Applied Comparative Genomics

Michael Schatz

January 30, 2018

Lecture 1: Course Overview

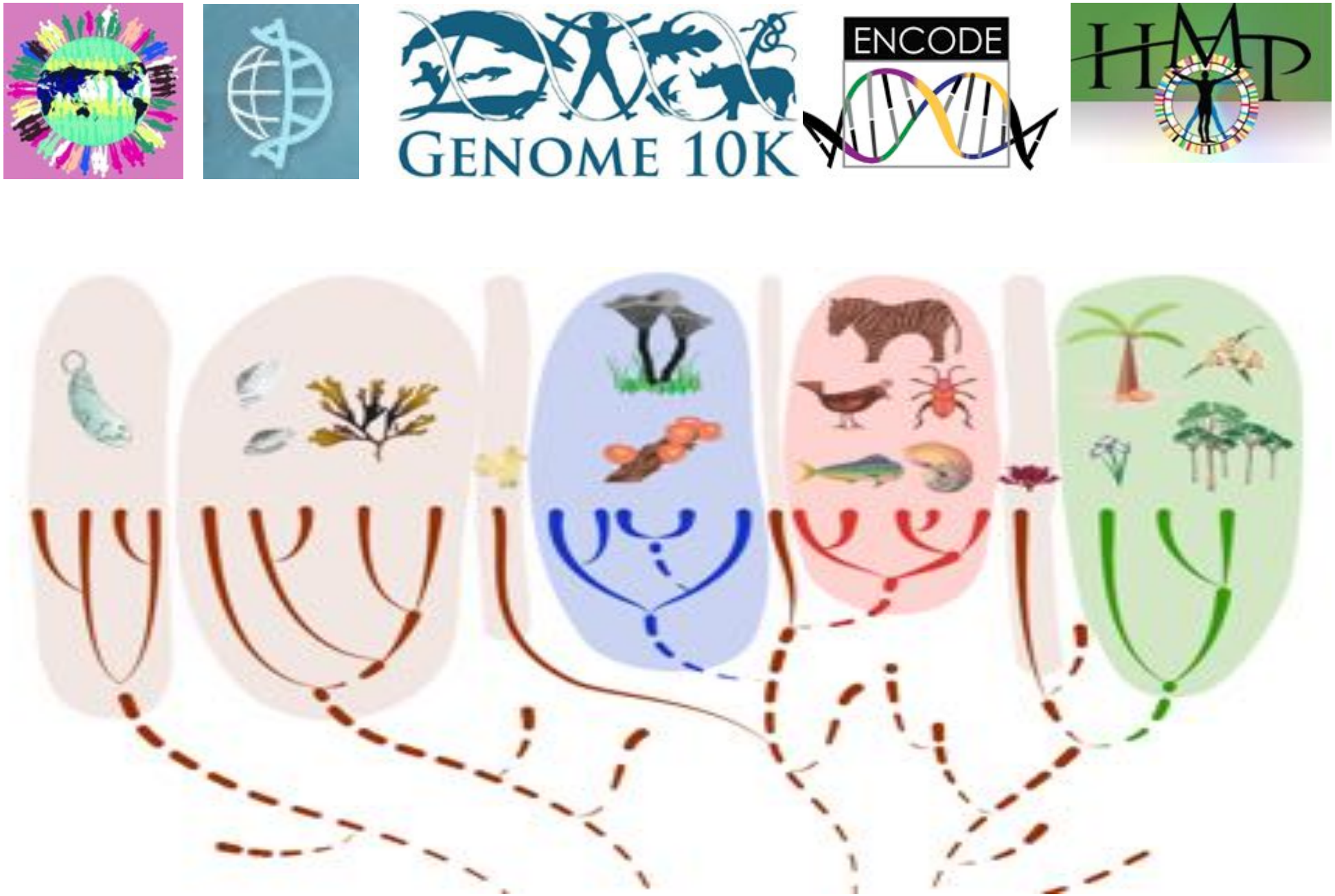
DNA: The secret of life

Your DNA, along with your environment and experiences, shapes who you are

- Height
- Hair, eye, skin color
- Broad/narrow, small/large features
- Susceptibility to disease
- Response to drug treatments
- Longevity and cognition

Physical traits tend to be strongly genetic, social characteristics tend to be strongly environmental, and everything else is a combination

Genomics across the tree of life



Unsolved Questions in Biology

- What is your genome sequence?
-
-
-
-
-
-
-
-
-
-
-
-
-
- ***Plus thousands and thousands more***

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?

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 genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

Course Webpage:	https://github.com/schatzlab/appliedgenomics2018
Course Discussions:	http://piazza.com
Class Hours:	Tues + Thurs @ 1:30p – 2:45p, Shaffer 304
Schatz Office Hours:	Tues + Thurs @ 3-4p and by appointment
Darby Office Hours:	Wed @ 4-5 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 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: Spring 2017
- https://github.com/quinlan-lab/applied-computational-genomics

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

Grading Policies

Assessments:

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

- 1 Exam: 30% In class (Tentatively 4/3)
  Assess your performance, focusing on the methods

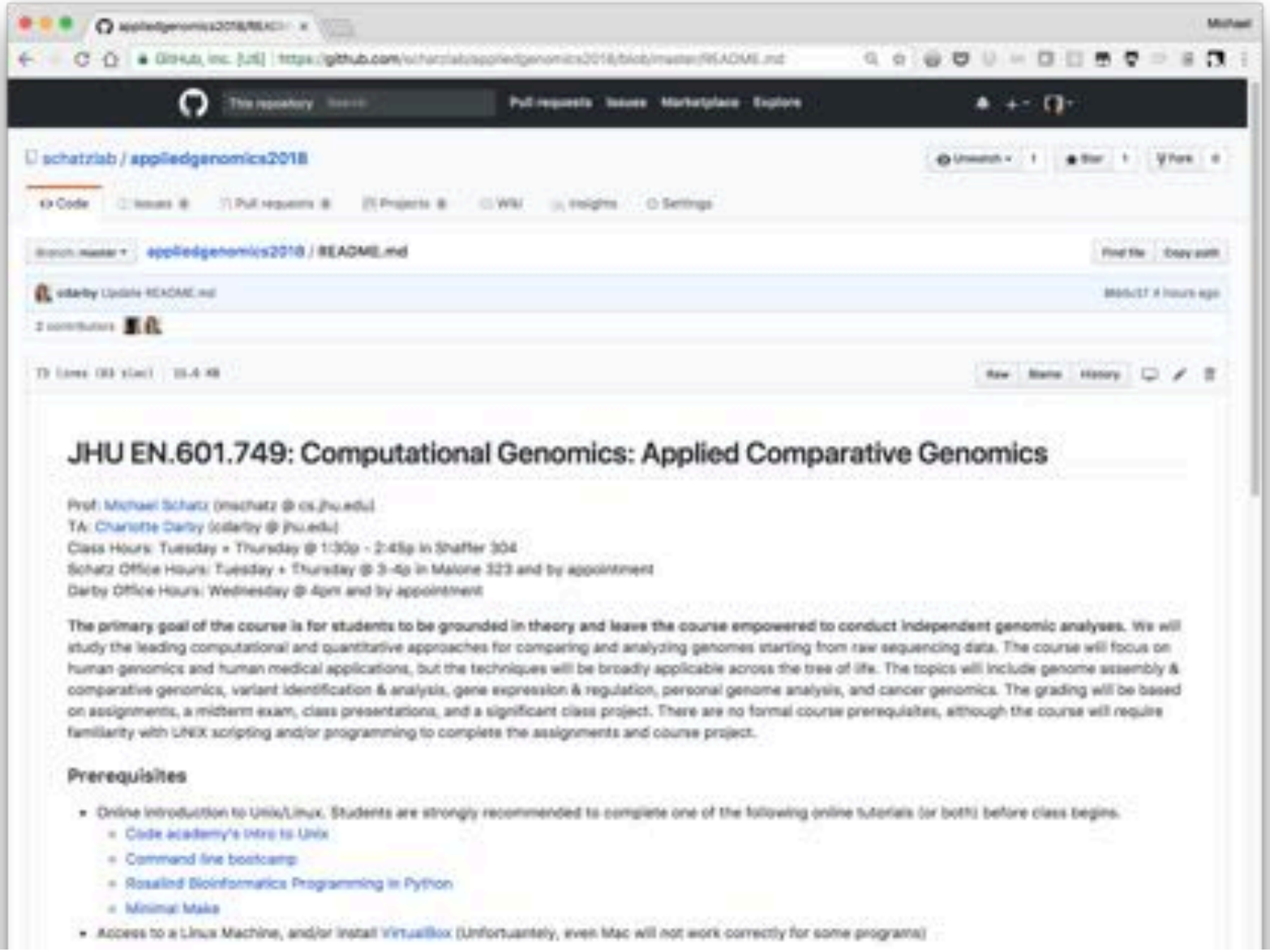
- 1 Class Project: 40% Presented last week of class
  Significant project developing a novel analysis/method

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

Policies:

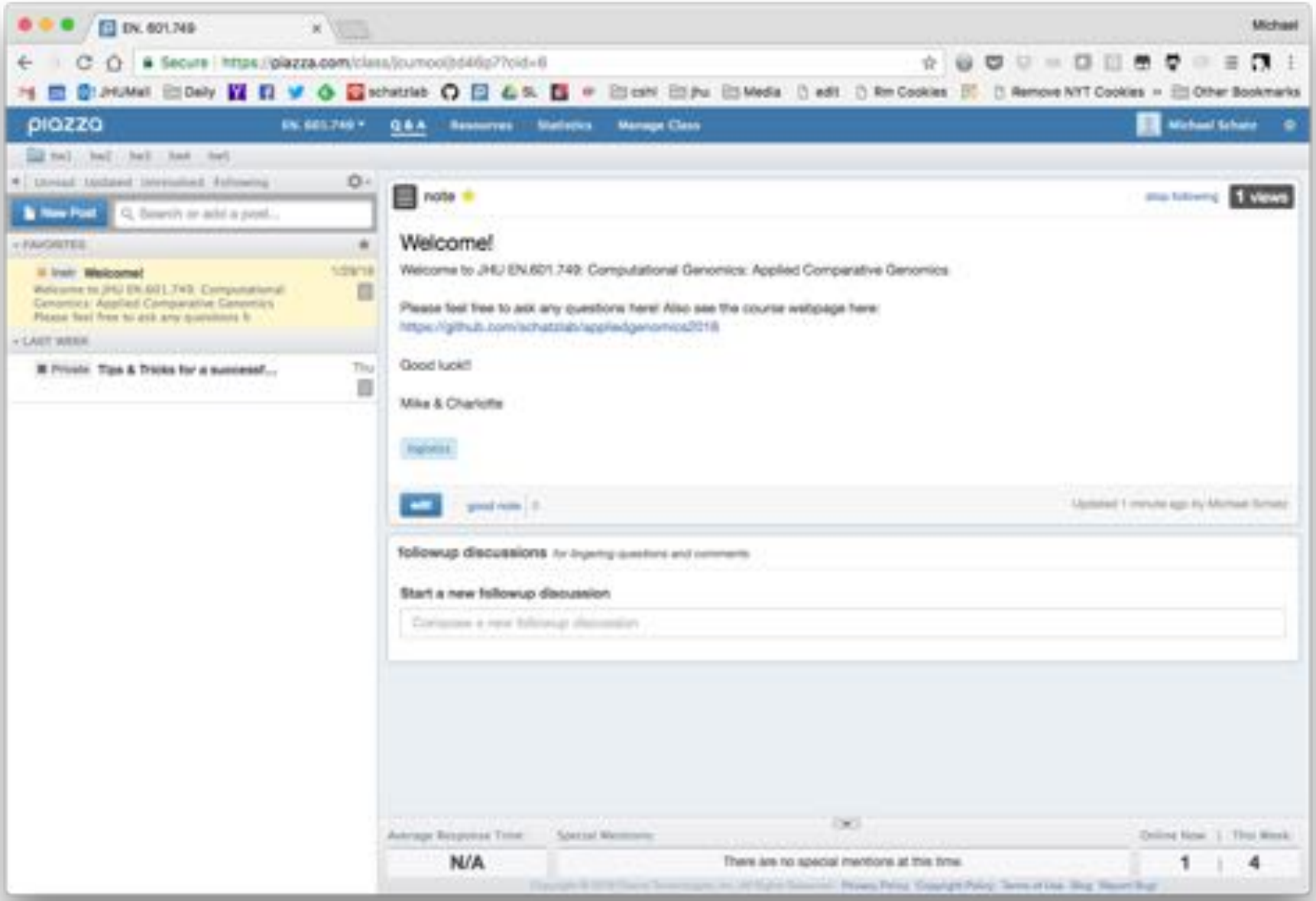
- Scores assigned relative to the highest points awarded
- Automated testing and grading of assignments
- ***Late Days:***
  - Four (4) chances to extend the deadline for assignments by 24 hours without any penalty

