

The Resurgence of Reference Quality Genome

Hayan Lee




Simons postdoctoral fellow
Prokaryote Super Program @ JGI

- **Background**
 - Third-Gen sequencing technology
- **The resurgence of reference quality genome (3Cs)**
 - Contiguity
 - The next version of Lander-Waterman Statistics
 - ✓ • How to model to predict de novo genome assembly performance
 - Support vector regression (SVR)
 - Completeness
 - Historical human genome quality by gene block analysis
 - Correctness
 - The effectiveness of long read sequencing technology in de novo assembly
- **Contributions**


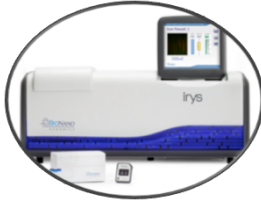
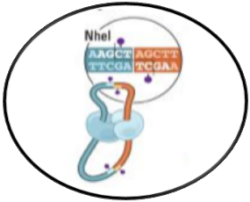
- **Sanger + BAC-by-BAC Era (1995 to 2007)**
 - Very high quality reference genomes for human, mouse, worm, fly, rice, Arabidopsis and a select few other high value species.
 - Contig sizes in the megabases, but costs in the 10s to 100s of millions of dollars
- **Next-Gen Era (2007 to current)**
 - Costs dropped, but genome quality suffered
 - Genome finishing was completely abandoned; “exon-sized” contigs
 - These low quality draft sequences are (1) missing important sequences, (2) lack context to discover regulatory elements or evolutionary patterns, and (3) contain many errors
- **Third-Gen Era (current)**
 - New biotechnologies (single molecule, chromatin assays, etc) and new algorithms (MHAP, LACHESIS, etc) are leading to the *Resurgence of Reference Quality Genomes*
 - *De novo* assemblies of human and other large genomes with contig sizes over 1Mbp.

Third-Gen Technology

- Long Read Sequencing: De novo assembly, SV analysis, phasing

Illumina/Moleculo  3-5kbp (Kuleshov et al. 2014)	Pacific Biosciences  10-15kbp (Berlin et al, 2014)	Oxford Nanopore  5-10kbp (Quick et al, 2014)
---	---	---

- Long Spanning Technology: Chromosome Scaffolding, SV analysis, phasing

Molecular Barcoding  30-60kbp (10Xgenomics.com)	Optical Mapping  25-100kbp (Putnam et al, 2015)	Chromatin Assays  100-150kbp (Cao et al, 2014)
---	---	--

Many Questions are raised but...

Given a target genome,

- How long should the read length be?
- What coverage should be used?

✓ Given the read length and coverage,

- **How long are contigs? <- Contiguity prediction**
- How many contigs?
- How many reads are in each contigs?
- How big are the gaps?

GENOMICS 2, 231-239 (1988)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

ERIC S. LANDER^{*†} AND MICHAEL S. WATERMAN[‡]

^{*}Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; [†]Harvard University, Cambridge, Massachusetts 02138; and [‡]Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

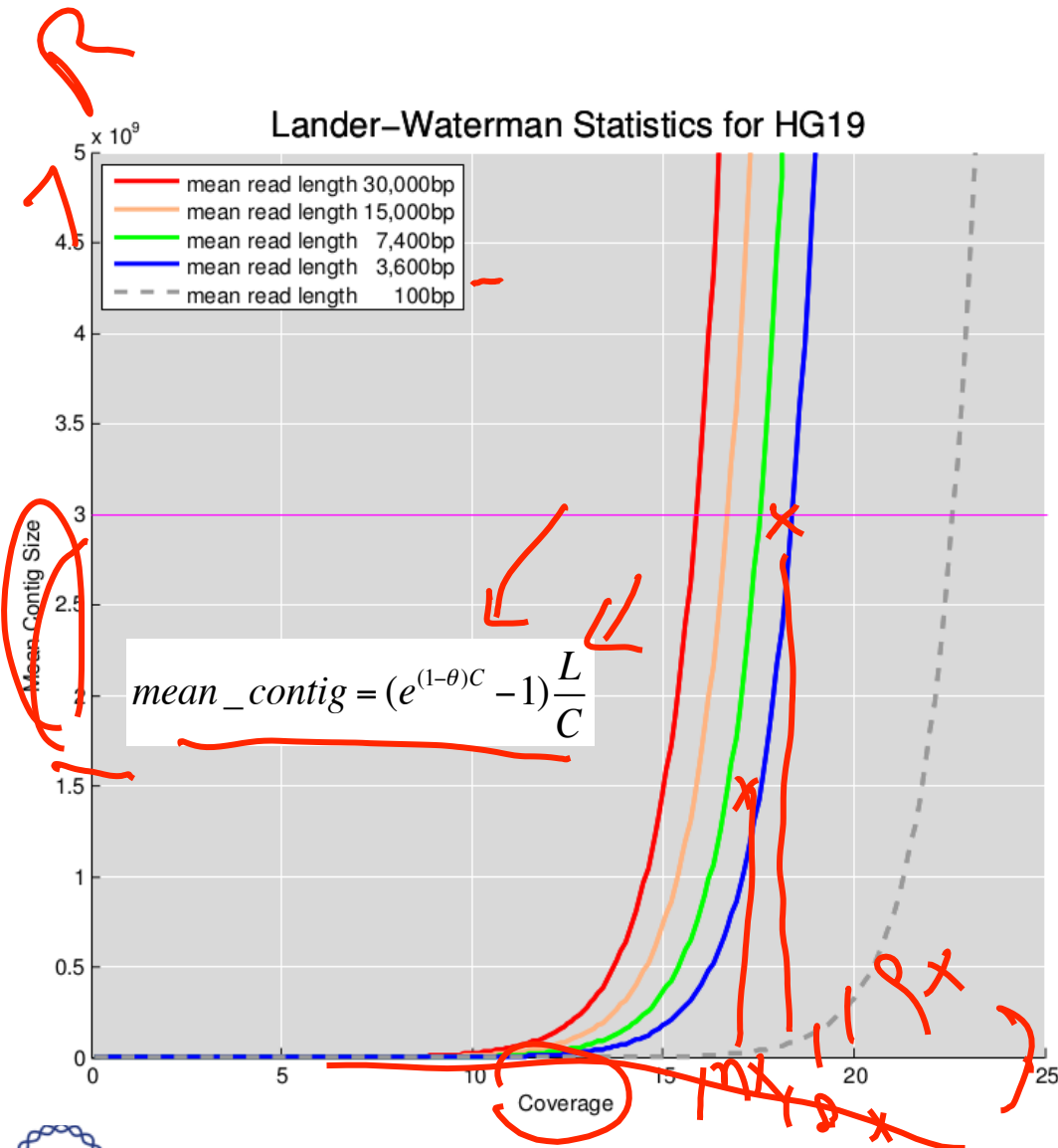
Received January 13, 1988; revised March 31, 1988

Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first “fingerprinting” a large number of clones chosen at random from a recombinant library and then inferring overlaps between clones with sufficiently similar fingerprints.

available region of up to several megabases and of studying its properties. In addition, the overlapping clones comprising the physical map would constitute the logical substrate for efforts to sequence an organism’s genome.

Recently, three pioneering efforts have investigated the feasibility of assembling physical maps by means of “fingerprinting” randomly chosen clones. The fingerprints consisted of information about restriction fragment lengths. Overlaps between clones were in-

HG19 Genome Assembly Performance by Lander-Waterman Statistics



Two key observations

1. Contig over genome size
2. Read Length vs. Coverage

Handwritten red notes:

$L < < <$

$lin < < exp$

Technology vs. Money

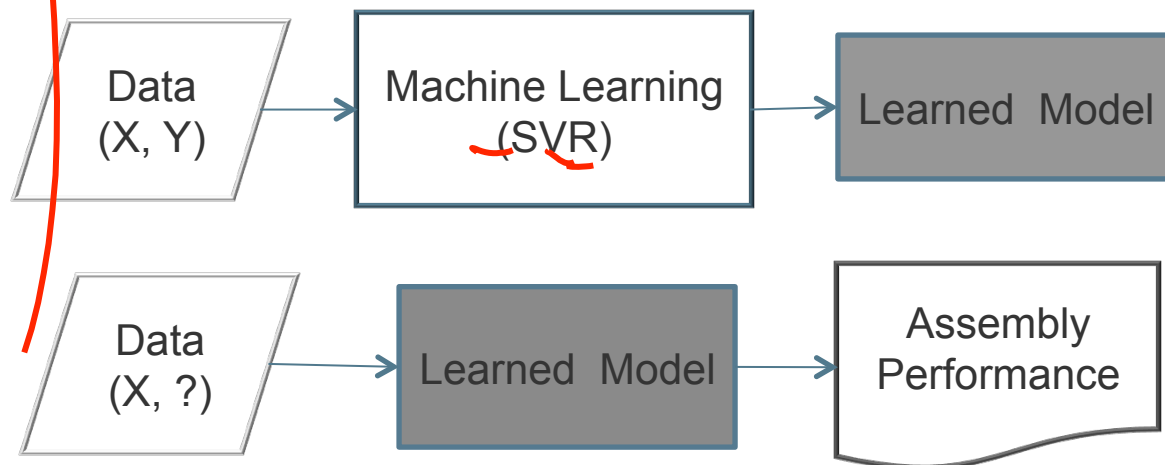
Handwritten red notes:

$d \sim 2^w$

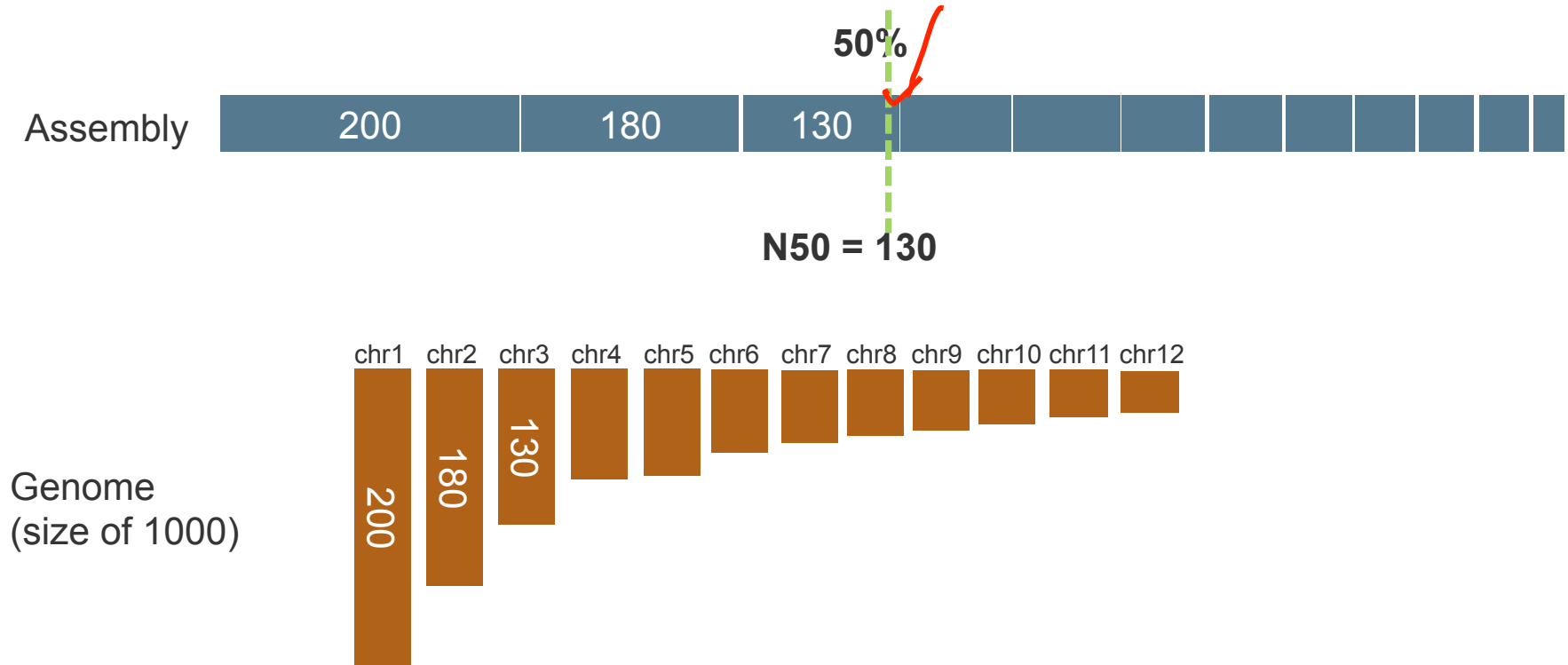
Empirical Data-driven Approach

Model Organism	ID	Genome Size
M.jannaschii	1	1,664,970
C.hydrogenoformans	2	2,401,520
E.coli	3	4,639,675
Y.pestis	4	4,653,728
B.anthraxis	5	5,227,293
A.mirum	6	8,248,144
yeast	7	12,157,105
Y.lipolytica	8	20,502,981
slime mold	9	34,338,145
Red bread mold	10	41,037,538
sea squirt	11	78,296,155
roundworm	12	100,272,276
green alga	13	112,305,447
arabidopsis	14	119,667,750
fruitfly	15	130,450,100
peach	16	227,252,106
rice	17	370,792,118
poplar	18	417,640,243
tomato	19	781,666,411
soybean	20	973,344,380
turkey	21	1,061,998,909
zebra fish	22	1,412,464,843
lizard	23	1,799,126,364
corn	24	2,066,432,718
mouse	25	2,654,895,218
human	26	3,095,693,983

- We carefully selected 26 species across tree of life and exhaustively analyzed their assemblies using simulated reads for 4 different length (6 for HG19) and 4 different coverage per species

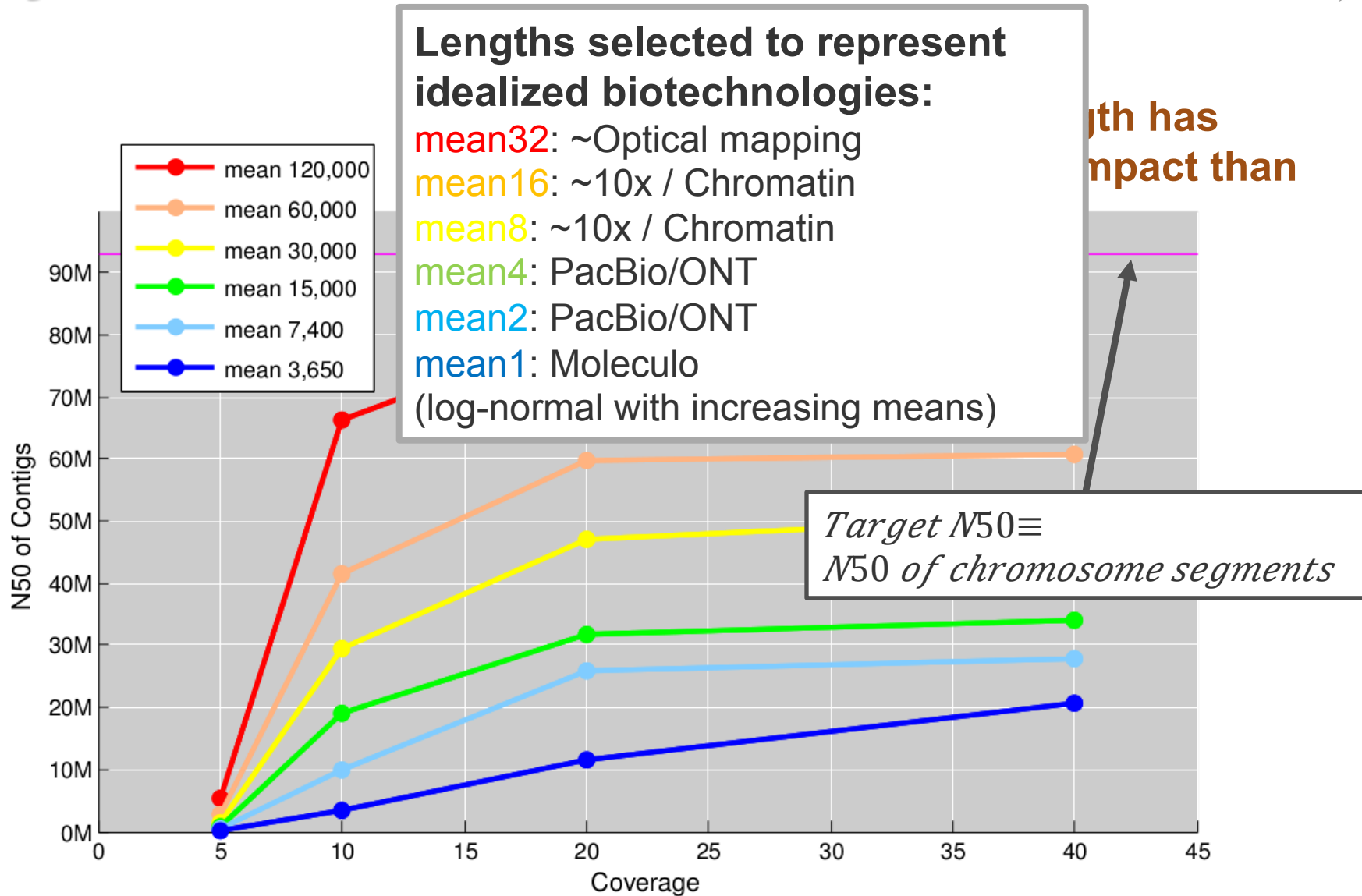


N50 : Contiguity Metric



- **N50 from assembly = 130**
- **N50 from chromosome segments (Target N50) = 130**
- **(Near) Perfect assembly**
 - N50 of assembly \approx N50 of chromosome segments

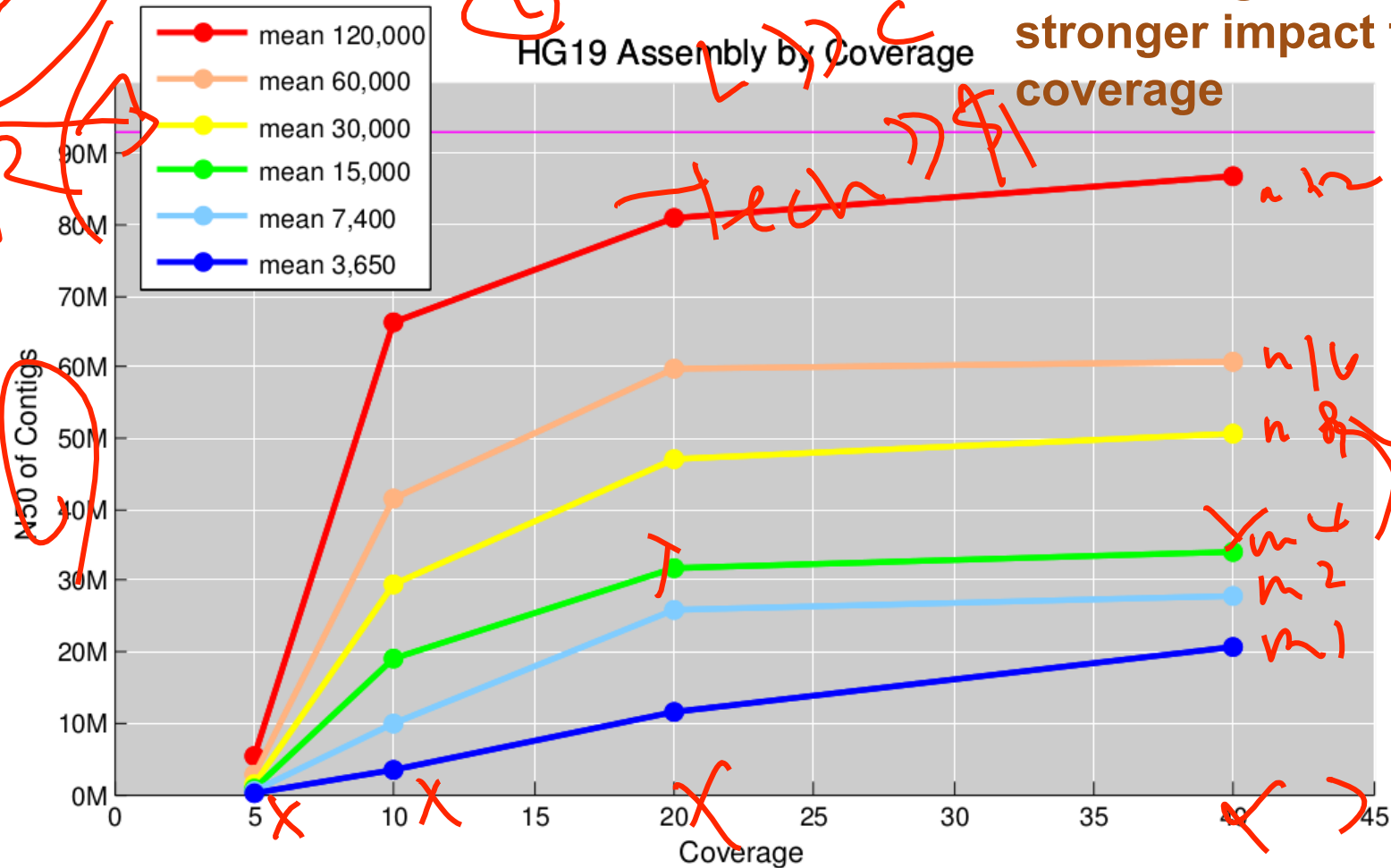
HG19 Genome Assembly Performance by Our Simulation



