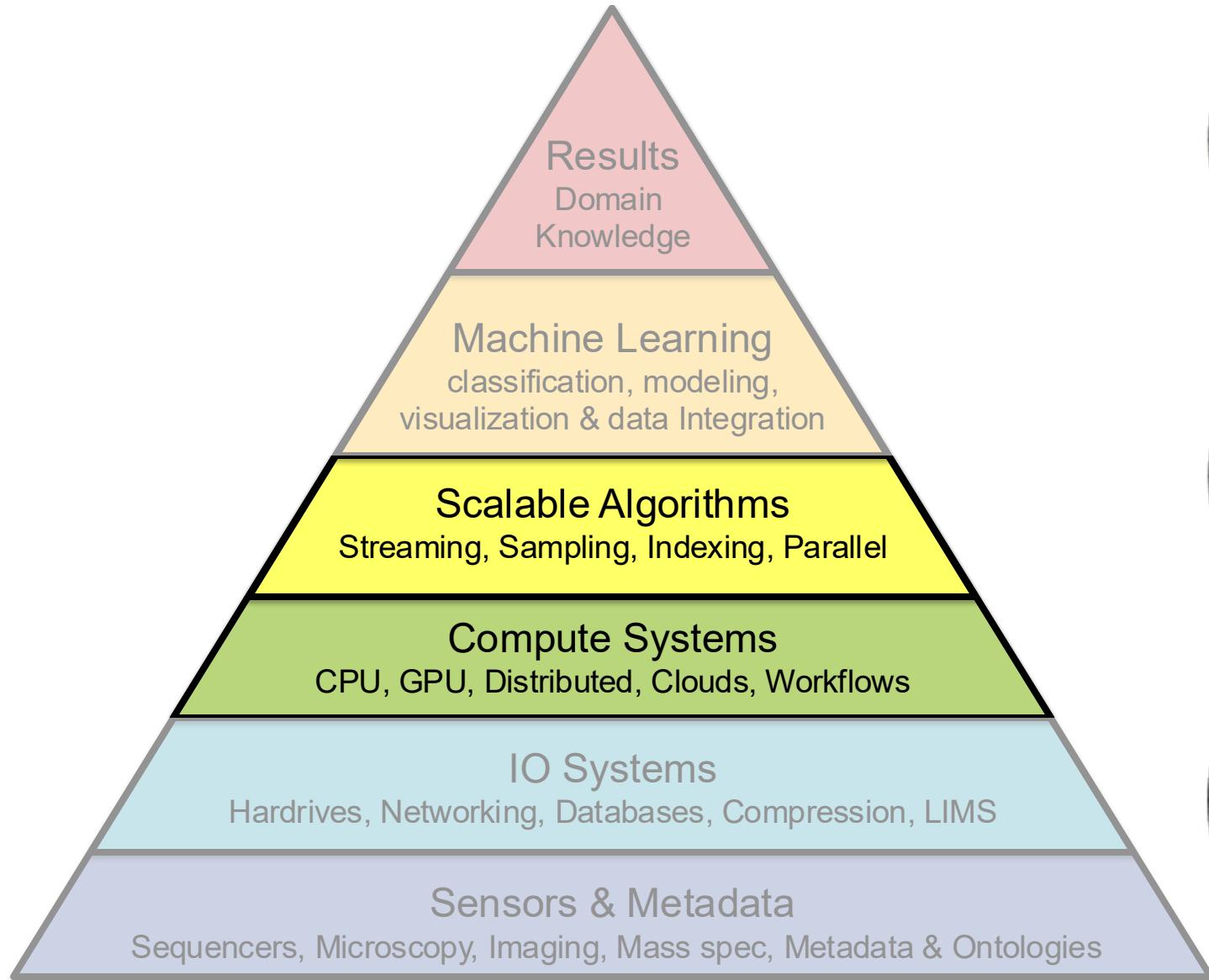






Comprehensive single-cell transcriptional profiling of a multicellular organism
 Cao, et al. (2017) Science. doi: 10.1126/science.aam8940