Course Webpage



https://github.com/schatzlab/appliedgenomics2018

Piazza



http://piazza.com/jhu/spring2018/en601749

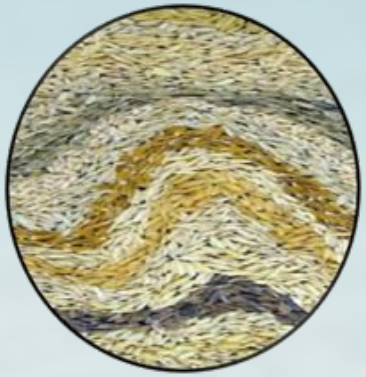
A Little About Me

Schatzlab Overview

Human Genetics

Role of mutations in disease

Feigin et al. (2017)
Fang *et al.* (2016)



Agricultural Genomics

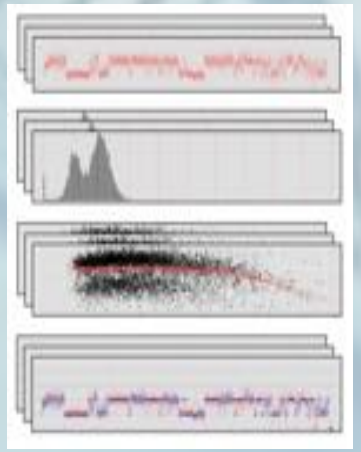
Genomes & Transcriptomes

Lemmon *et al.* (2016)
Ming *et al.* (2015)

Algorithmics & Systems Research

Ultra-large scale biocomputing

Fang et al. (2018)
Stevens *et al.* (2015)



Biotechnology Development

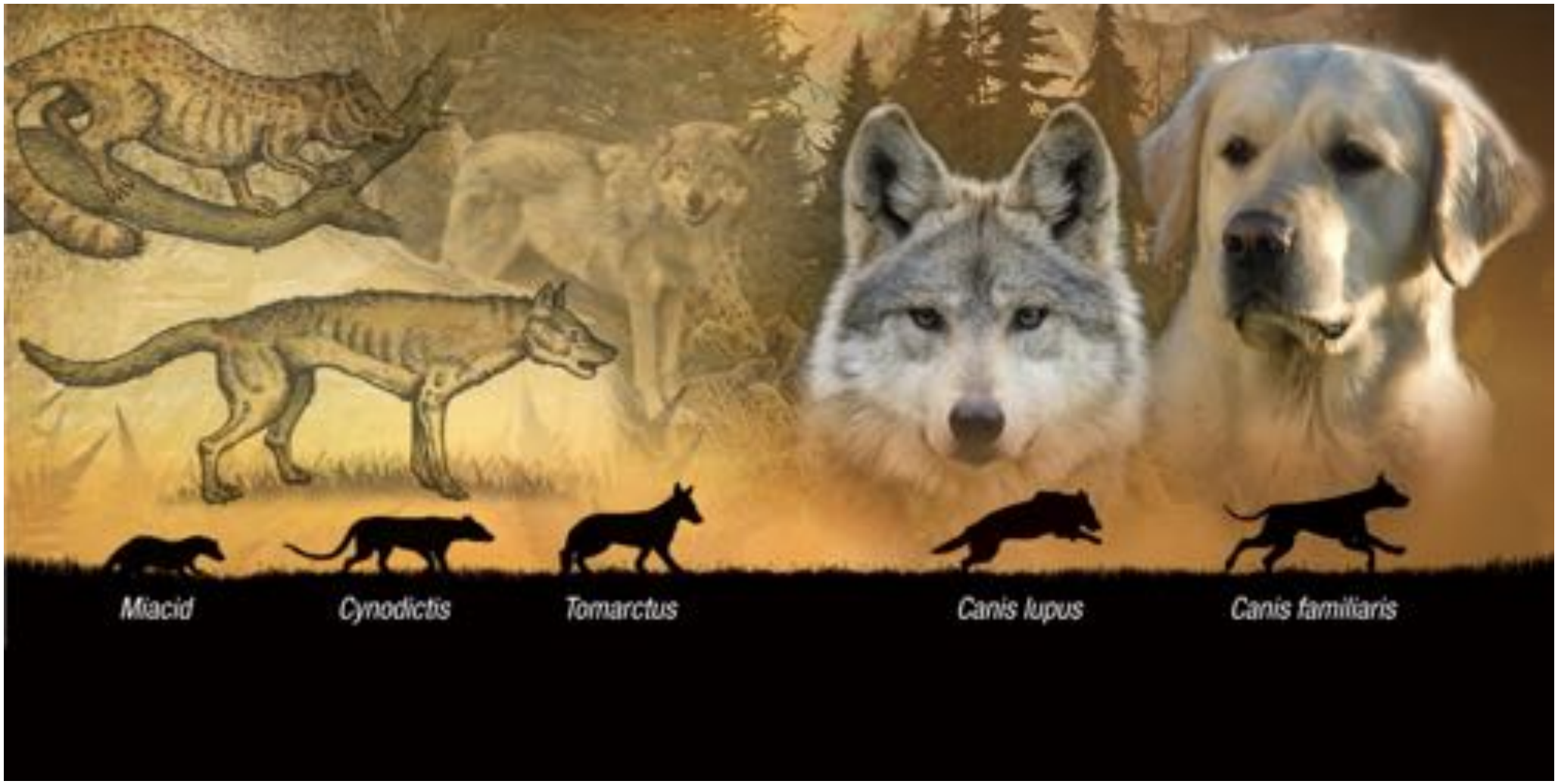
Single Cell + Single Molecule Sequencing

Chin *et al.* (2016)
Garvin *et al.* (2015)

Earliest Genomics

Any Guesses?

Earliest Genomics

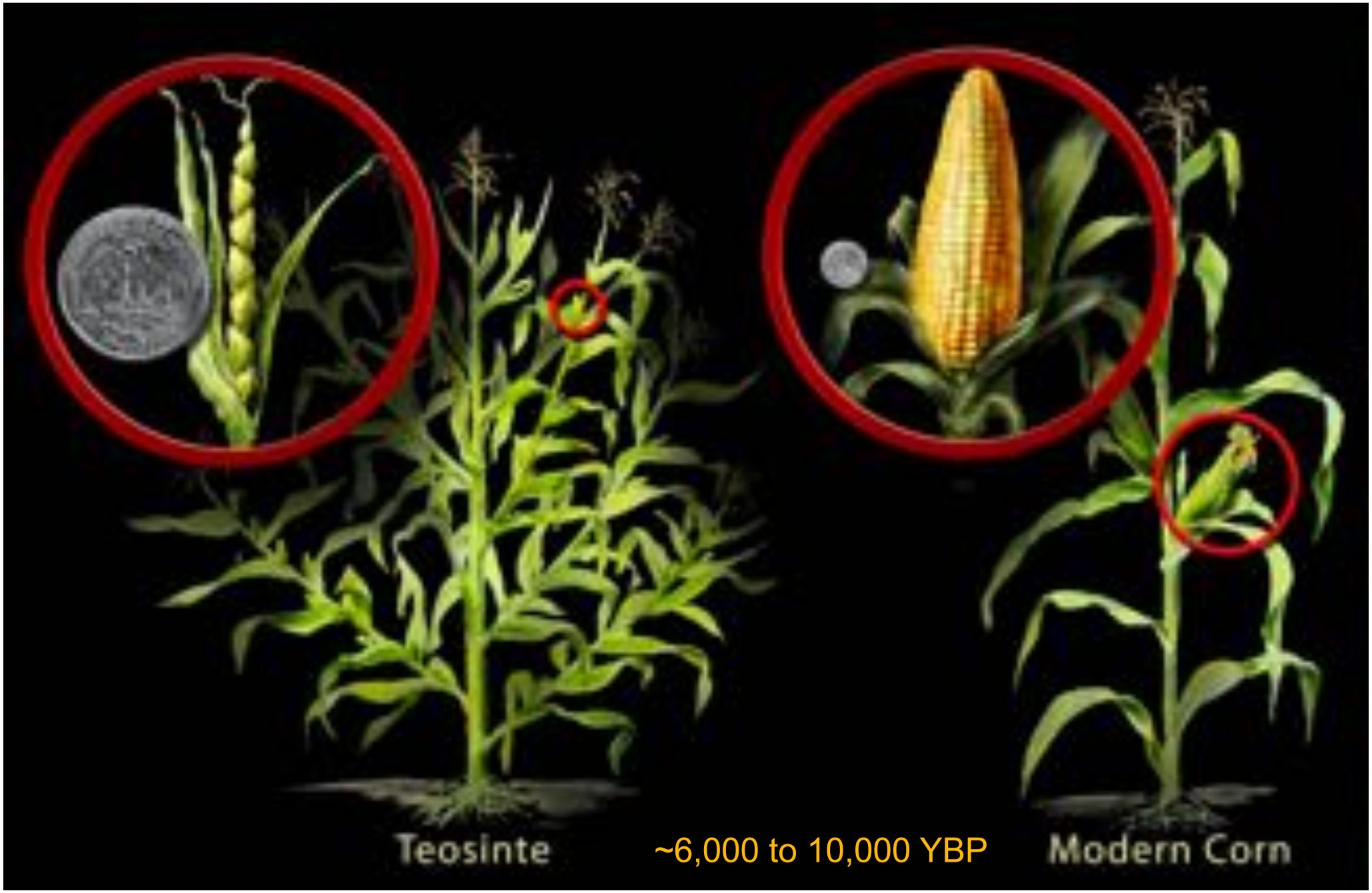


15,000 to 35,000 YBP

Earliest Genomics

~1,000 to 10,000 YBP

Earliest Genomics



Discovery of Chromosomes

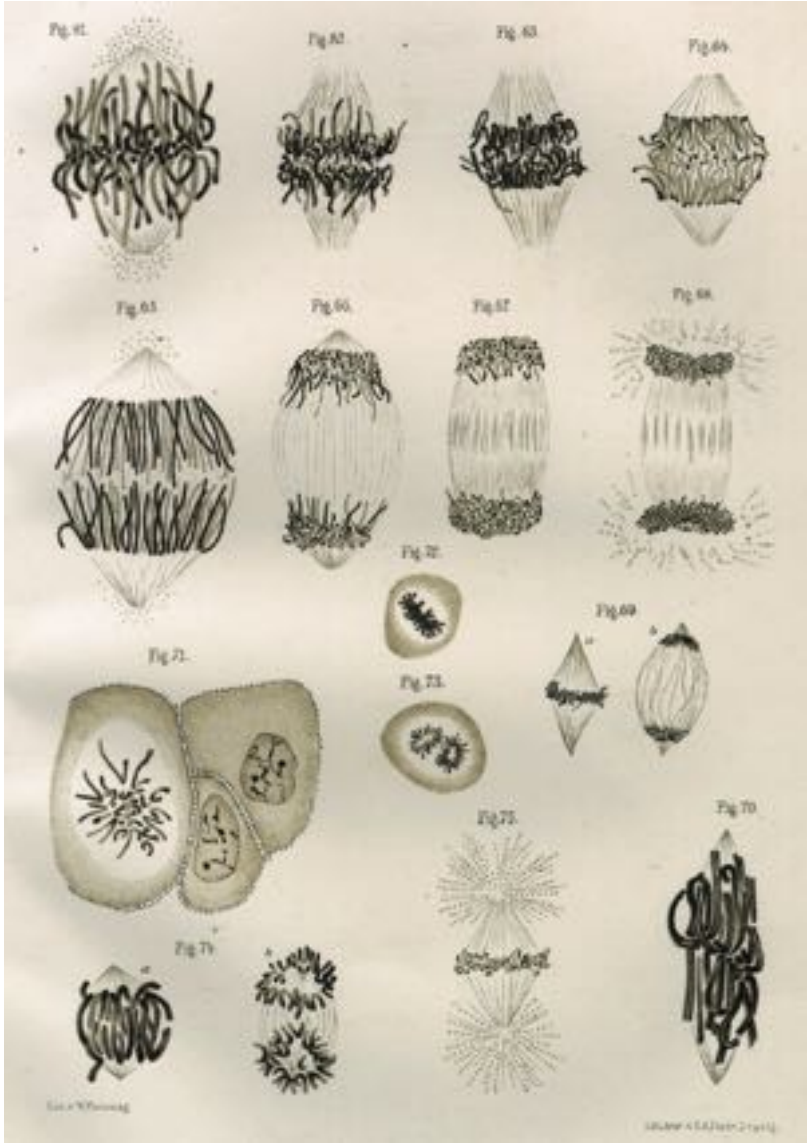
By the mid-1800s, microscopes were powerful enough to observe the presence of unusual structures called “chromosomes” that seemed to play an important role during cell division.

