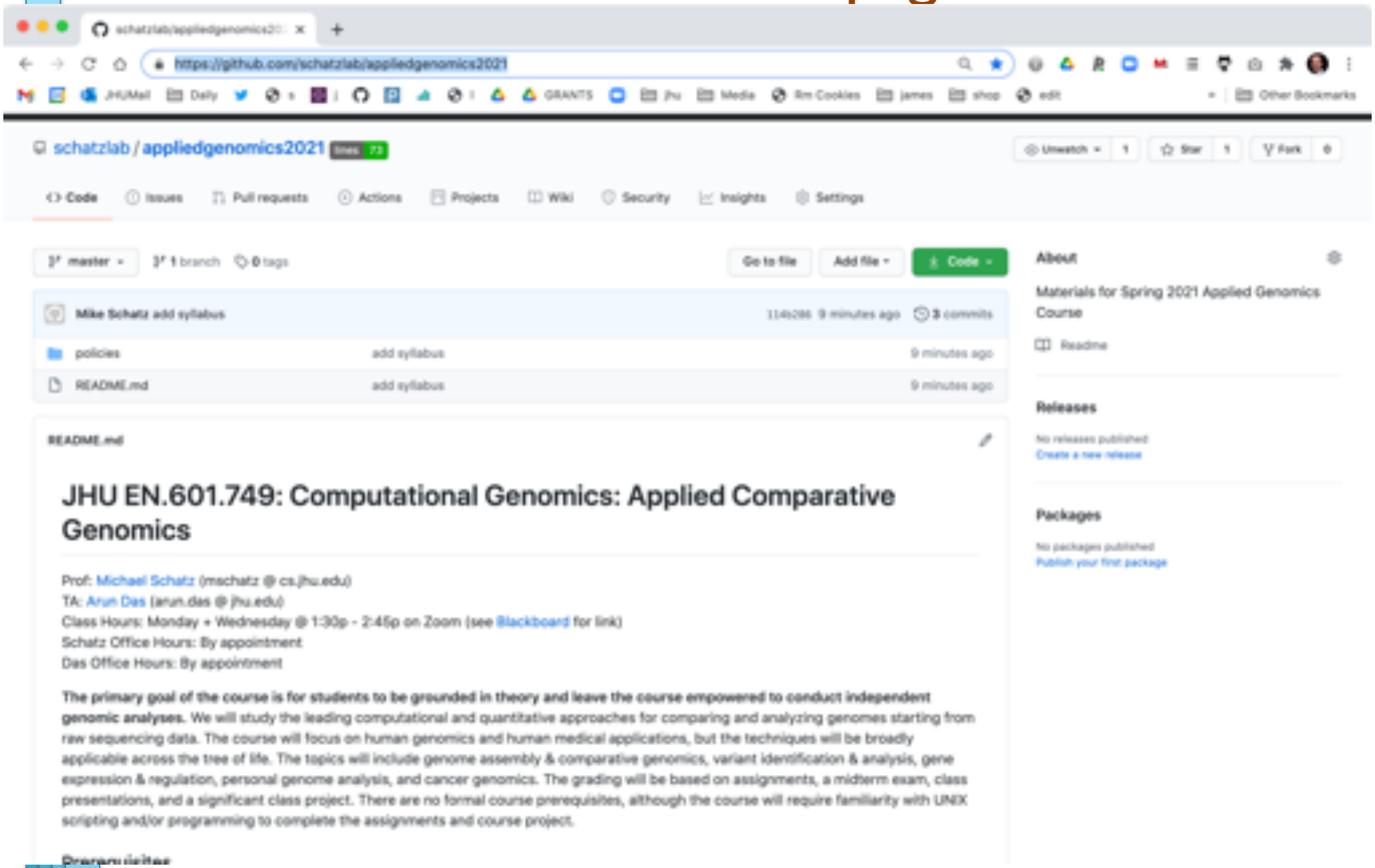
Genomic Technologies

Michael Schatz

January 27, 2020

Lecture 2: Applied Comparative Genomics

Course Webpage

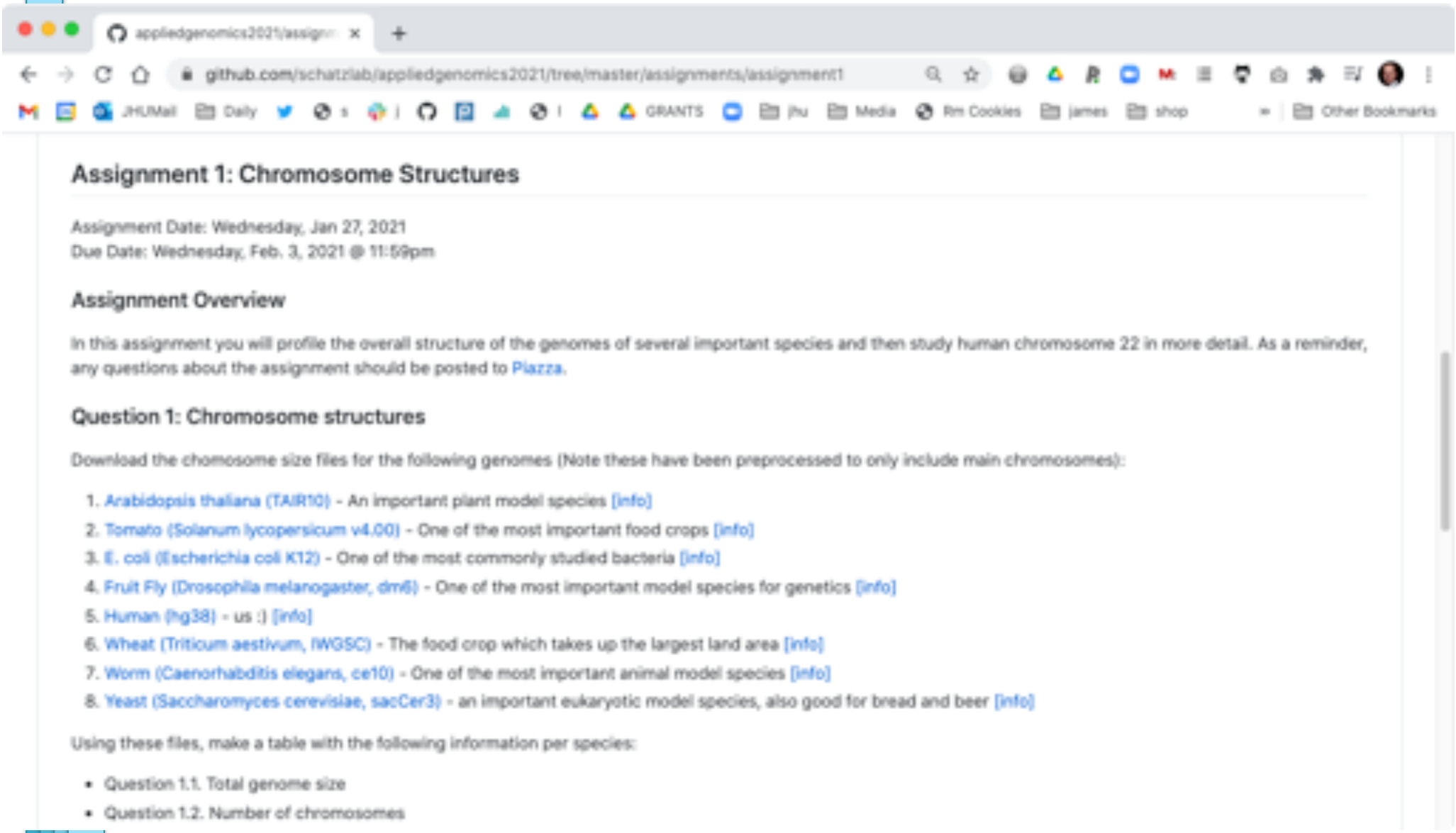


https://github.com/schatzlab/appliedgenomics2021

TA: Arun Das

Check Piazza for Office Hours Poll

Assignment 1



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, Jan 27, 2021
Due Date: Wednesday, Feb. 3, 2021 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then study human chromosome 22 in more detail. As a reminder, any questions about the assignment should be posted to Piazza.

Question 1: Chromosome structures

Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

1. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
2. Tomato (Solanum lycopersicum v4.00) - One of the most important food crops [info]
3. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
4. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
5. Human (hg38) - us :) [info]
6. Wheat (Triticum aestivum, IWGSC) - The food crop which takes up the largest land area [info]
7. Worm (Caenorhabditis elegans, ce10) - One of the most important animal model species [info]
8. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes

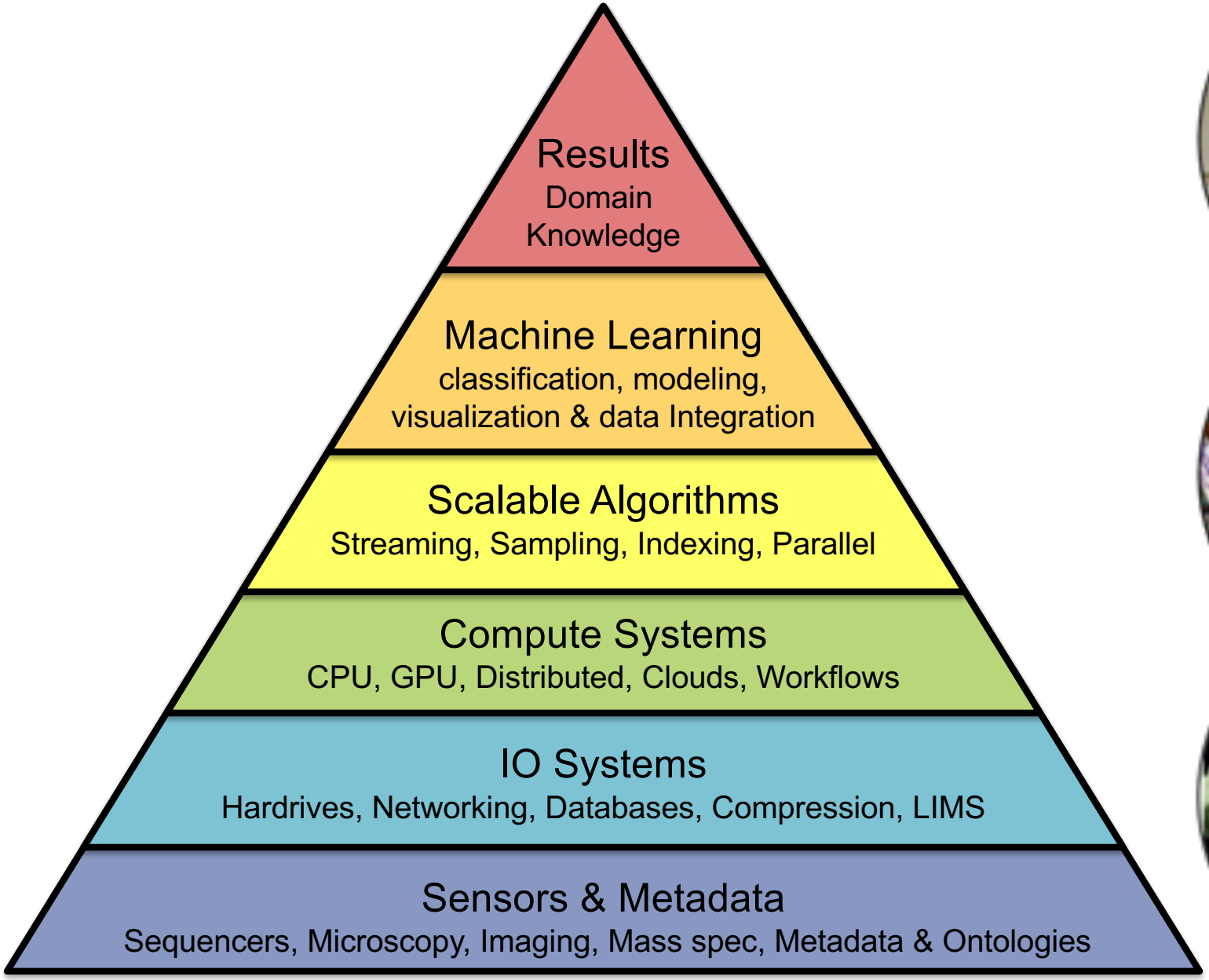
https://github.com/schatzlab/appliedgenomics2021/tree/master/assignments/assignment1