Applied Genomics

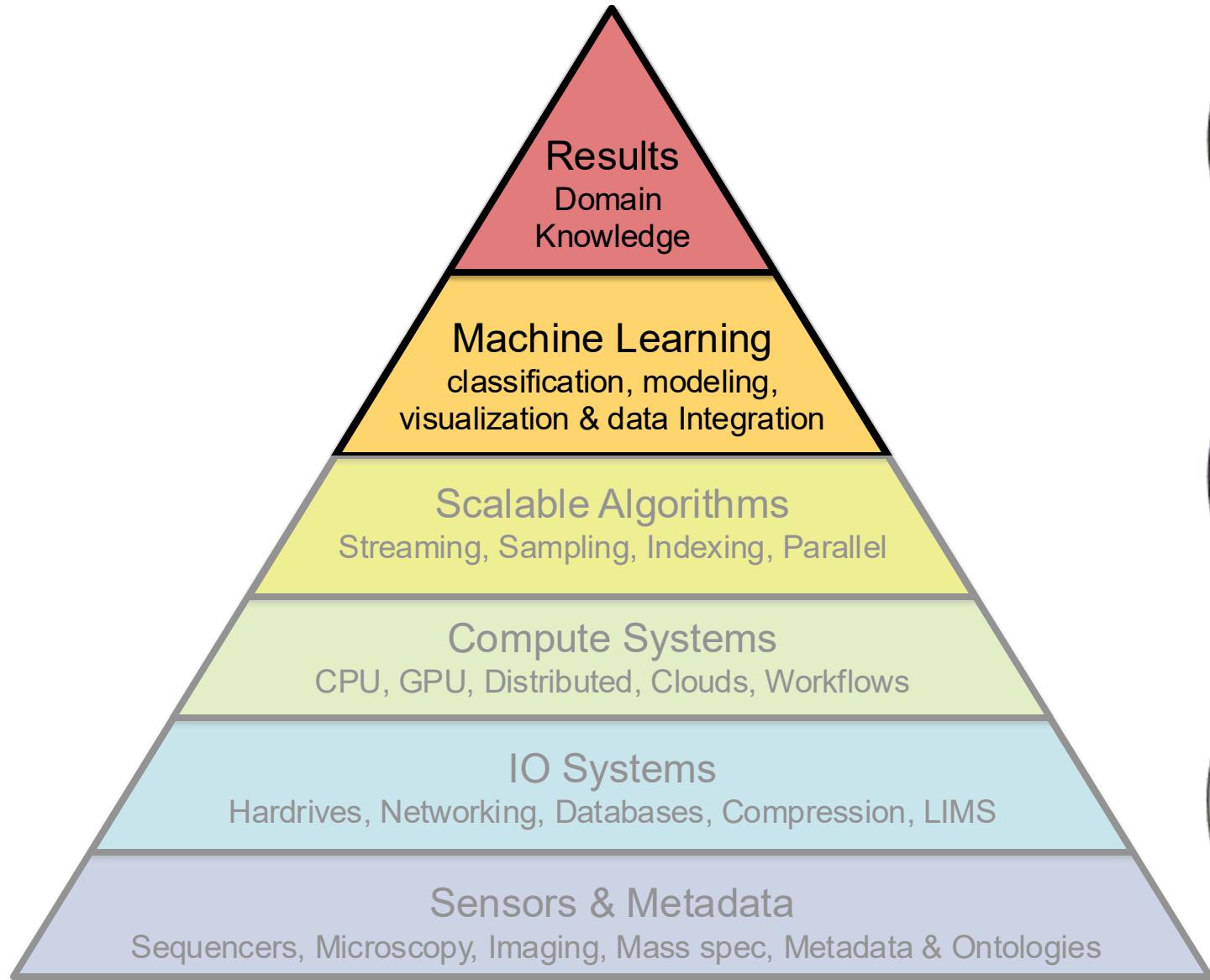


Potential Topics

- 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

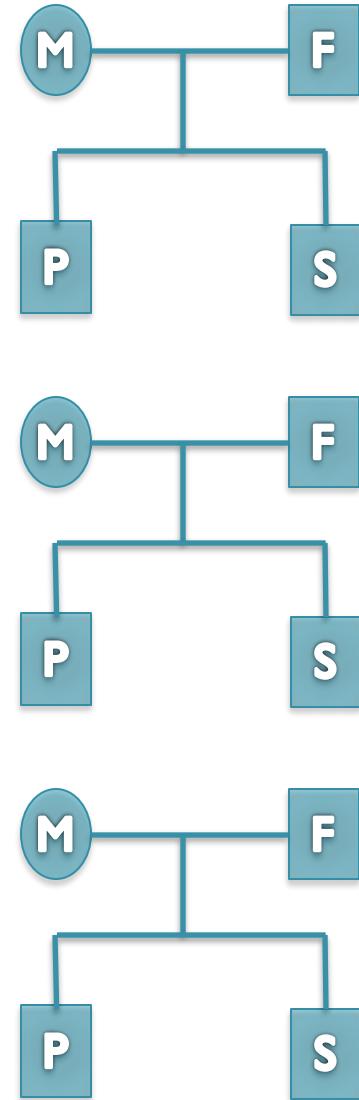
- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

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

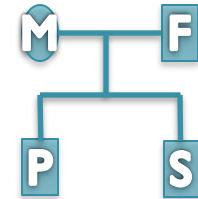
Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?