HG19 Genome Assembly Performance by Our Simulation

Read Length has stronger impact than coverage

HG19 Assembly by Coverage



Lander-Waterman Statistics

- Assumptions!!!
- If genome is a random sequence, it will work
- It works only in low coverage 3-5x
- It works for small genomes (< yeast)

Our Approach

- We tried to assume as little as possible.
- Instead of building on top of assumptions, we let the model learn from the data
- Empirical data-driven approach

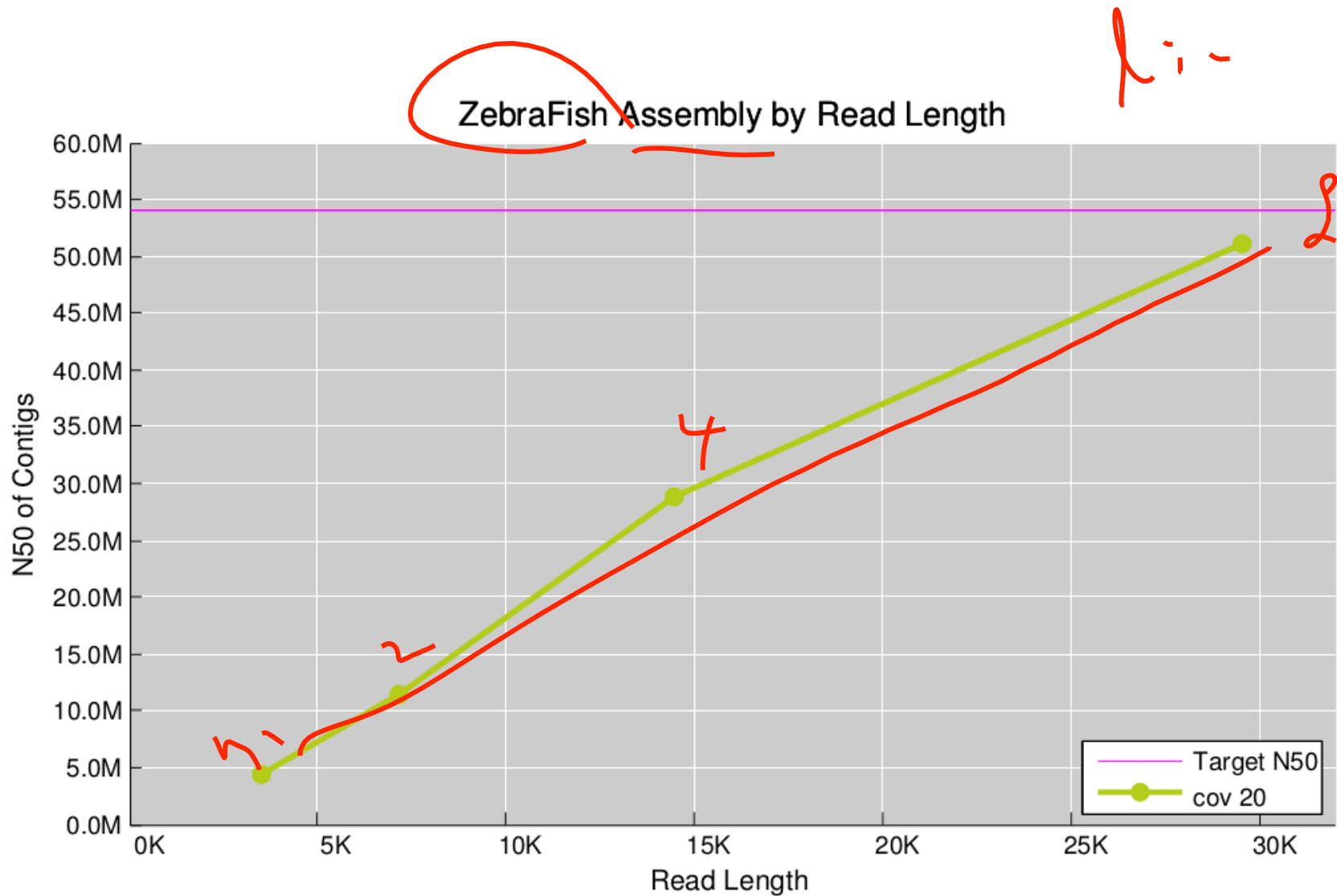
- To predict genome assembly contiguity

$$\text{Performance}(\%) \equiv \frac{N50 \text{ from Assembly}}{N50 \text{ from Chromosome Segments}} \times 100$$

$$\approx f \left(\begin{array}{c} \text{Read Length} \\ \text{Coverage} \\ \text{Genome Size} \\ \text{Repeat} \end{array} \right)$$

Read Length

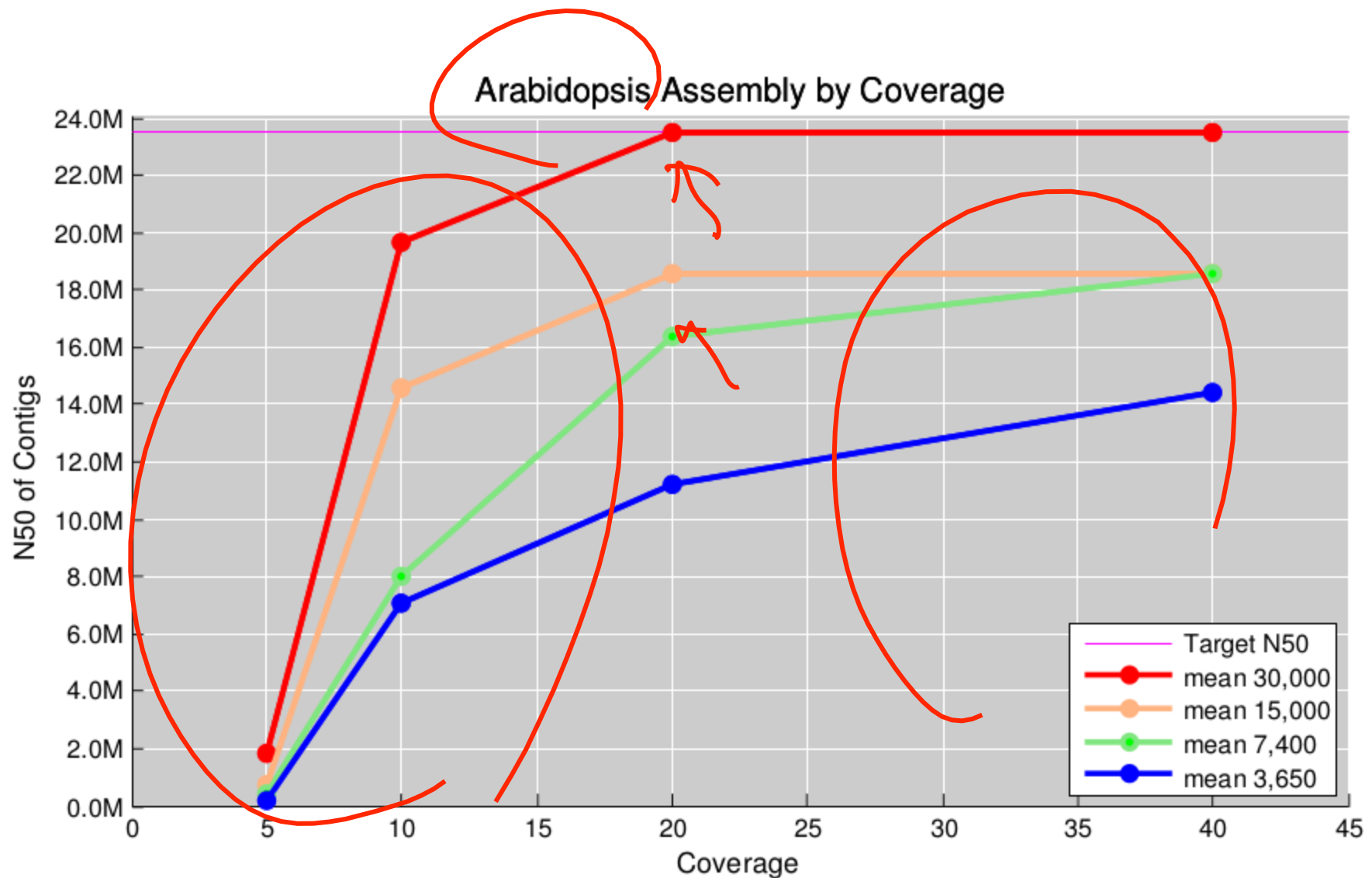
- Read length is very important
- A matter of technology
- The longer is the better
- Quality was important but can be corrected
 - PacBio produces long reads, but low quality (~15% error rate)
 - Error correction pipeline are developed
 - Errors are corrected very accurately up to 99%



Coverage

- A matter of money
- Using perfect reads, assembly performance increased for most genomes : Lower bound
- Using real reads, overall performance line will shift to the higher coverage
- The higher is the better (?)
- But still it suggests that there would be a threshold that can maximize your return on investment (ROI)