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

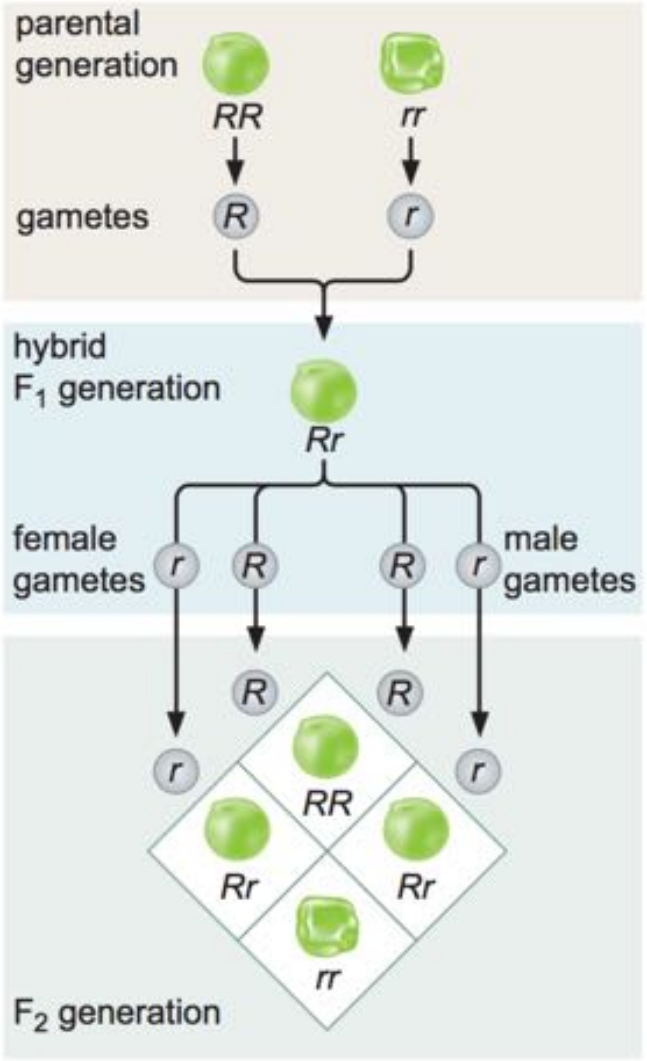


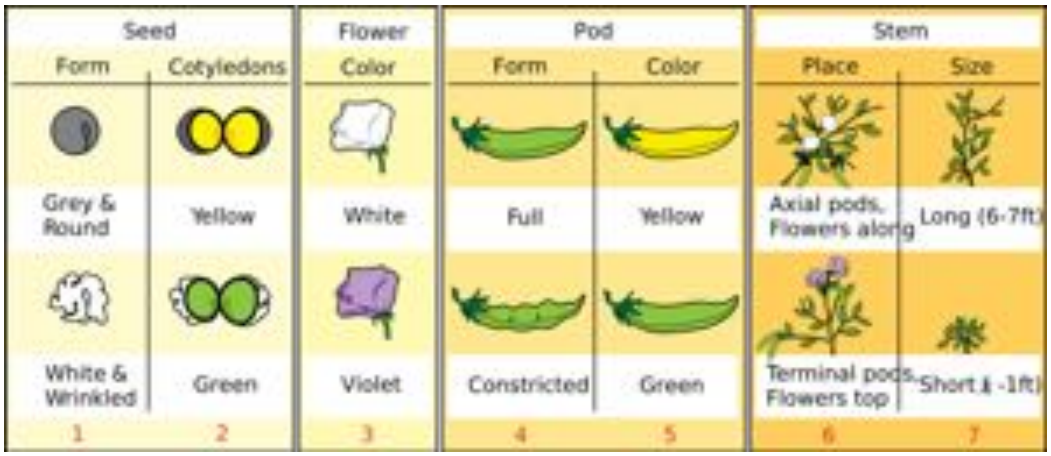
Drawing of mitosis by Walther Flemming.
Flemming, W. Zellsubstanz, Kern und Zelltheilung (F. C. W. Vogel, Leipzig, 1882).

The “first” quantitative biologist

Any Guesses?

Laws of Inheritance





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).

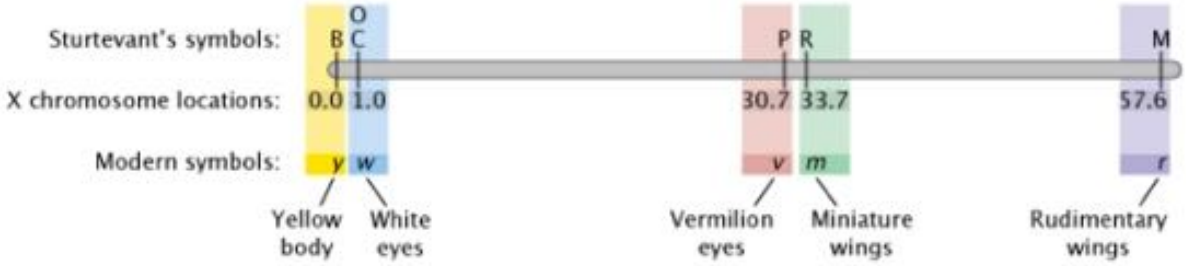
The first genetic map

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

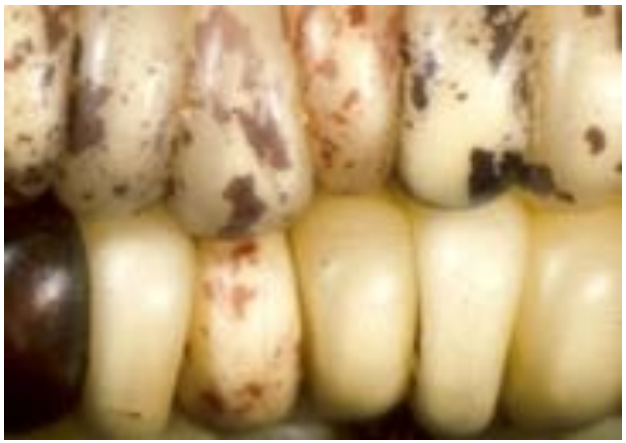


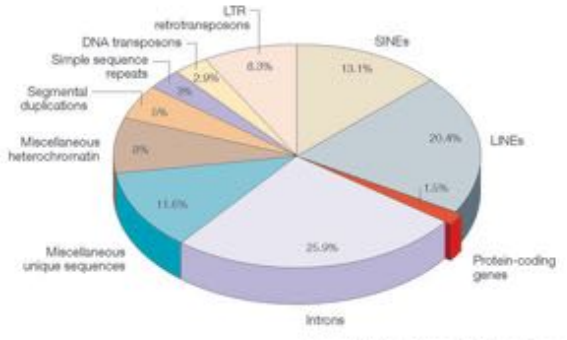
The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) *Journal of Experimental Zoology*, 14: 43-59

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

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.

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[2]. They kindly made their manuscript available to us in advance of publication. Their model consists of three interwined 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.

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[3] 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[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.

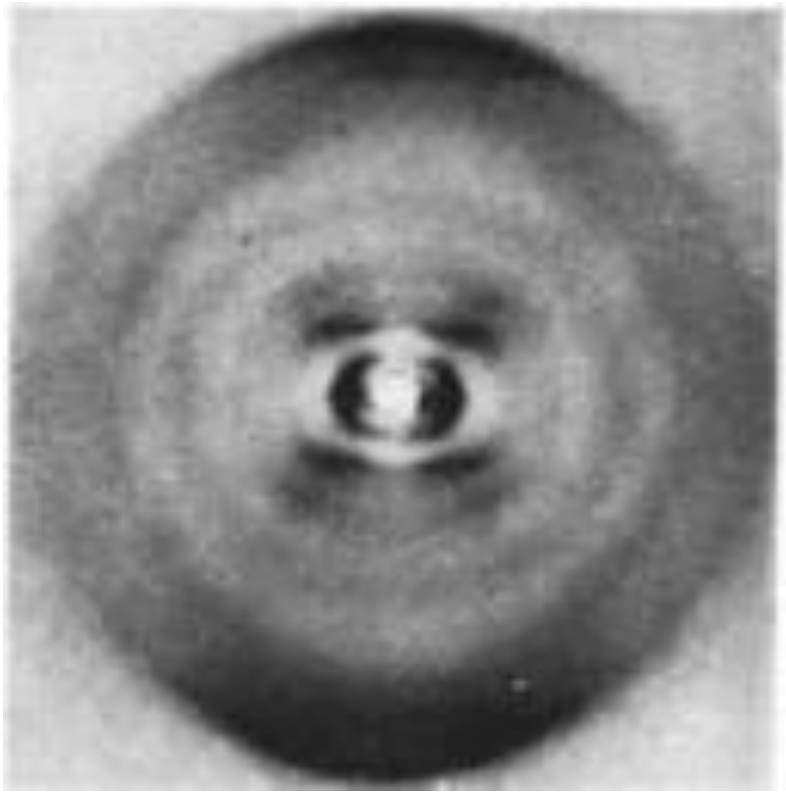
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[4,5] 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-

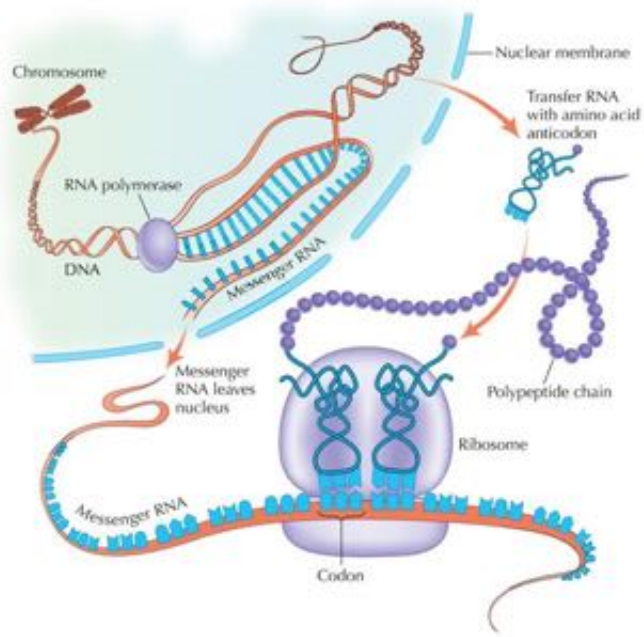
Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid

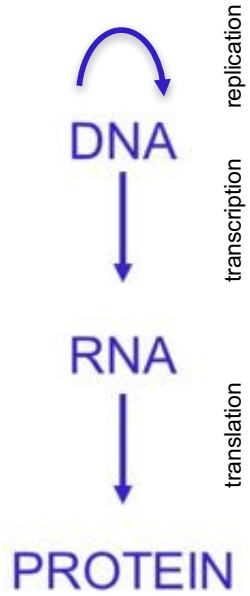
Watson JD, Crick FH (1953). *Nature* 171: 737–738.

