

Welcome to Biomedical Research!

Michael Schatz

August 29, 2017 – Lecture I

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research



Welcome!

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components:

1. **Lectures** on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing
2. **Research presentations** from distinguished faculty on their active research projects
3. **A major research project** with in-class research labs;
Satisfies the CS TEAM requirement

Course Webpage:

<https://github.com/schatzlab/biomedicalresearch2019>

Course Discussions:

<http://piazza.com>

Class Hours:

Mon + Wed @ 3p – 3:50p Hodson 311

Office Hours:

Monday @ 4-5p and by appointment
Please try Piazza first!



Prerequisites and Resources

Prerequisites

- No formal course requirements
- Access to an Apple or Linux Machine, or Install VirtualBox
- Familiarity with a major programming language will be needed
 - C/C++, Java, R, Perl, Python, JavaScript, others?

Primary Texts

- None! We will be studying primary research papers

Other Resources:

- Google, SEQanswers, Biostars, StackOverflow
- Applied Computational Genomics Course
 - <https://github.com/schatzlab/appliedgenomics2019>
- Ben Langmead's teaching materials:
 - <http://www.langmead-lab.org/teaching-materials/>



Grading Policies

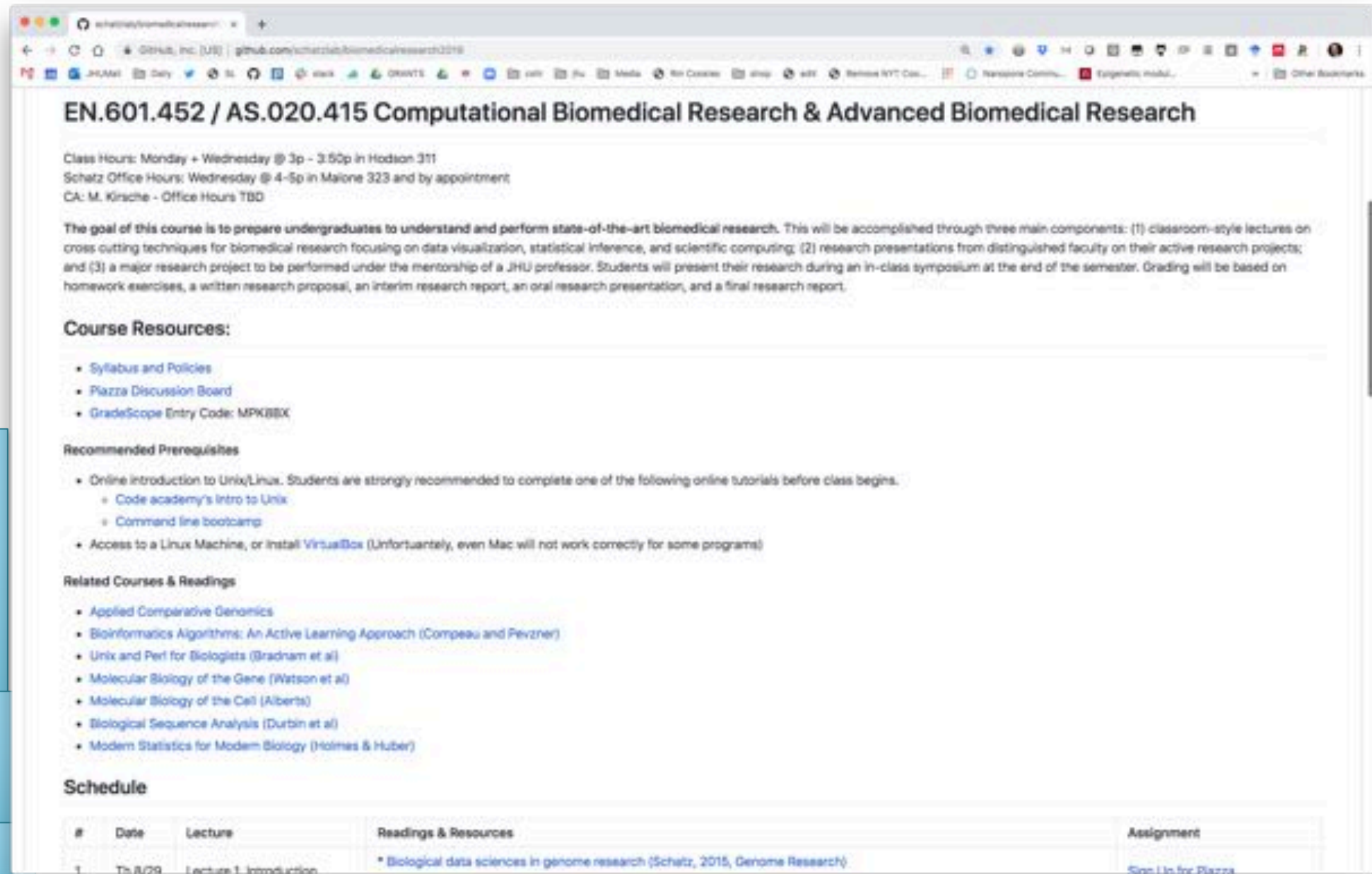
Assessments:

- ~4 HW Exercises: 10% Due at 11:59pm a week later
- Research Proposal: 10% ~1 page write up + oral presentation
- Interim Report: 10% ~3 page progress report
- Project Presentation: 30% Presented last week of class
- Final Report: 30% Due last week of semester
- In-class Participation: 10% Please ask questions!

Policies:

- Scores assigned relative to the highest points awarded
- **Late Days:**
 - 120 late hours without any penalty, then 25% deduction per day

Course Webpage



EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research

Class Hours: Monday + Wednesday @ 3p - 3:50p in Hodson 311
Schatz Office Hours: Wednesday @ 4-5p in Malone 323 and by appointment
CA: M. Kirsche - Office Hours TBD

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components: (1) classroom-style lectures on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing; (2) research presentations from distinguished faculty on their active research projects; and (3) a major research project to be performed under the mentorship of a JHU professor. Students will present their research during an in-class symposium at the end of the semester. Grading will be based on homework exercises, a written research proposal, an interim research report, an oral research presentation, and a final research report.

Course Resources:

- [Syllabus and Policies](#)
- [Piazza Discussion Board](#)
- [GradeScope Entry Code: MPKBBX](#)

Recommended Prerequisites

- Online introduction to Unix/Linux. Students are strongly recommended to complete one of the following online tutorials before class begins.
 - [Code academy's Intro to Unix](#)
 - [Command line bootcamp](#)
- Access to a Linux Machine, or install [VirtualBox](#) (Unfortunately, even Mac will not work correctly for some programs)

Related Courses & Readings

- [Applied Comparative Genomics](#)
- [Bioinformatics Algorithms: An Active Learning Approach](#) (Compeau and Pevzner)
- [Unix and Perl for Biologists](#) (Bradnam et al)
- [Molecular Biology of the Gene](#) (Watson et al)
- [Molecular Biology of the Cell](#) (Alberts)
- [Biological Sequence Analysis](#) (Durbin et al)
- [Modern Statistics for Modern Biology](#) (Holmes & Huber)

Schedule

#	Date	Lecture	Readings & Resources	Assignment
1	Th, 8/29	Lecture 1: Introduction	• Biological data sciences in genome research (Schatz, 2015, Genome Research)	Sign Up for Piazza

<https://github.com/schatzlab/biomedicalresearch2019>

Piazza

The screenshot displays the Piazza Q&A platform interface for the course EN 601.452. The top navigation bar includes links for 'Q & A', 'Resources', 'Statistics', and 'Manage Class'. The left sidebar features a 'New Post' button and a list of pinned and today's posts. The main content area shows a 'Welcome to Piazza!' note with the following tips:

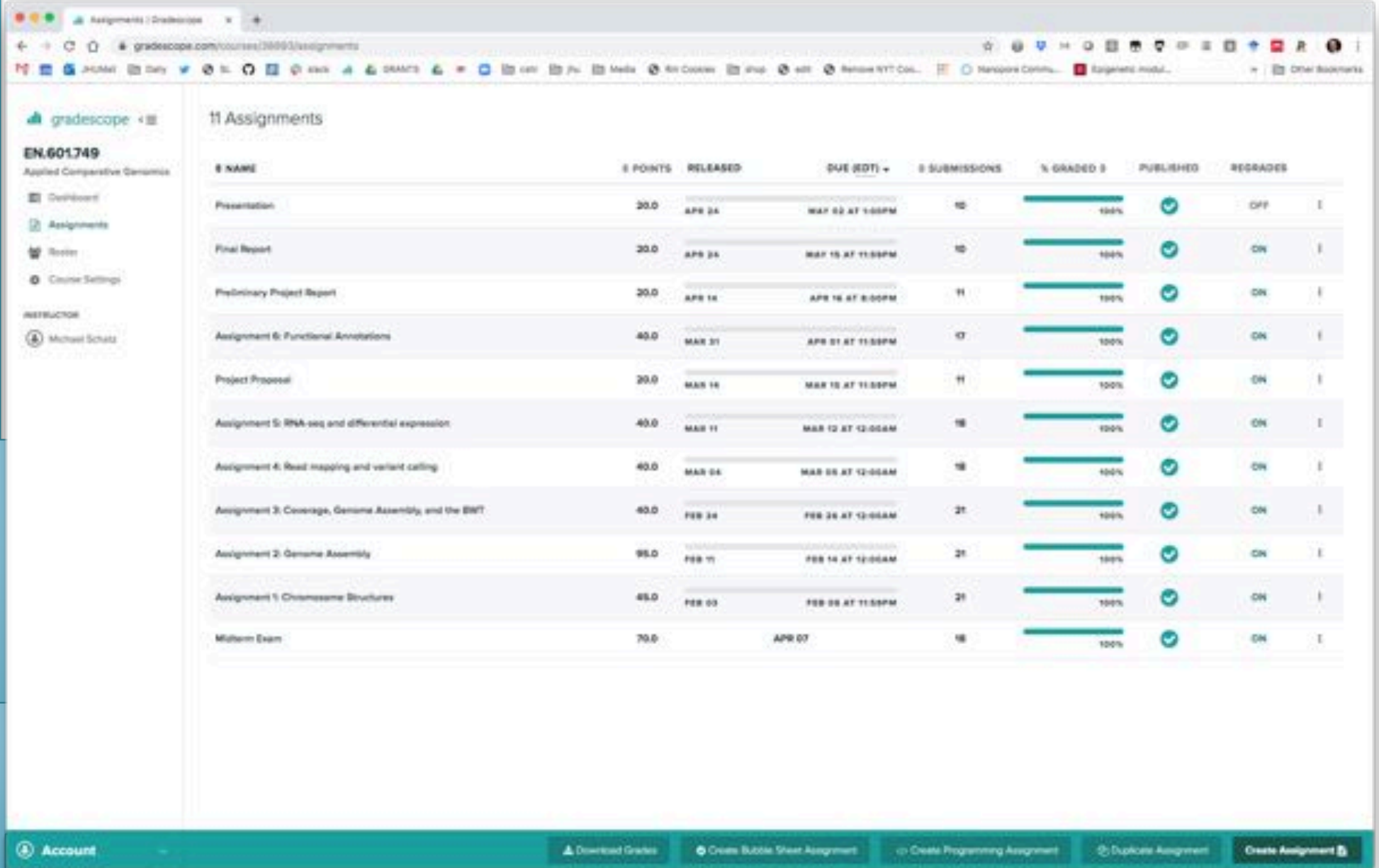
- 1. Ask questions!**
The best way to get answers is to ask questions! Ask questions on Piazza rather than emailing your teaching staff so everyone can benefit from the response (and so you can get answers from classmates who are up as late as you are).
- 2. Edit questions and answers wiki-style.**
Think of Piazza as a Q&A wiki for your class. Every question has just a single **students' answer** that students can edit collectively (and a single **instructors' answer** for instructors).
- 3. Add a followup to comment or ask further questions.**
To comment on or ask further questions about a post, start a **followup discussion**. Mark it resolved when the issue has been addressed, and add any relevant information back into the Q&A above.
- 4. Go anonymous.**
Shy? No problem. You can always opt to post or edit anonymously.
- 5. Tag your posts.**
It's far more convenient to find all posts about your Homework 3 or Midterm 1 when the posts are tagged. Type a "t" before a key word to tag. Click a blue tag in a post or the question feed to filter for all posts that share that tag.
- 6. Format code and equations.**
Adding a code snippet? Click the **pre** or **tt** button in the question editor to add pre-formatted or inline teletype text. Mathematical equation? Click the **fx** button to access the LaTeX editor to build a nicely formatted equation.
- 7. View and download class details and resources.**
Click the **Course Page** button in your top bar to access the class syllabus, staff contact information, office hours details, and course resources—all in one place!

Contact the Piazza Team anytime with questions or comments at team@piazza.com. We love feedback!

The bottom status bar shows: Average Response Time: N/A, Special Mentions: There are no special mentions at this time., and Online Now: 1 | This Week: 1.

<http://piazza.com/jhu/fall2019/en601452>

GradeScope



gradescope

EN.601.749
Applied Comparative Genomics

Dashboard
Assignments
Router
Course Settings

INSTRUCTOR
Michael Schatz

11 Assignments

NAME	POINTS	RELEASED	DUE (EDT)	SUBMISSIONS	% GRADED	PUBLISHED	REGRADES
Presentation	20.0	APR 24	MAY 22 AT 1:00PM	10	100%	✓	OFF
Final Report	20.0	APR 24	MAY 15 AT 11:00PM	10	100%	✓	ON
Preliminary Project Report	20.0	APR 14	APR 16 AT 8:00PM	11	100%	✓	ON
Assignment 6: Functional Annotations	40.0	MAR 31	APR 01 AT 11:00PM	17	100%	✓	ON
Project Proposal	20.0	MAR 14	MAR 15 AT 11:00PM	11	100%	✓	ON
Assignment 5: RNA-seq and differential expression	40.0	MAR 11	MAR 12 AT 12:00AM	18	100%	✓	ON
Assignment 4: Read mapping and variant calling	40.0	MAR 04	MAR 05 AT 12:00AM	18	100%	✓	ON
Assignment 3: Coverage, Genome Assembly, and the BWT	40.0	FEB 24	FEB 25 AT 12:00AM	21	100%	✓	ON
Assignment 2: Genome Assembly	95.0	FEB 11	FEB 14 AT 12:00AM	21	100%	✓	ON
Assignment 1: Chromosome Structures	45.0	FEB 03	FEB 05 AT 11:00PM	21	100%	✓	ON
Midterm Exam	70.0		APR 07	18	100%	✓	ON

Account
Download Grades
Create Bubble Sheet Assignment
Create Programming Assignment
Duplicate Assignment
Create Assignment

<https://www.gradescope.com/courses/60230> Entry Code: MPK8BX

Schatzlab Overview



Human Genetics

Role of mutations in disease

Nattestad et al. (2018)
Feigin et al. (2017)



Agricultural Genomics

Genomes & Transcriptomes

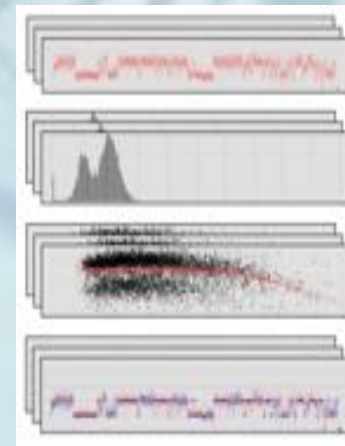
Soyk et al. (2019)
Zheng et al. (2018)



Algorithmics & Systems Research

Ultra-large scale biocomputing

Chen et al. (2019)
Parsana et al. (2019)



Single Cell & Single Molecule

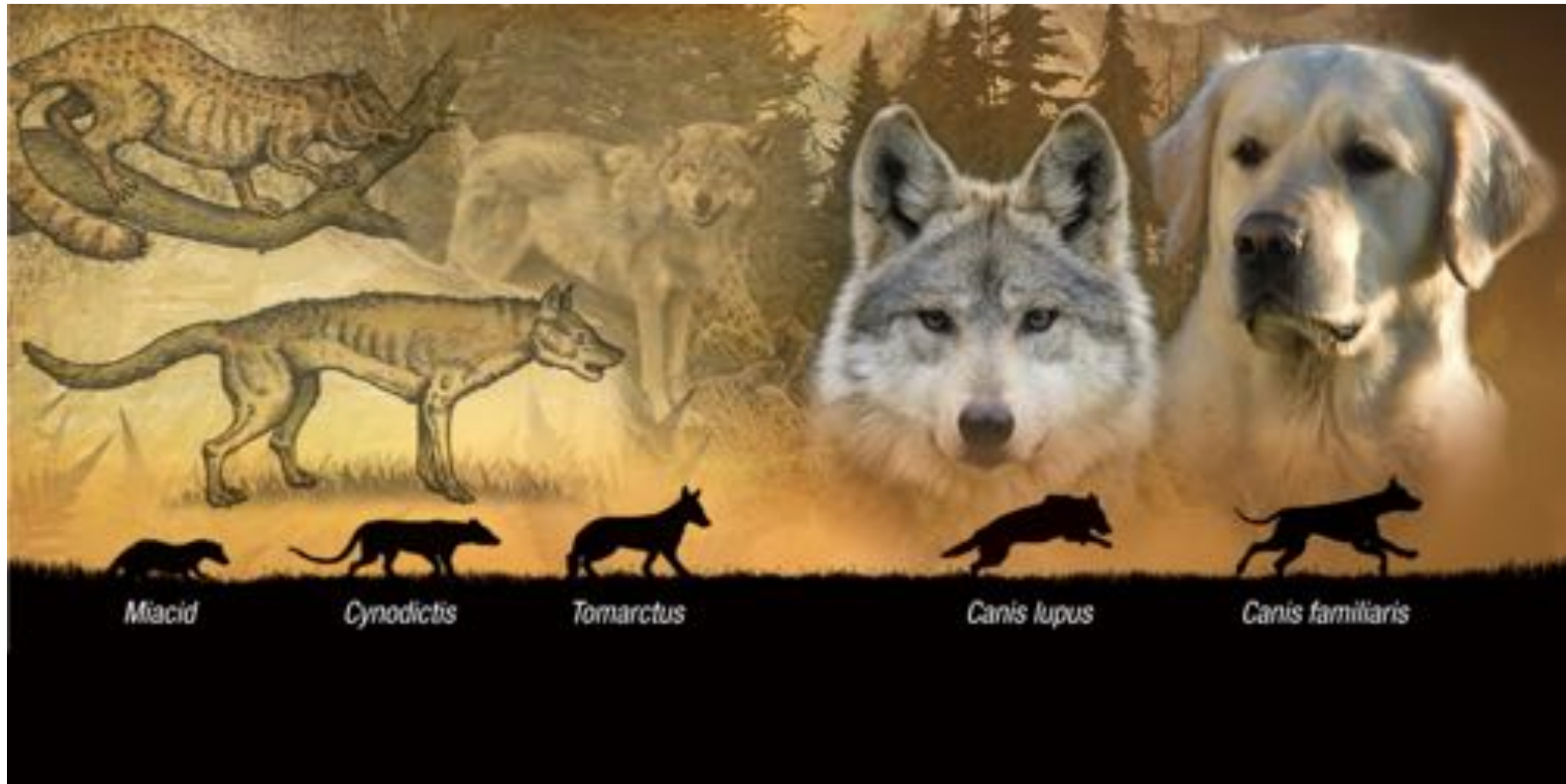
CNVs, SVs, & Cell Phylogenetics

Luo et al. (2019)
Sedlazeck et al. (2018)

Earliest Genomics

Any Thoughts?

