BERKELEY LAB
Lawrence Berkeley National Laboratory

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Long read sequencing technologies and its applications

Hayan Lee
Ph.D. Computer Science
Simons postdoctoral fellow @ Lawrence Berkeley National Laboratory
Research fellow @ Simons Institute of the theory of computing, UC Berkeley

Outline

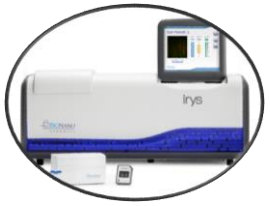
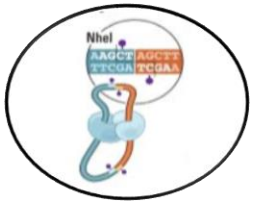
- **Background**
  - Long read sequencing technology
- **Mappability - The limitations of short read mapping illustrated by Genome Mappability Score (GMS)**
- **Assemblability - The Resurgence of reference quality genome (3Cs)**
  - The next version of Lander-Waterman Statistics (Contiguity)
  - Historical human genome quality by gene block analysis (Completeness)
  - The effectiveness of long reads in de novo assembly (Correctness)
  - De novo genome assembly for highly heterozygous complex genomes
    - Pineapple de novo assembly (1~2% heterozygous genome)
    - Sugarcane de novo assembly (~10% heterozygous and polyploid/aneuploid genome)
- **Detectability**
  - Long range structural variations of HER2 in BRSK3 using PacBio reads
- **The illusion of short read deep sequencing**

Third-Gen Technology

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

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

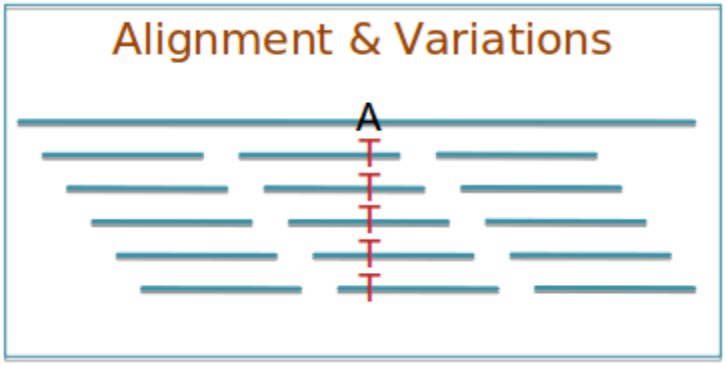
- **Long Mapping Technology: Chromosome Scaffolding, SV analysis, phasing**

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

Mappability

How well is the genome mapped by any read?

Short read mapping
(Resequencing)



- **Discovering genome variations**
- **Investigating the relationship between variations and phenotypes**
- **Profiling epigenetic activations and inactivations**
- **Measuring transcription rates**

1000 Genomes
A Deep Catalog of Human Genetic Variation

ENCODE

HMP

TCGA

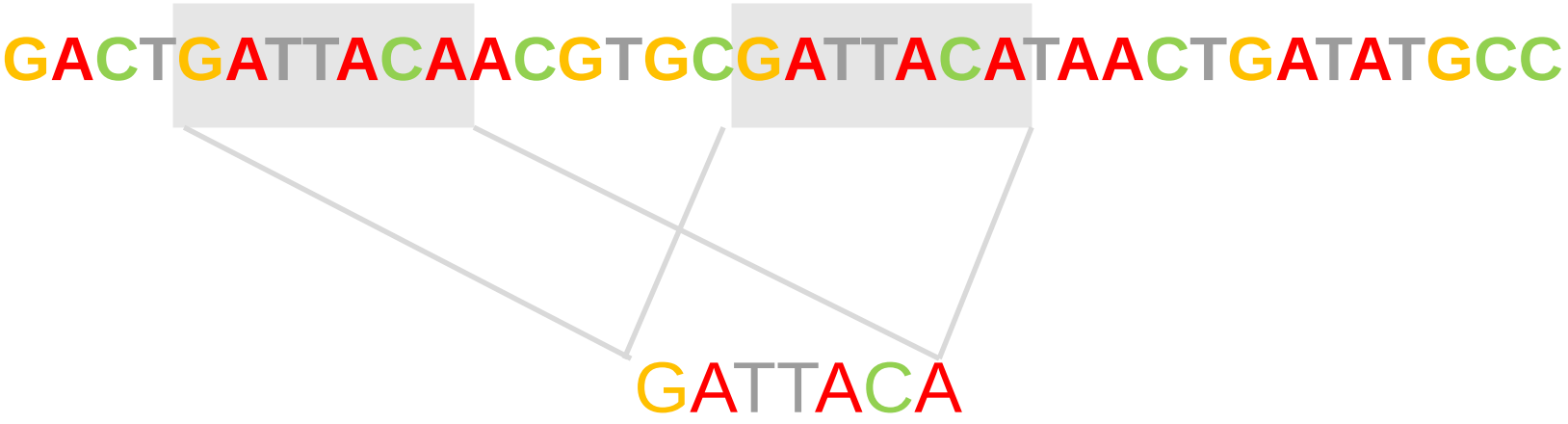
NCI | NHGRI

Cold Spring Harbor Laboratory

Simons Institute for the Theory of Computing

Berkeley UNIVERSITY OF CALIFORNIA

Repeats



Read Quality Score - MAQ(1)

$$Q_s = -10\log_{10}[Pr(read\ is\ wrongly\ mapped)]$$

$$Q_s = -10\log_{10}[1 - p_s(u \mid x, z)]$$

position

read

reference

$$P_s(u \mid x, z) = \frac{P(z \mid x, u)}{\sum_{v=1}^{L-l+1} P\,(z \mid x, v)}$$

The mapping quality score Q_S of a given alignment is typically written in Phred-scale

L = |x| the length of reference genome x,

l = |z| is a length of a read z

P(z|x, u), the probability of observing the particular read alignment

The posterior probability P_S is minimized when the alignment with the fewest mismatches is selected.

Q_S will be lower for reads that could be mapped to multiple locations with nearly the same number of mismatches and Q_S will be zero if there are multiple positions with the same minimum number of mismatches weighted by quality value.

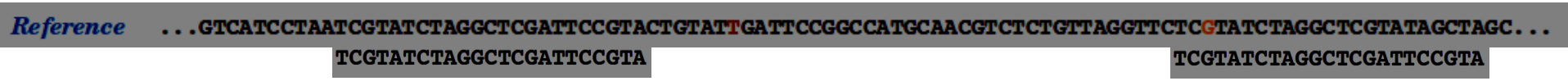
Read Quality Score – MAQ (2)

$$Q_s = -10\log_{10}[1 - p_s(u \mid x, z)]$$

$$P_s(u \mid x, z) = \frac{P(z \mid x, u)}{\sum_{v=1}^{L-l+1} P\,(z \mid x, v)}$$

- **X is a reference**
- **Z is a read**
- **U is a position**
- **L = |x| the length of reference genome x,**
- **l = |z| is a length of a read z**
  - Position u has 2 mismatches
  - Base quality scores are 20 for C, 10 for A
  - Error probability of C is 1%, A is 10%
  - Correctly mapped probability of position U is 0.1 %
- **Q: If a read z is (almost) uniquely mapped?**

Read Quality Score – MAQ (3)