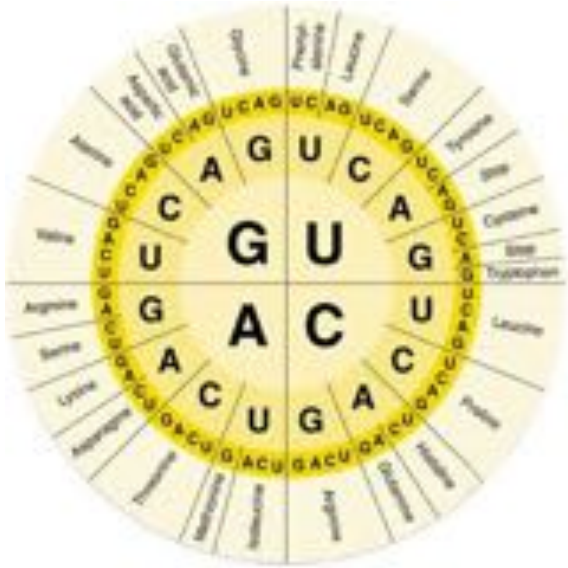
Nobel Prize in Physiology or Medicine in 1962

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"

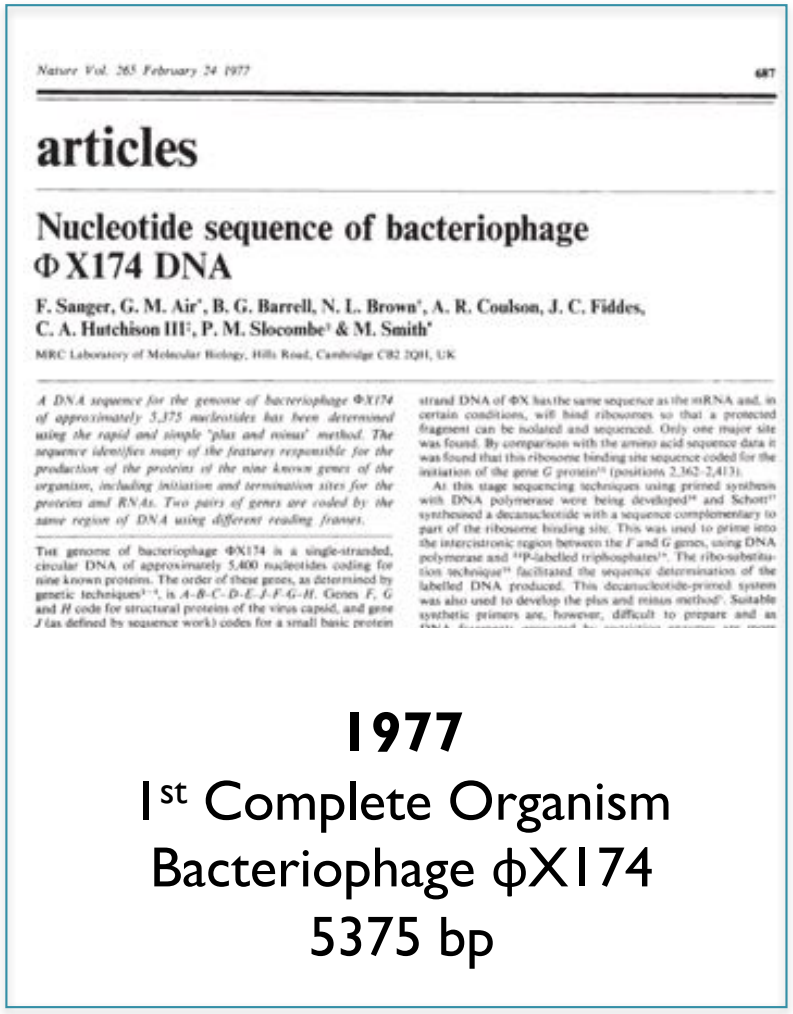






On Protein Synthesis
Crick, F.H.C. (1958). Symposia of the Society for Experimental Biology pp. 138–163.

Milestones in Genomics: Zeroth Generation Sequencing



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 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 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 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 ΦX174 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[1–3], is *A–B–C–D–E–J–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 ΦX 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* protein[10] (positions 2,362–2,413).

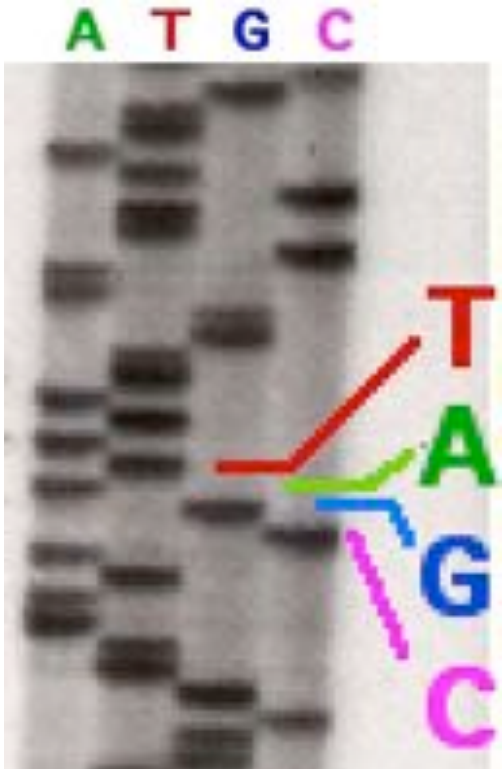
At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[14] and Schott[15] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[16]. The ribo-substitution technique[17] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[5]. Suitable synthetic primers are, however, difficult to prepare and as

1977

1st Complete Organism

Bacteriophage ϕX174

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 φX174 DNA

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

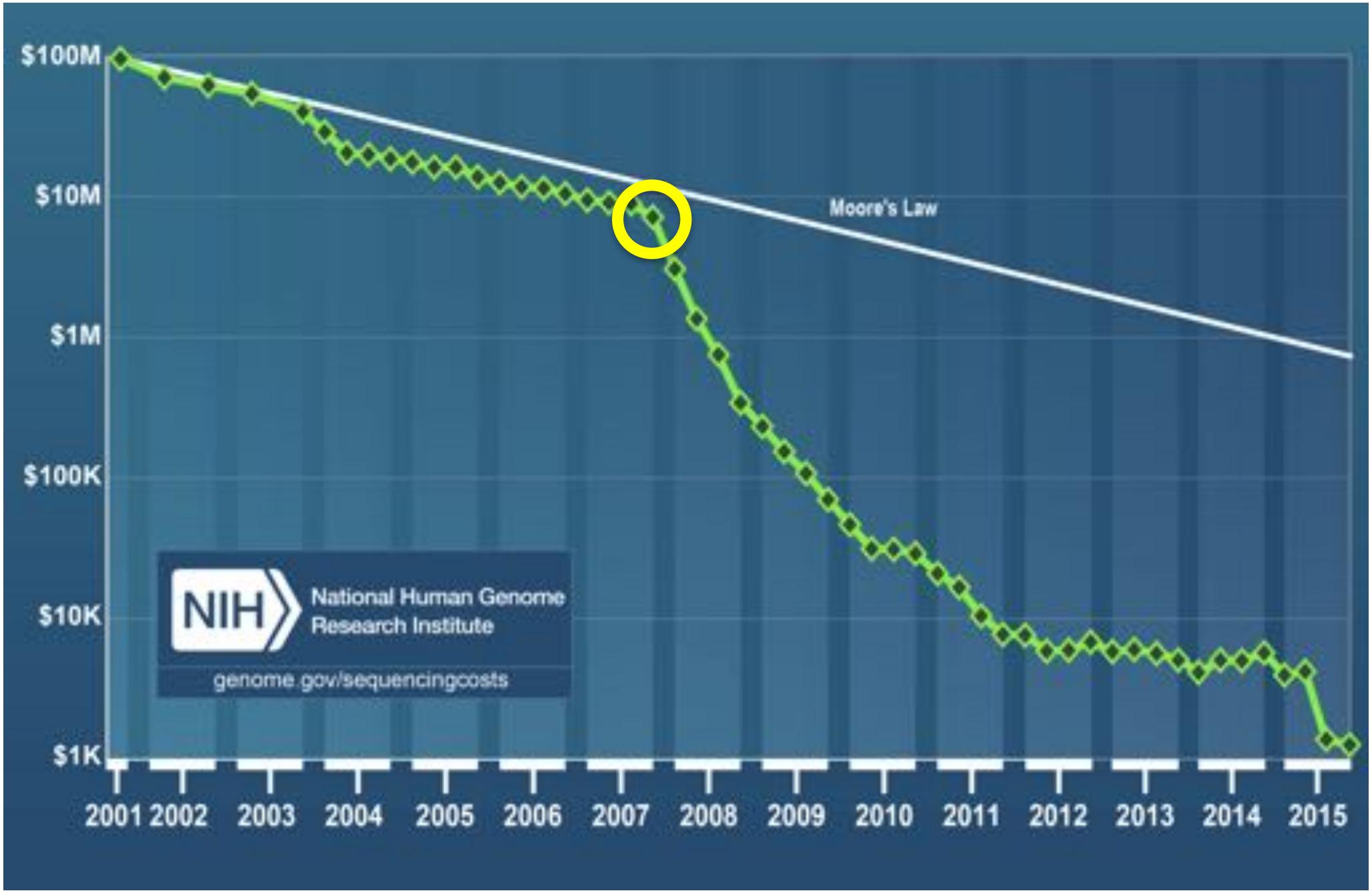
Nobel Prize in Chemistry in 1980

Milestones in DNA Sequencing

(TIGR/Celera, 1995-2001)

The most wondrous map…

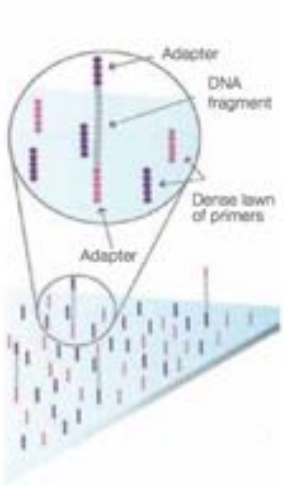
Cost per Genome



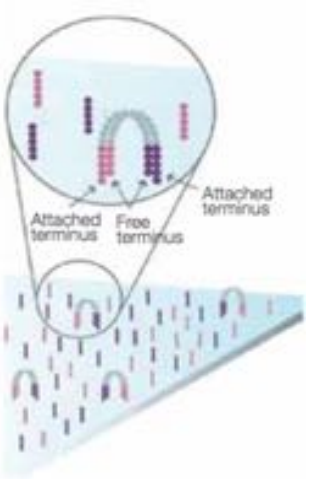
Second Generation Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