Coverage



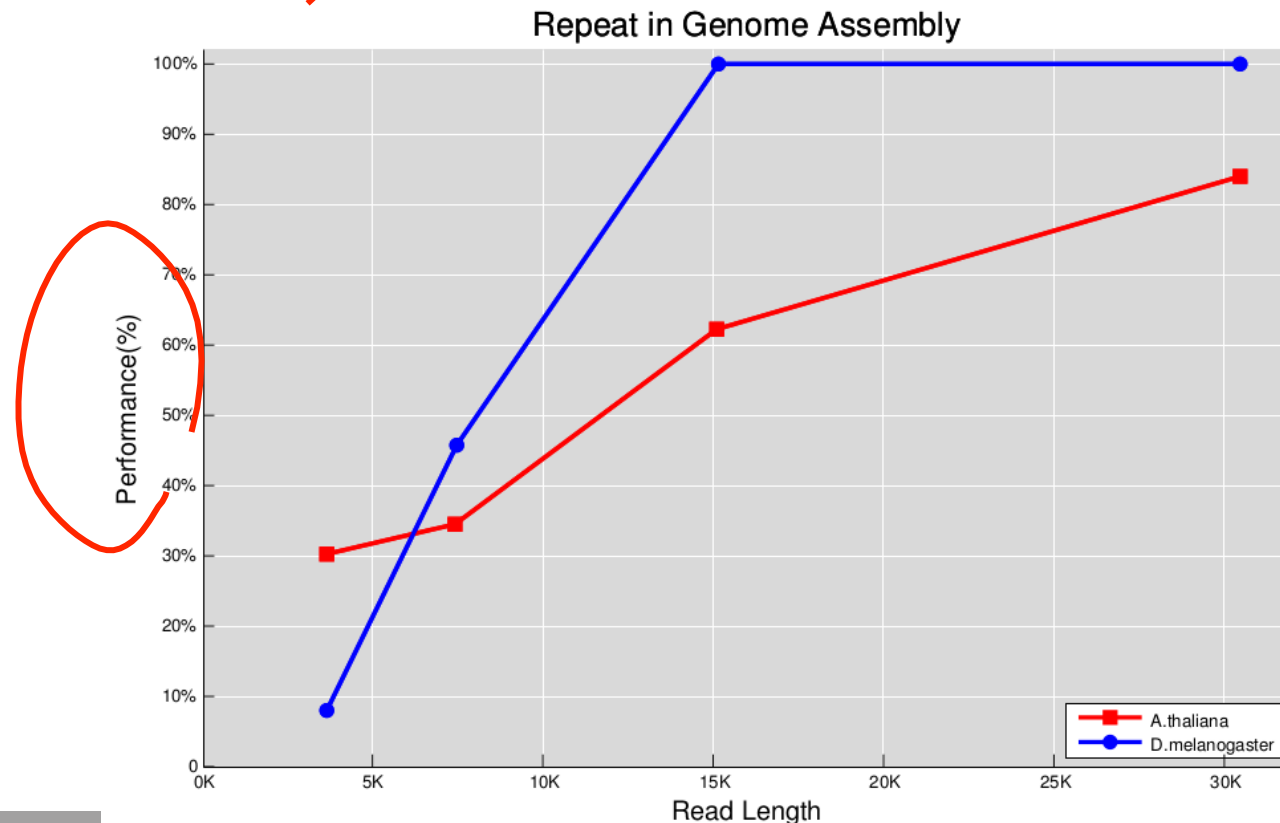
Repeats

- Genome is not a random sequence
- Repeat hurts genome assembly performance
- Isolating the impact of repeats is not trivial
- Quantifying repeat characteristics is not trivial as well
 - The longest repeat size
 - # of repeats > read length

Assembly Challenge (3)

Repeats

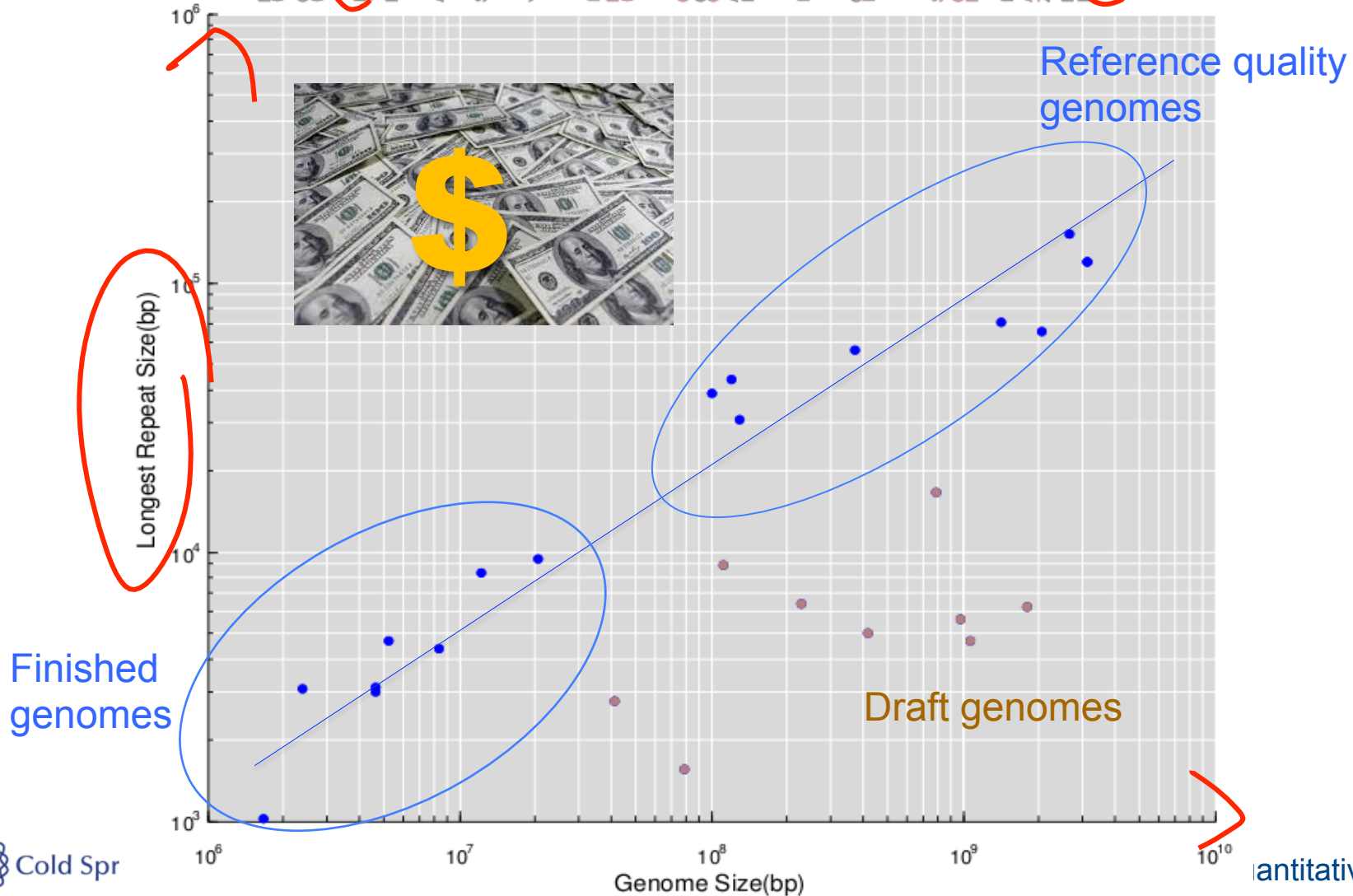
	Arabidopsis (120M) Longest repeat: 44kbp	Fruit fly (130M) Longest repeat: 30kbp
Mean Read Length	# of repeats > read length	# of repeats > read length
3,650	210	5564
7,400	112	394
15,000	44	8
30,000	14	2



X

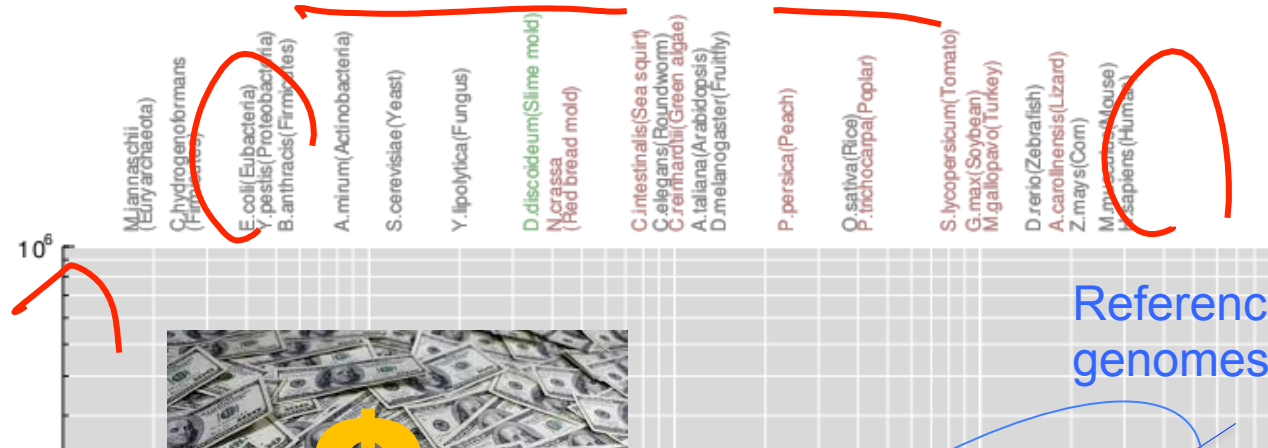
Longest Repeat Size and Genome Size

M. jannaschii (Euryarchaeota)
C. hydrogeniformans (Firmicutes)
E. coli (Eubacteria)
Y. pestis (Proteobacteria)
B. anthracis (Firmicutes)
A. minum (Actinobacteria)
S. cerevisiae (Yeast)
Y. lipolytica (Fungus)
D. discoideum (Slime mold)
N. crassa (Red bread mold)
C. intestinalis (Sea squirt)
C. elegans (Roundworm)
C. reinhardtii (Green alga)
A. thaliana (Arabidopsis)
D. melanogaster (Fruitfly)
P. persica (Peach)
O. sativa (Rice)
P. trichocarpa (Poplar)
S. lycopersicum (Tomato)
G. max (Soybean)
M. gallopavo (Turkey)
D. rerio (Zebrafish)
A. carolinensis (Lizard)
Z. mays (Corn)
M. musculus (Mouse)
H. sapiens (Human)