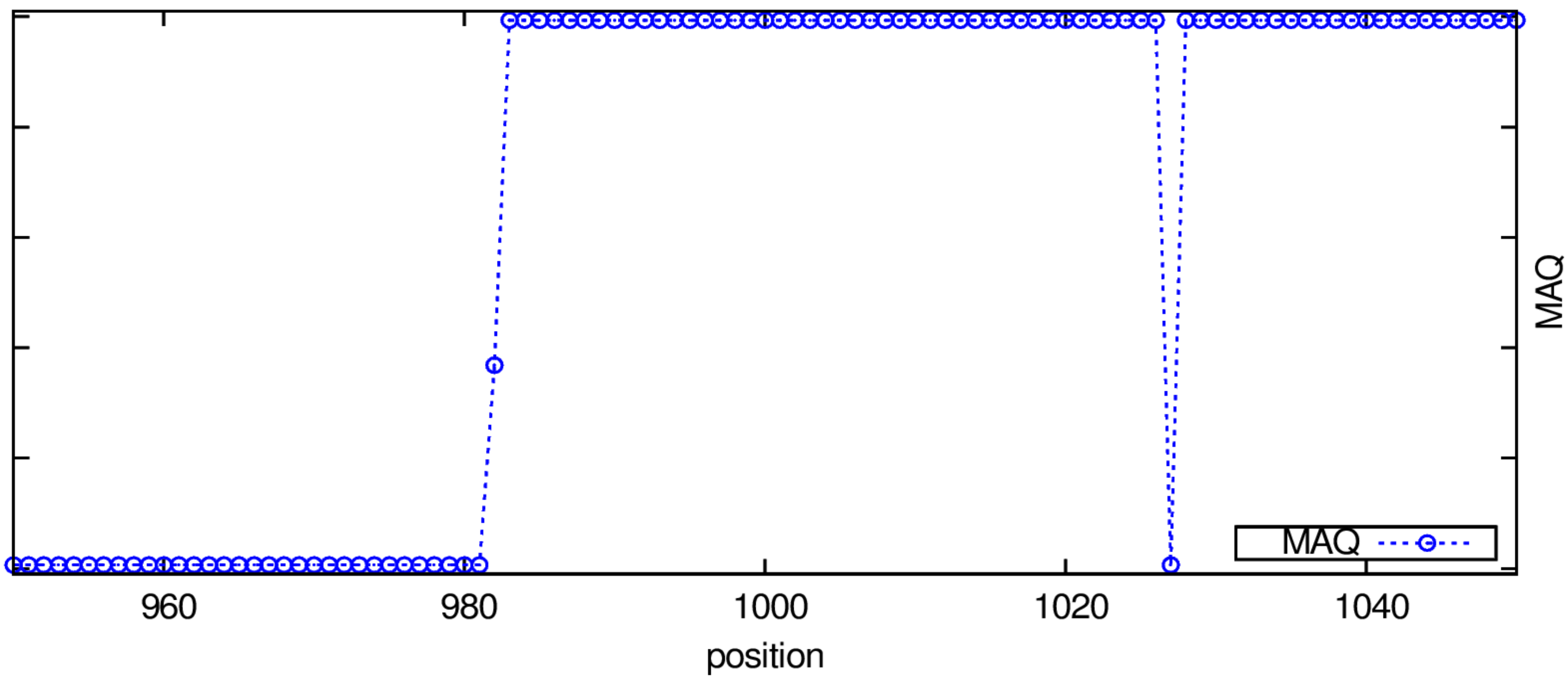
$$Q_s = -10\log_{10}[1 - p_s(u \mid x, z)]$$

$$P_s(u \mid x, z) = \frac{P(z \mid x, u)}{\sum_{v=1}^{L-l+1} P\,(z \mid x, v)}$$

- **X is a reference**
- **Z is a read**
- **U is a position**
- **L = |x| the length of reference genome x,**
- **l = |z| is a length of a read z**
- $P(z \mid x, u)$
  - Position u has 2 mismatches
  - Base quality scores are 20 for C, 10 for A
  - Error probability of C is 1%, A is 10%
  - Correctly mapped probability of position U is 0.1 %
- **Q: If a read z is (almost) uniquely mapped?**
- **Q: If a read z is mapped to many positions?**
- **Q: What is the reliability of a specific position?**
- **Q: Do we have a metric to measure such reliability in a consistent view?**

Read Quality Score – MAQ

Sensitivity of Read Mapping Score



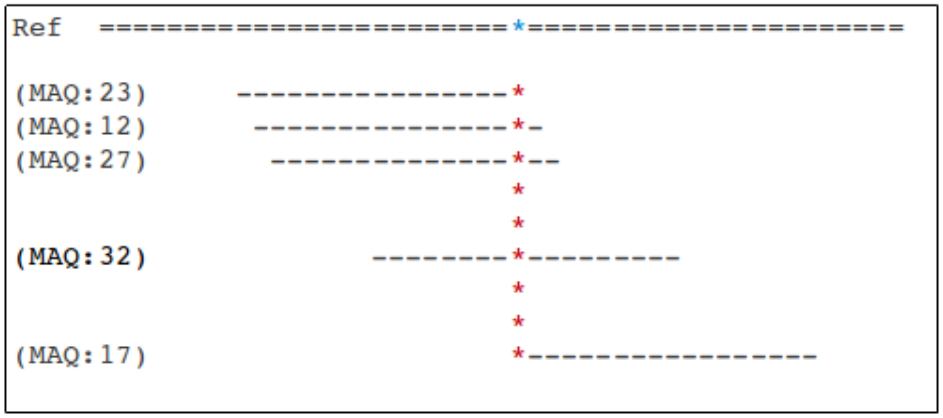
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA CTAGCATGCCCTAAGCCCGTAATGCAGTCATACTAGCTATCCTCGCCCTCTCCGTCAAGCTAC

The Global View
(GPS for a genome)

- **Challenges**
  - There is inherent uncertainty to mapping
  - Read quality score is very sensitive to a minute change
  - Base quality score is useful only inside a single read
  - Read quality score is assigned to each read not a position of a genome, thus provides only local view
  - However, there is no tool to measure the reliability of each position of reference genome in a global perspective.

- **Our approach**
  - We need more stable “GPS” in genome
  - All possible reads should be considered

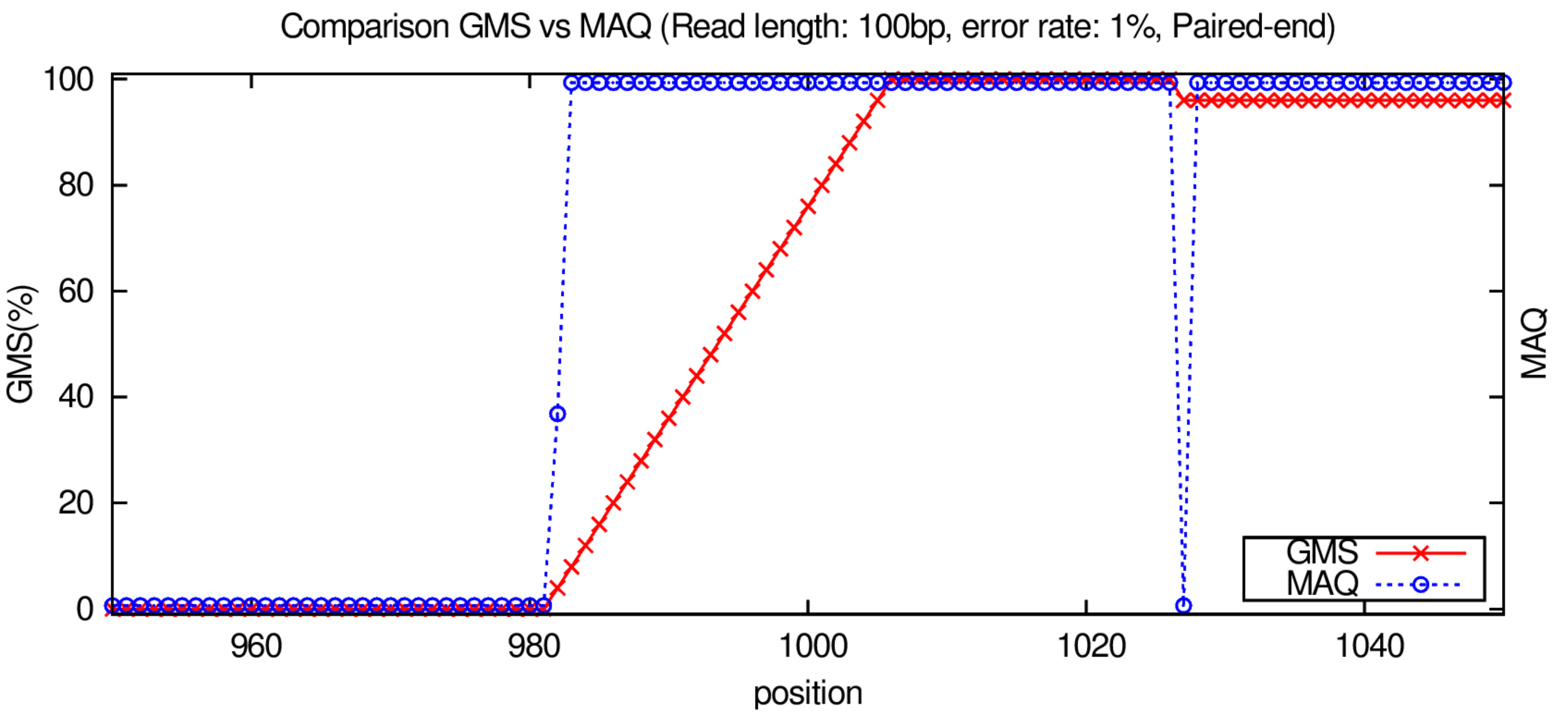
Genome Mappability Score (GMS)