Due end of day on Feb 3 (right before midnight)

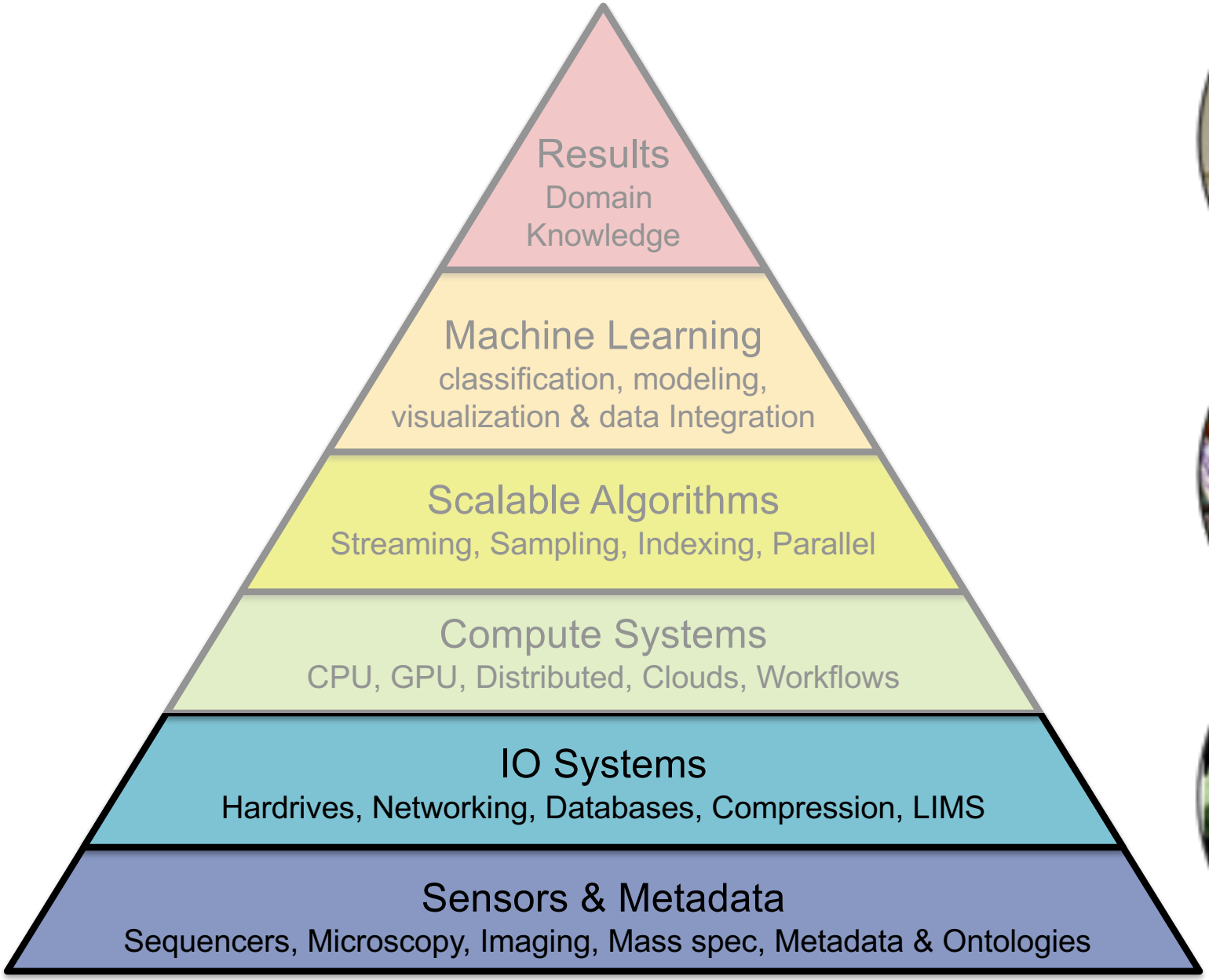
Unsolved Questions in Biology

- 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***

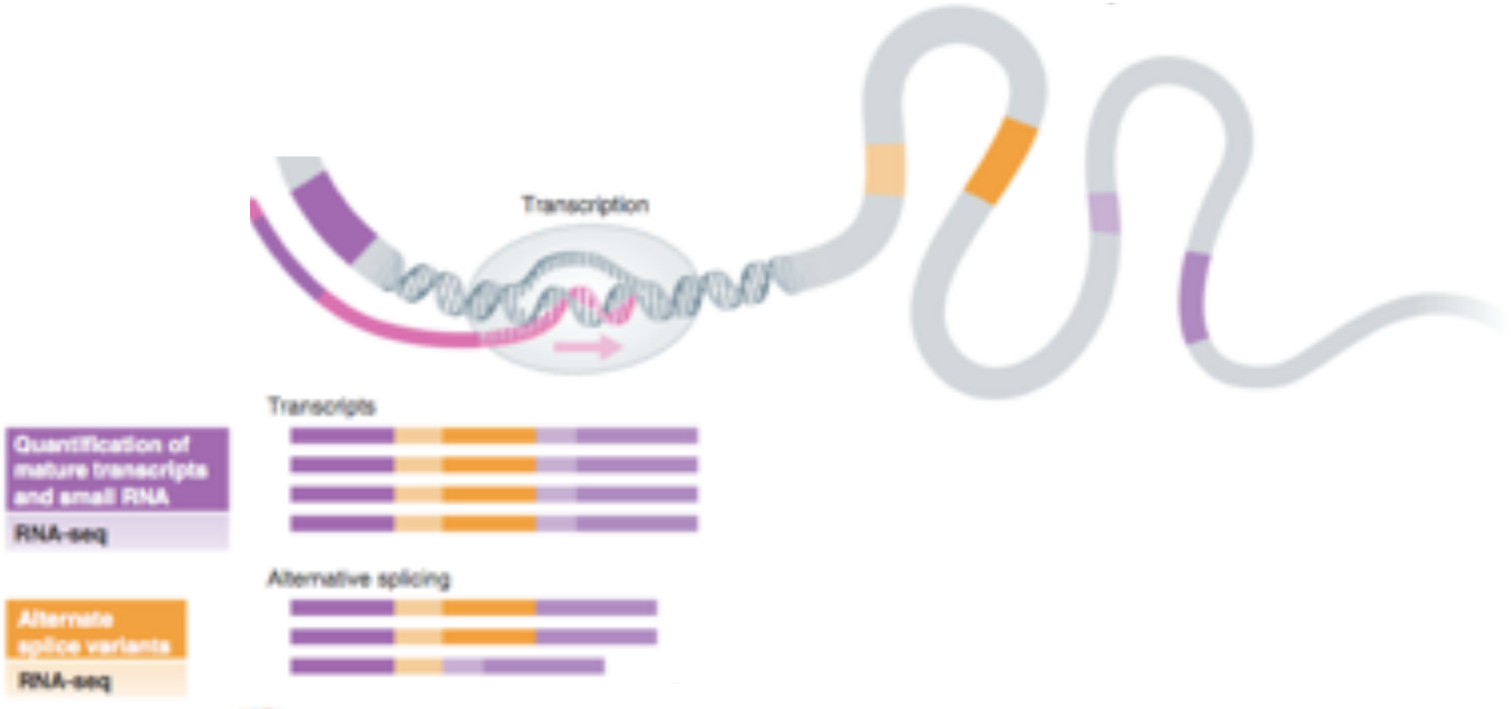
Comparative Genomics Technologies



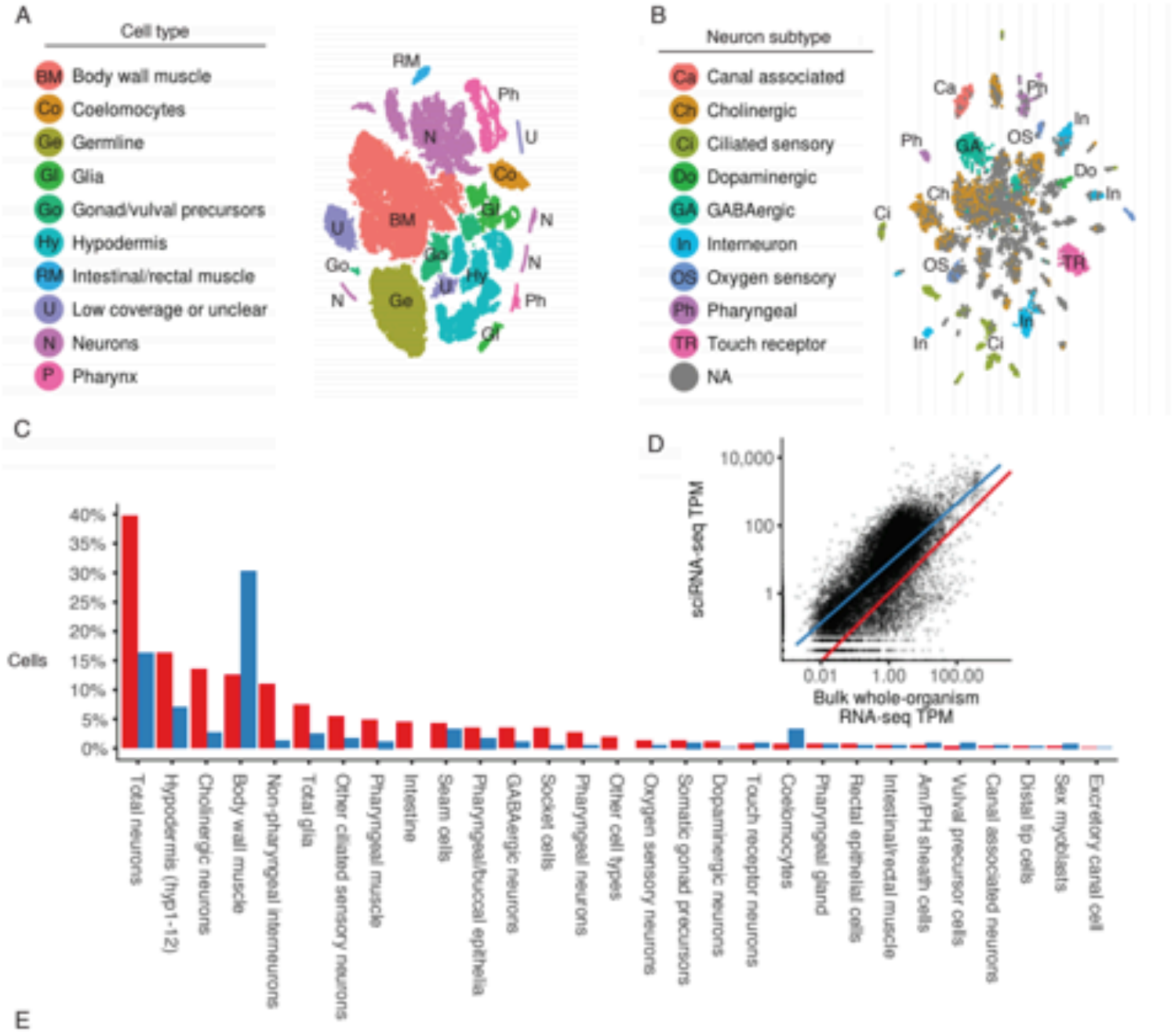