De novo mutation discovery and validation

De novo mutations:

Sequences not inherited from your parents.



Reference: . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Father (1) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Father (2) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Mother (1) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Mother (2) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Sibling (1) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Sibling (2) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

Proband (1) : . . . TCAAATCCTTTAATAAAAGAAGAGCTGACA . . .

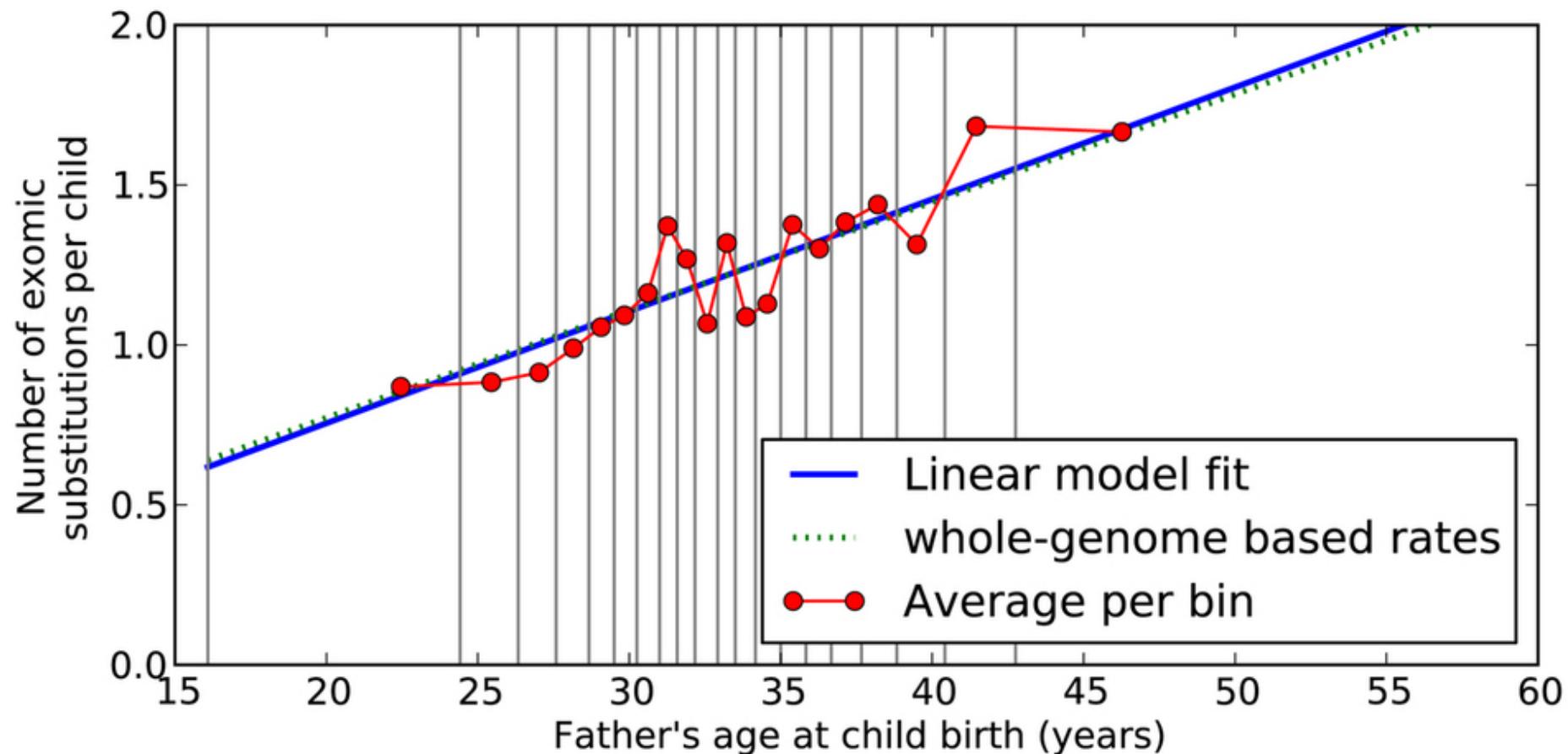
Proband (2) : . . . TCAAATCCTTTAAT****AAGAGCTGACA . . .