$$GMS(u) = \frac{100}{|z|} \sum_{\forall z \ni u} p_s(u|x,z) = \frac{100}{l} \sum_{\forall z \ni u} (1 - 10^{-\frac{Q_s(u|x,z)}{10}})$$

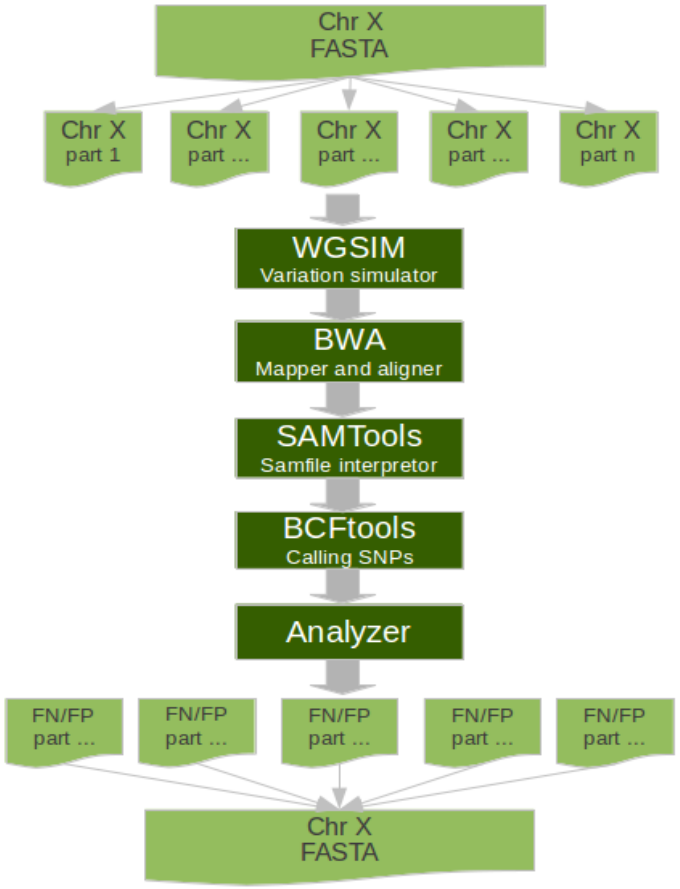
- u is a position
- x is a reference
- z is a read
- l is read length

GMS vs. MAQ
Sensitivity of Read Mapping Score



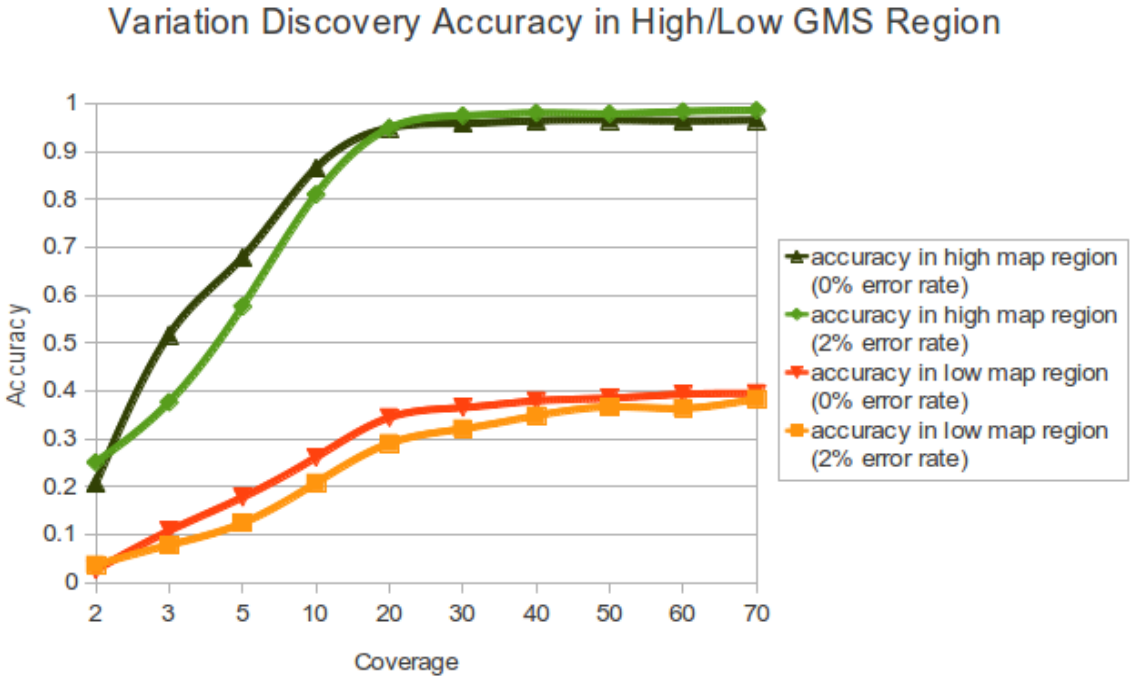
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA CTAGCATGCCCTAAGCCCGTAATGCAGTCATACTAGCTATCCTCGCCCTCTCCGTCAAGCTAC

Variation Accuracy Simulator



- Simulation of resequencing experiments to measure the accuracy of variation detection

Genomic Dark Matter



- Unlike false negatives in high GMS region that can be discovered in high coverage (>=20-fold), false negatives in low GMS regions cannot be discovered, because variation calling program will not use poorly mapped reads

JGI
JOINT GENOME INSTITUTE

BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 16 2012, pages 2097–2105
doi:10.1093/bioinformatics/bts330

Genome analysis

Advance Access publication June 4, 2012

Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score

Hayan Lee[1,2,*] and Michael C. Schatz[1,2]

[1]...rook, NY, USA and [2]Simons Center for ...bor, NY, USA

Kim *et al. Genome Biology* 2013, **14**:R90
http://genomebiology.com/2013/14/8/R90

Genome Biology

METHOD

Virmid: accurate detection of so... with sample impurity inference

Sangwoo Kim[1*†], Kyowon Jeong[2†], Kunal Bhutani[1], Jeong Ho Lee[3,6], Anar...
Hayan Lee[5], Joseph G Gleeson[3] and Vineet Bafna[1*]

Abstract

Detection of somatic variation using sequence from disease-control match... many cases including cancer, however, it is hard to isolate pure disease tis... mutation analysis by disrupting overall allele frequencies. Here, we propos... determines the level of impurity in the sample, and uses it for improved d... tests on simulated and real sequencing data from breast cancer and hemi... of our model. A software implementation of our method is available at ht...