Earliest Genomics



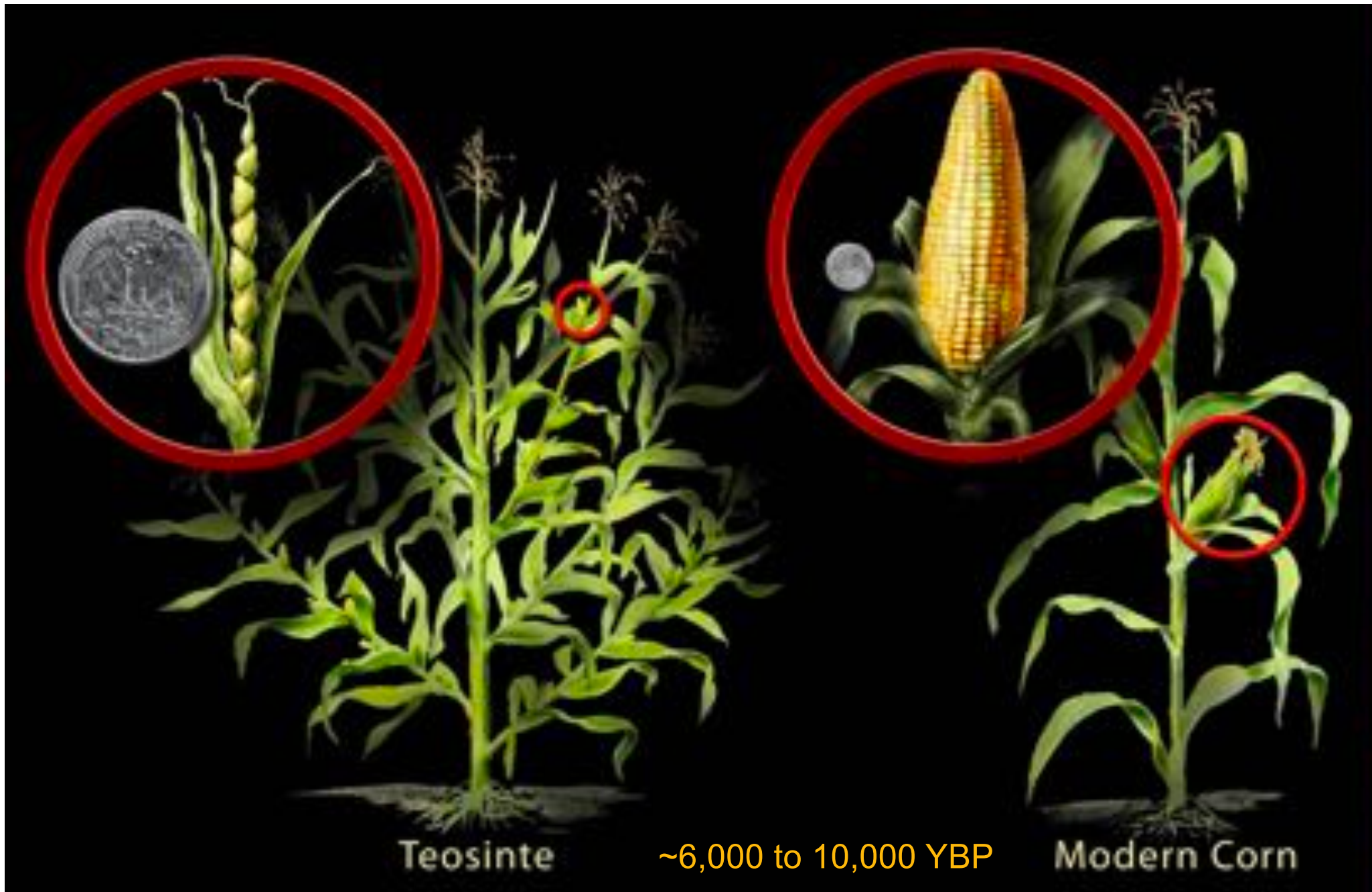
15,000 to 35,000 YBP

Earliest Genomics



~1,000 to 10,000 YBP

Earliest Genomics



Discovery of the Double Helix

NO. 4322 April 25, 1953 NATURE 737

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

¹Young, F. B., Corvett, E., and Jenson, W., *Phil. Mag.*, **46**, 141 (1928).

²Langer, H., M. S., *Ann. Ent. Soc. Amer.*, **35**, 104 (1942).

³See also, G. P., *Woods Hole Papers in Phys. Geology*, **11**, 11 (1934).

⁴Elmer, T. W., *Ann. Ent. Soc. Amer.*, **35**, 104 (1942).

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 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 Foweraker² 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 the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining 5-p-deoxy-ribose 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 resembling Purrberg's³ 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 Purrberg'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 *i*-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^{4,5} 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.

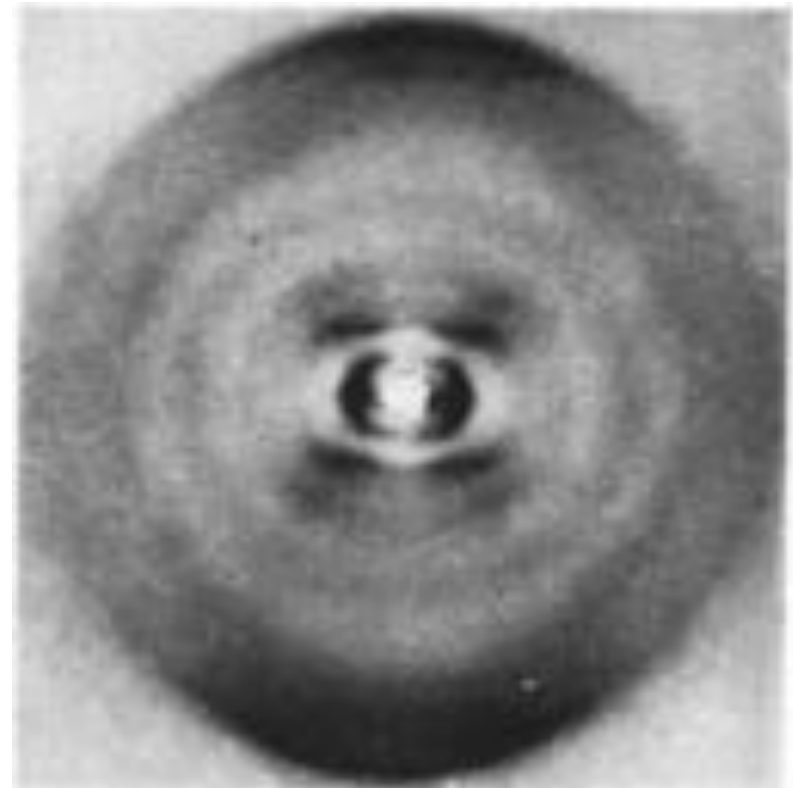
The previously published X-ray data^{6,7} on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following conversations. 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 *i*-co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Doreau for constant advice and criticism, especially on inter-atomic 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