Comparative Genomics Technologies



Genomics Arsenal in the year 2021

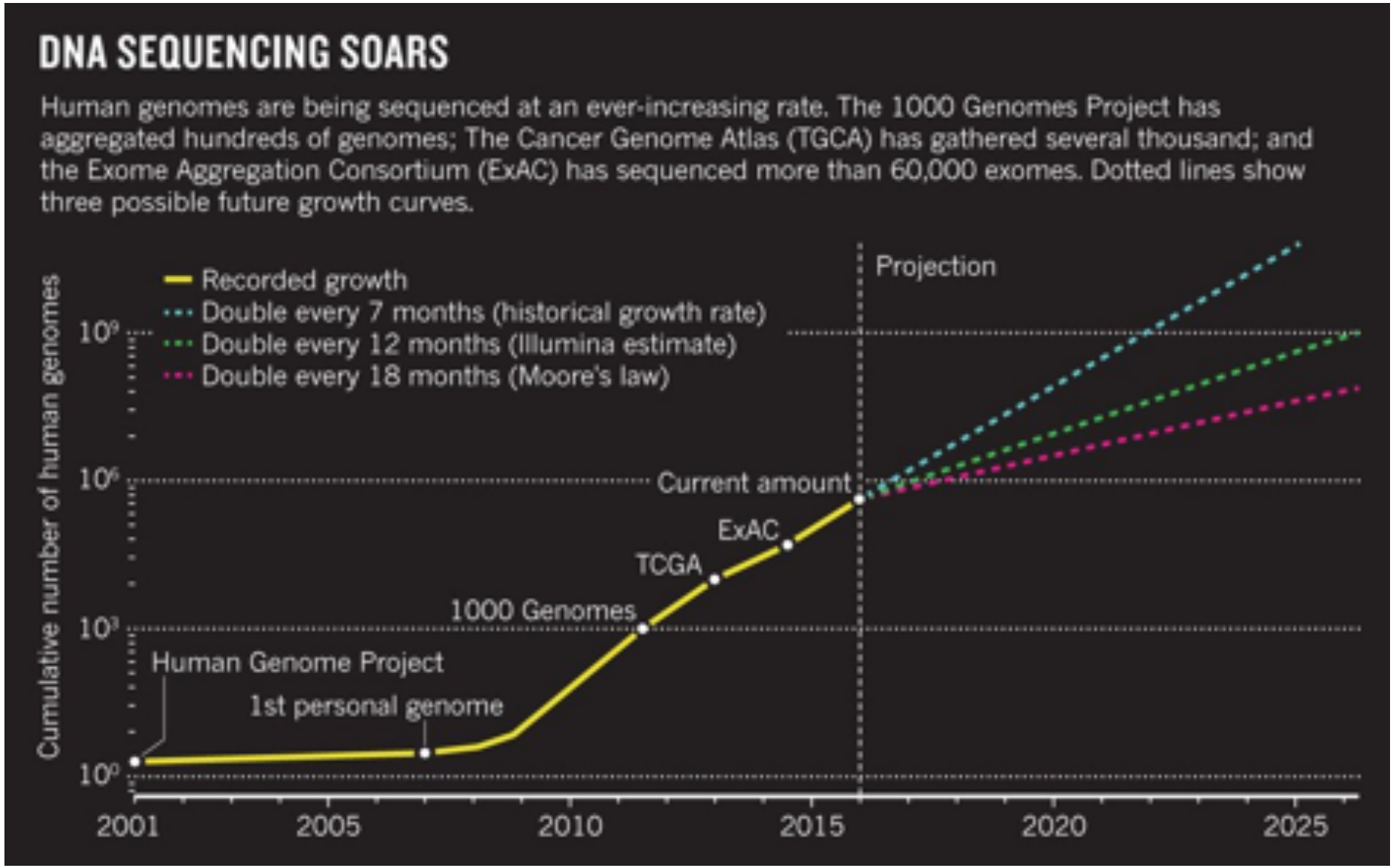


Soon et al., Molecular Systems Biology, 2013



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

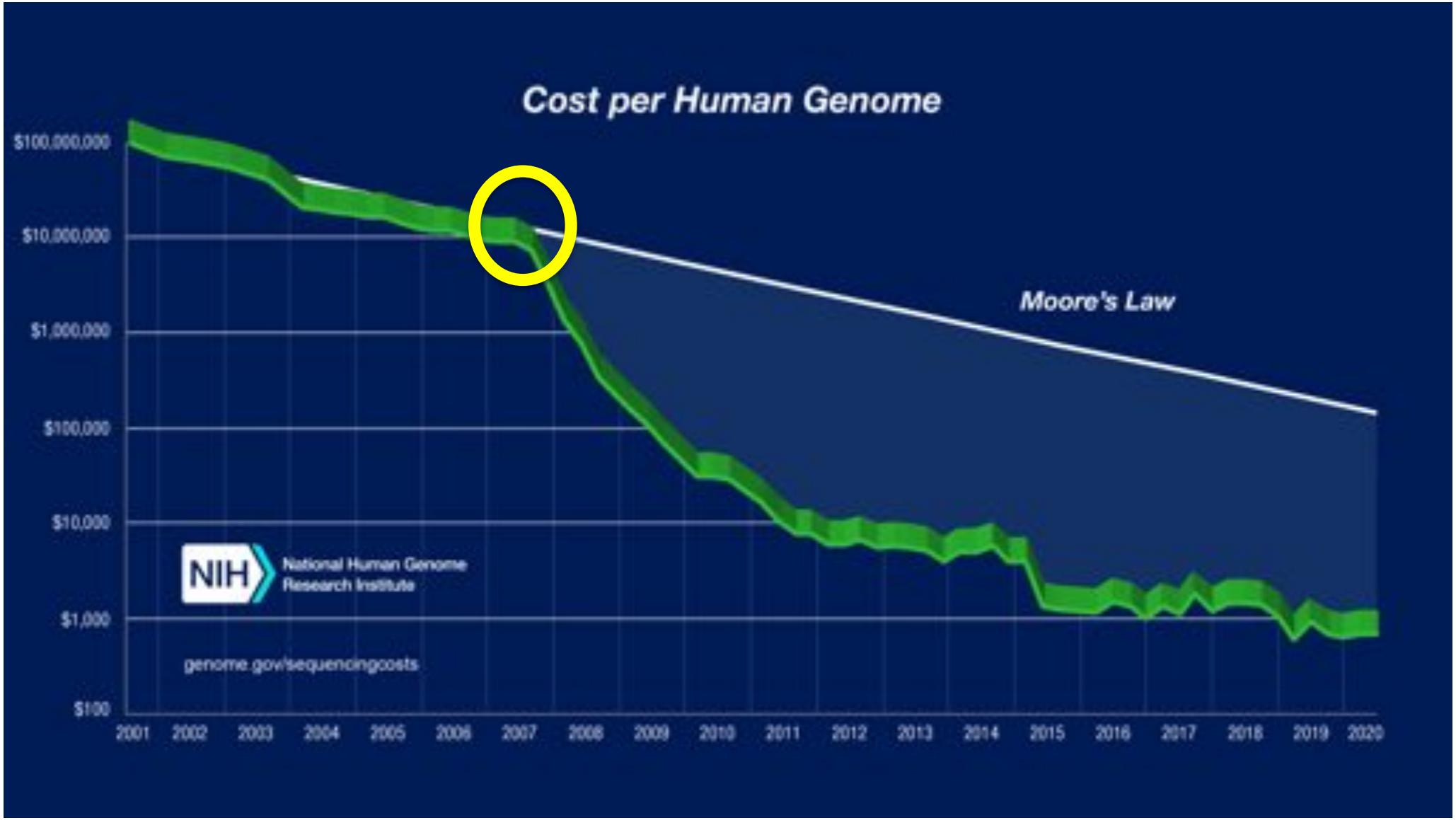
Sequencing Capacity



Big Data: Astronomical or Genomical?