Background

Identifying mutations relevant to a specific phenotype is one of the primary goals in sequence analysis. With the advent of massively parallel sequencing technologies, we can produce an immense amount of genomic information to estimate the landscape of sequence variations. However, the error rates for base-call and read alignment still remain much higher than the empirical frequencies of single nucleotide variations (SNVs) and *de novo* mutations [1]. Many statistical methods have been proposed to strengthen mutation discovery in the presence of confounding errors [2-4].

Finding somatic mutations is one particular type of

mutations fro... relations betw... probabilistic ... exome seque... potential *de n...* schizophrenia ...

However, th... covery might ... impurity and ... example, gastr... tain large nu... acquisition o... More importa...

LETTER

doi:10.1038/nature13907

Resolving the complexity of the human genome using single-molecule sequencing

Mark J. P. Chaisson[1], John Huddleston[1,2], Megan Y. Dennis[1], Peter H. Sudmant[1], Maika Malig[1], Fereydoun Hormozdiari[1], Francesca Antonacci[3], Urvashi Surti[4], Richard Sandstrom[1], Matthew Boitano[5], Jane M. Landolin[5], John A. Stamatoyannopoulos[1], Michael W. Hunkapiller[5], Jonas Korlach[5] & Evan E. Eichler[1,2]

The human genome is arguably the most complete mammalian reference assembly[1–3], yet more than 160 euchromatic gaps remain[4–6] and aspects of its structural variation remain poorly understood ten years after its completion[7–9]. To identify missing sequence and genetic variation, here we sequence and analyse a haploid human genome (CHM1) using single-molecule, real-time DNA sequencing[10]. We close or extend 55% of the remaining interstitial gaps in the human GRCh37 reference genome—78% of which carried long runs of degenerate short tandem repeats, often several kilobases in length, embedded within (G+C)-rich genomic regions. We resolve the complete sequence of 26,079 euchromatic structural variants at the base-pair level, including inversions, complex insertions and long tracts of tandem repeats. Most have not been previously reported, with the greatest increases in sensitivity occurring for events less than 5 kilobases in size. Compared to the human reference, we find a significant insertional bias

for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were enriched for simple repeats, long tandem repeats, and high (G+C) content (Fig. 1) but also included novel exons (Supplementary Table 20) and putative regulatory sequences based on DNase I hypersensitivity and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) analysis (Supplementary Information). We identified a significant 15-fold enrichment of short tandem repeats (STRs) when compared to a random sample ($P < 0.00001$) (Fig. 1a). A total of 78% (39 out of 50) of the closed gap sequences were composed of 10% or more of STRs. The STRs were frequently embedded in longer, more complex, tandem arrays of degenerate repeats reaching up to 8,000 bp in length (Extended Data Fig. 1a–c), some of which

Assemblablity

How well is the genome assembled given condition?

Background

- **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.

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?**

Lander-Waterman Statistics

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

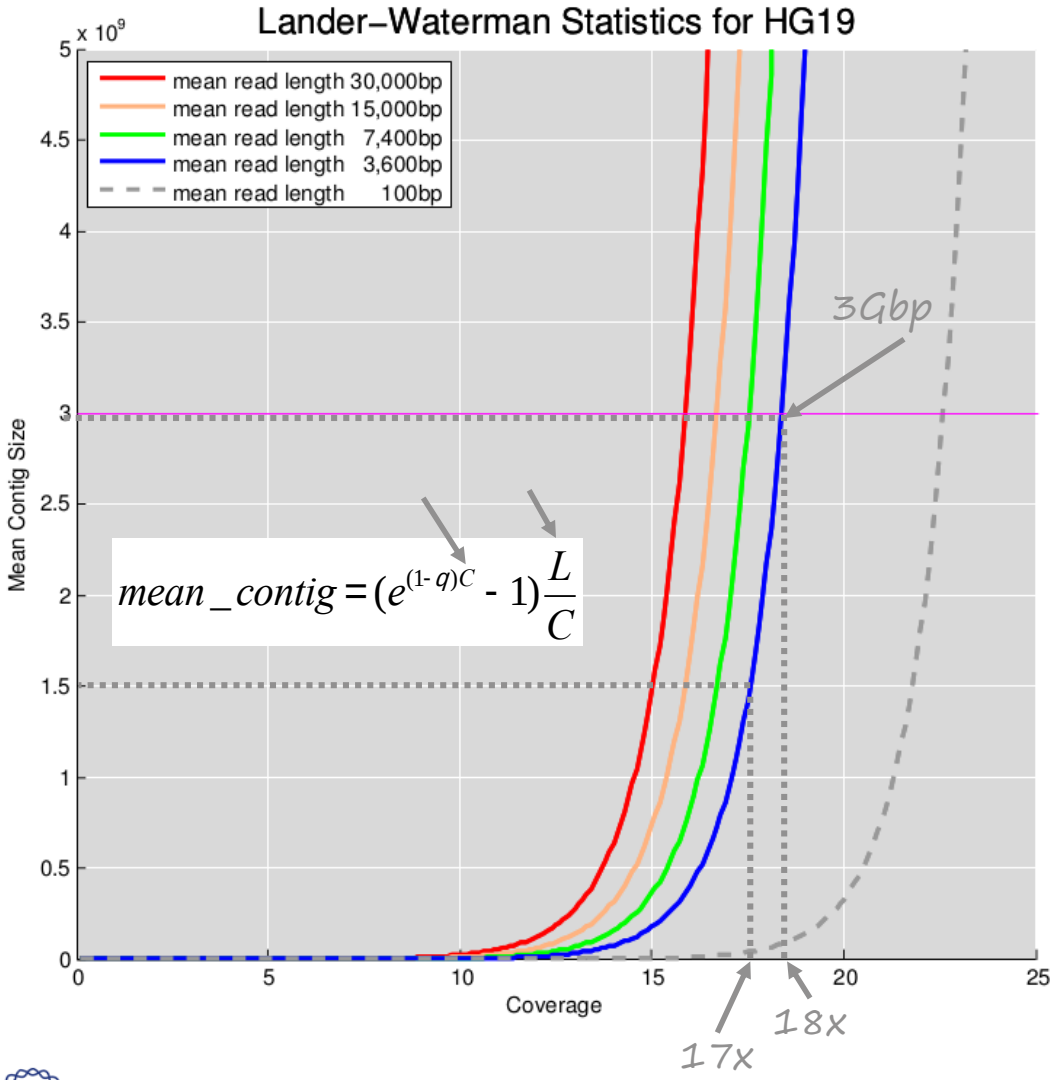
Received January 13, 1988; revised March 31, 1988

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

Linear vs. Exponential

Technology vs. Money

Cold Spring Harbor Laboratory

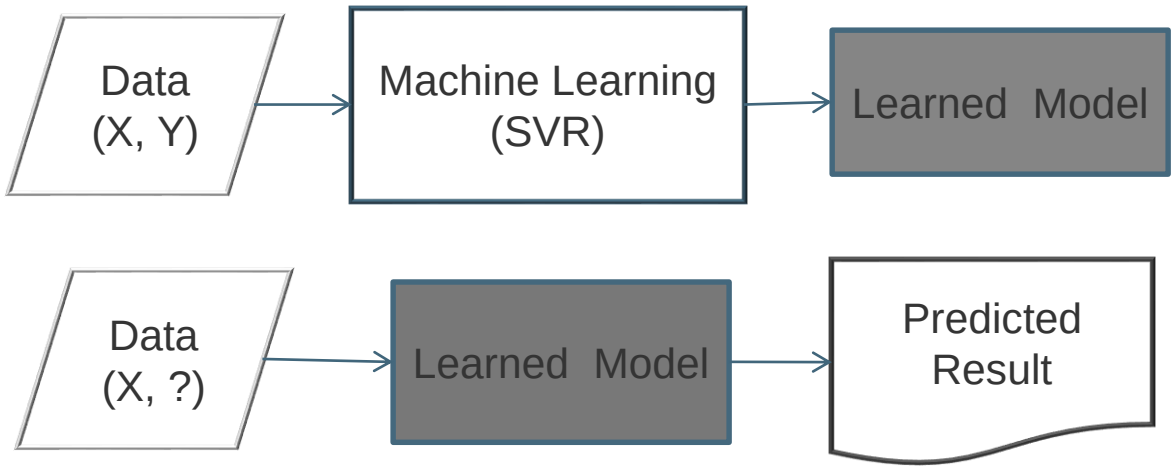
Simons Institute for the Theory of Computing