4bp heterozygous deletion at chr15:93524061 CHD2

De novo Genetics of Autism

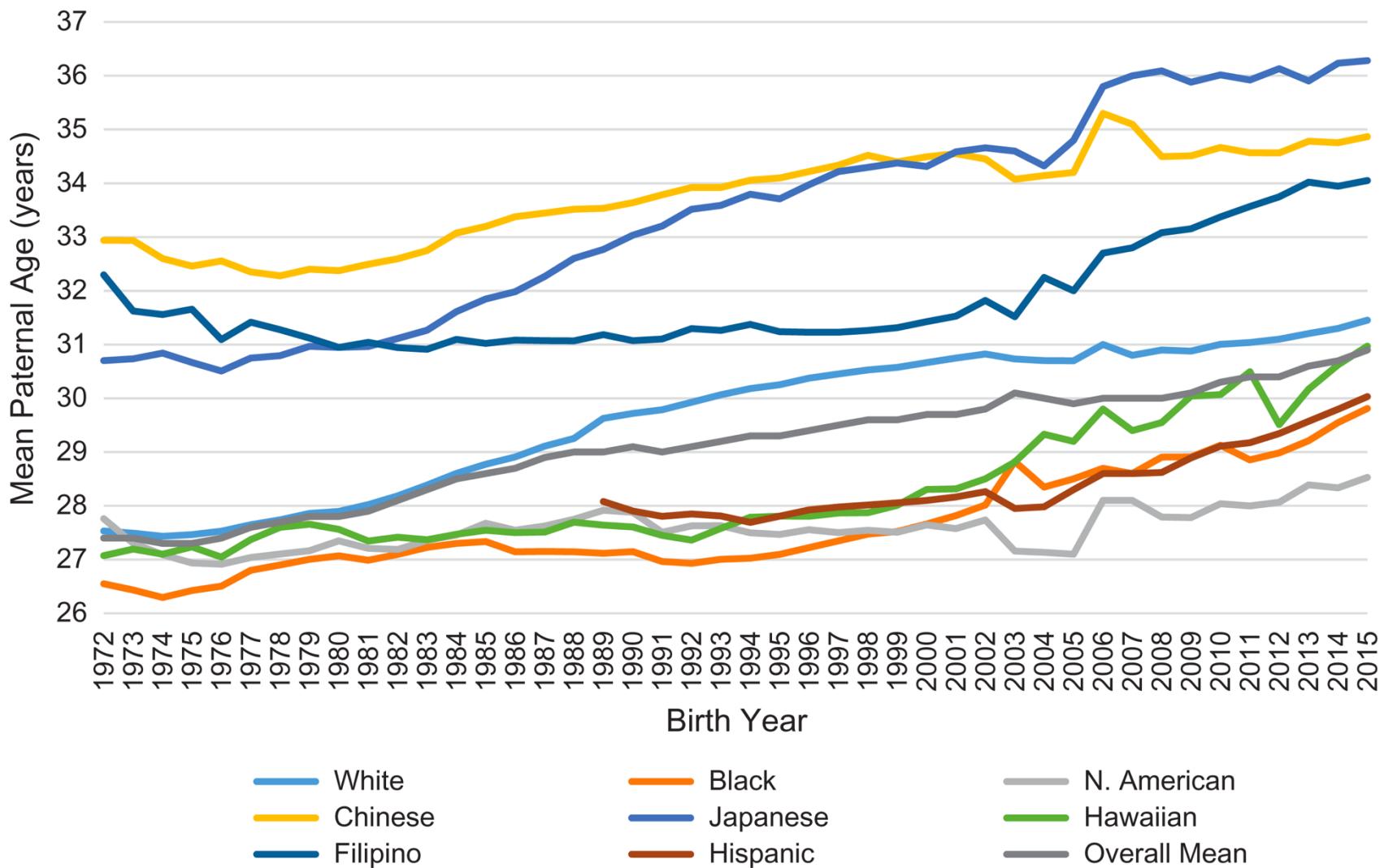
- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
 - Overall rate basically 1:1
 - 2:1 enrichment in nonsense mutations
 - 2:1 enrichment in frameshift indels
 - 4:1 enrichment in splice-site mutations
 - Most de novo originate in the 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