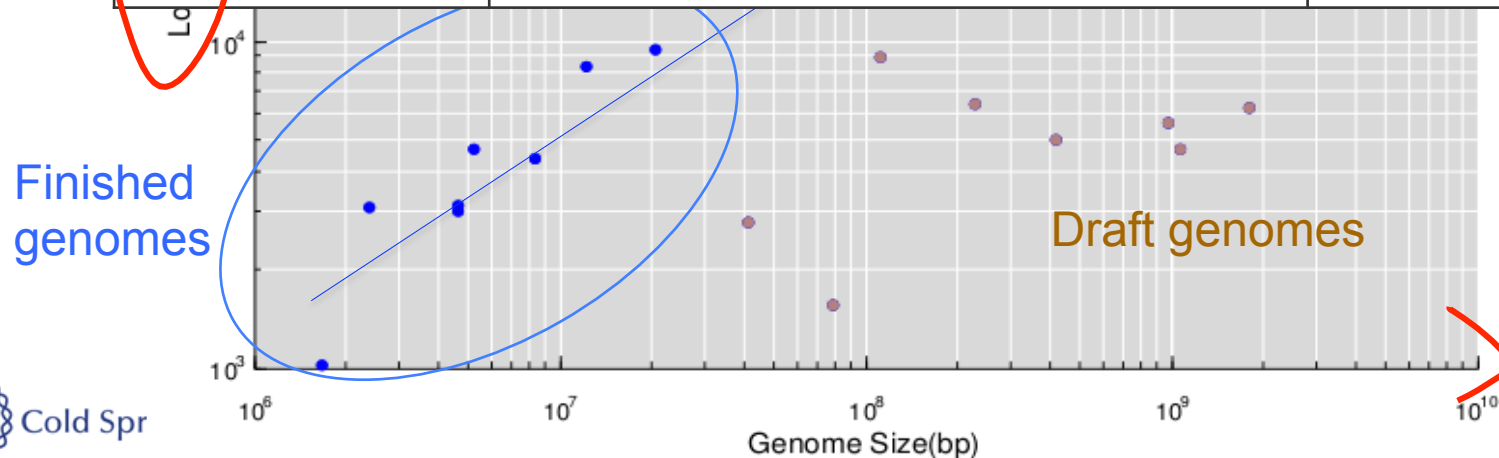


X

Longest Repeat Size and Genome Size



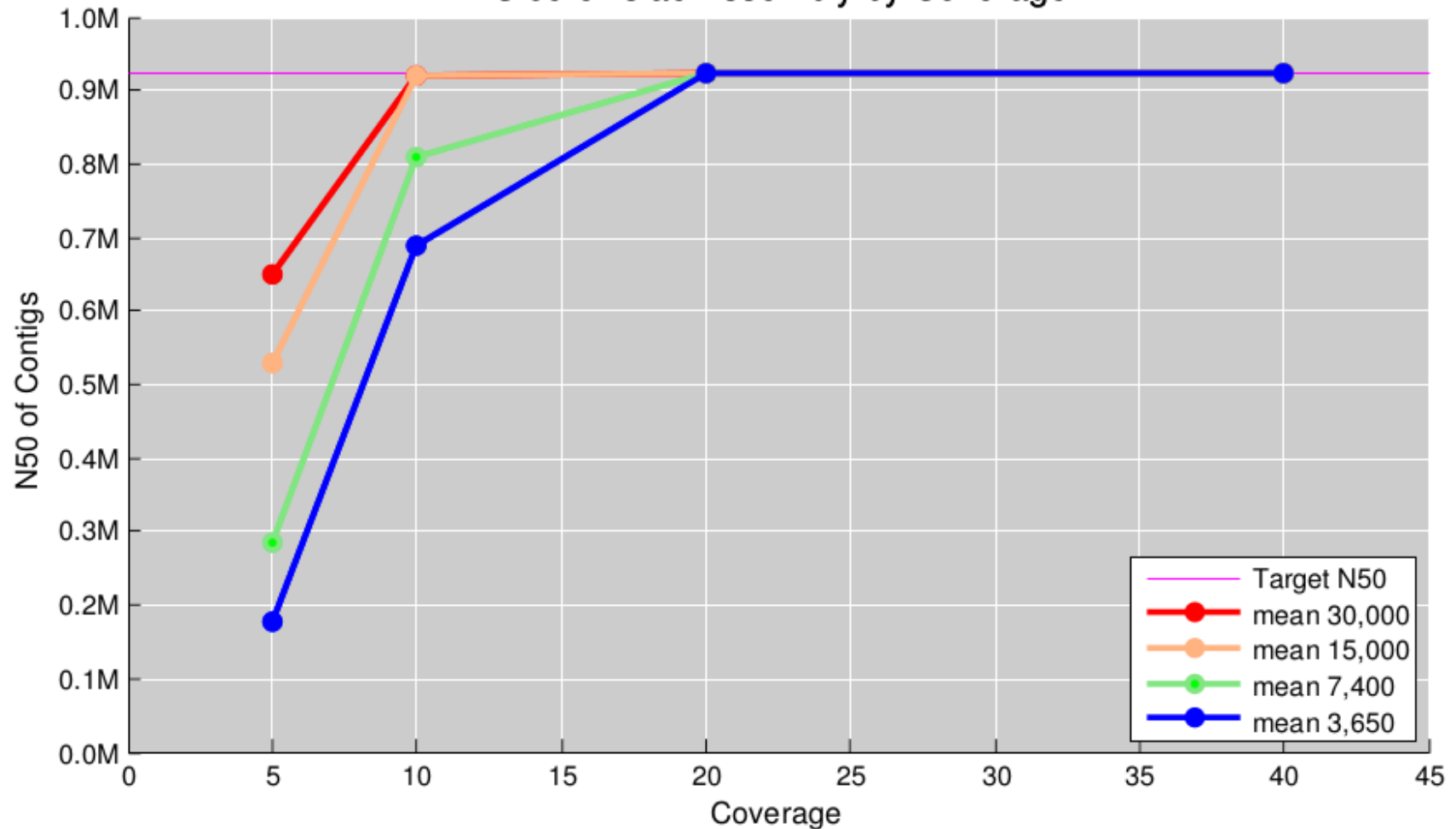
Category	Description	Examples
Finished genome	All (or almost) bases are resolved with high confidence Quality is guaranteed as well as quantity.	E.coli, Yeast
Reference genome	Quantity is well achieved but quality need to be improved (% of Ns, gene order etc.)	Human
Draft genome	Even quality needs to be improved, short contigs Hard to expect quality. Gene are still found but unlikely to identify regulation networks.	Poplar, Turkey, Tomato, Lizard etc.



Genome Size

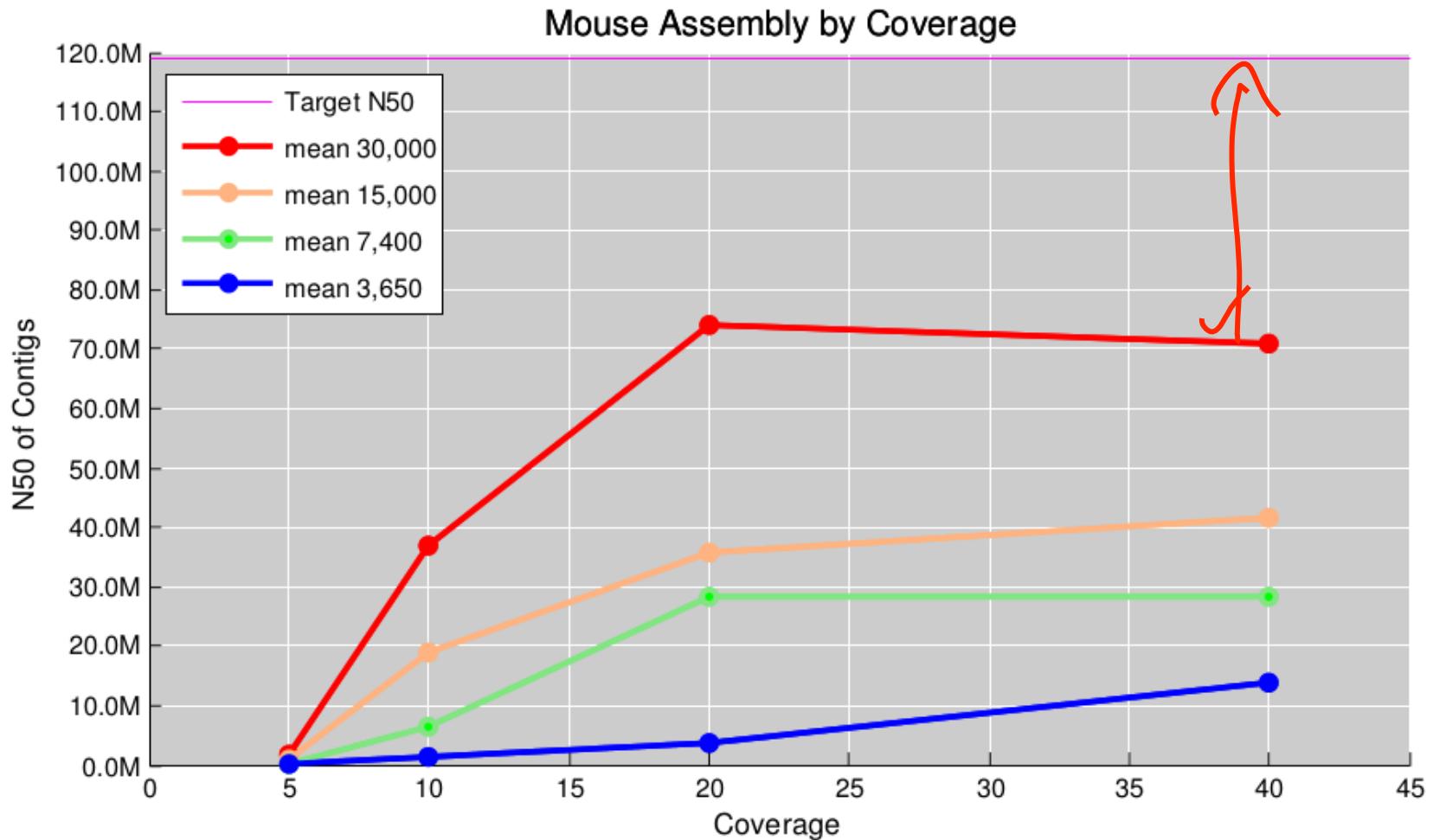
- Increase the assembly complexity
- Make a hard problem harder.