Berkeley University of California

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.anthracis	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**

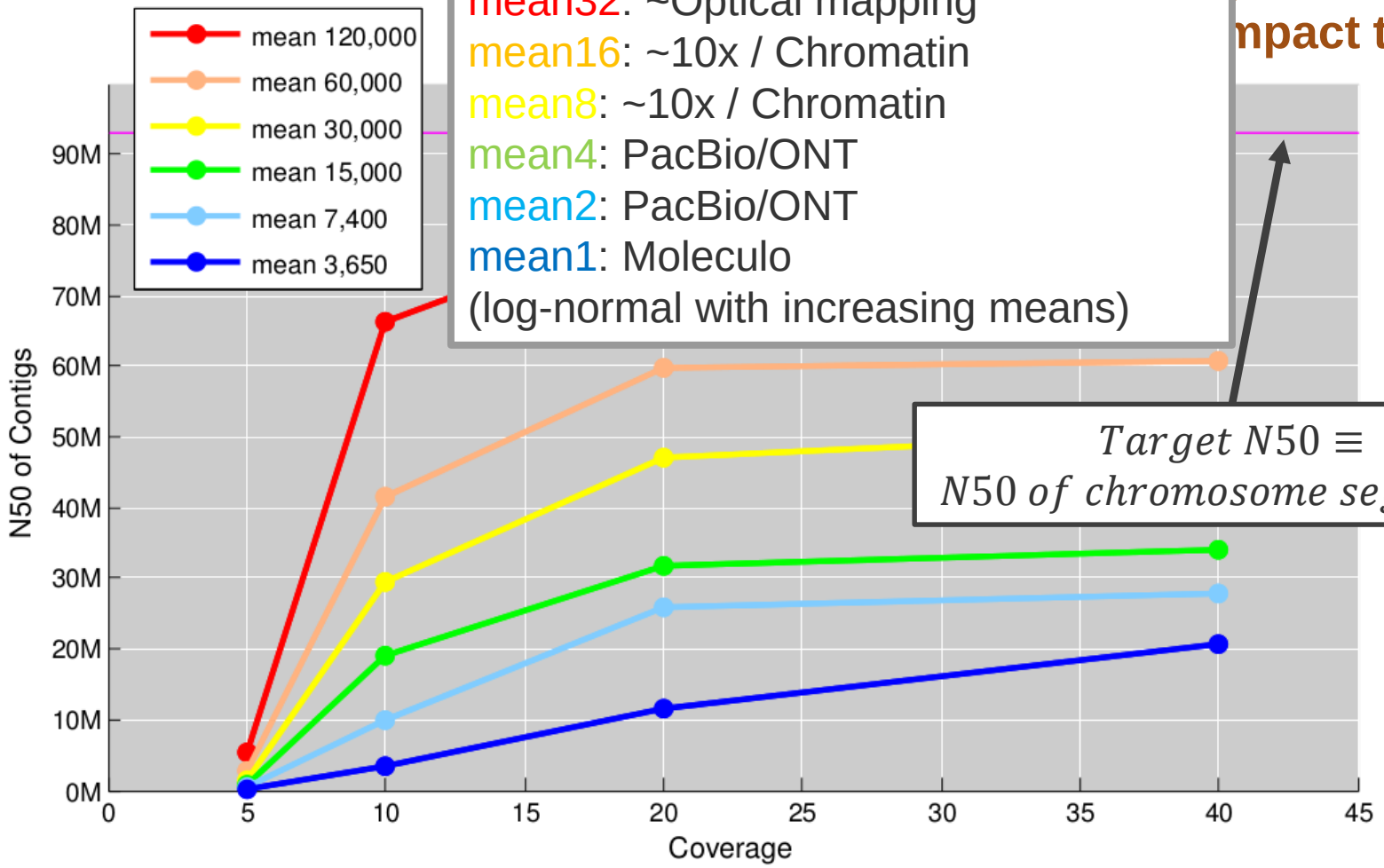


HG19 Genome Assembly Performance by Our Simulation

Lengths selected to represent idealized biotechnologies:

mean32: ~Optical mapping
mean16: ~10x / Chromatin
mean8: ~10x / Chromatin
mean4: PacBio/ONT
mean2: PacBio/ONT
mean1: Moleculo
(log-normal with increasing means)

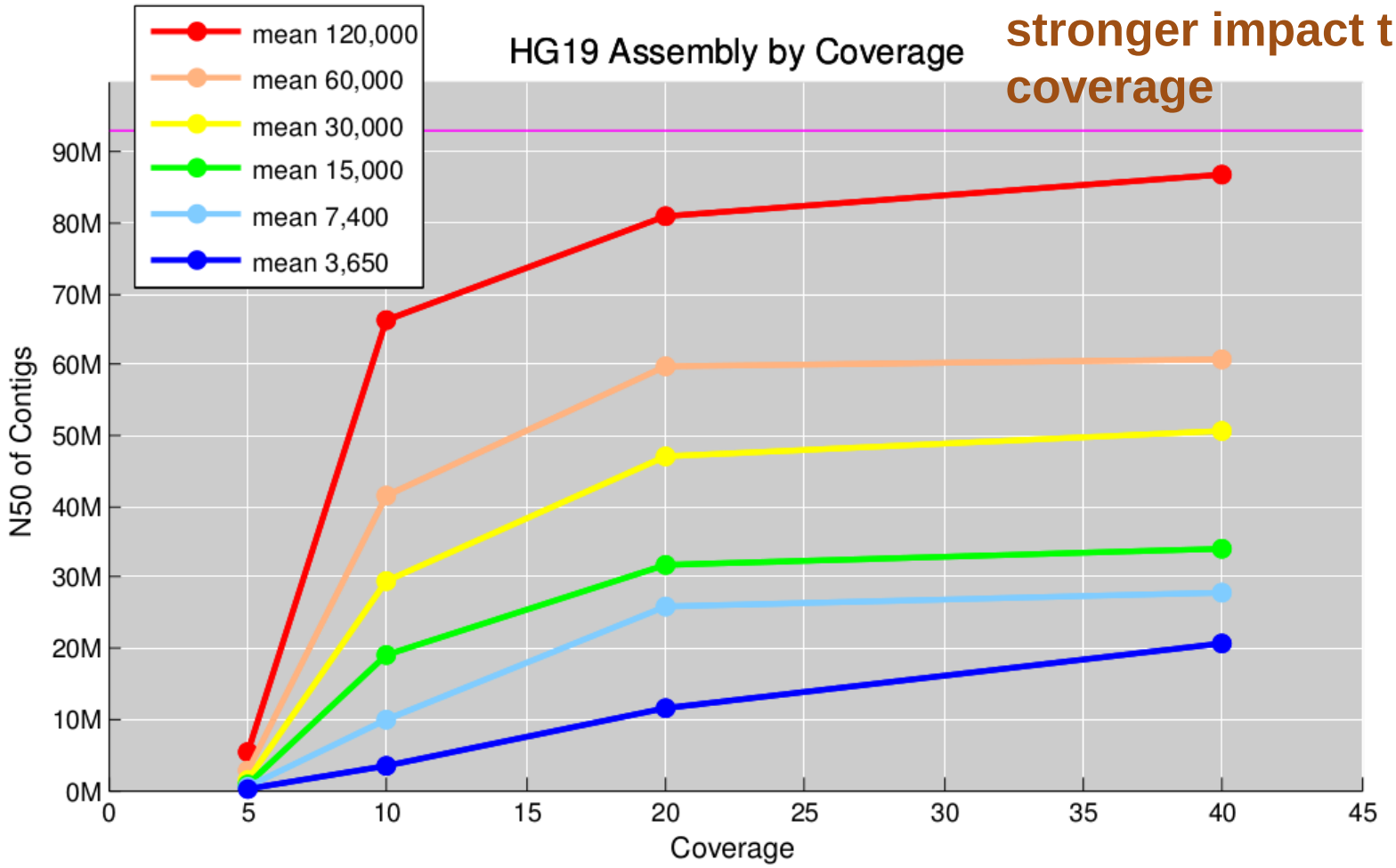
gth has mpact than



$$Target\ N50 \equiv N50\ of\ chromosome\ segments$$

HG19 Genome Assembly Performance by Our Simulation

Read Length has stronger impact than coverage



Why?