De novo Mutations in Men



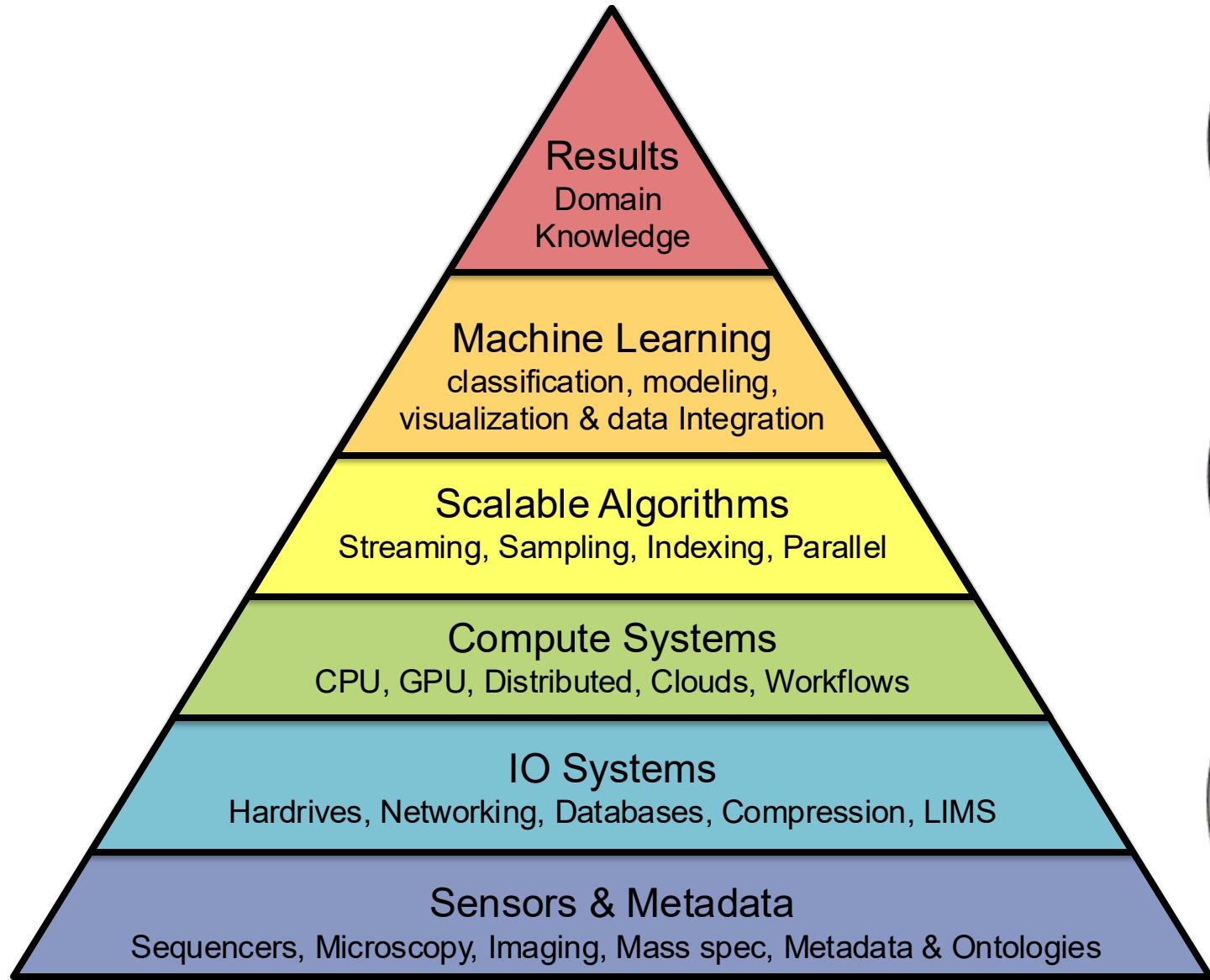
The contribution of de novo coding mutations to autism spectrum disorder
Iossifov et al (2014) Nature. doi:10.1038/nature13908

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 I
 1. Set up conda
 2. Set up Dropbox for yourself!
 3. Get comfortable on the command line