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

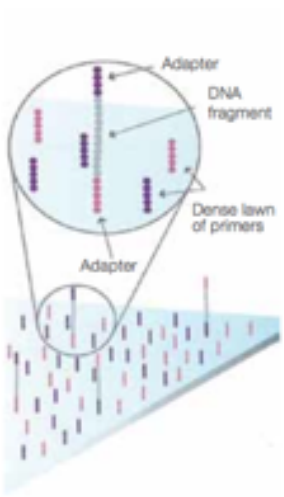
Cost per Genome



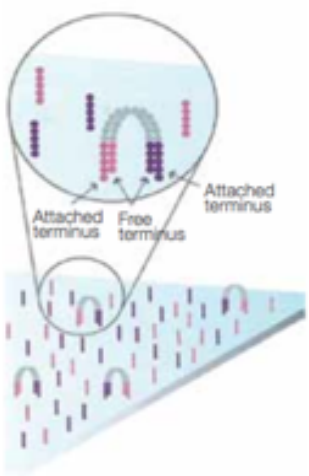
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

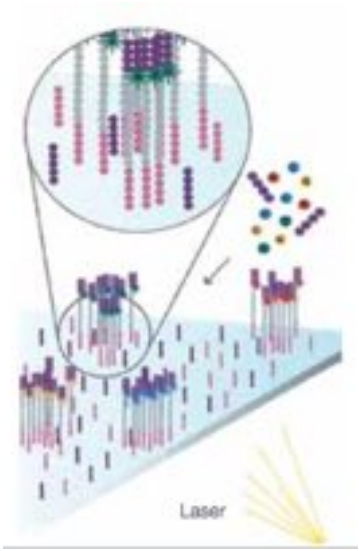
>3Tbp / day



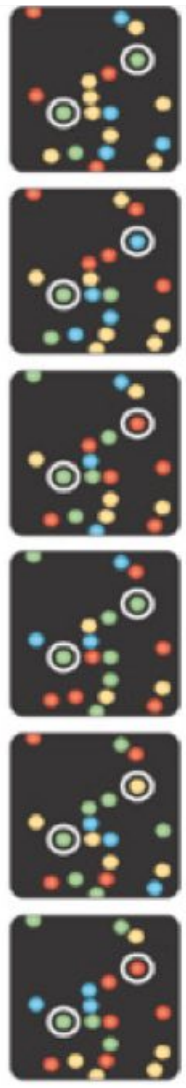
1. Attach



2. Amplify



3. Image

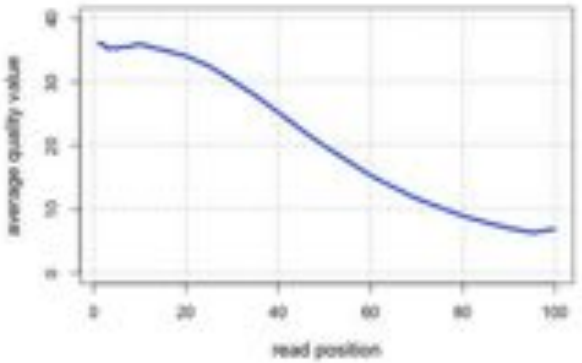


Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

$$Q_{\text{sanger}} = -10 \log_{10} p$$

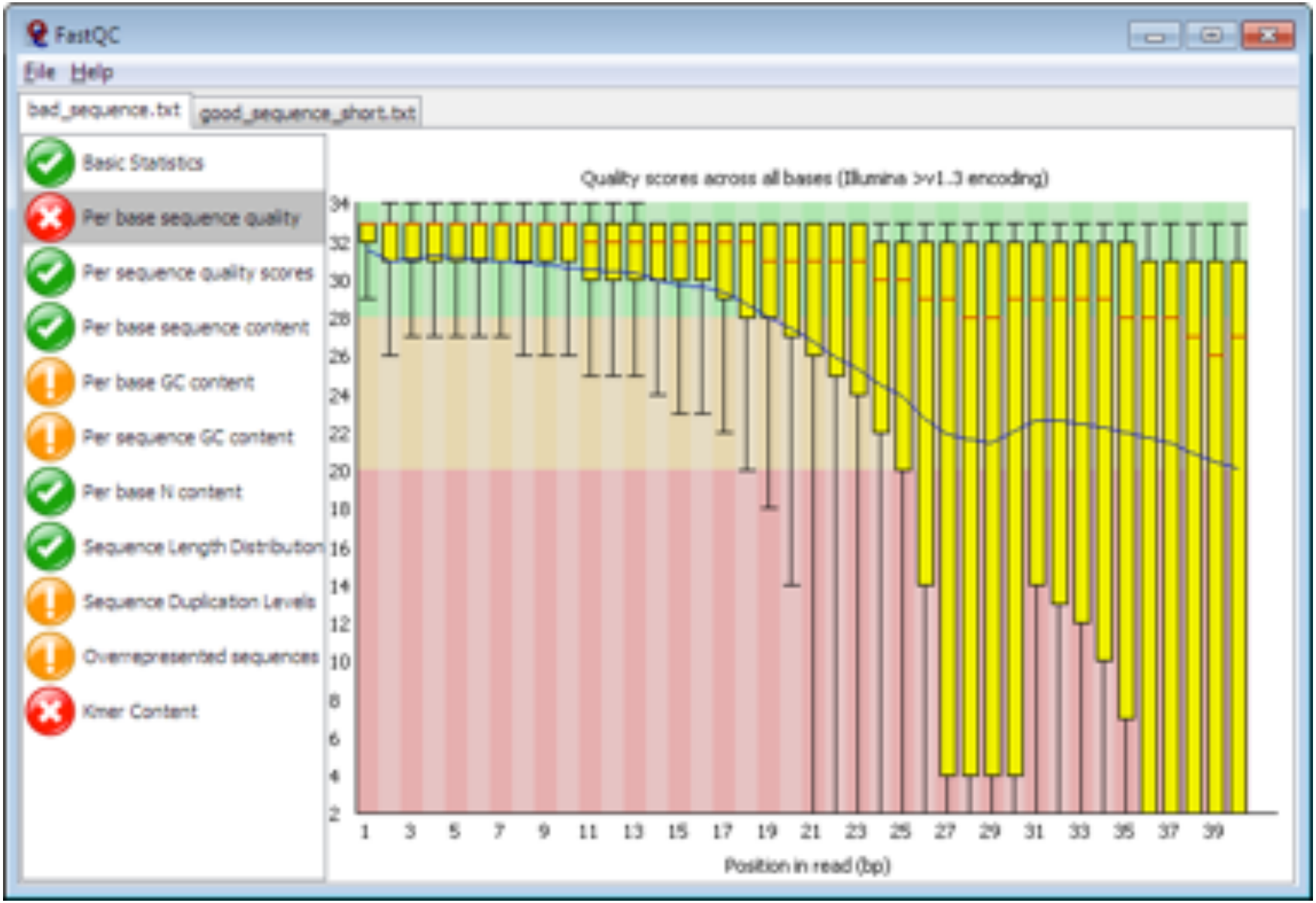



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..........................
  ..................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..........................
  ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..........................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL..................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

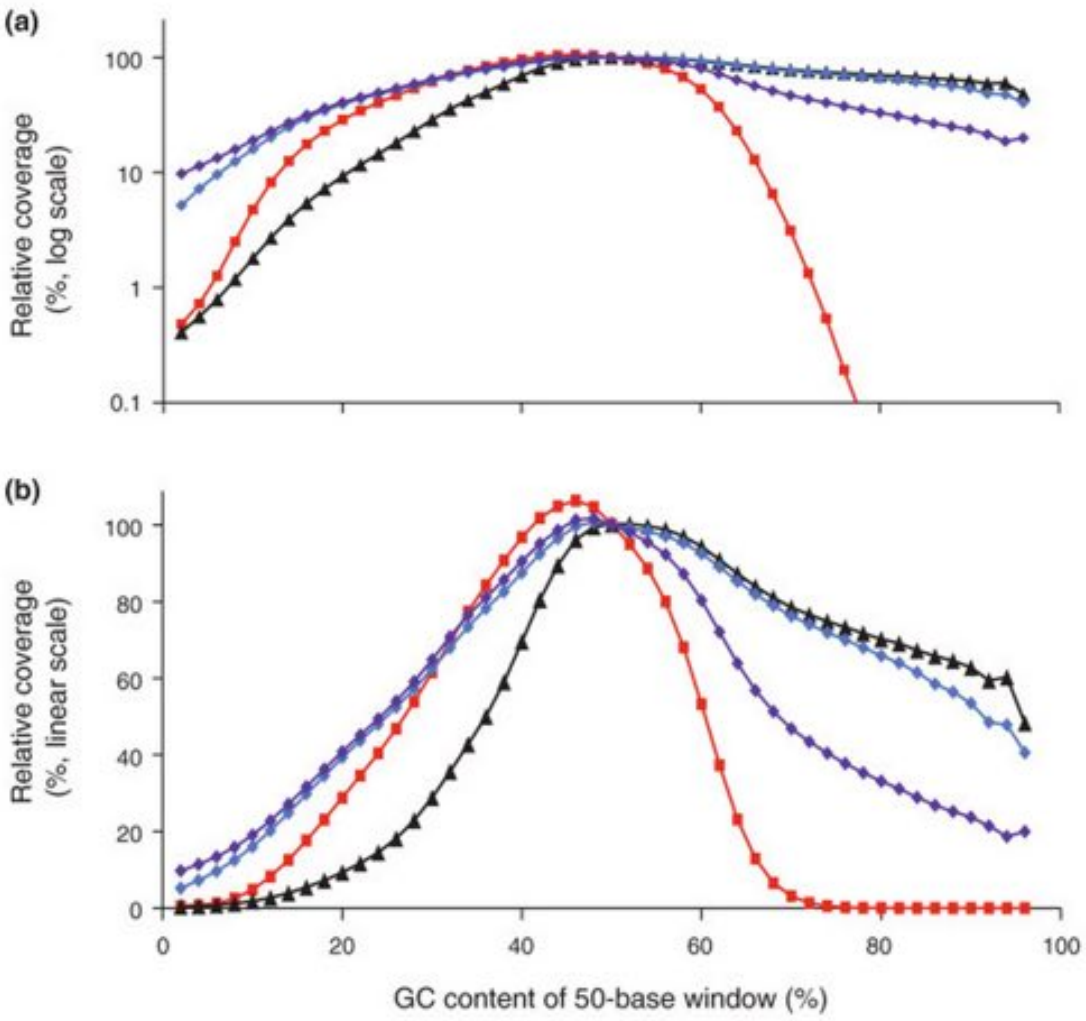
http://en.wikipedia.org/wiki/FASTQ_format

FASTQC: Is my data any good?