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 a model on top of assumptions, we let the model learn from the data
- Empirical data-driven approach

Our Goal

- **To predict genome assembly contiguity**

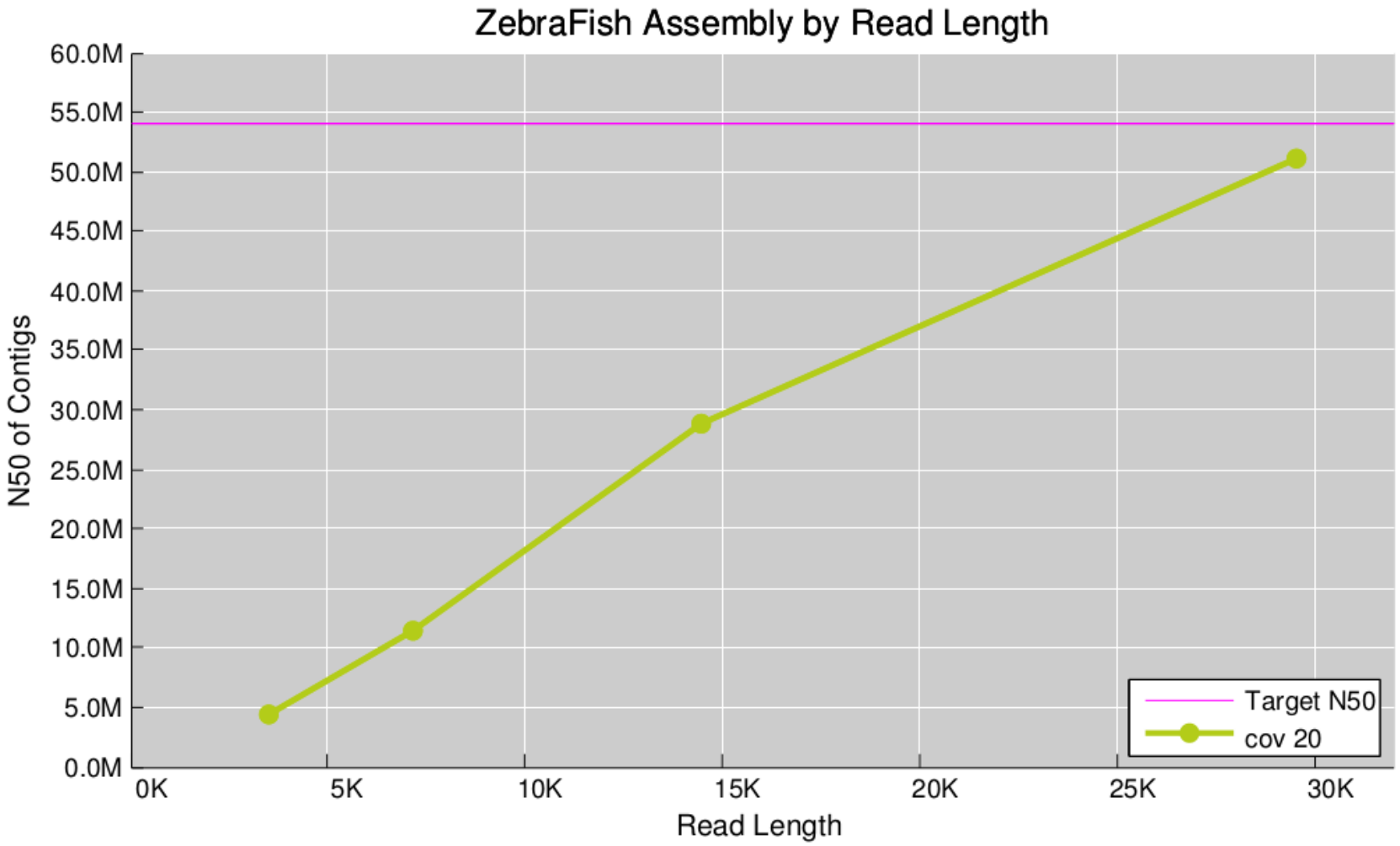
$$Performance(\%) \propto \frac{N50\ from Assembly}{N50\ from Chromosome Segments} \times 100$$

$$\propto f\begin{pmatrix} Read\ Length \\ Coverage \\ Genome\ Size \\ Repeat \end{pmatrix}$$