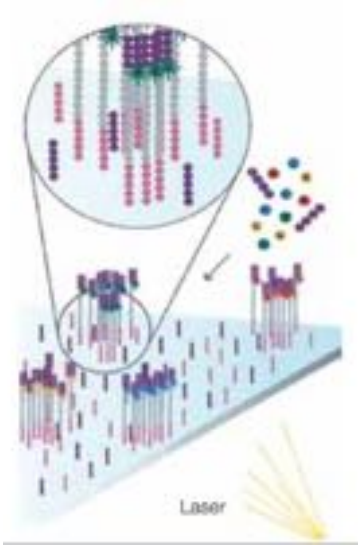
>60Gbp / day



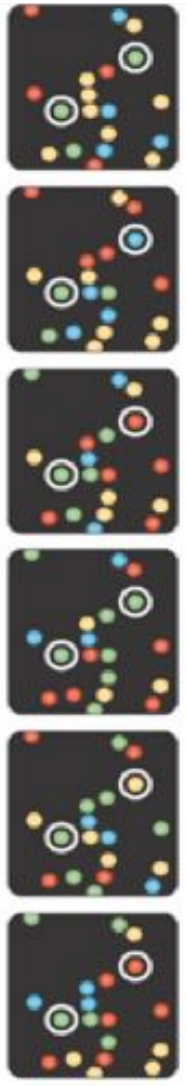
1. Attach



2. Amplify



3. Image



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

HiSeq X Ten

BROAD INSTITUTE

GARVAN INSTITUTE

NEW YORK GENOME CENTER

deCODE genetics

MACROGEN

Axeq Technologies

Novogene 诺禾致源

药明康德 WuXi PharmaTech

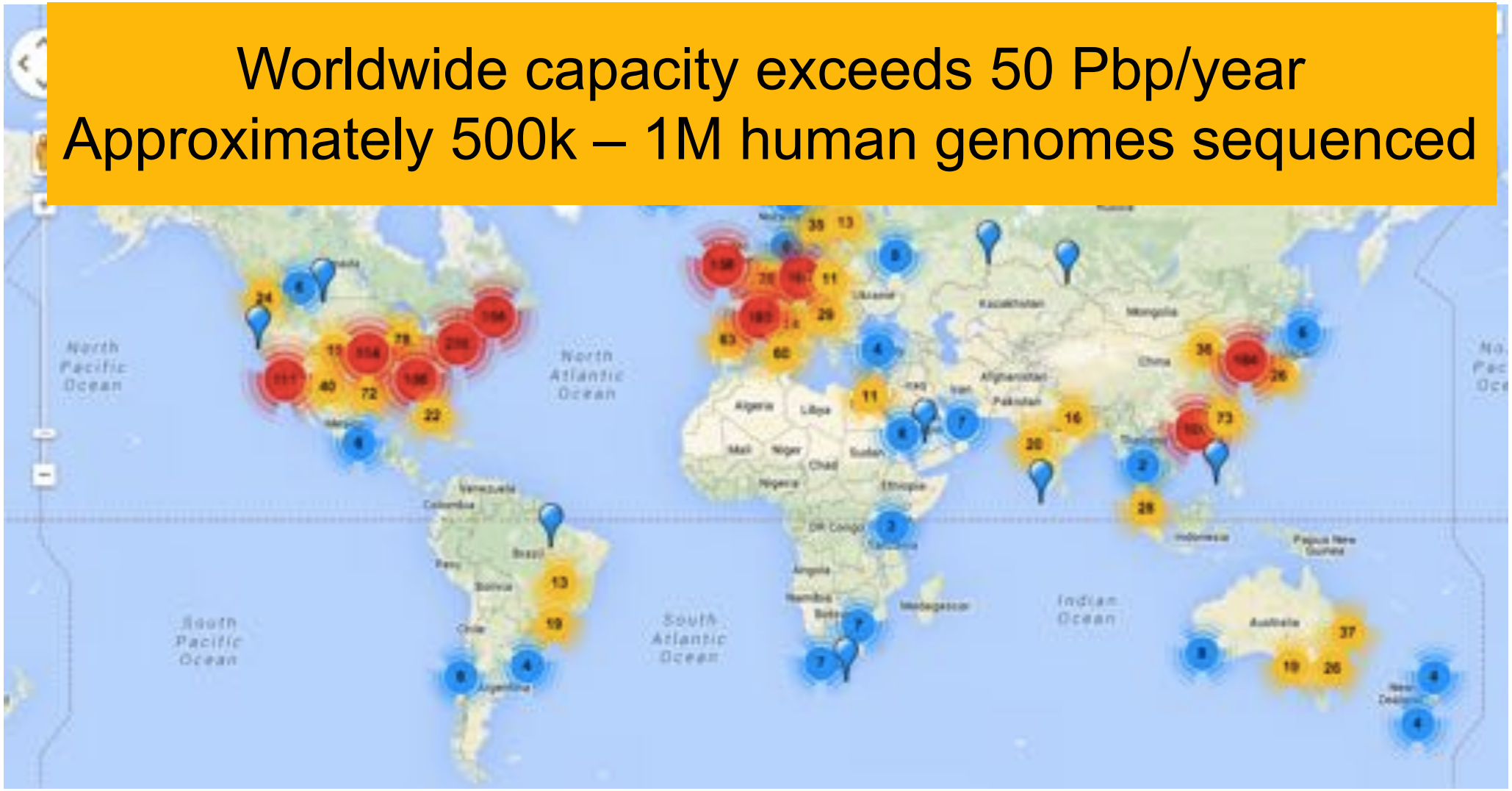
wellcome trust sanger institute

HUMAN LONGEVITY, INC.

320 genomes per week / 18,000 genomes per year
$1000 per genome / ~$10 M per instrument

Sequencing Centers

Worldwide capacity exceeds 50 Pbp/year
Approximately 500k – 1M human genomes sequenced



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

How much is a petabyte?

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

*Technically a kilobyte is 2^{10} and a petabyte is 2^{50}

How much is a petabyte?

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

X

10,000 Genomes

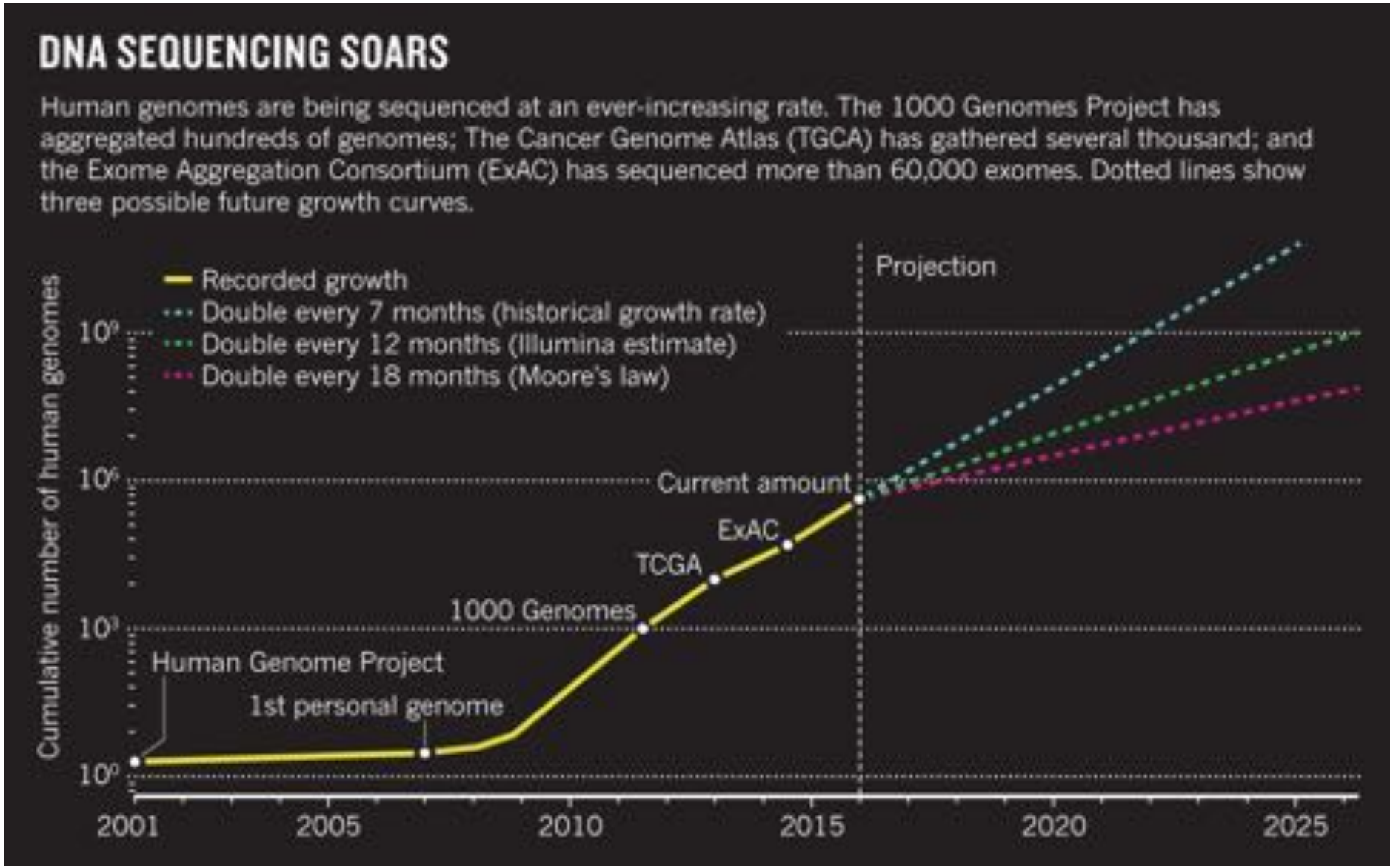
=

1PB Data
200,000 DVDs

787 feet of DVDs
~1/6 of a mile tall

500 2 TB drives
$100k

Sequencing Capacity



Big Data: Astronomical or Genomical?

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

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

=

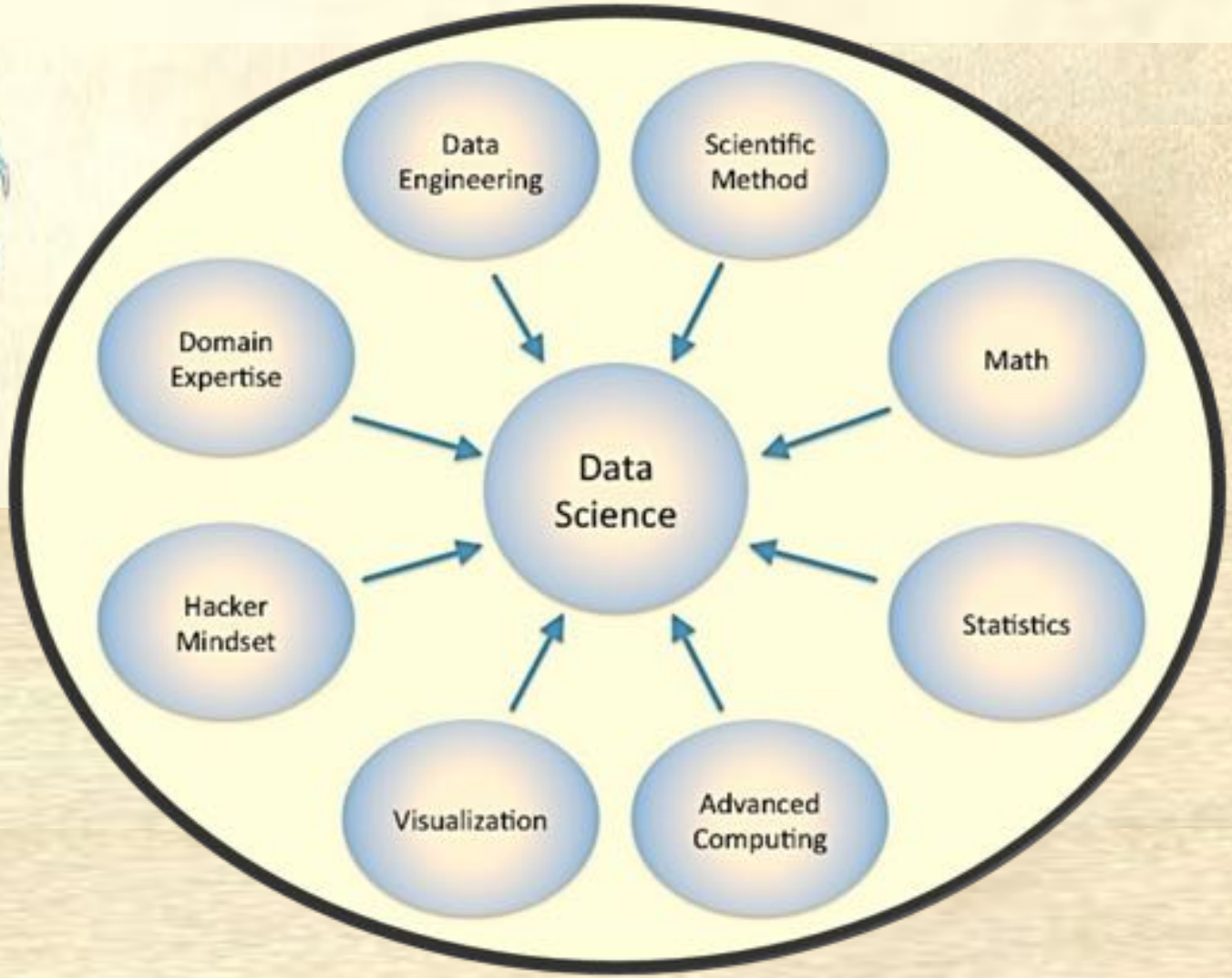
1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