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Beware of GC Biases

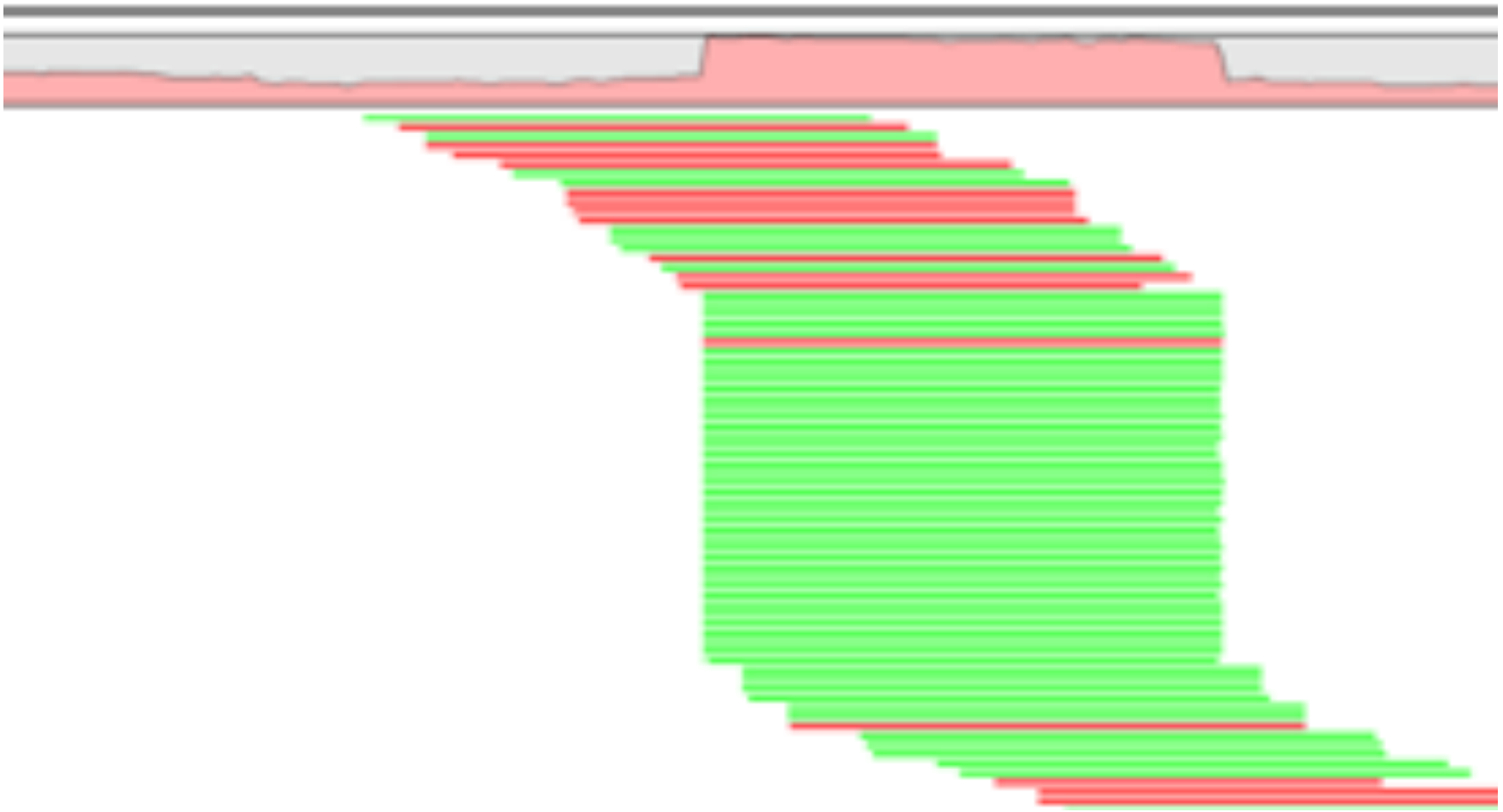


Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

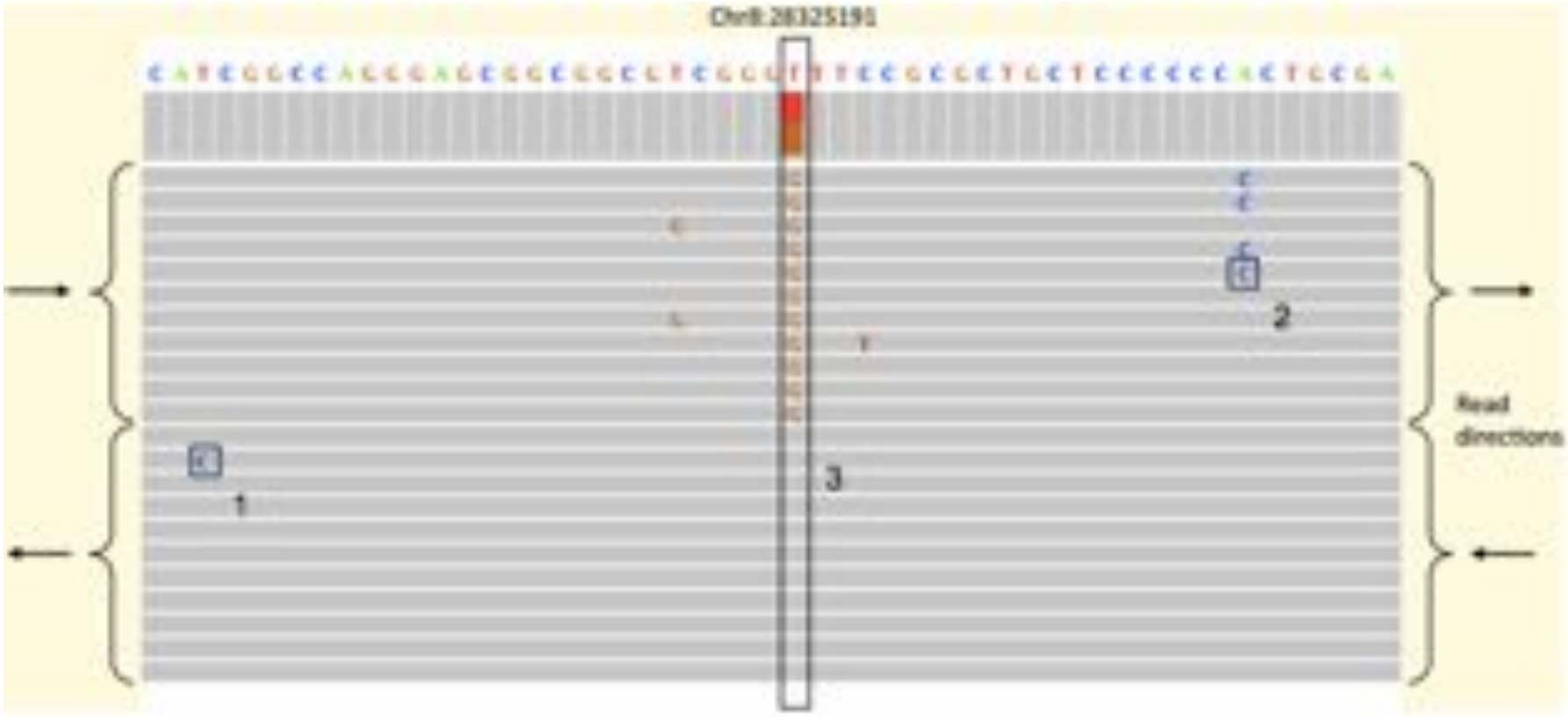
Beware of Duplicate Reads



The Sequence alignment/map (SAM) format and SAMtools.
Li et al. (2009) *Bioinformatics.* 25:2078-9

Picard: http://picard.sourceforge.net

Beware of (Systematic) Errors



Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics.* 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

Question?

We would love to generate
longer and longer reads with this technology

What can we do?

Illumina Hacking

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 12 2013, pages 1492–1497
doi:10.1093/bioinformatics/btt178

Genome analysis

Advance Access publication May 22, 2013

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D. Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A. Taylor[1], Macaire Man Saint Yuen[4], Christopher I. Keeling[4], Dana Brand[1], Benjamin P. Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A. Moore[1], Yongjun Zhao[1], Andrew J. Mungall[1], Barry Jaquish[5], Alvin Yanchuk[6], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven J.M. Jones[1,2,9]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada, [2]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada, [3]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [5]British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2, Canada, [6]Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [7]Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1V 0A6, Canada, [8]Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada and [9]Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Michael Brudno

ABSTRACT

White spruce (*Picea glauca*) is a dominant conifer of the boreal forests of North America, and providing genomics resources for this commercially valuable tree will help improve forest management and conservation efforts. Sequencing and assembling the large and highly repetitive spruce genome though pushes the boundaries of the current technology. Here, we describe a whole-genome shotgun sequencing strategy using two Illumina sequencing platforms and an assembly approach using the ABySS software. We report a 20.8 giga base pairs draft genome in 4.9 million scaffolds, with a scaffold N50 of 20 356 bp. We demonstrate how recent improvements in the sequencing technology, especially increasing read lengths and paired end reads from longer fragments have a major impact on the assembly contiguity. We also note that scalable bioinformatics tools are instrumental in providing rapid draft assemblies.

Availability: The *Picea glauca* genome sequencing and assembly data are available through NCBI (Accession#: ALWZ01000000000 PID: PRJNA83435). http://www.ncbi.nlm.nih.gov/bioproject/83435.

Contact: ibirol@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 10, 2013; accepted on April 11, 2013

1 INTRODUCTION

The assembly of short reads to develop genomic resources for non-model species remains an active area of development (Schatz *et al.*, 2012). The feasibility of the approach and its scalability to large genomes was demonstrated b (Simpson *et al.*, 2009) using human and was later used to assemble the SOAPdenovo tool (Li *et al.*, 2010). high quality results, as demonstrate *et al.*, 2012; Ladner *et al.*, 2013; Rib been successfully applied numerous ti omes (Chan *et al.*, 2011; Chu *et al.*, 2 2011; Godel *et al.*, 2012; Swart *et al.*,

Estimated at 20 giga base pairs (Gi cing and assembly of the genome of t the pine (*Pinaceae*) family present uni generation end, those challenges inclu whole-genome shotgun sequencing da ing reduced representation resources tude of the problem. On the bioin massive sequencing datasets is extrem ing cycles, memory usage, storage req programming implementations on co

We addressed the data representati and sequencing multiple whole-genom HiSeq 2000 and MiSeq sequencers (CA, USA). Compared with localized as building and sequencing fosmi approach of isolating ~10 kb DNA s sequencing fragments in high through CA, USA), a shotgun only sequencing sequence data effectively covering th that can be an order of magnitude le especially substantial when sequencin

In this work, we demonstrate that t at this scale remains viable and pro

*To whom correspondence should be addressed.

assemble the spruce genome, we used the ABySS algorithm (Simpson *et al.*, 2009), which captures a representation of read-to-read overlaps by a distributed de Bruijn graph and uses parallel computations to build the target genome. The modular nature of the tool allowed us to execute a large number of tests to tune the message passing interface for a successful execution, train the assembly parameters for an optimal assembly and quantify the utility of long reads for large genome assemblies. To the best of our knowledge, the ABySS algorithm is unique in its ability to enable genome assemblies of this scale using whole-genome shotgun sequencing data.