S.cerevisiae Assembly by Coverage



Assembly Challenge (4)

Genome Size



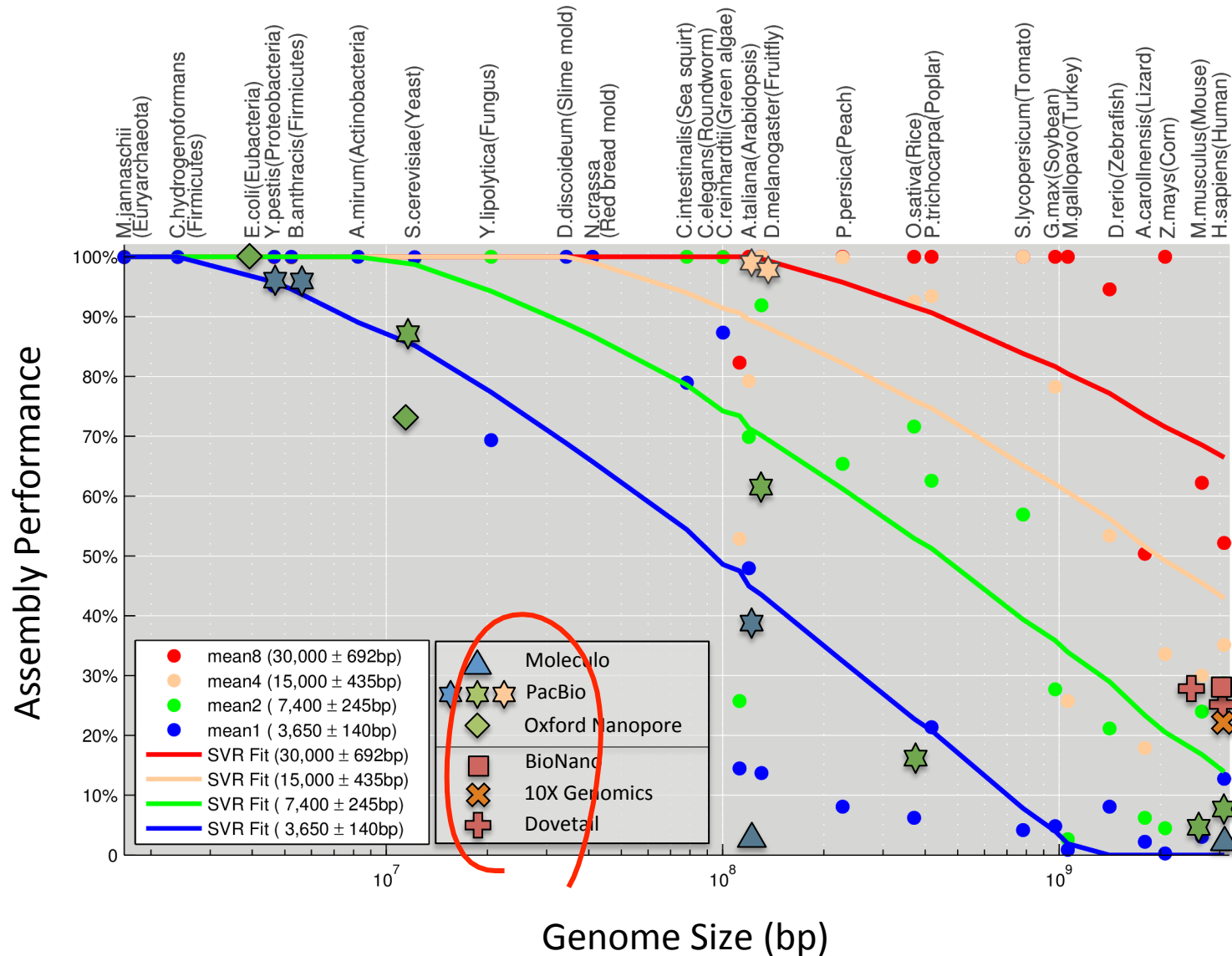
- **Correlation Coefficient**

- Performance vs. genome size
 - $R = -0.38$
- Performance vs. read length
 - $R = 0.2$
- Performance and *log* (genome size)
 - $R = -0.49$
- Performance and *log* (read length)
 - $R = 0.32$

- **Inputs for Support Vector Regression**

- Performance and *log* (genome size)/ *log* (read length)
 - $R = 0.6$
- Performance and *log* (coverage)
 - $R = 0.58$
- Performance and *log* (# of repeats longer than read length)
 - $R = -0.44$