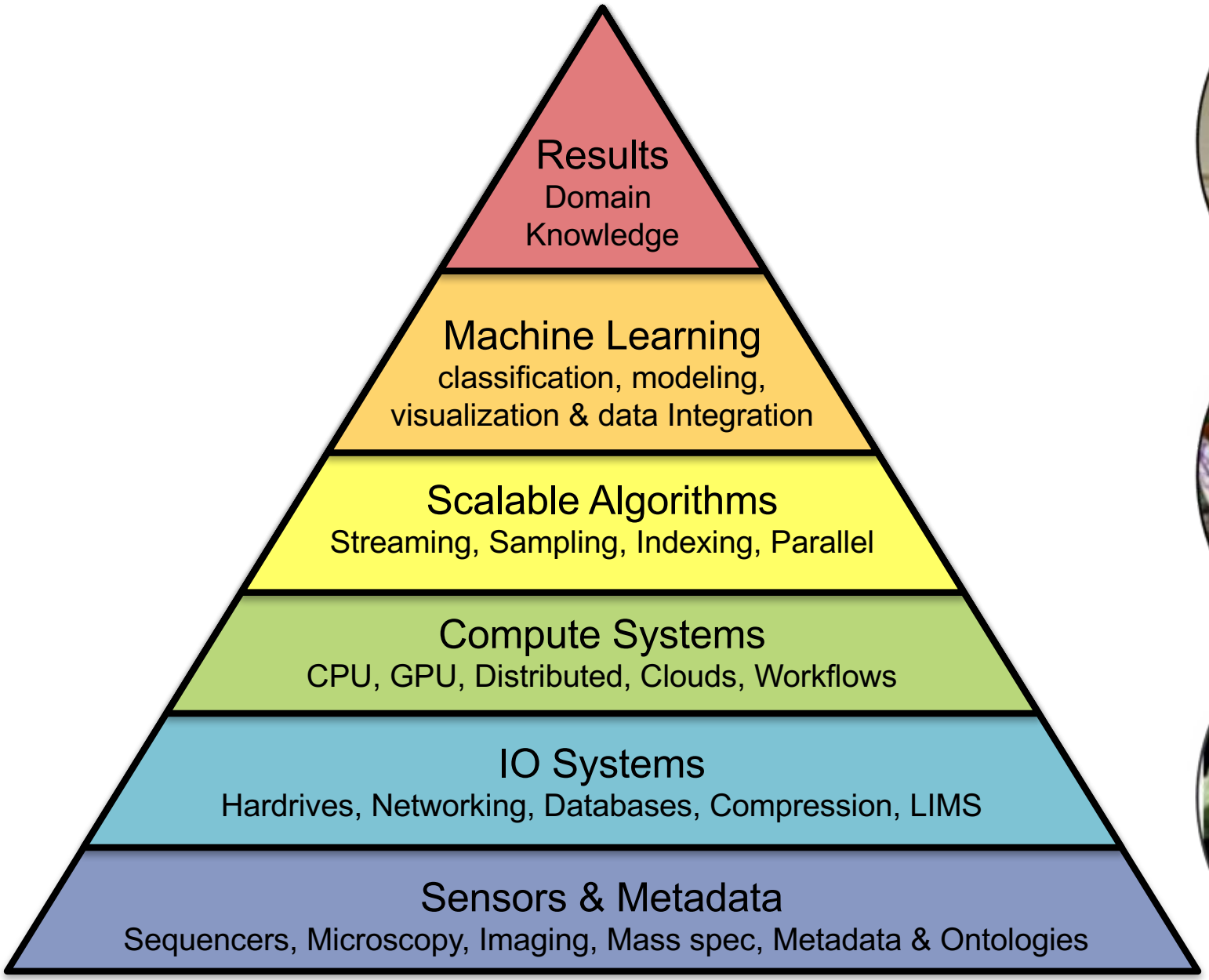
Both currently ~100Pb
And growing exponentially

Who is a Data Scientist?