2 METHODS

2.1 Sample collection

Apical shoot tissues were collected in April 2006 from a single white spruce (*Picea glauca*, genotype PG29) tree at the Kalamalka Research Station of the British Columbia Ministry of Forests and Ranges, Vernon, British Columbia, Canada. Genomic DNA was extracted from 60 gm tissue by BioS&T (http://www.biost.com), Montreal, QC, Canada) using an organelle exclusion method yielding 300 µg of high quality purified nuclear DNA.

2.2 Library preparation and sequencing

DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 s using an E210 sonicator (Covaris) and then analysed on 8% PAGE gels. The 200–300 bp (for libraries with 250 bp insert size) or 450–550 bp (for libraries with 500 bp insert size) DNA size fractions were excised and eluted from the gel slices overnight at 4°C in 300 µl of elution buffer [5:1 (vol/vol) LoTE buffer [3 mM Tris–HCl (pH 7.5), 0.2 mM EDTA]/7.5 M ammonium acetate] and was purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adapters (with 5′ overhangs). Adapter-ligated products were purified on QiAquick spin columns (Qiagen) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen).

The mate pair (MPET, a.k.a. jumping) libraries were constructed using 4 µg of genomic DNA with the Illumina Nexlera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin containing mate pair junction adapter, which left a short sequence gap in the tagmented DNA. The gap was filled by a strand displacement reaction using a polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9 kb and 9–13 kb fractions, which were purified using a Zymoclean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by a digestion to remove any linear molecules and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by Agilent DNA 1000 series II assay and KAPA Library Quantification protocol. The two fractions were pooled for sequencing paired end 100 bp using Illumina HiSeq2000.

2.3 MiSeq modification

In sequencing the spruce genome, we generated longer re modifying the MiSeq platform. The MiSeq uses a clamshell (Supplementary Fig. S1A) to hold reagent tubes in an arr cessed by the MiSeq's sippers. Most of the reagents are length independent steps such as denaturation and cluster g three reagents, the Scan, Cleavage and Incorporation sumed at each cycle. Although the MiSeq allows any read specified in the control software, the reagent cartridge carou during the run without stopping it. Increasing the read len requires increasing the quantity of the length-dependent cartridge. This led to the solution of combining the leng reagents of two kits into one.

A tool was designed that opens the snap-hook latches cartridge together (Supplementary Figs S1B and S2), giv the reagent tubes, yet allowing the cartridge to be put b without damage to its components (Supplementary Fi 40 ml, the stock length-dependent reagent containers allo imum of ~650 cycles in total. To maximize the potential of kit approach, a new reagent tray with 70 ml wells was placed in a modified clamshell base.

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A Taylor[1], Macaire Man Saint Yuen[4], Christopher I Keeling[4], Dana Brand[1], Benjamin P Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A Moore[1], Yongjun Zhao[1], Andrew J Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven JM Jones[1,2,9]

[1] British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6
[2] University of British Columbia, Department of Medical Genetics, Vancouver, BC V6H 3N1
[3] Simon Fraser University, School of Computing Science, Burnaby, BC V5A 1S6
[4] University of British Columbia, Michael Smith Laboratories, Vancouver, BC V6T 1Z4
[5] British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2
[6] University of British Columbia, Department of Forest Sciences, Vancouver, BC V6T 1Z4
[7] Université Laval, Institute for Systems and Integrative Biology, Québec, QC G1V 0A6
[8] Université Laval, Department of Wood and Forest Sciences, Québec, QC G1V 0A6
[9] Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC V5A 1S6

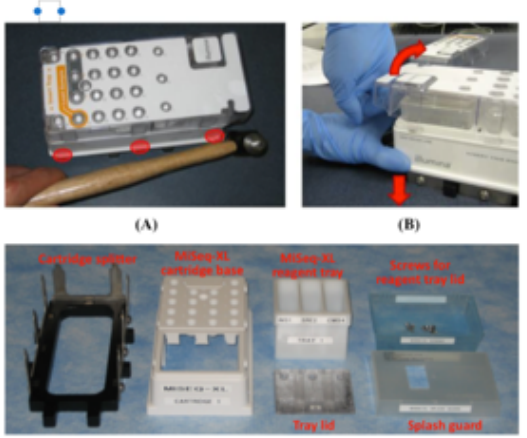


Figure S1. Modification of the MiSeq cartridge. MiSeq reagent cartridge was modified to allow for longer read lengths. **(A, B)** Opening of the clamshell style cartridge. **(C)** Contents of the modified cartridge. This was initially used to combine two PE150 kits for PE300 runs. When Illumina introduced the P250 kit, the same apparatus was used to enable PE500 runs.

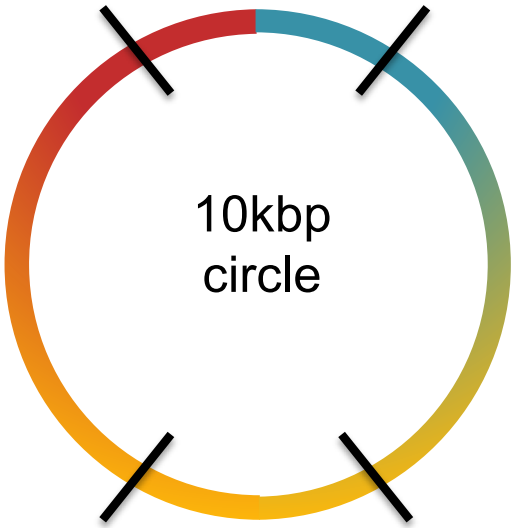
Paired-end and Mate-pairs

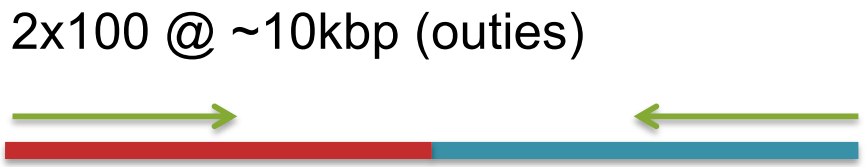
Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

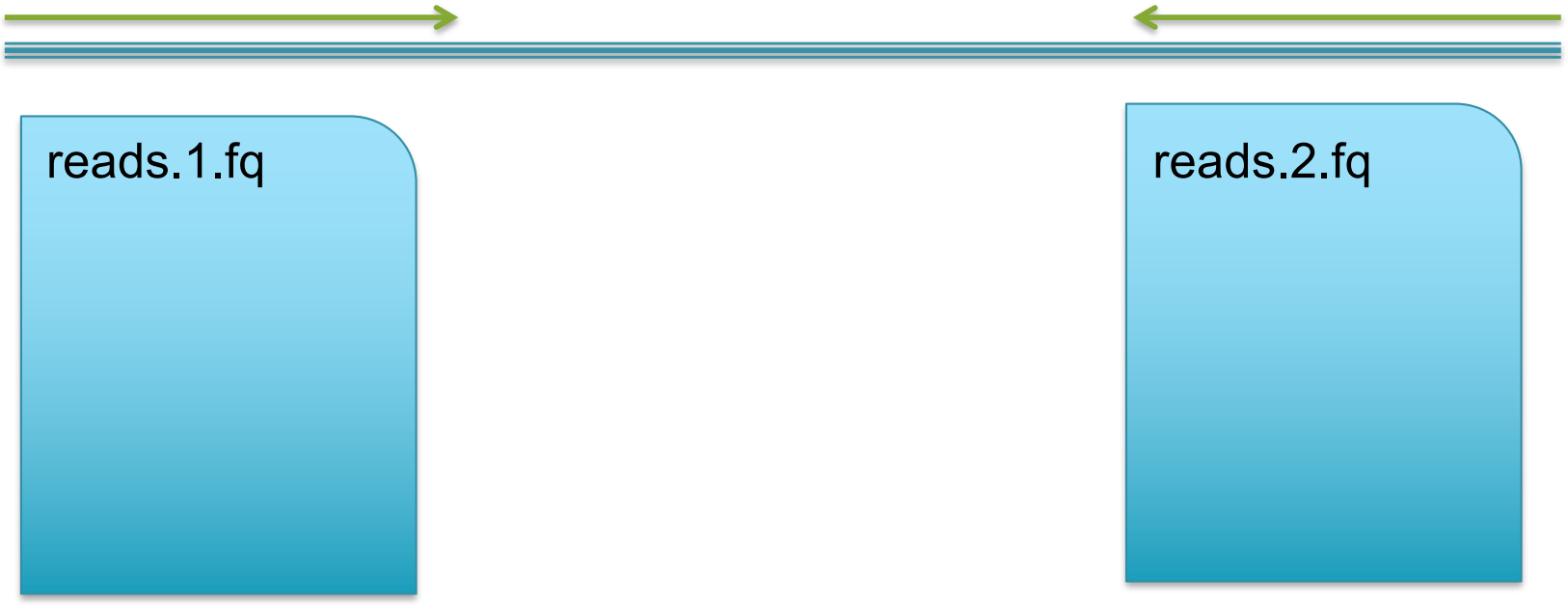
Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads





FASTQ Files




```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

@Identifier
Sequence
+Separator
Quality Values
…

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

Illumina Sequencing Summary

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)
- Some runs are error prone
- Requires amplification, sequences a population of molecules

Illumina HiSeq
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or “rapid run” ~90Gb in 1-2 days)

Illumina X Ten / NovaSeq
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or “rapid run” ~90Gb in 1-2 days)

Illumina NextSeq
One human genome in **<30 hours**

Market Summary > **Illumina, Inc.**
NASDAQ: ILMN

+ Follow

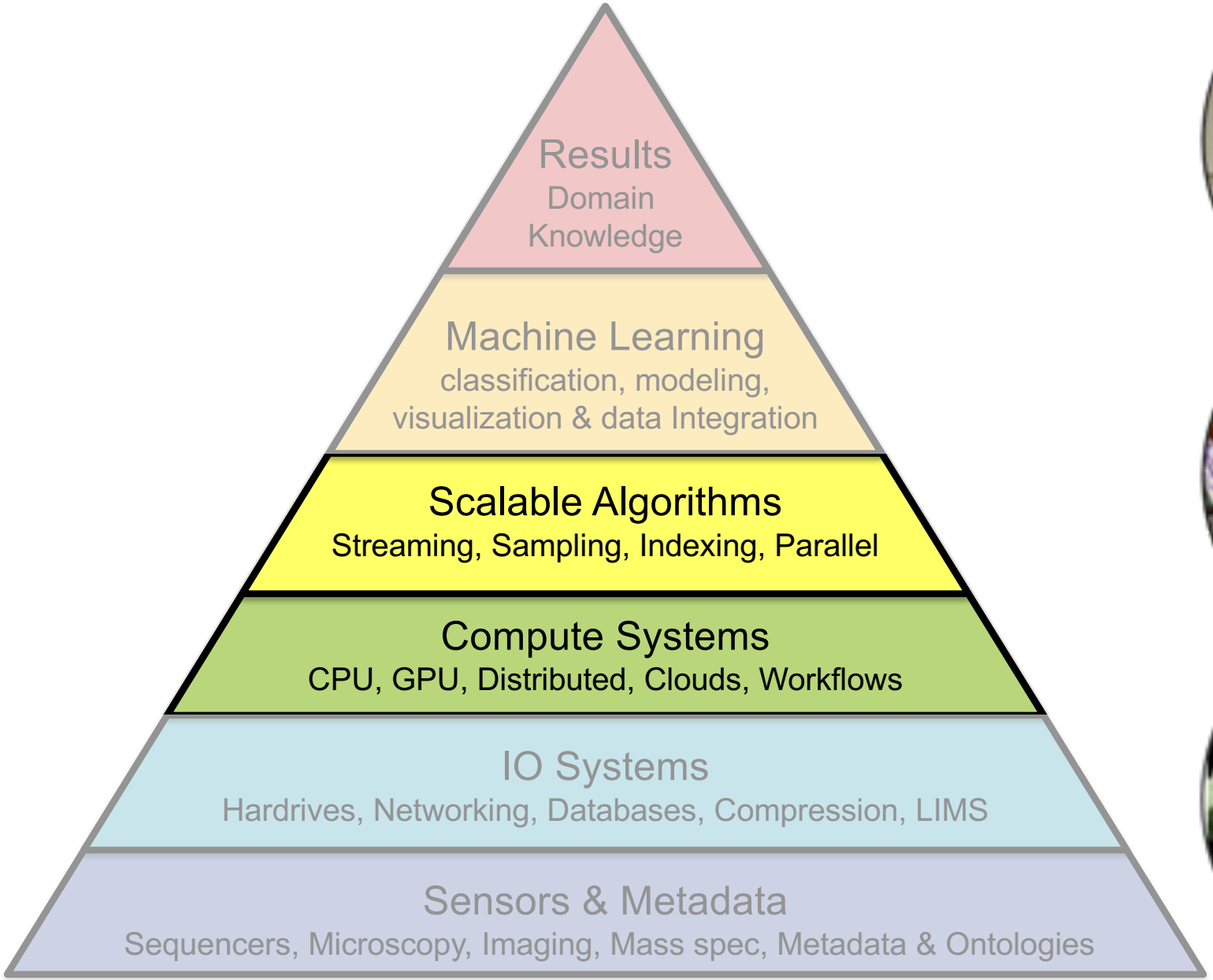
420.35 USD −0.65 (0.15%) ↓
Jan 26, 3:56 PM EST · Disclaimer

1 day | 5 days | 1 month | 6 months | YTD | 1 year | 5 years | **Max**

Open	418.95	Div yield	-
High	425.92	Prev close	421.00
Low	416.29	52-wk high	425.92
Mkt cap	61.33B	52-wk low	196.78
P/E ratio	97.70		

→ More about Illumina, Inc.

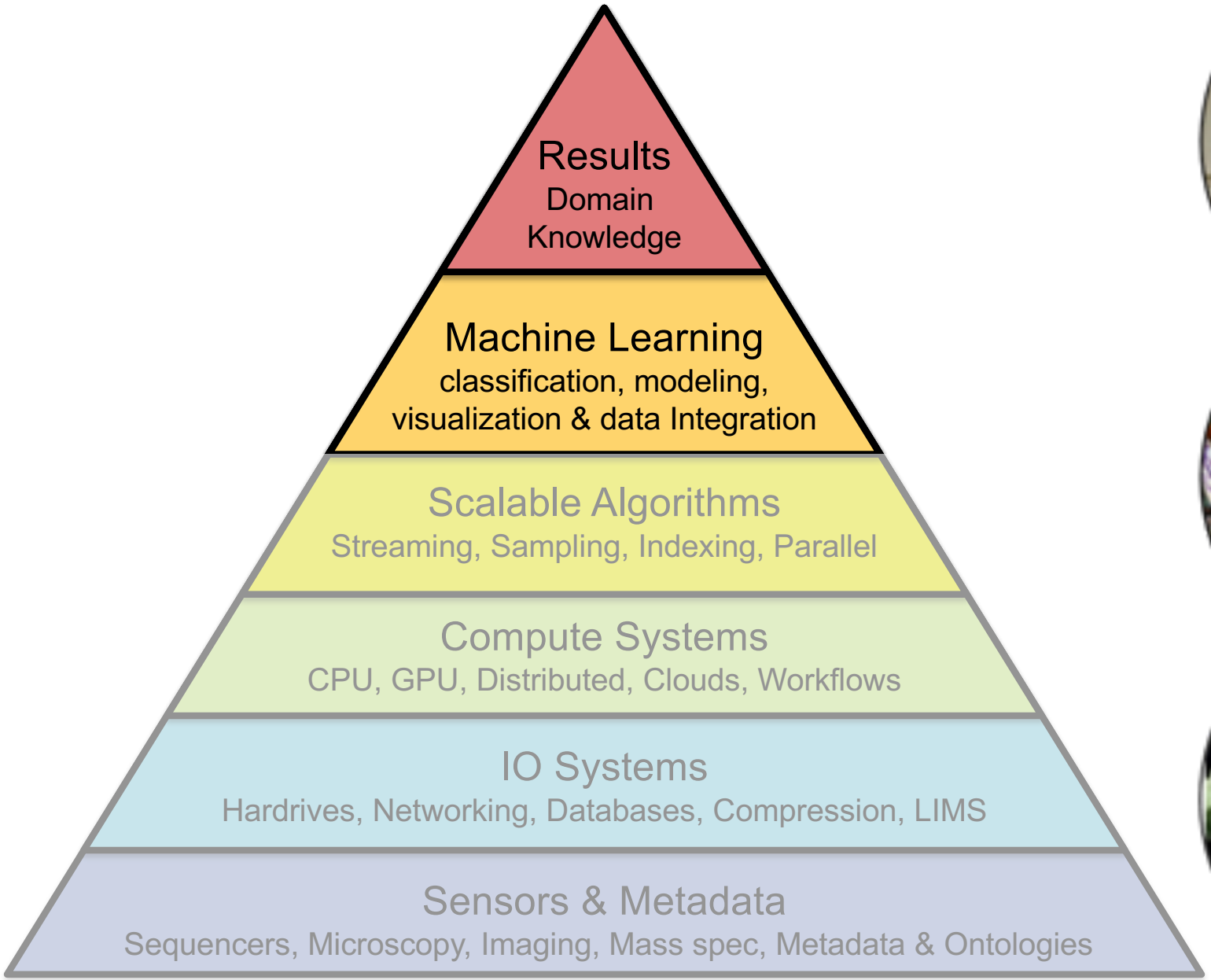
Comparative Genomics Technologies