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



Conclusion 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-

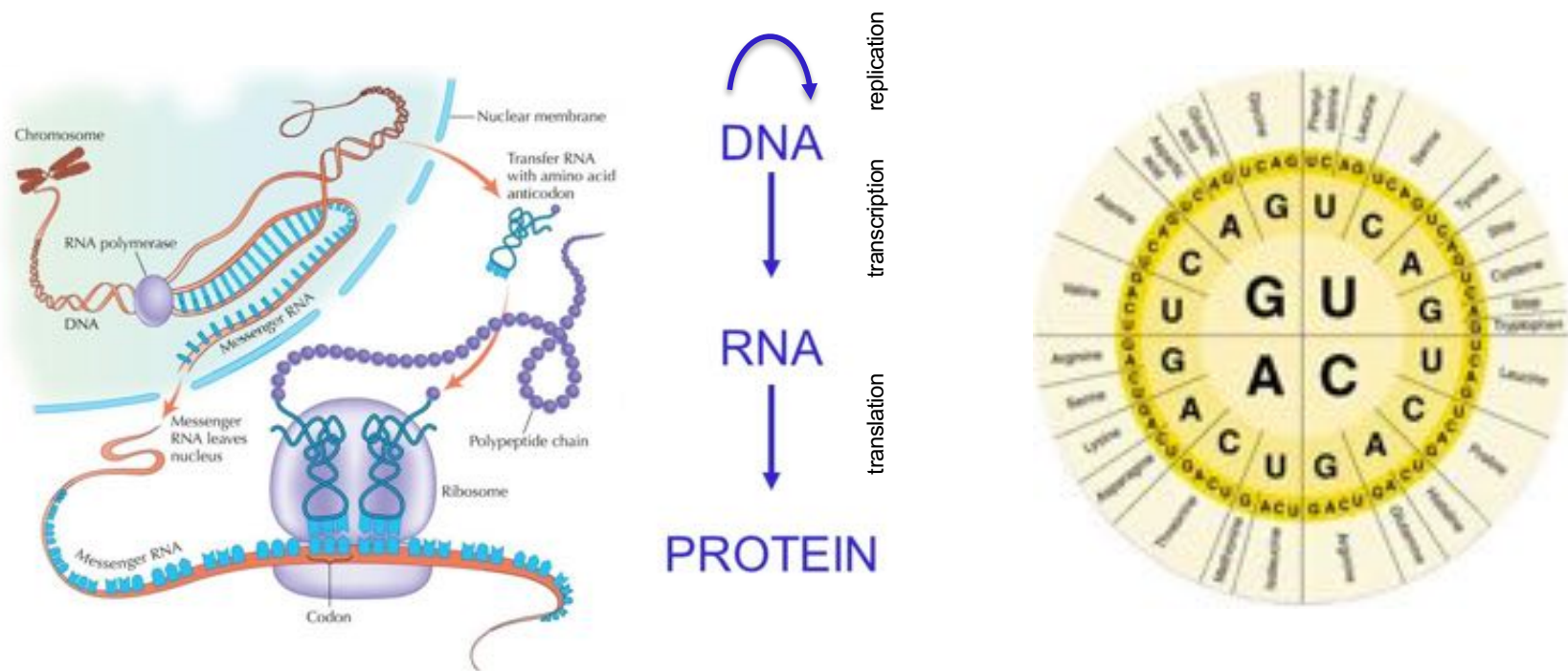
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

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”

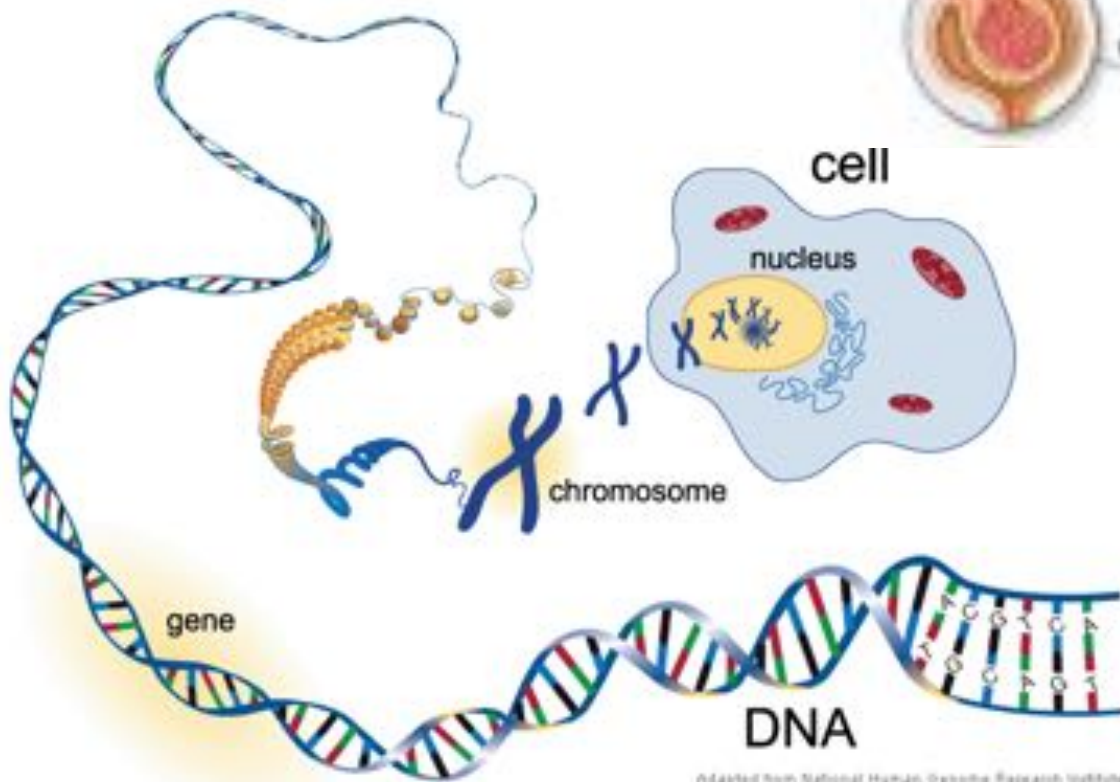


On Protein Synthesis

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

The Origins of DNA Sequencing

Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage ϕ X174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown*, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III*, P. M. Slocombe* & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QF, UK

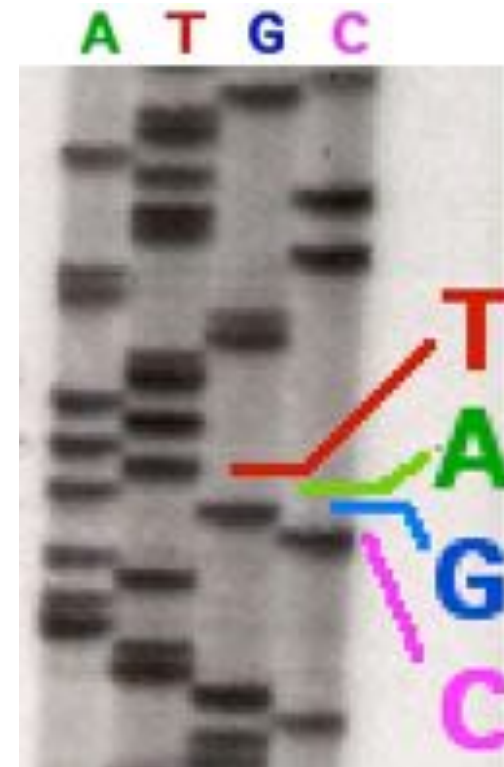
A DNA sequence for the genome of bacteriophage ϕ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

The genome of bacteriophage ϕ X174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques^{1,2}, is A-B-C-D-E-F-G-H. Genes F, G and H code for structural proteins of the virus capsid, and gene J (as defined by sequence work) codes for a small basic protein.

strand DNA of ϕ X has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene G protein³ (positions 2,362-2,411).

At this stage sequencing techniques using primer synthesis with DNA polymerase were being developed^{4,5} and Schmitt⁶ synthesized a dinucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intergenic region between the F and G genes, using DNA polymerase and ³²P-labelled triphosphates⁷. The ribonutrition technique^{8,9} facilitated the sequence determination of the labelled DNA produced. This dinucleotide-primed system was also used to develop the plus and minus method¹⁰. Suitable synthetic primers are, however, difficult to prepare and so

1977
1st Complete Organism
Bacteriophage ϕ X174
5375 bp



Radioactive Chain Termination
5000bp / week / person

<http://en.wikipedia.org/wiki/File:Sequencing.jpg>
<http://www.answers.com/topic/automated-sequencer>

Nucleotide sequence of bacteriophage ϕ X174 DNA

Sanger, F. et al. (1977) *Nature*. 265: 687 - 695

Milestones in DNA Sequencing



(TIGR/Celera, 1995-2001)