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%

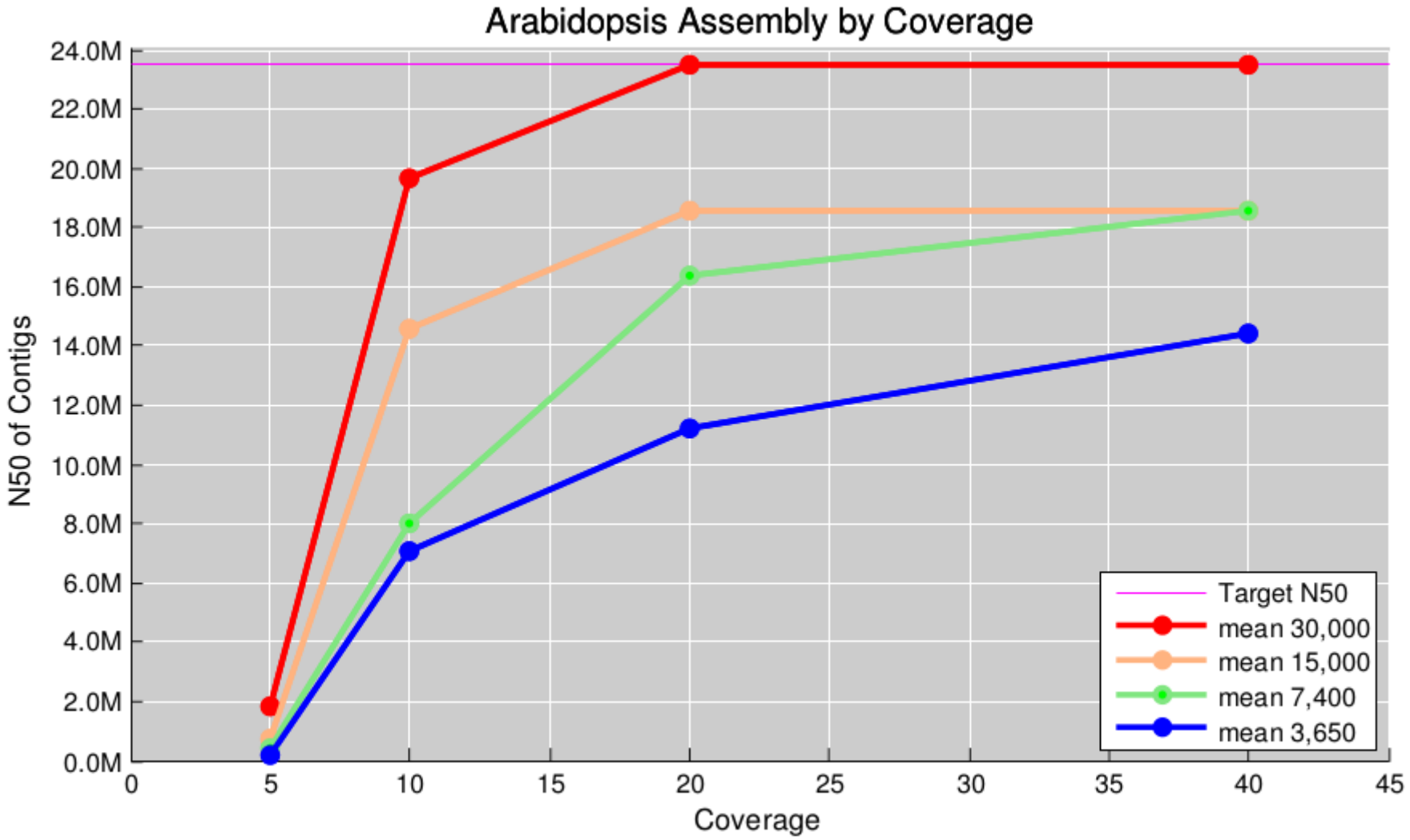
Read Length


ZebraFish Assembly by Read Length
N50 of Contigs
60.0M
55.0M
50.0M
45.0M
40.0M
35.0M
30.0M
25.0M
20.0M
15.0M
10.0M
5.0M
0.0M
0K
5K
10K
15K
20K
25K
30K
Read Length
Target N50
cov 20

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

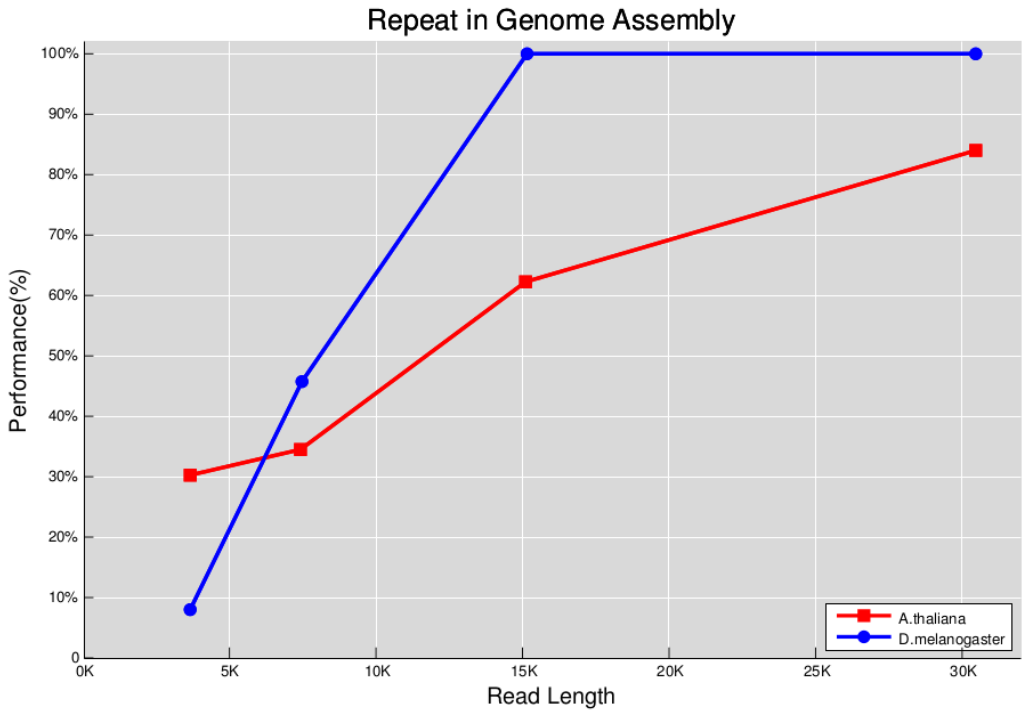

Arabidopsis Assembly by Coverage
N50 of Contigs
24.0M
22.0M
20.0M
18.0M
16.0M
14.0M
12.0M
10.0M
8.0M
6.0M
4.0M
2.0M
0.0M
0
5
10
15
20
25
30
35
40
45
Coverage
Target N50
mean 30,000
mean 15,000
mean 7,400
mean 3,650