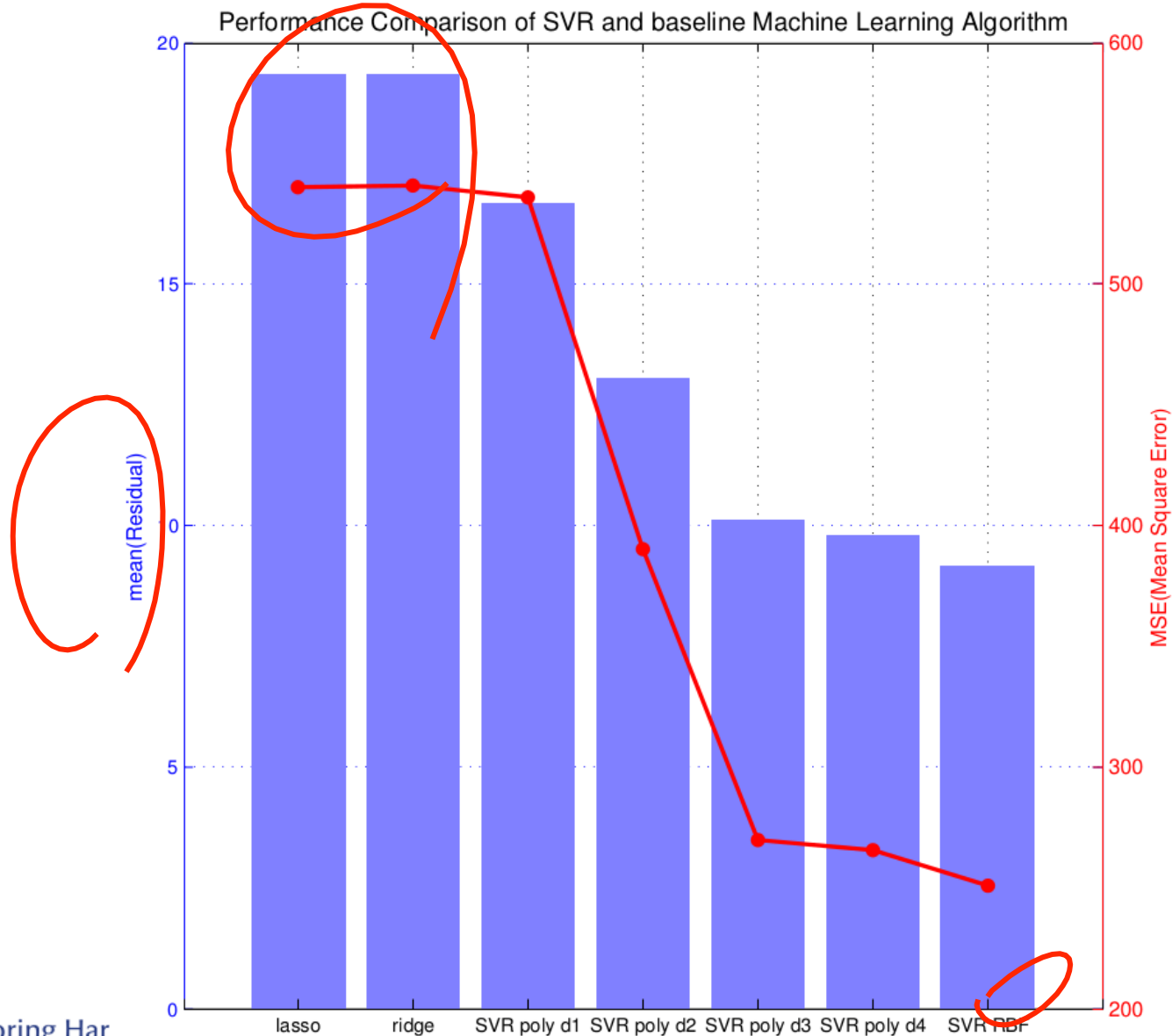
Reference Genome Quality



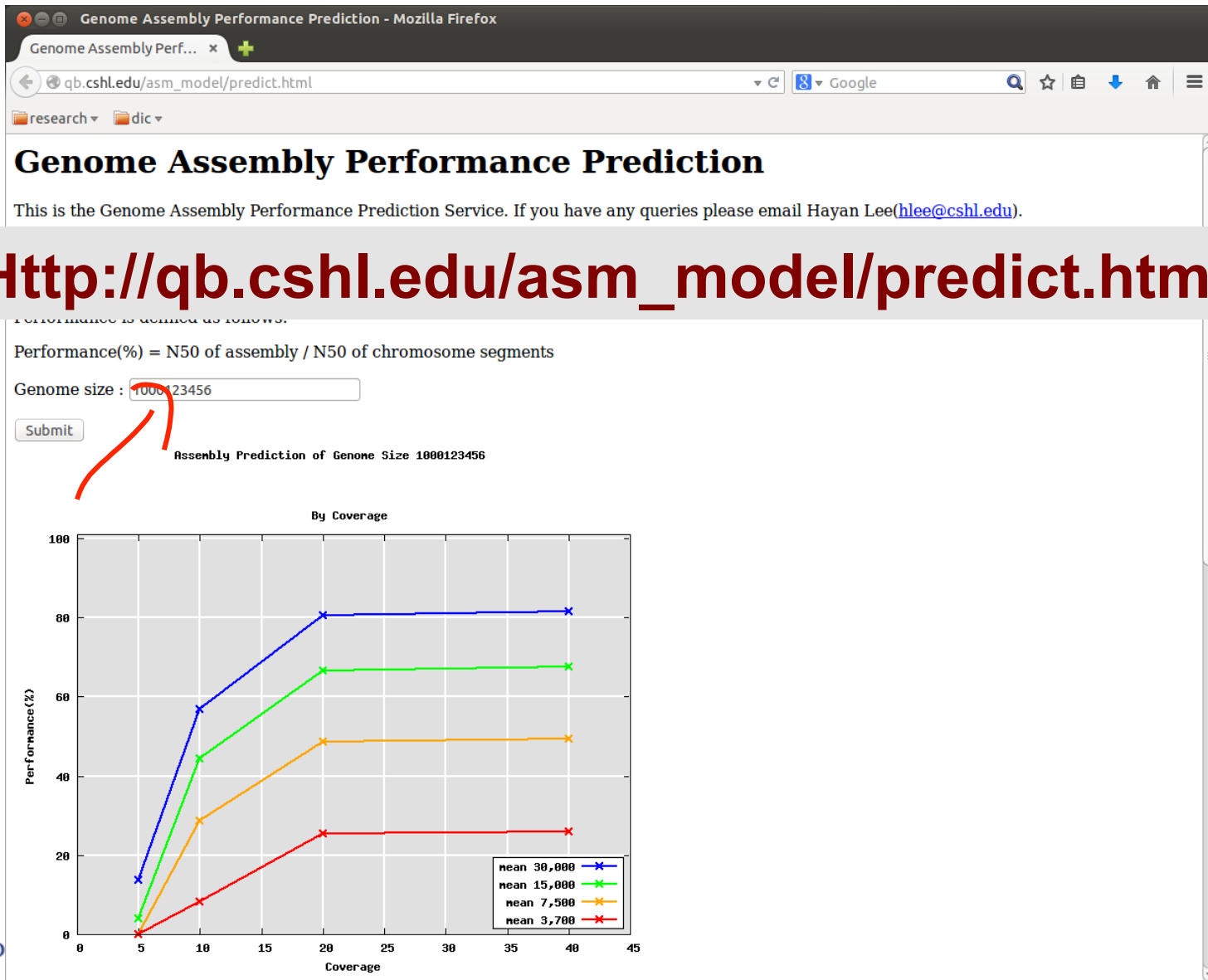
- **K-fold Cross Validation**
- **A variation of Leave-One-Out Cross Validation (LOOCV)**
- **Leave one species out approach (LOSO) <- Our approach**
 - A variation of Leave-One-Out Cross Validation (LOOCV)
 - Use 25 species as training data, test 1 species to measure predictive power
 - Avoid overfitting
- **Model selection by predictive power**



Prediction Performance

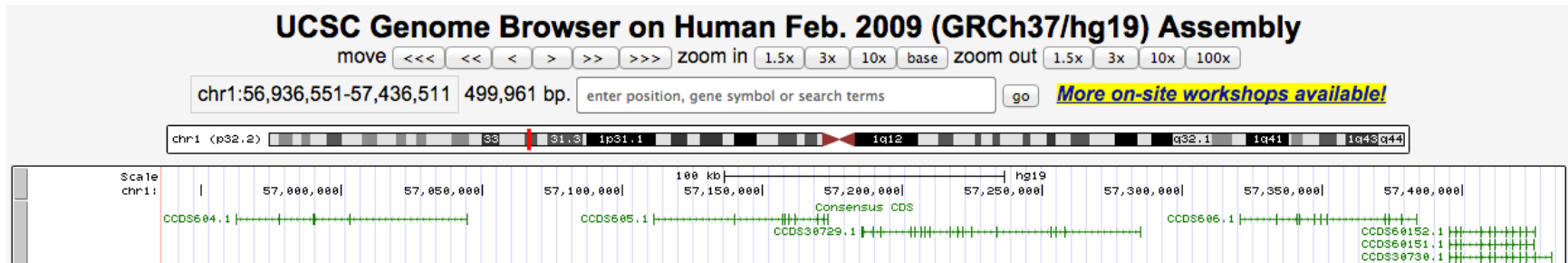


Web Service for Contiguity Prediction



Completeness

Human Reference Genome Quality by gene block analysis



gene1

gene2

gene1 - Gene

gene2

gene5

gene10

gene20

gene50

gene100

gene200

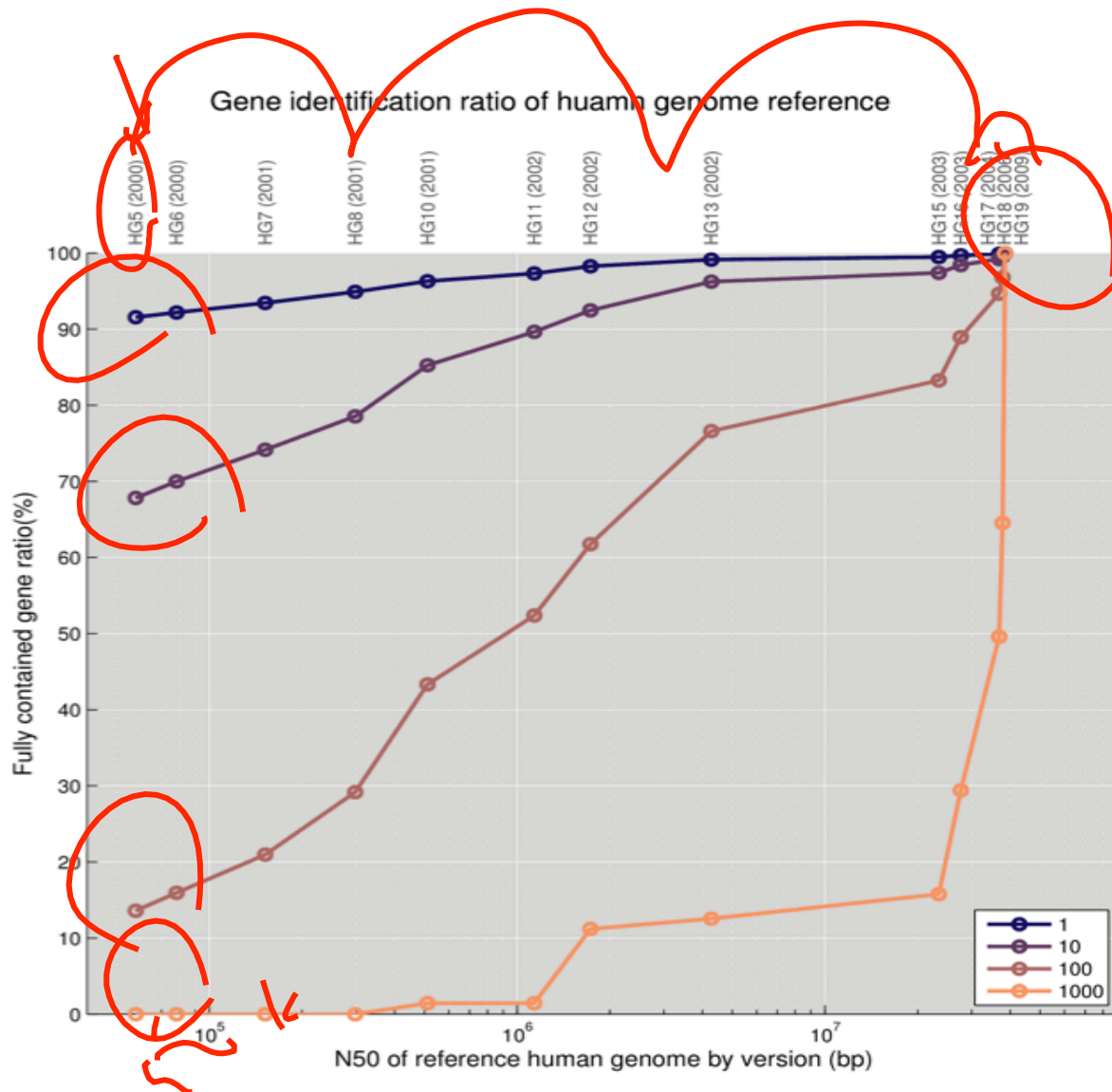
gene500

gene1000 - Chromosome structure

Regulatory elements

Synteny blocks

Completeness Human Reference Genome Quality by gene block analysis



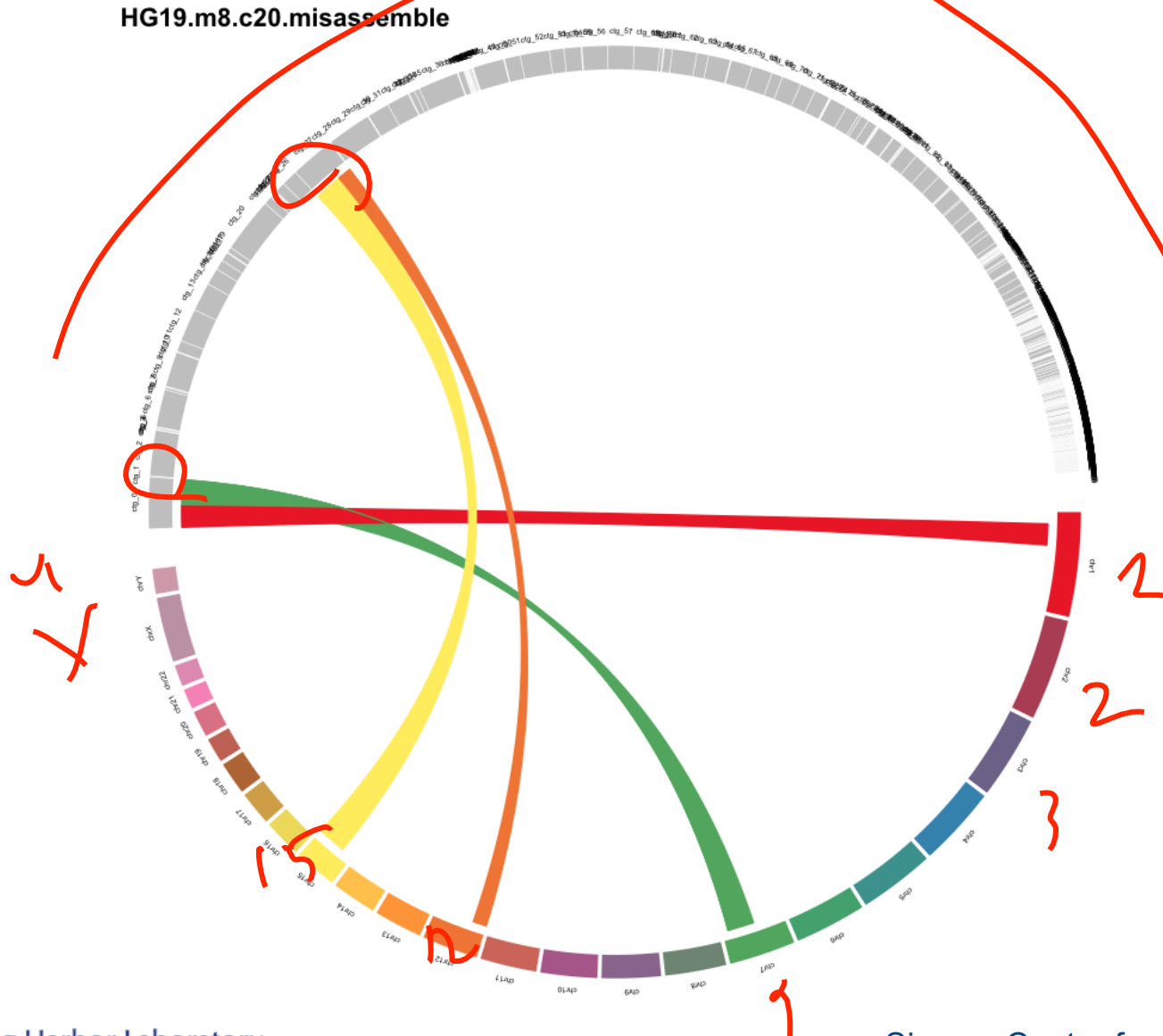
Larger contigs and scaffolds empowers analysis at every possible level.