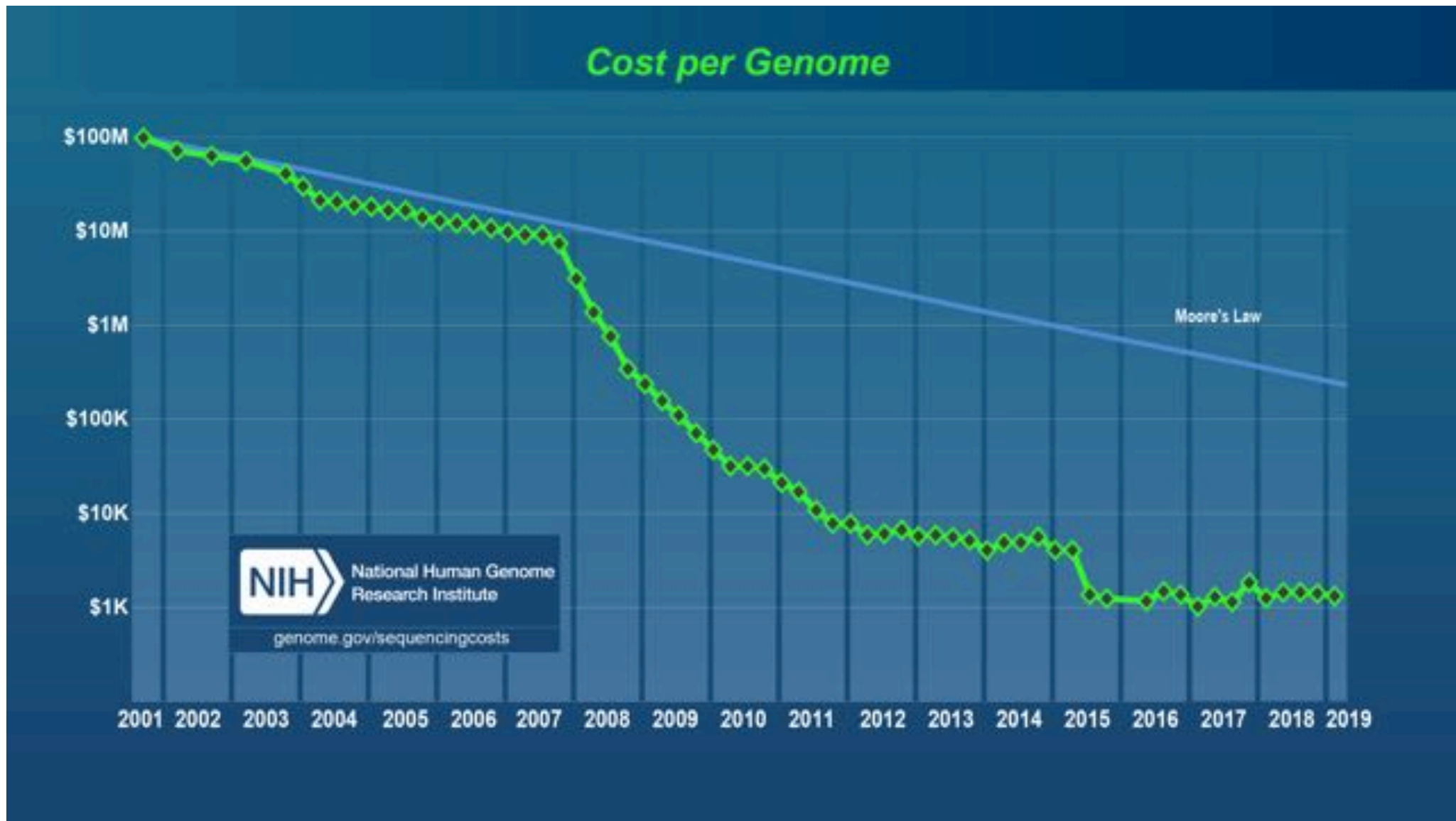
The most wondrous map...



“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”

*Bill Clinton
June 26, 2000*

Cost per Genome

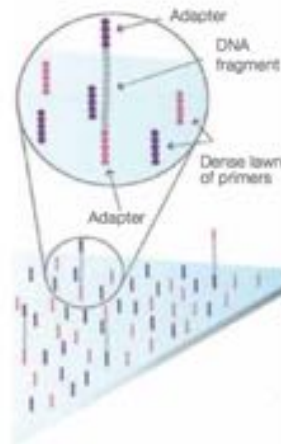


Next Generation Sequencing

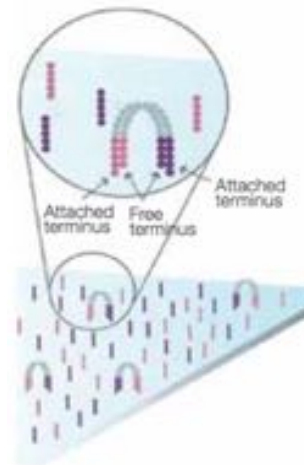


Illumina NovaSeq 6000
Sequencing by Synthesis

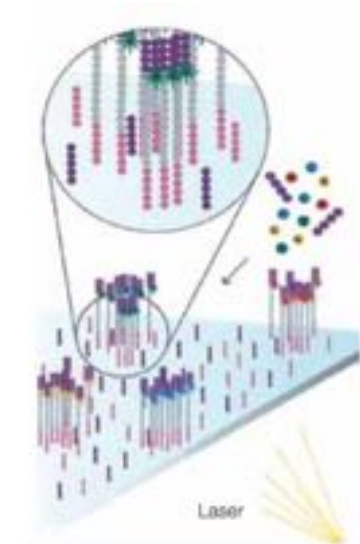
>3Tbp / day



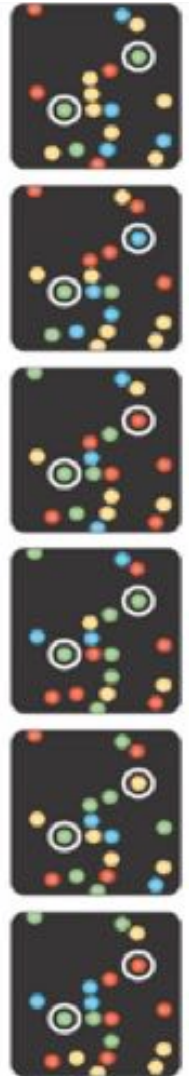
1. Attach



2. Amplify



3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46
<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

Sequencing Centers

Worldwide capacity exceeds 50 Pbp/year
Over 1.5M human genomes sequenced

On track to exceed over 1 Zbp / year by 2030

A zetta-what?

Next Generation Genomics: World Map of High-throughput Sequencers

<http://omicsmaps.com>

How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?



100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

1ZB Data
200,000,000,000 DVDs



150,000 miles of DVDs
~ ½ distance to moon



Both currently ~100Pb
And growing exponentially

Unsolved Questions in Biology

- What is your genome sequence?

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

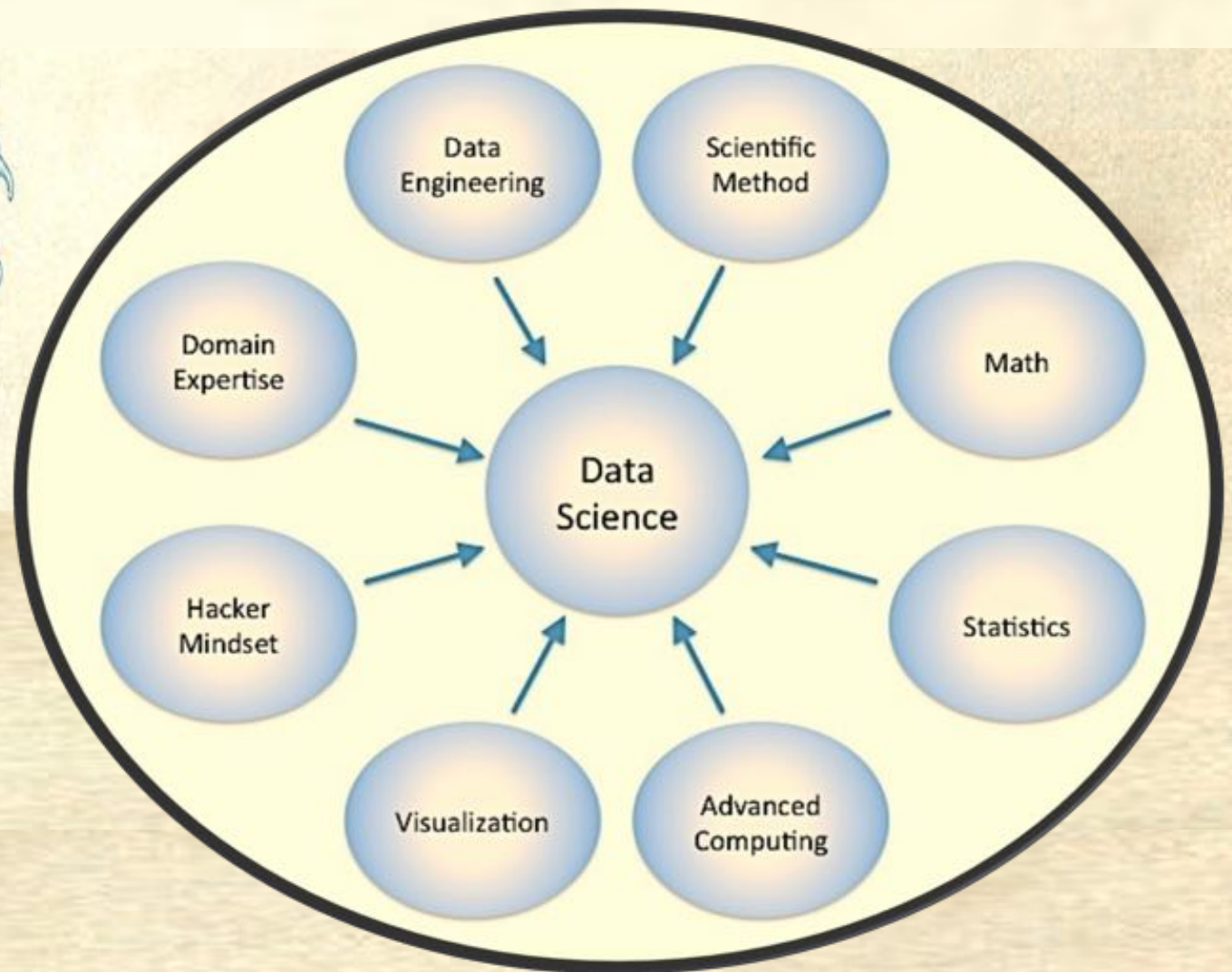
What software and systems will?

And who will create them?