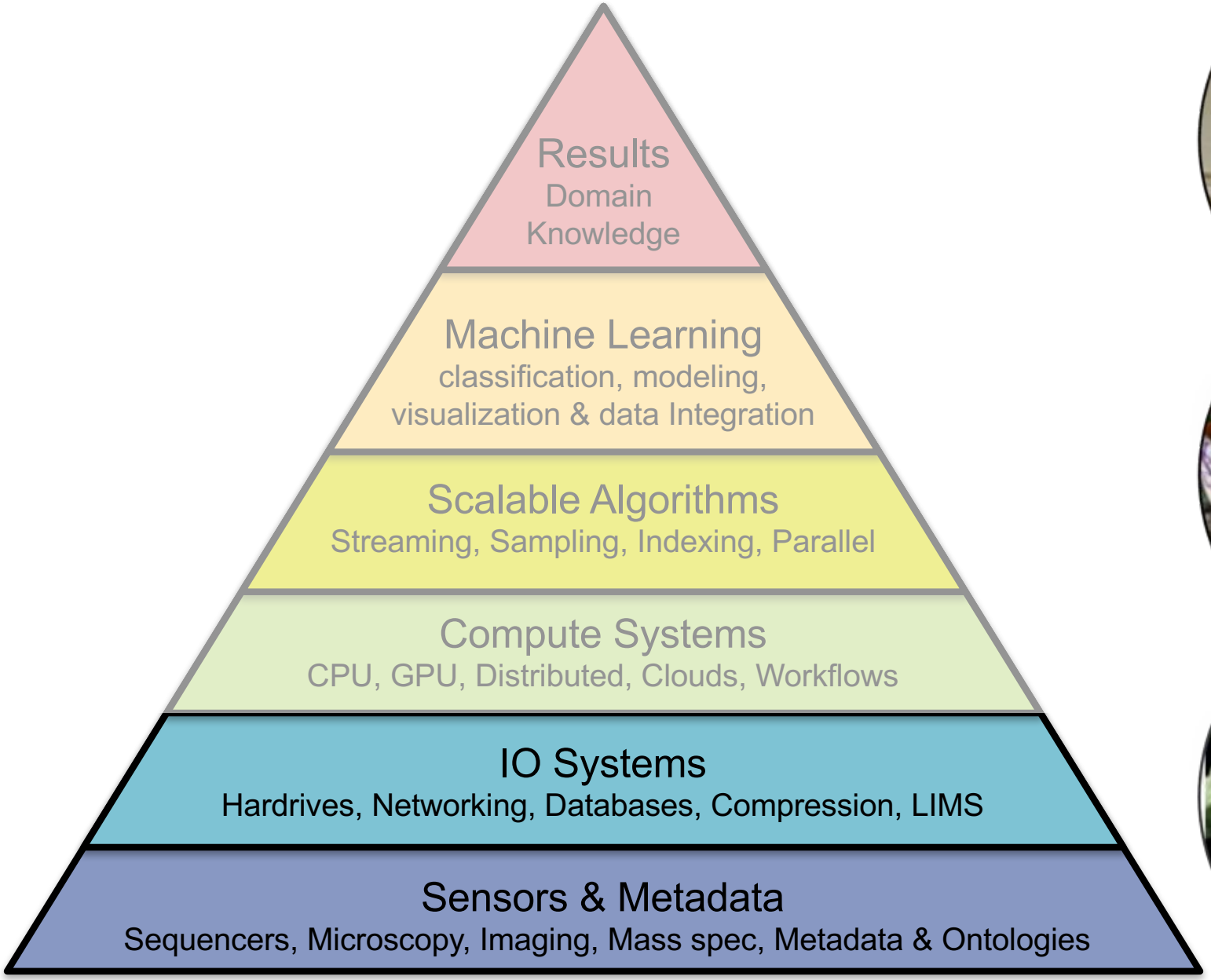


http://en.wikipedia.org/wiki/Data_science

Comparative Genomics Technologies



Comparative Genomics Technologies

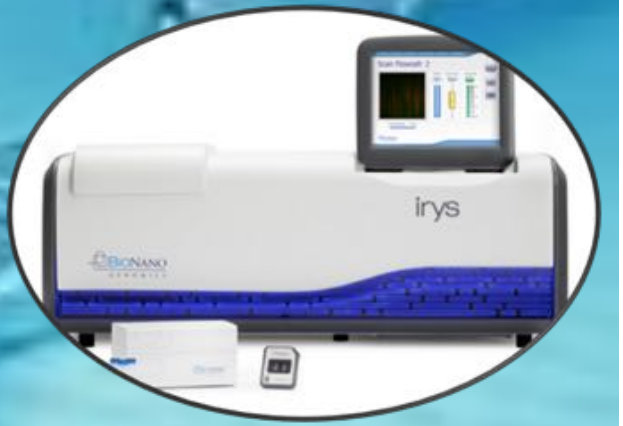


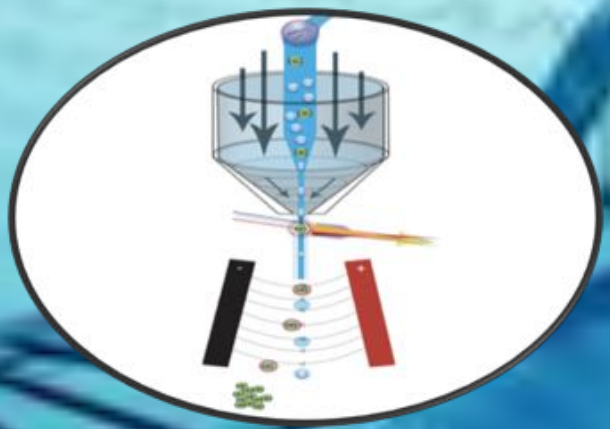
Genomics Arsenal in the year 2018

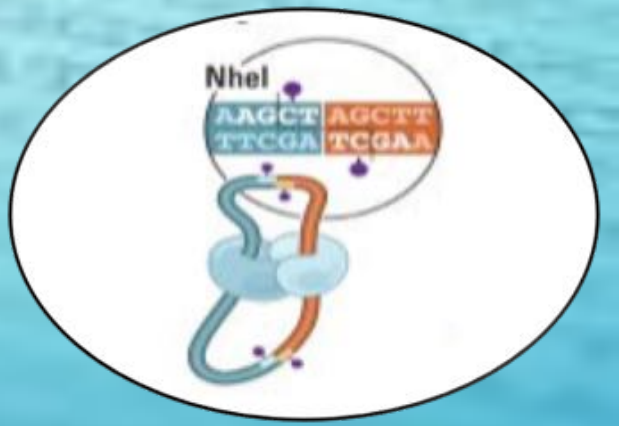
Sample Preparation

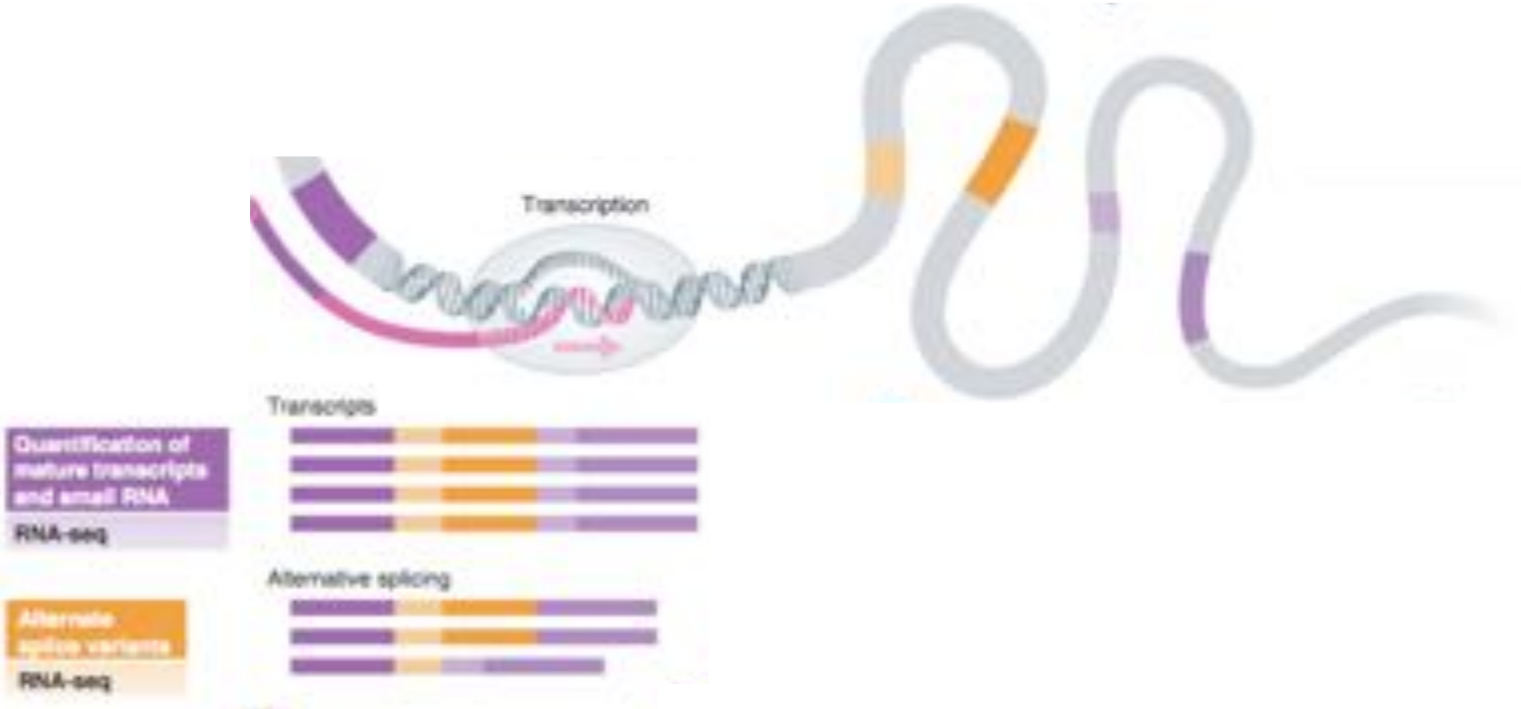
Sequencing

Chromosome Mapping

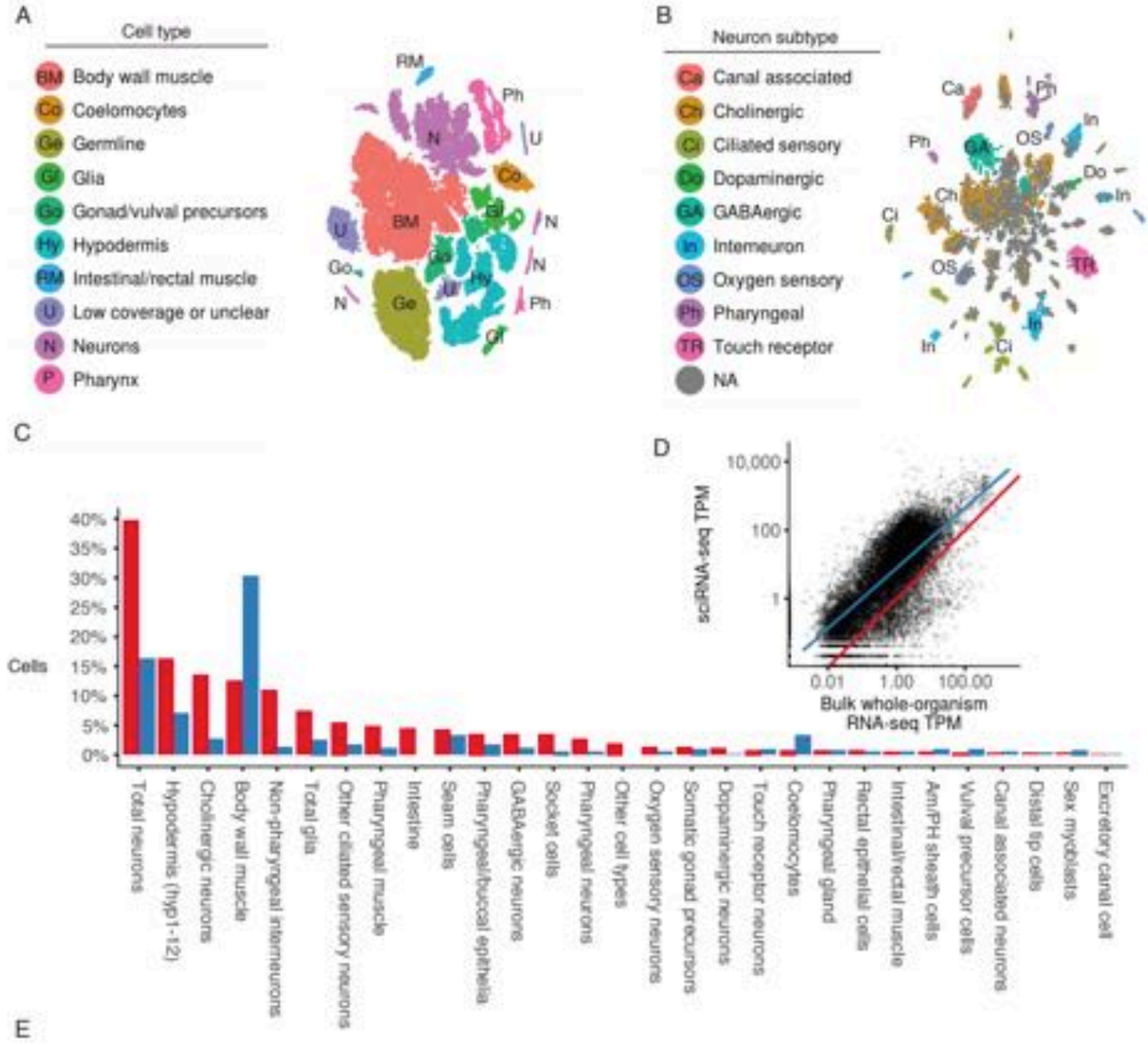






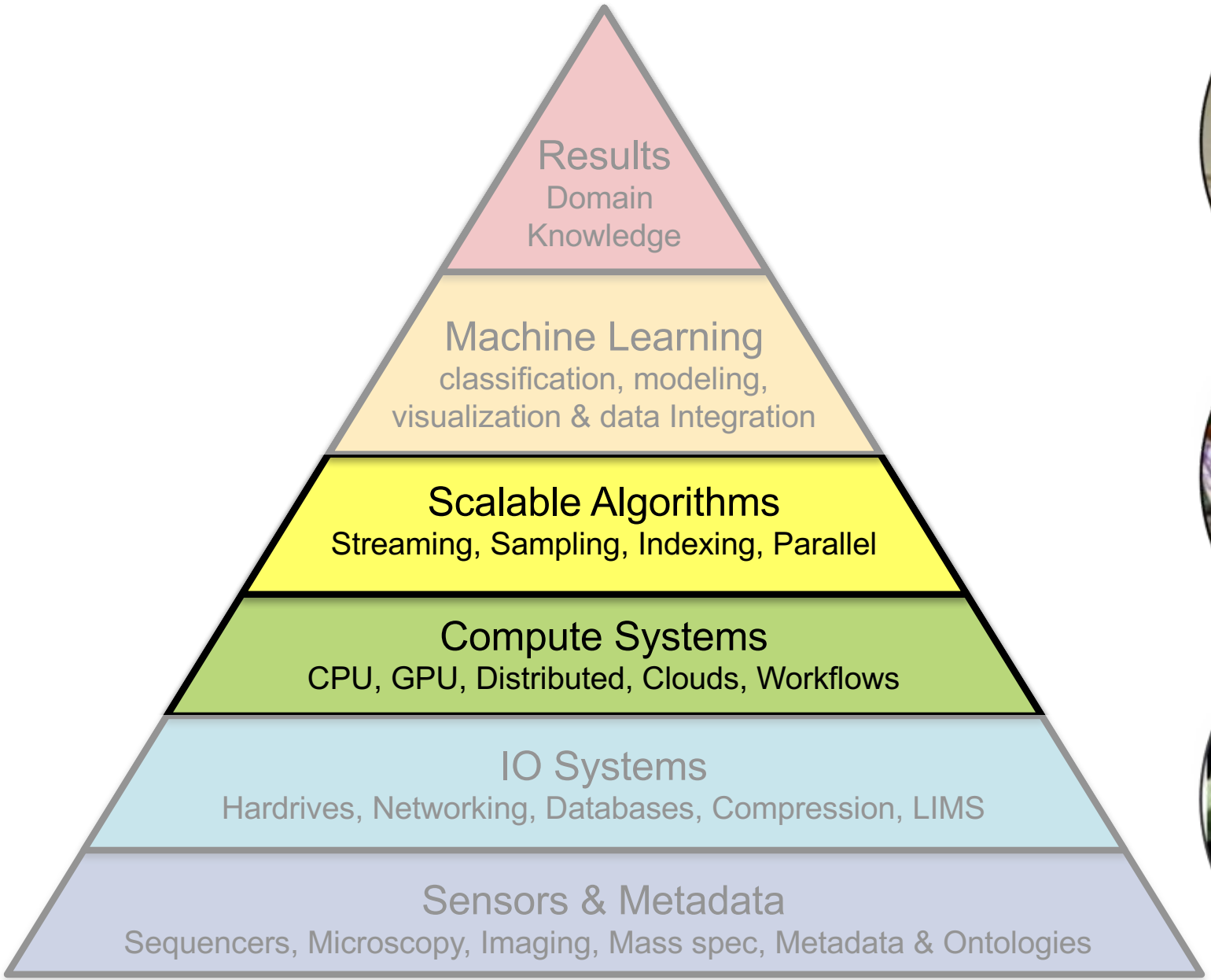


Soon et al., Molecular Systems Biology, 2013



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

Comparative Genomics Technologies



Sequencing Centers 2018

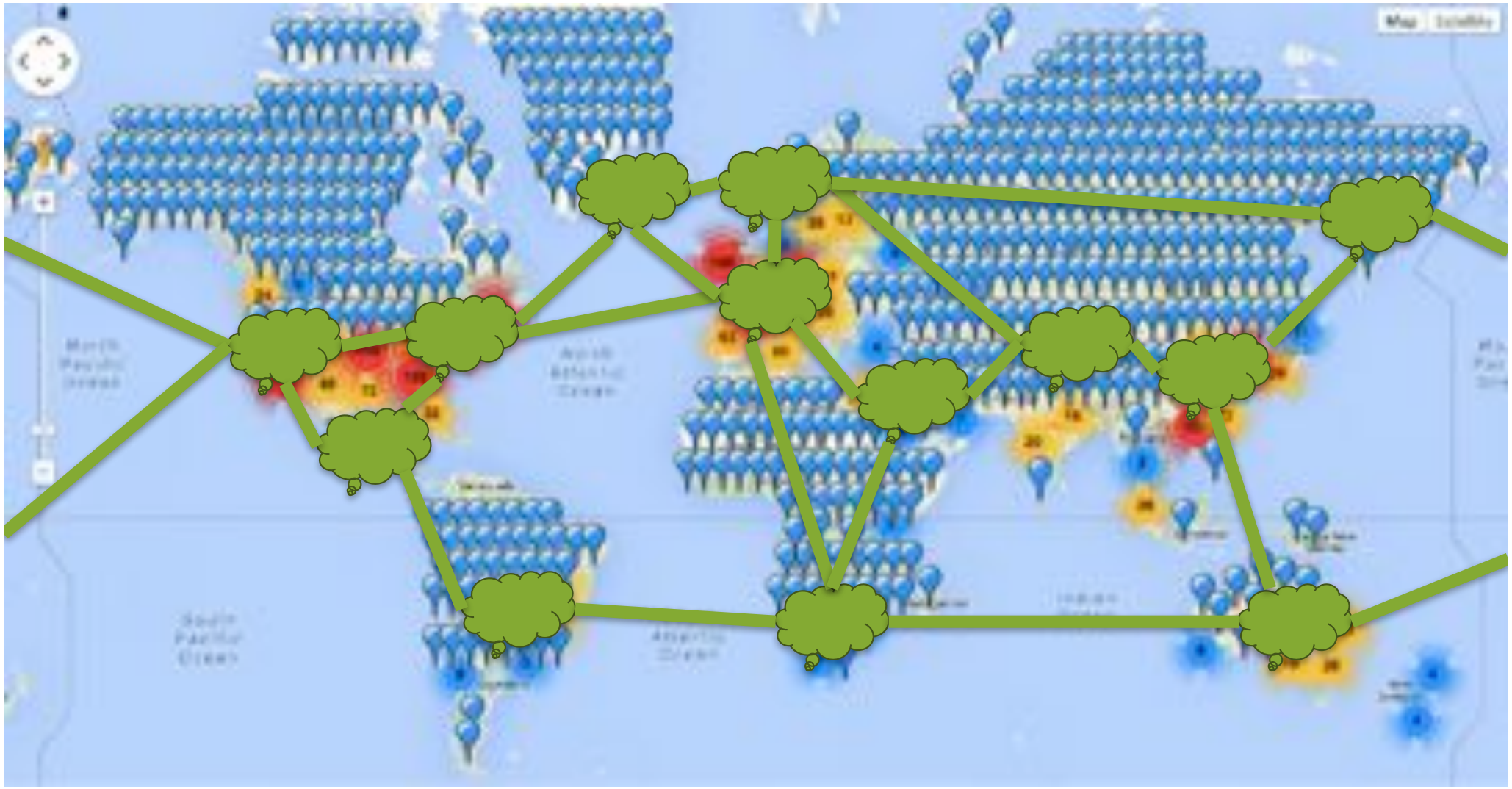
Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Sequencing Centers 2028