Selected 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

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

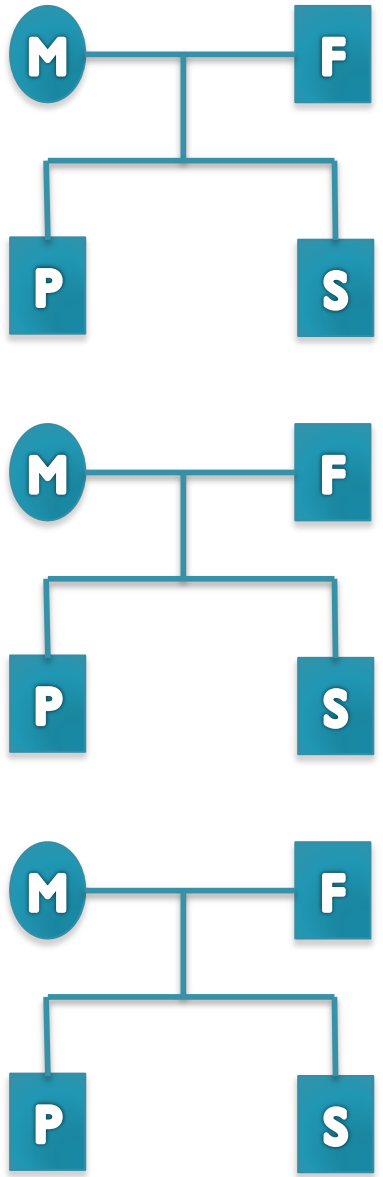
What is Autism?

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

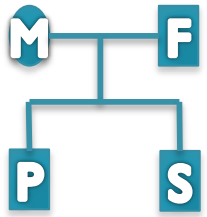


Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?

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

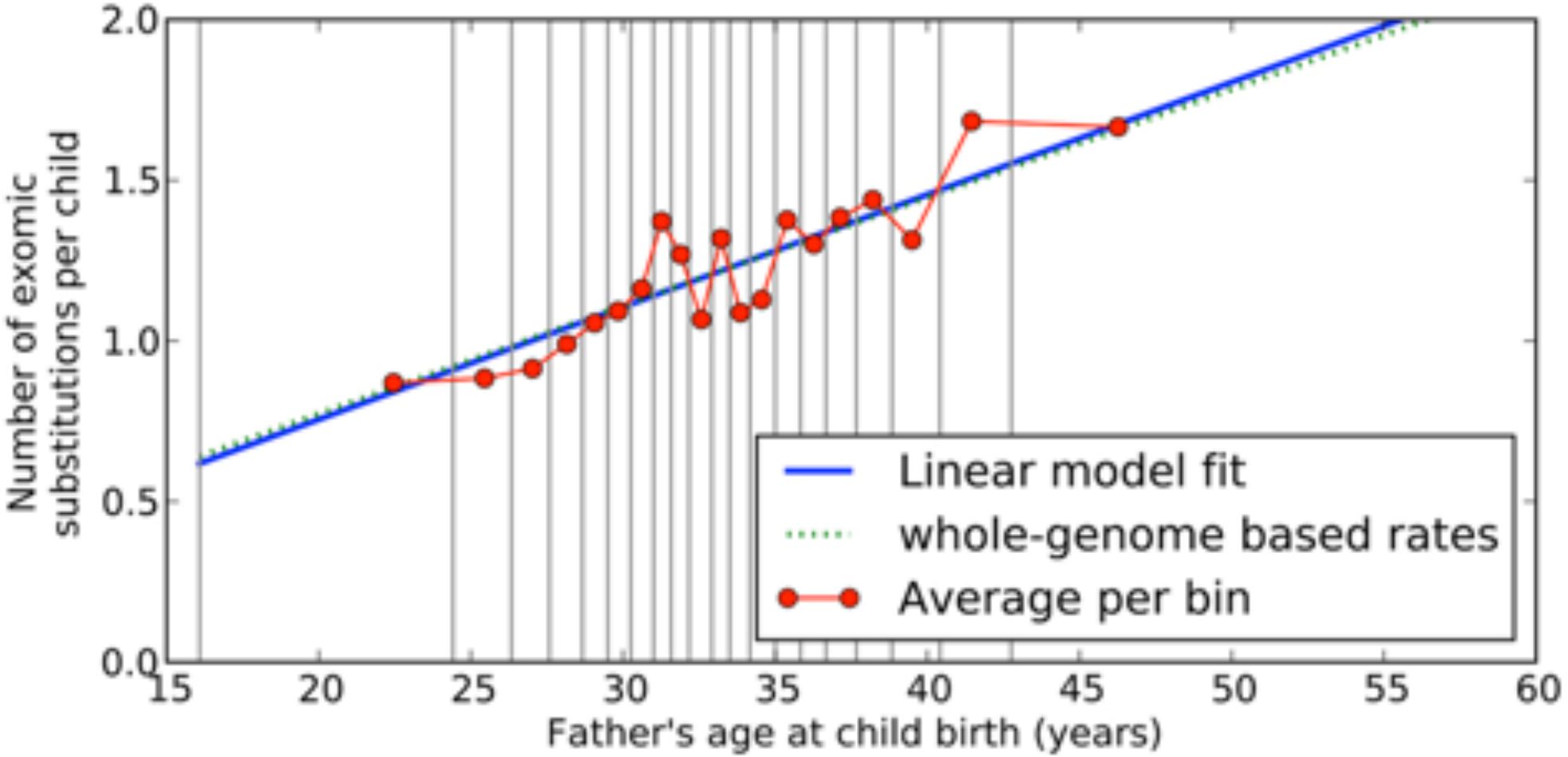
De novo Genetics of Autism

- 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

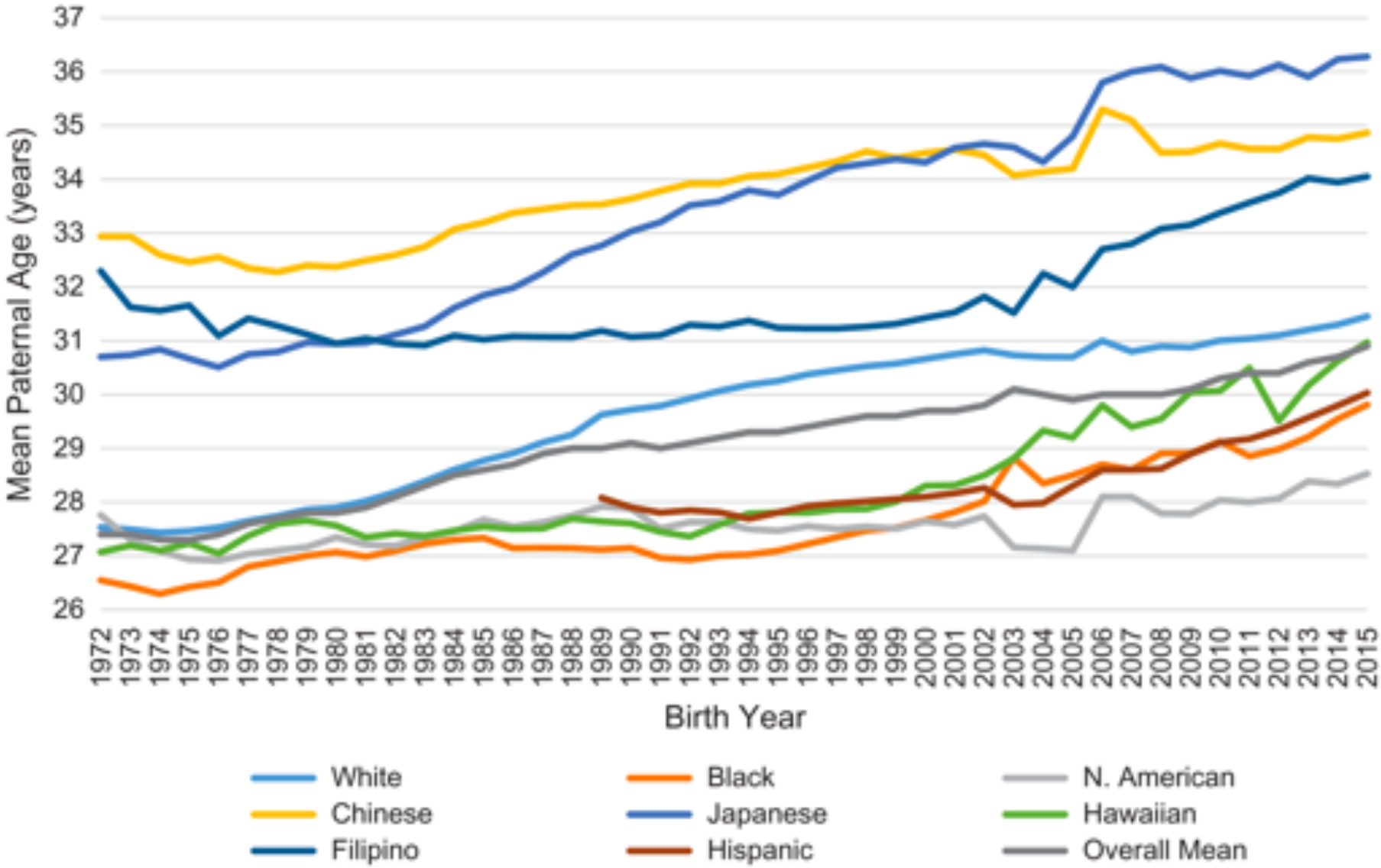
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

De novo Mutations in Men



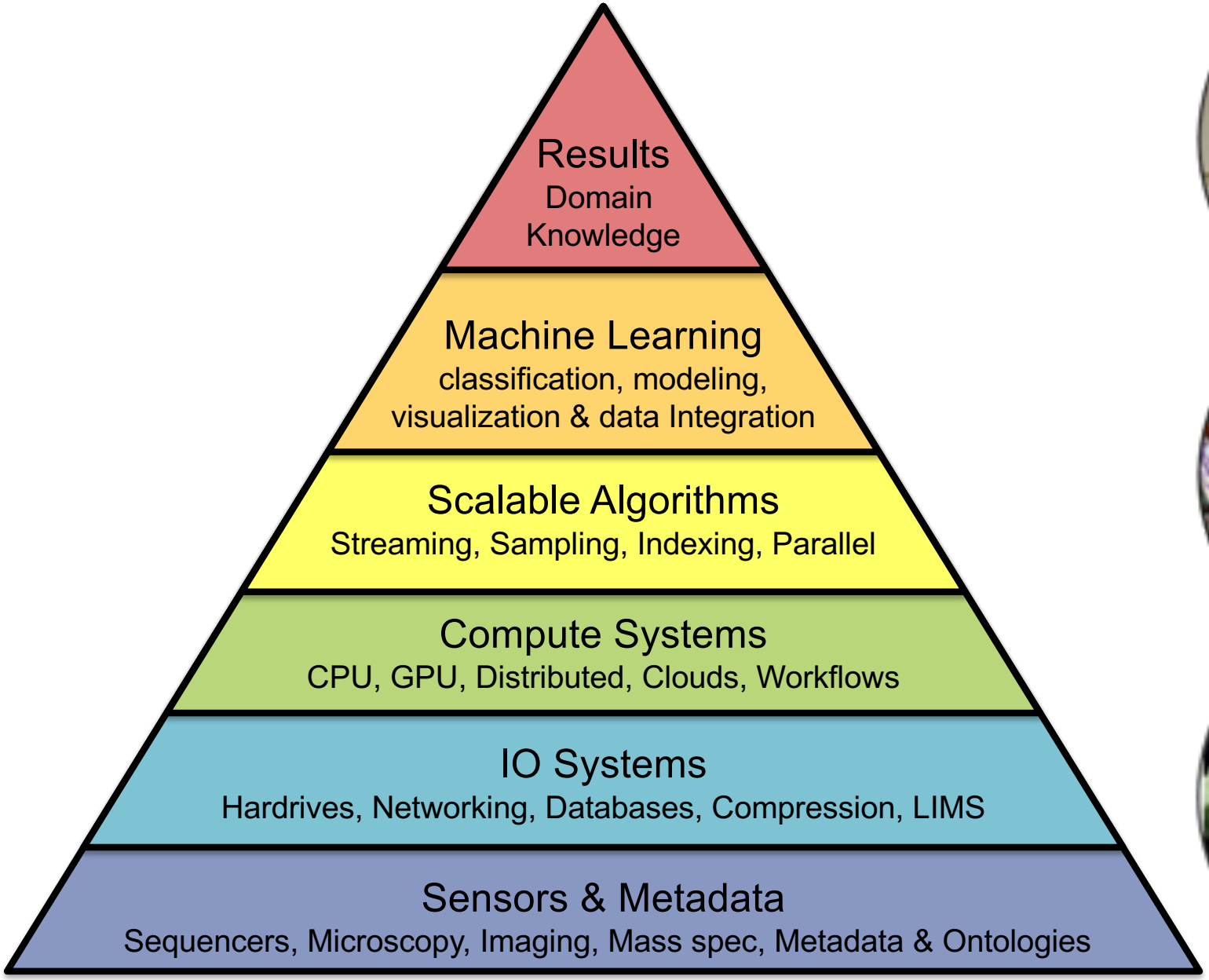
The contribution of de novo coding mutations to autism spectrum disorder
Iossifov *et al* (2014) *Nature*. doi:10.1038/nature13908

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*

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. Work on Assignment 1
    1. Set up Linux, set up Virtual Machine, set up Ubuntu
    2. Set up Dropbox for yourself!
    3. Get comfortable on the command line