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

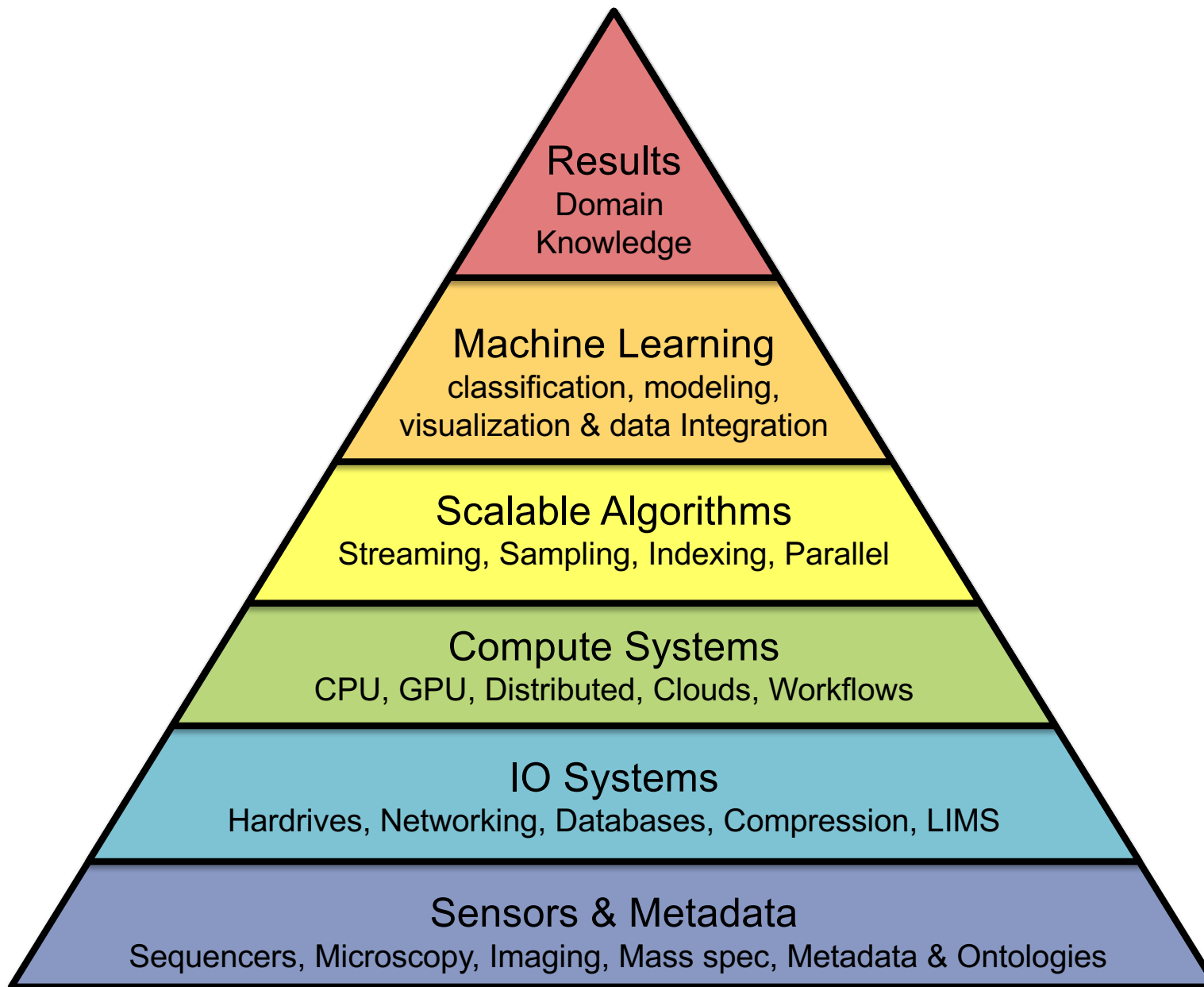


Who is a Data Scientist?

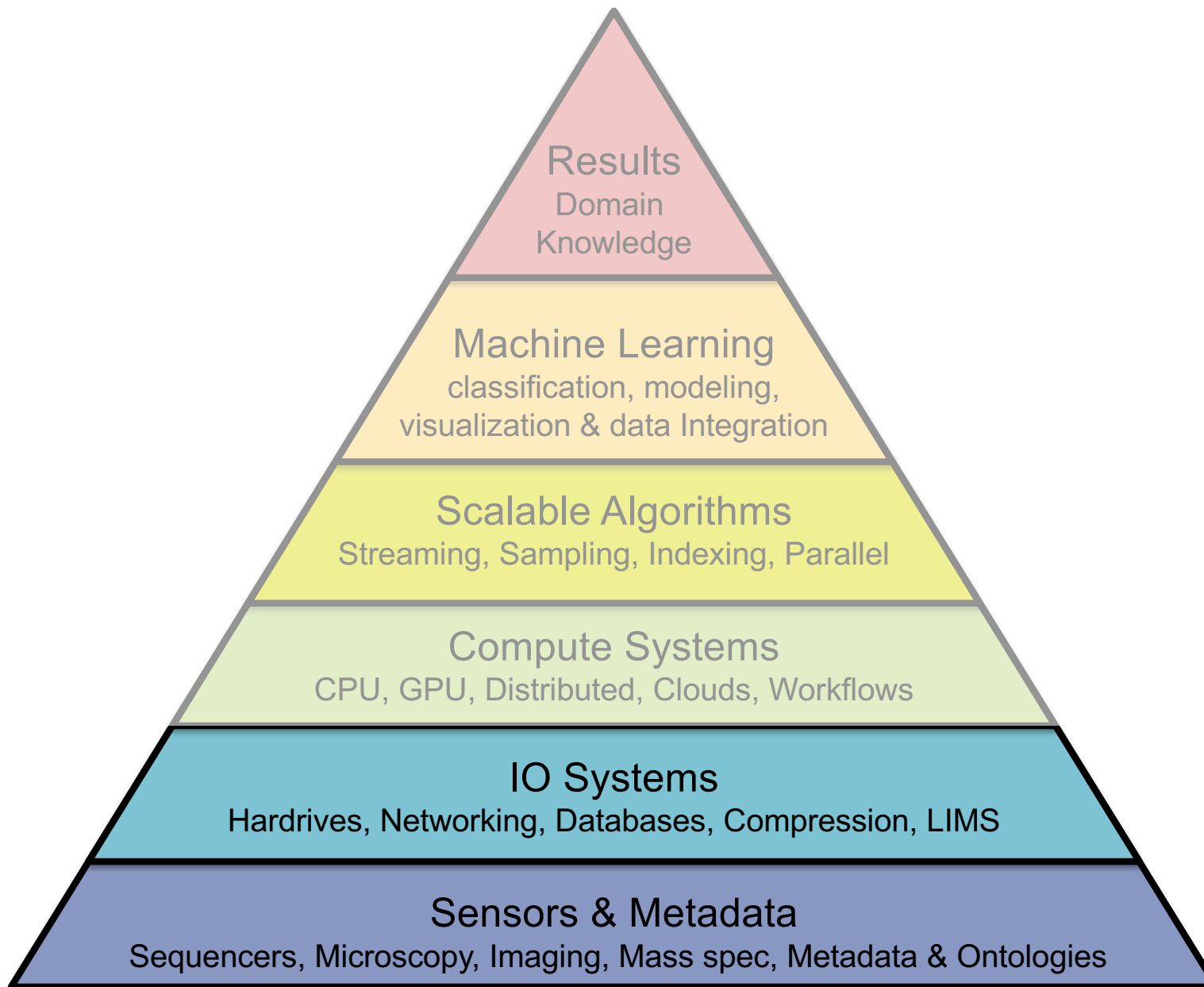


http://en.wikipedia.org/wiki/Data_science

Biomedical Research Technologies



Biomedical Research Technologies



Genomics Arsenal in the year 2019

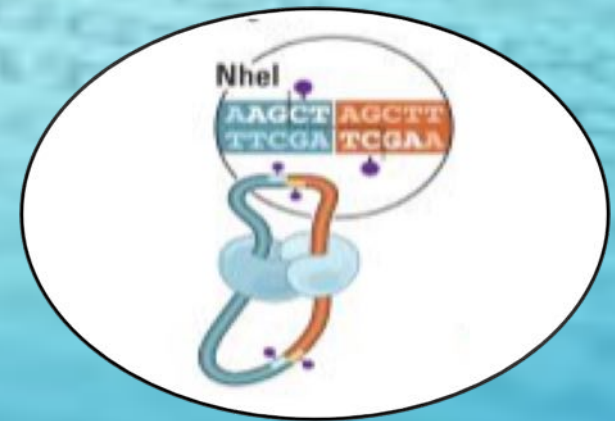
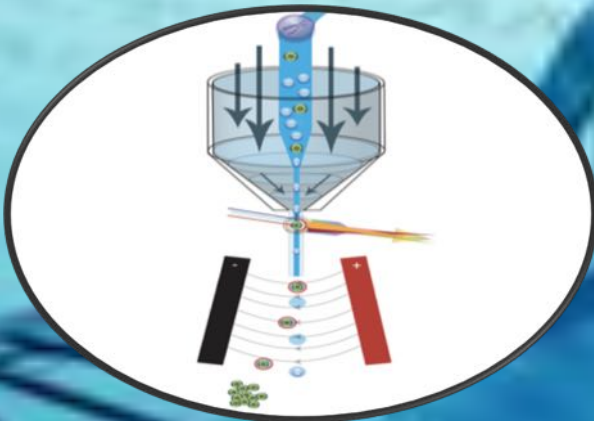
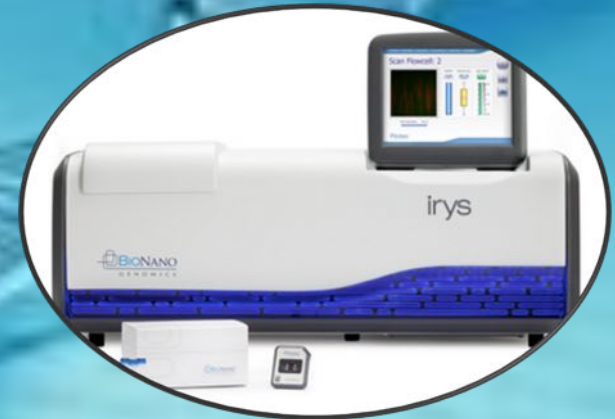
Sample Preparation