Informatics Centers 2028

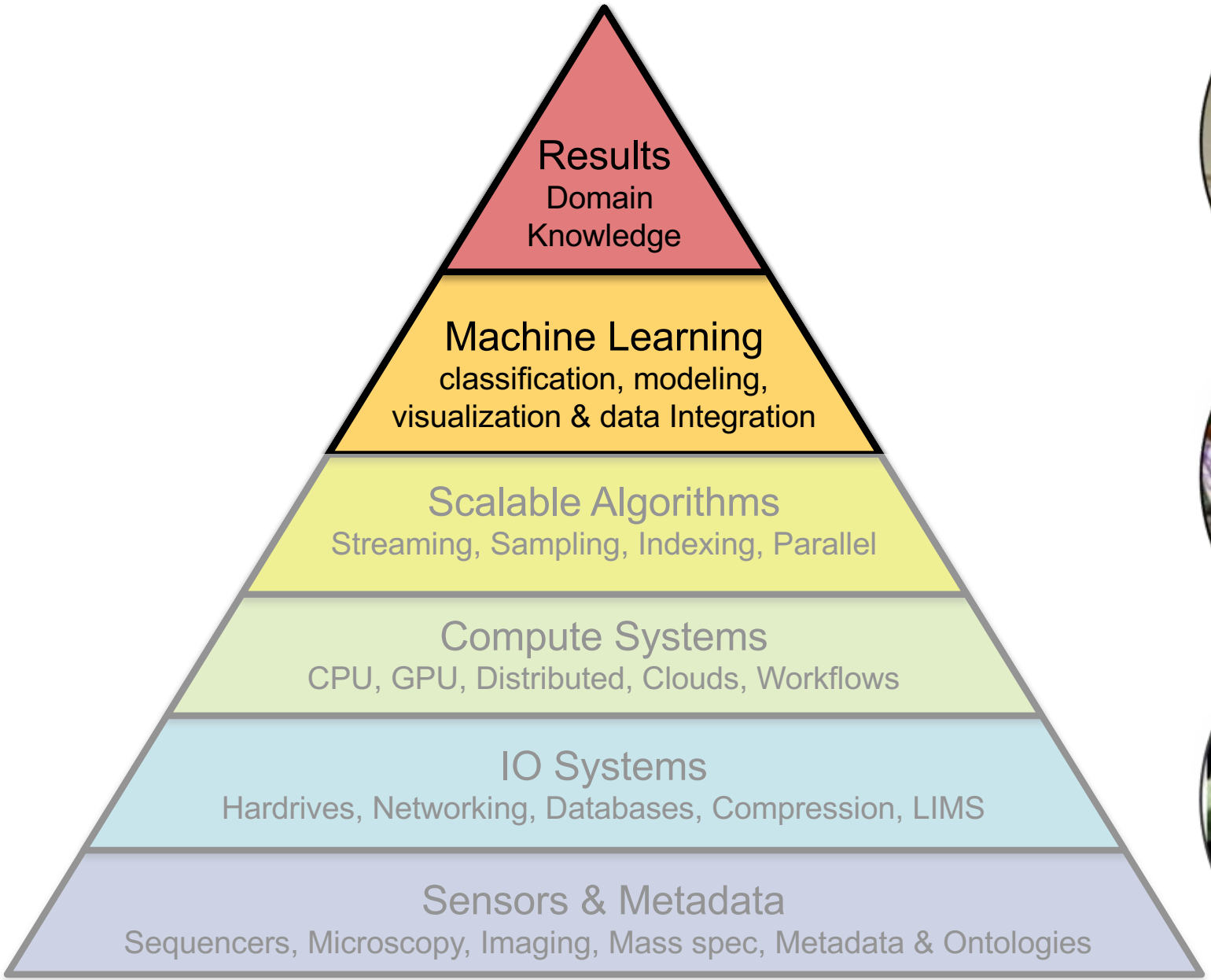
The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Informatics Centers 2028



The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Comparative Genomics 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

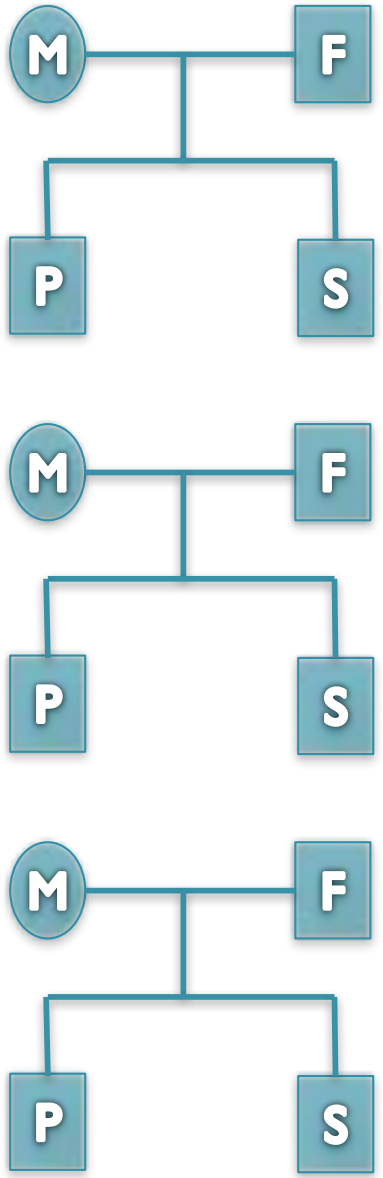
- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

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

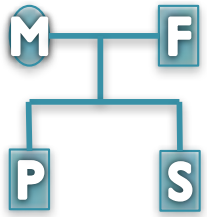


Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?

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

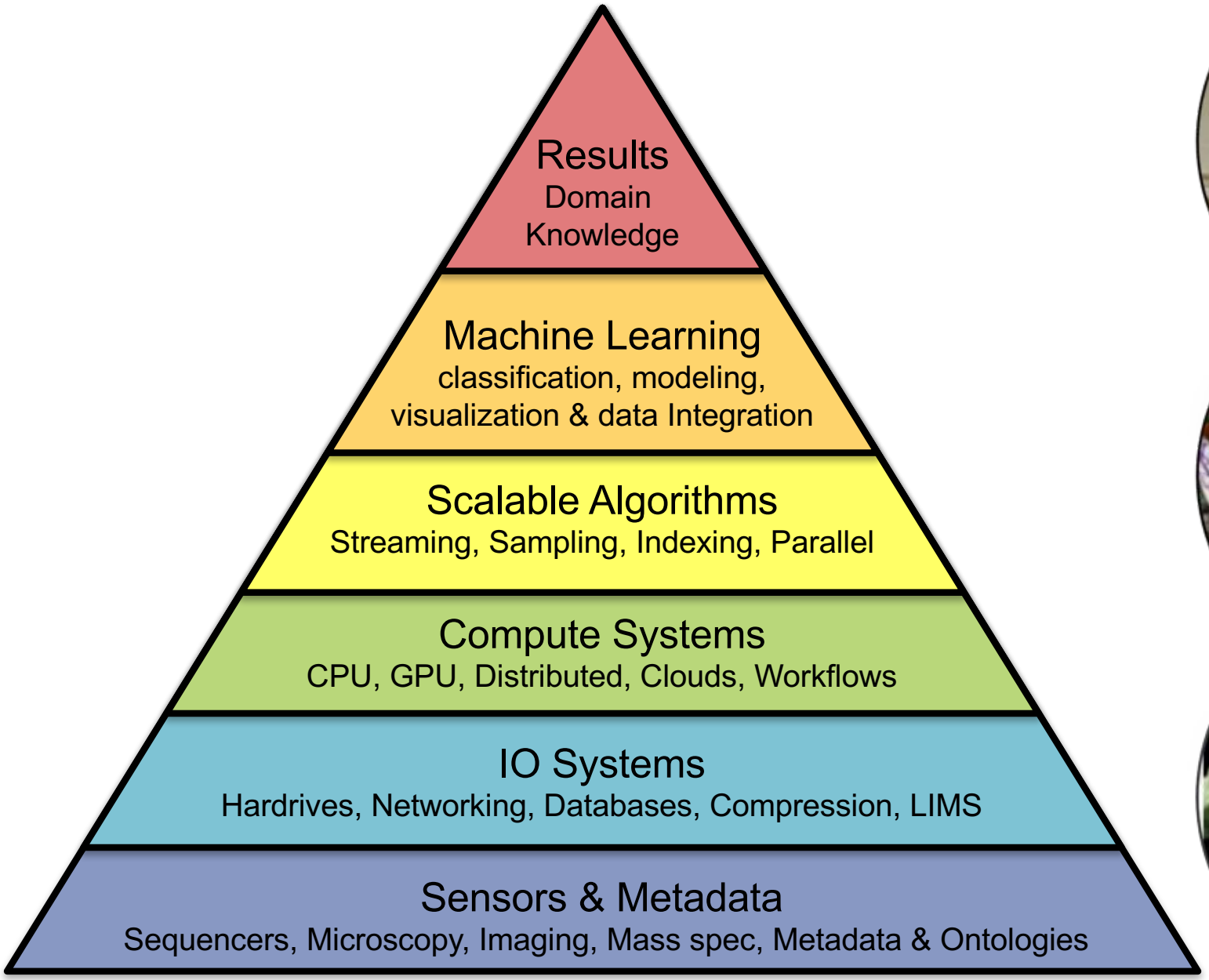
De novo Genetics of Autism

- 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

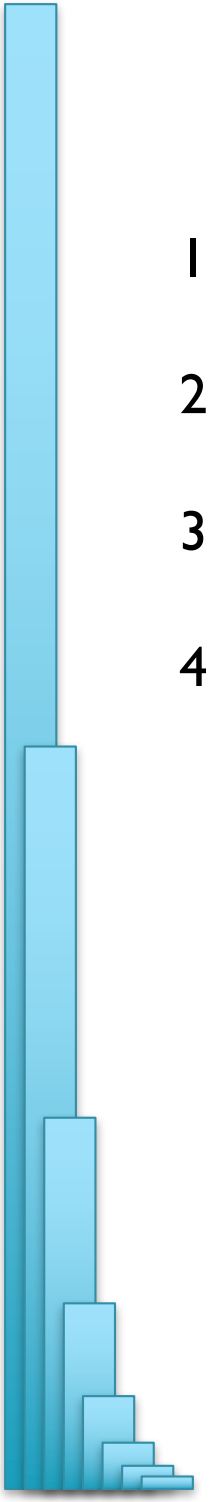
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

Comparative Genomics Technologies



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 Linux, set up Virtual Machine
    2. Set up Dropbox for yourself!
    3. Get comfortable on the command line




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics2018

Questions?