- SNPs (~10k clinically relevant)
- Genes
- Regulatory elements
- Synteny blocks
- Chromosome structure

Gene
Regulatory elements
Synteny blocks
Chromosome structure

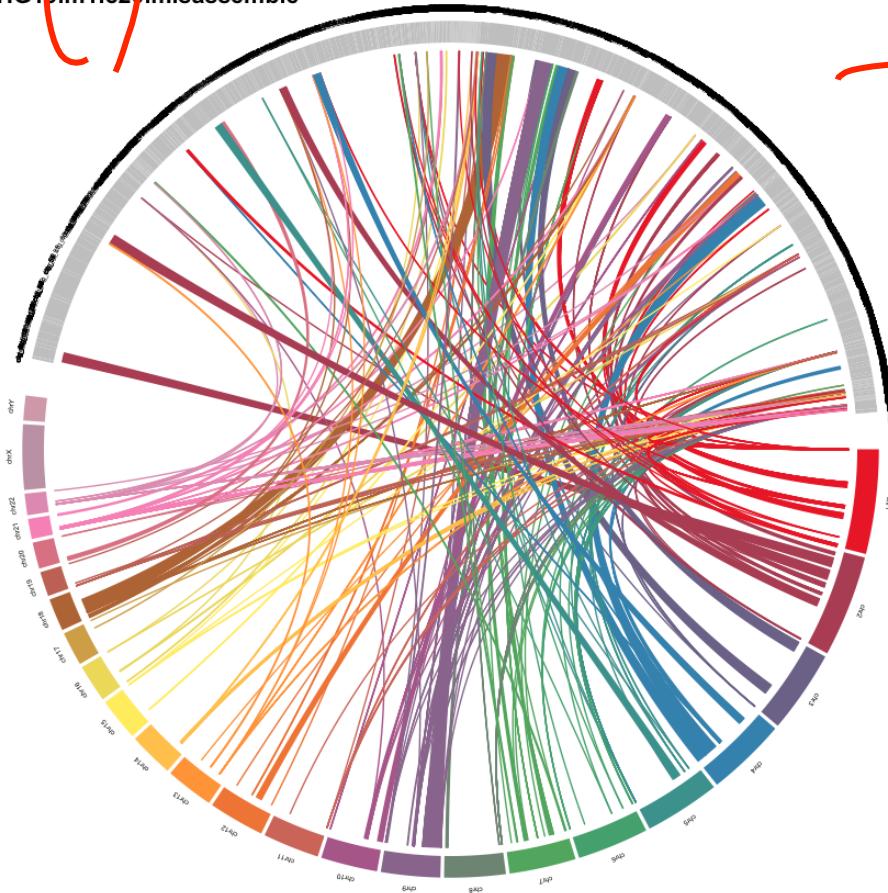
Correctness

Misassembly - A critical error in de novo assembly

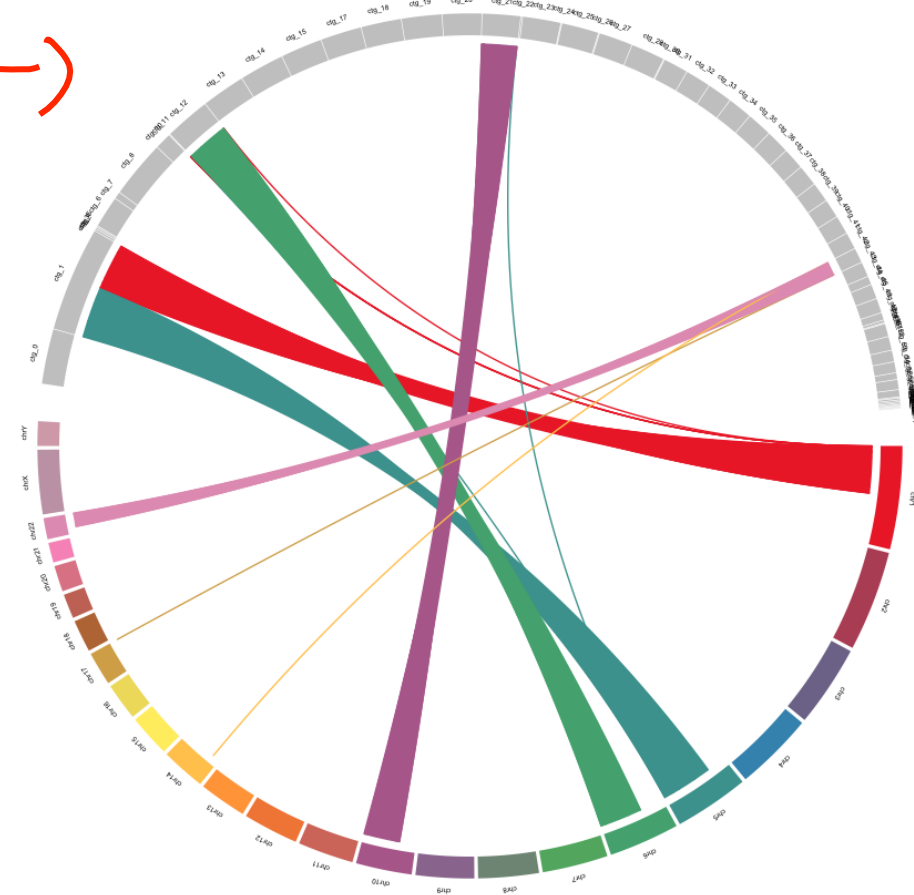


Misassembly Analysis in HG19

HG19.m1.c20.misassemble



HG19.m32.c20.misassemble



Long read sequencing technology helps to reduce both misassembly and breaks thus increase correctness of de novo genome assembly

	Lander-Waterman Statistics	Lee-Schatz Model
Features	Read Length (L) Coverage (C)	Read Length (L) Coverage (C) Genome Size (G) Repeats (R)
Methodology	Hypothesis driven	Data driven
Algorithm	Poisson distribution	Support Vector Regression

The resurgence of reference quality genomes

- New long read sequencing and long span technologies are dramatically improving de novo genome assemblies

We can predict the new genome assembly performance in 15% of error residual boundary

- Read length, coverage and genome size used explicitly
- Repeats are included implicitly

$$(e^{(1-\theta)C} - 1) \frac{L}{C} \propto L \cdot e^C$$

$$L \cdot e^C$$

$$L \cdot C$$

$$L \cdot \log C$$

$$L \cdot \log C \cdot f(G)$$

$$L \cdot \log C \cdot f(G) \cdot g(R)$$

$$\underline{SVR(L, C, G, R)}$$

Acknowledgements



Schatz Lab

Michael Schatz
Fritz Sedlazeck

James Gurtowski
Sri Ramakrishnan
Han fang
Maria Nattestad
Rob Aboukhalil
Tyler Garvin
Mohammad Amin
Shoshana Marcus

McCombie Lab

Dick McCombie
Sara Goodwin



✓ Shinjae Yoo



Stony Brook University



Microsoft®

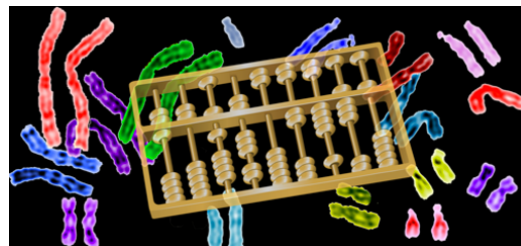
Research

Ravi Pandya
Bob Davidson
David Heckerman



University of São Paulo

Gabriel R.A. Margarido
Jonas W. Gaiarsa
Carolina G. Lembke
Marie-Anne Van Sluys
Glaucia M. Souza



Berkeley
UNIVERSITY OF CALIFORNIA

Algorithmic Challenges in Genomics
Jan. 11 – May 13, 2016

Thank You

Q & A

The Resurgence of Reference Quality Genomes

Hayan Lee^{1,2}, James Gurtowski¹, Shinjae Yoo³, Maria Nattestad⁵, Shoshana Marcus⁴, Sara Goodwin¹, W. Richard McCombie¹, and Michael C. Schatz^{1,4*}

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

²Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794

³Computational Science Center, Brookhaven National Laboratory, Upton, NY, 11973

⁴Department of Mathematics and Computer Science, Kingsborough Community College, City University of New York, Brooklyn, NY 11234

⁵Watson School of Biological Sciences, Cold Spring Harbor, NY, 11724