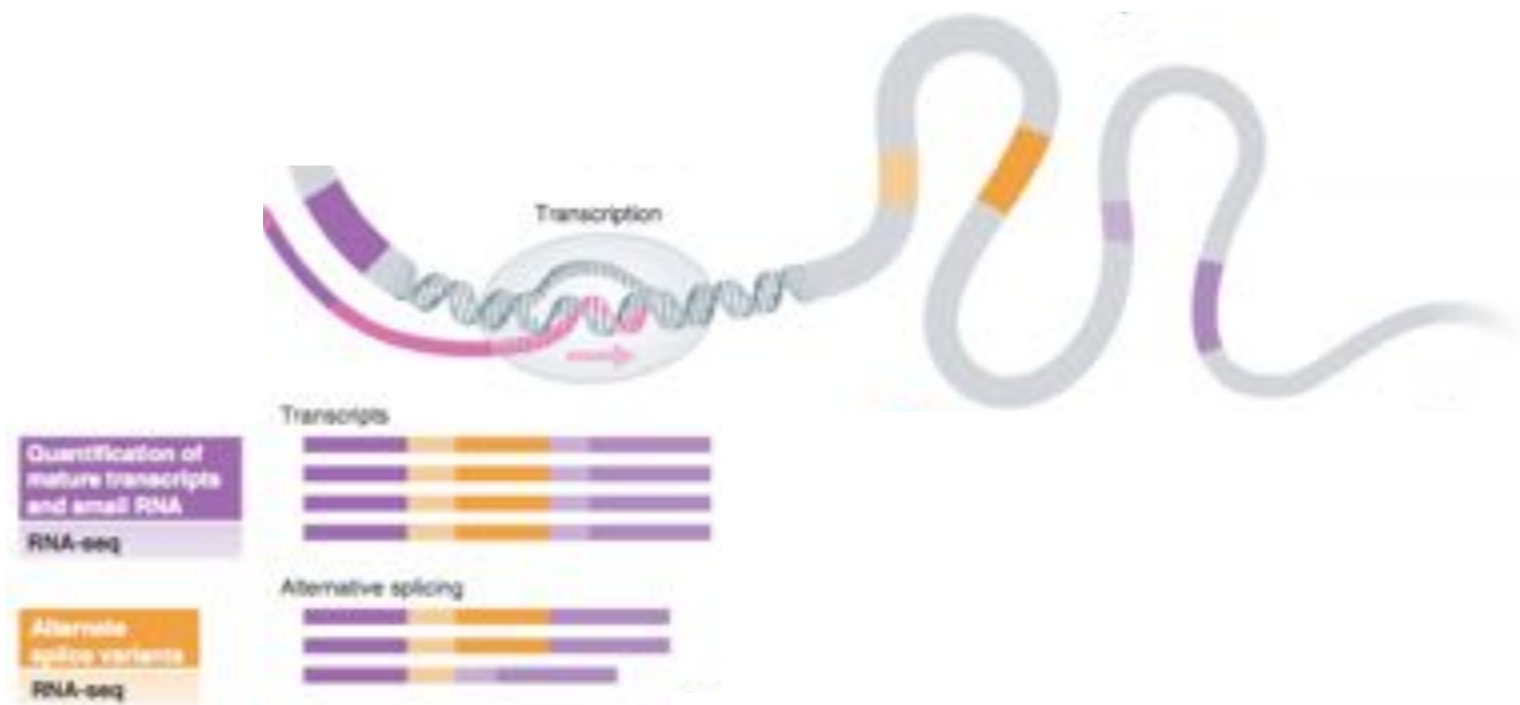


Sequencing

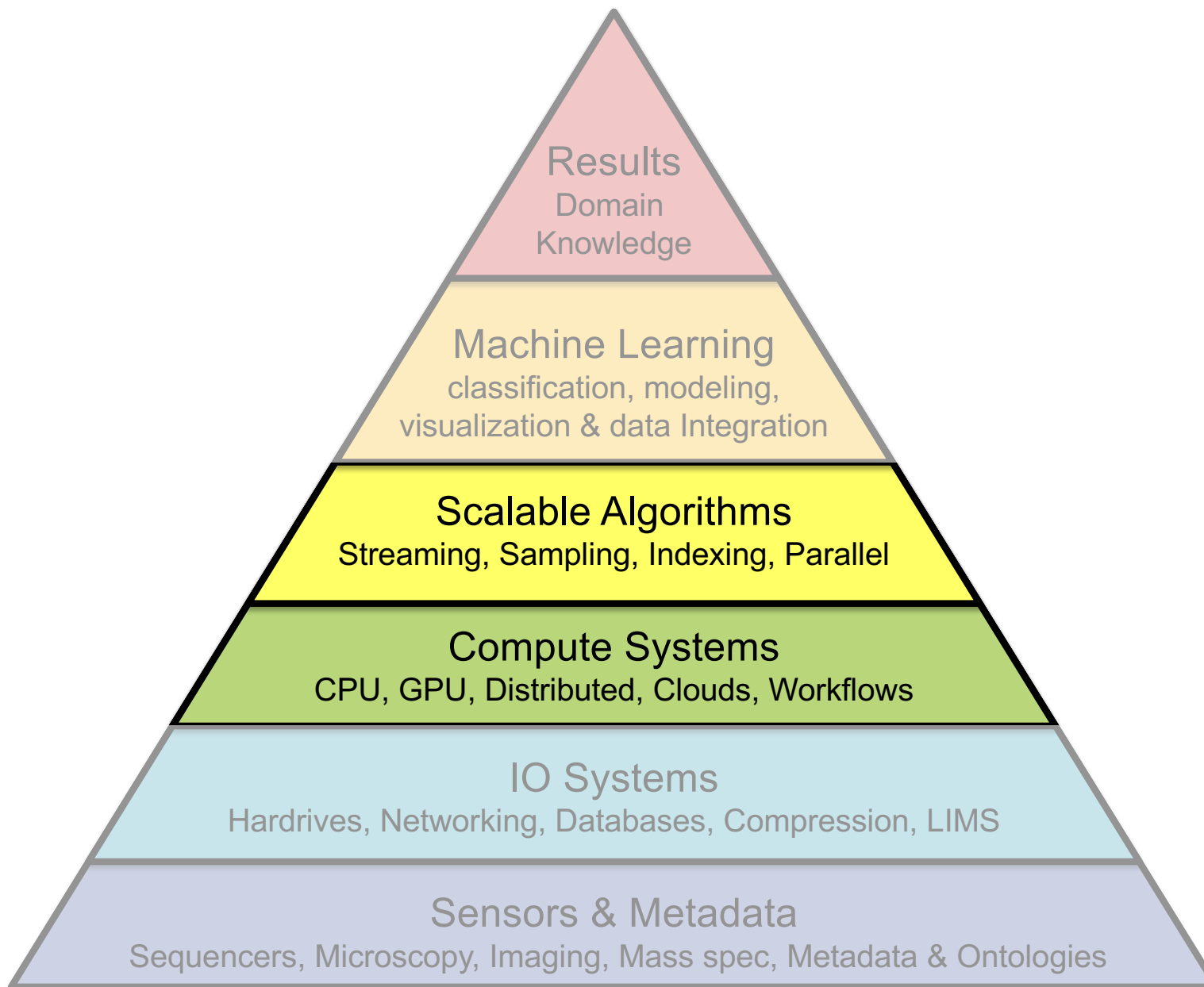


Chromosome Mapping





Biomedical Research Technologies

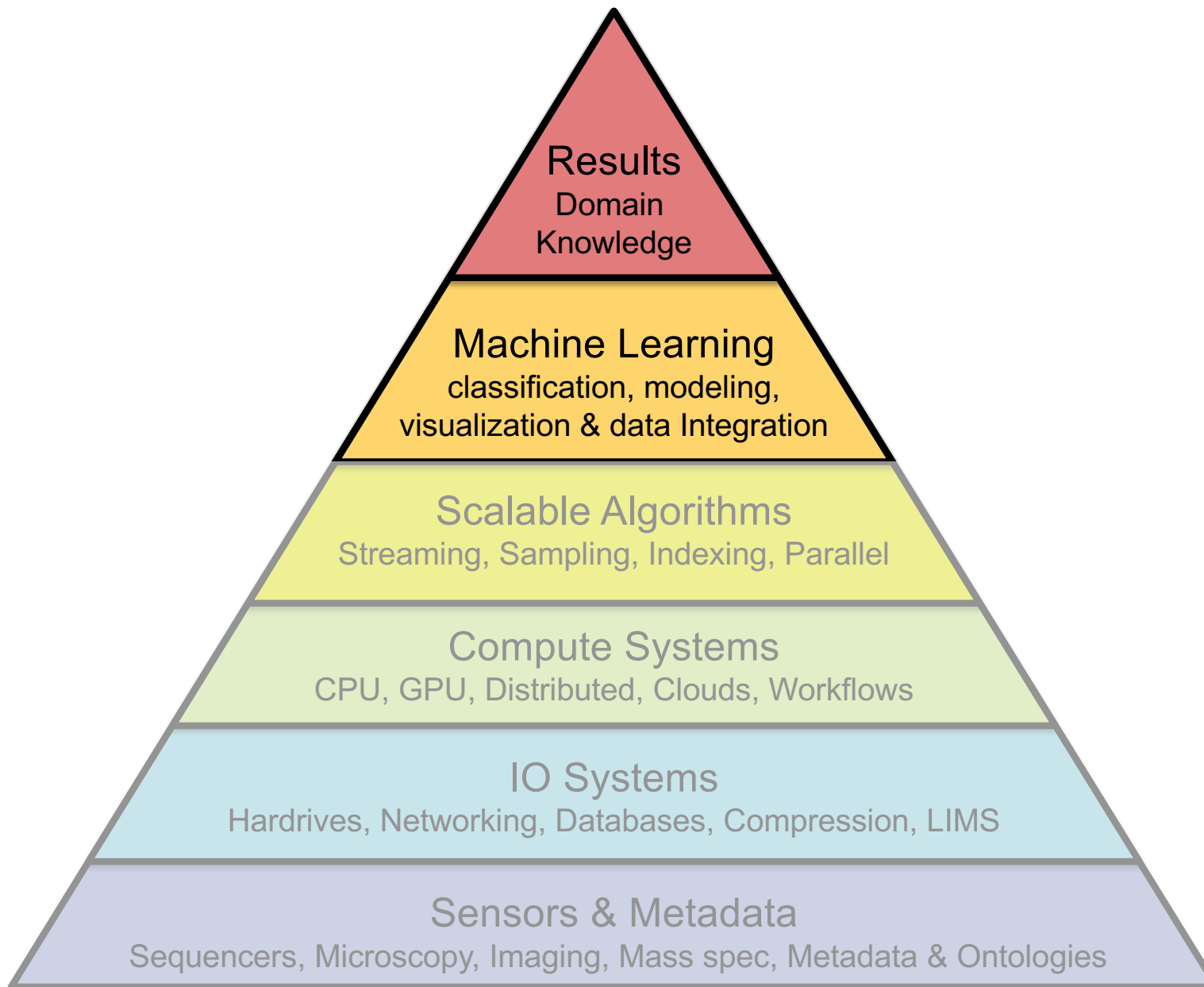


Course Topics

- 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



Biomedical Research Technologies



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?

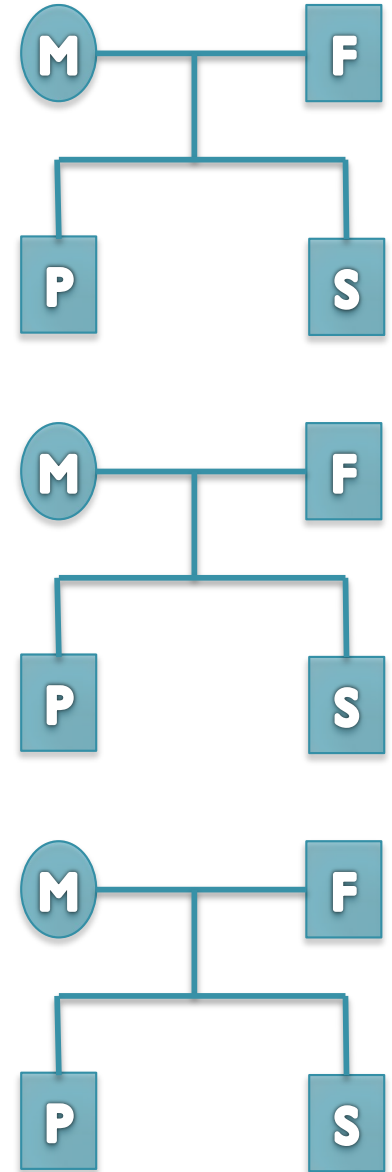
<http://www.autismspeaks.org/what-autism>

Searching for the genetic risk factors

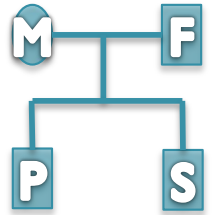
Search Strategy

- 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?



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 . . .

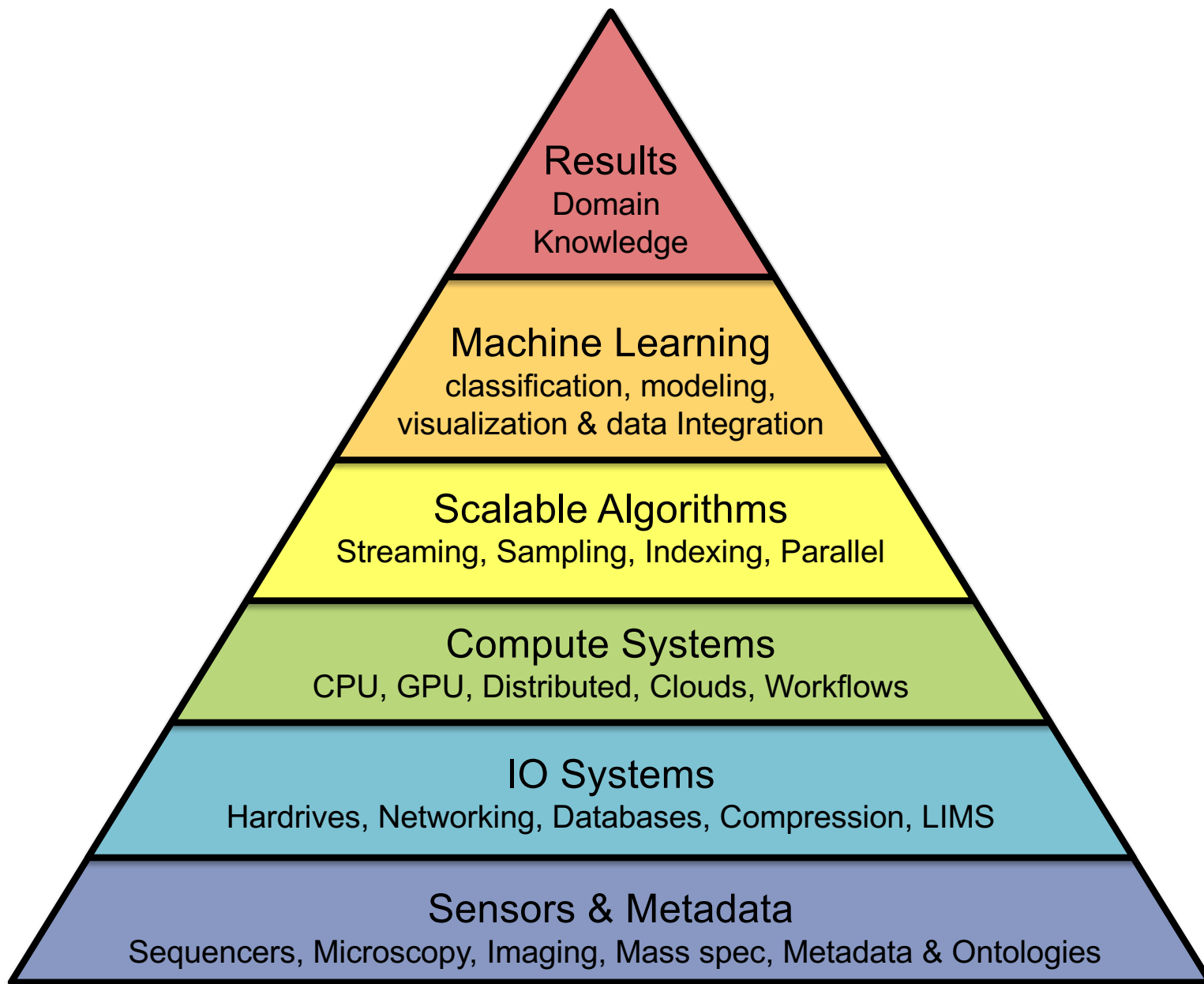
Proband(2): . . . TCAAATCCTTTTAAT****AAGAGCTGACA . . .

4bp heterozygous deletion at chr15:9352406 | 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 FMRP
 - Related to neuron development and synaptic plasticity
 - Also strong overlap with chromatin remodelers

Biomedical Research Technologies





Next Steps

1. Reflect on the magic and power of DNA 😊
2. Check out the course webpage
3. Register on Piazza & say hello!
4. Set up Dropbox for yourself!