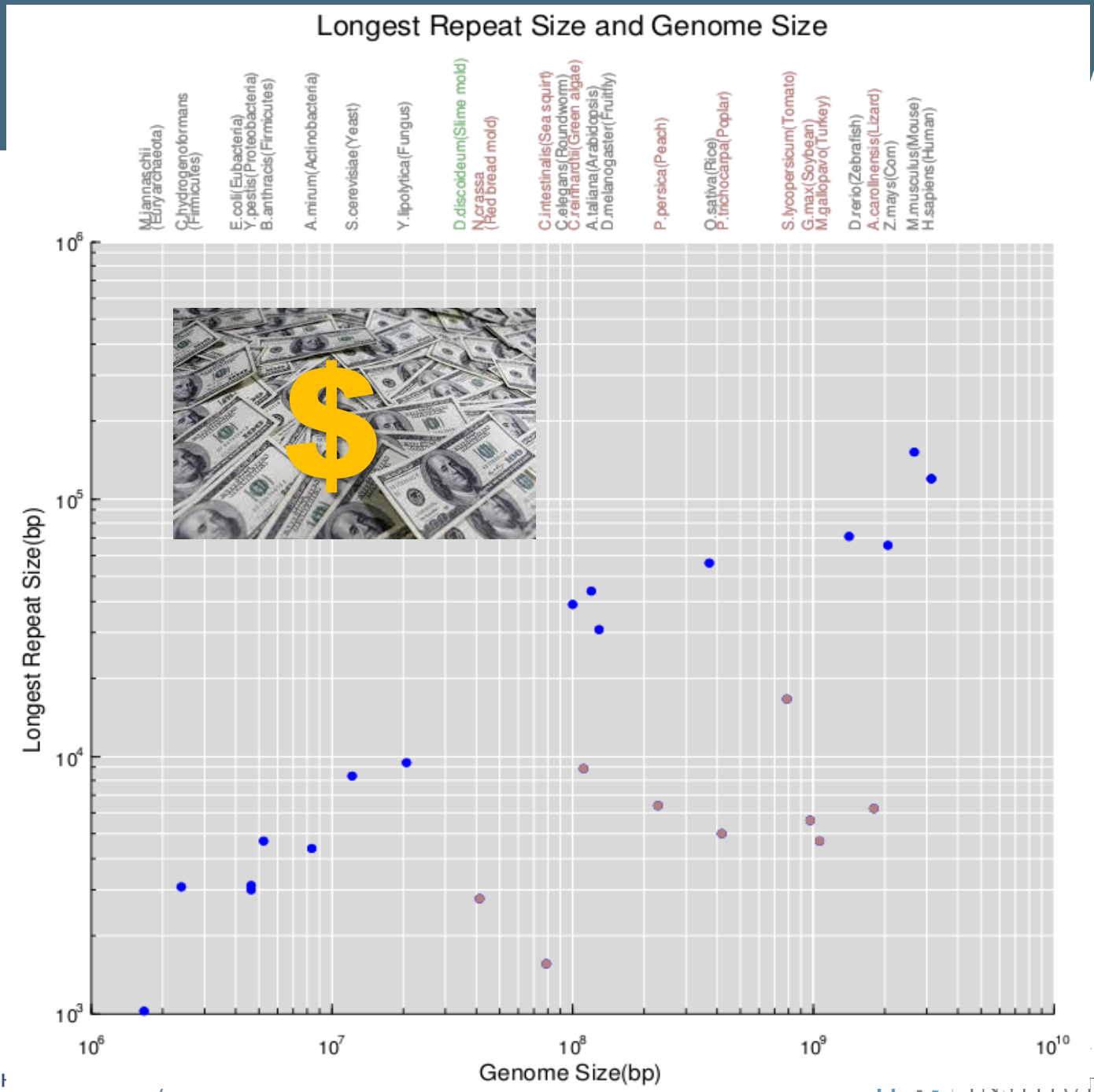
Repeats

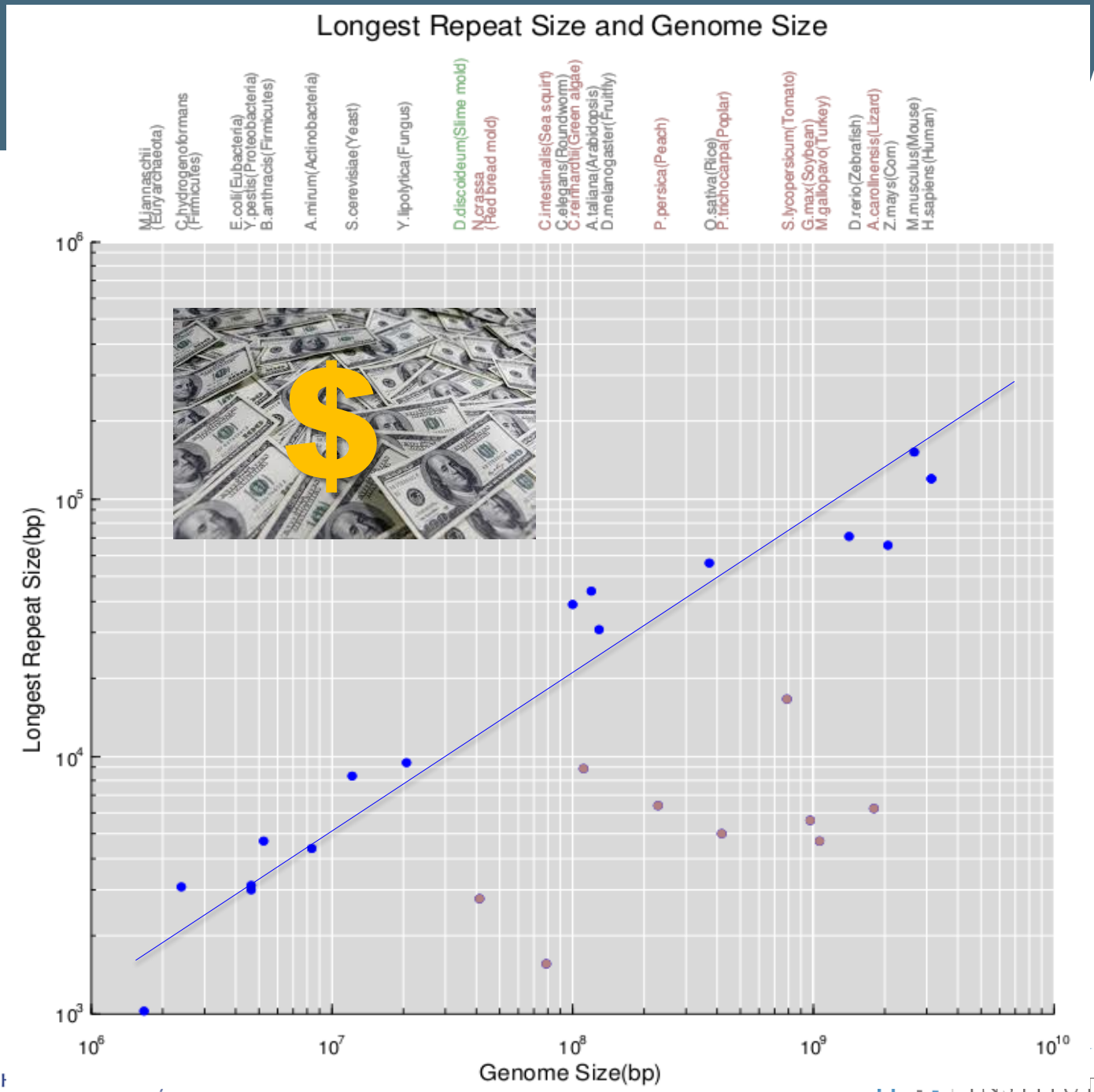
- **Genome is not a random sequence**
- **Repeat hurts genome assembly performance**
- **Isolating the impact of repeats is not trivial**
- **Quantifying repeat characteristics is not trivial as well**
  - The longest repeat size
  - # of repeats > read length

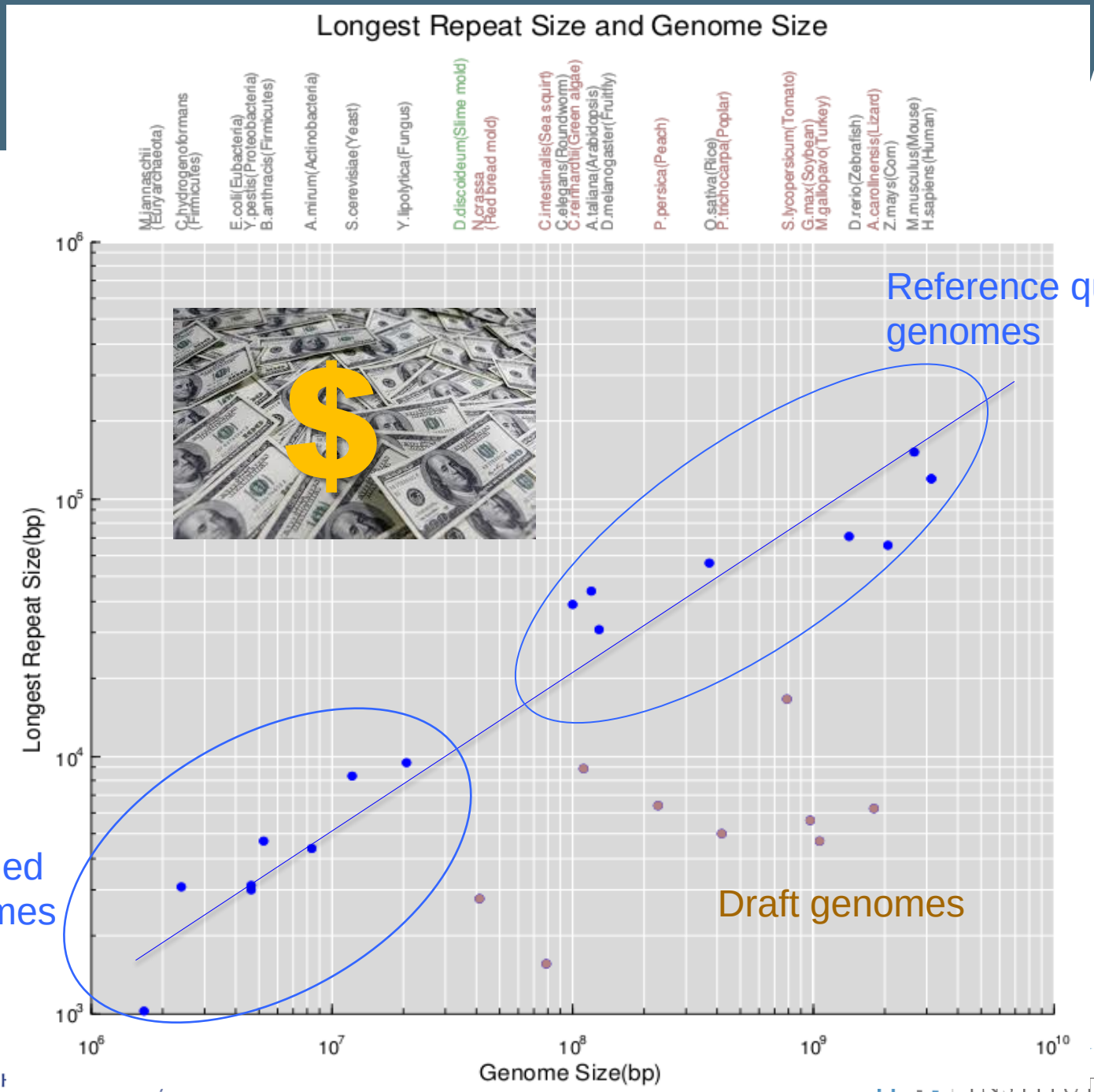
Repeats

	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





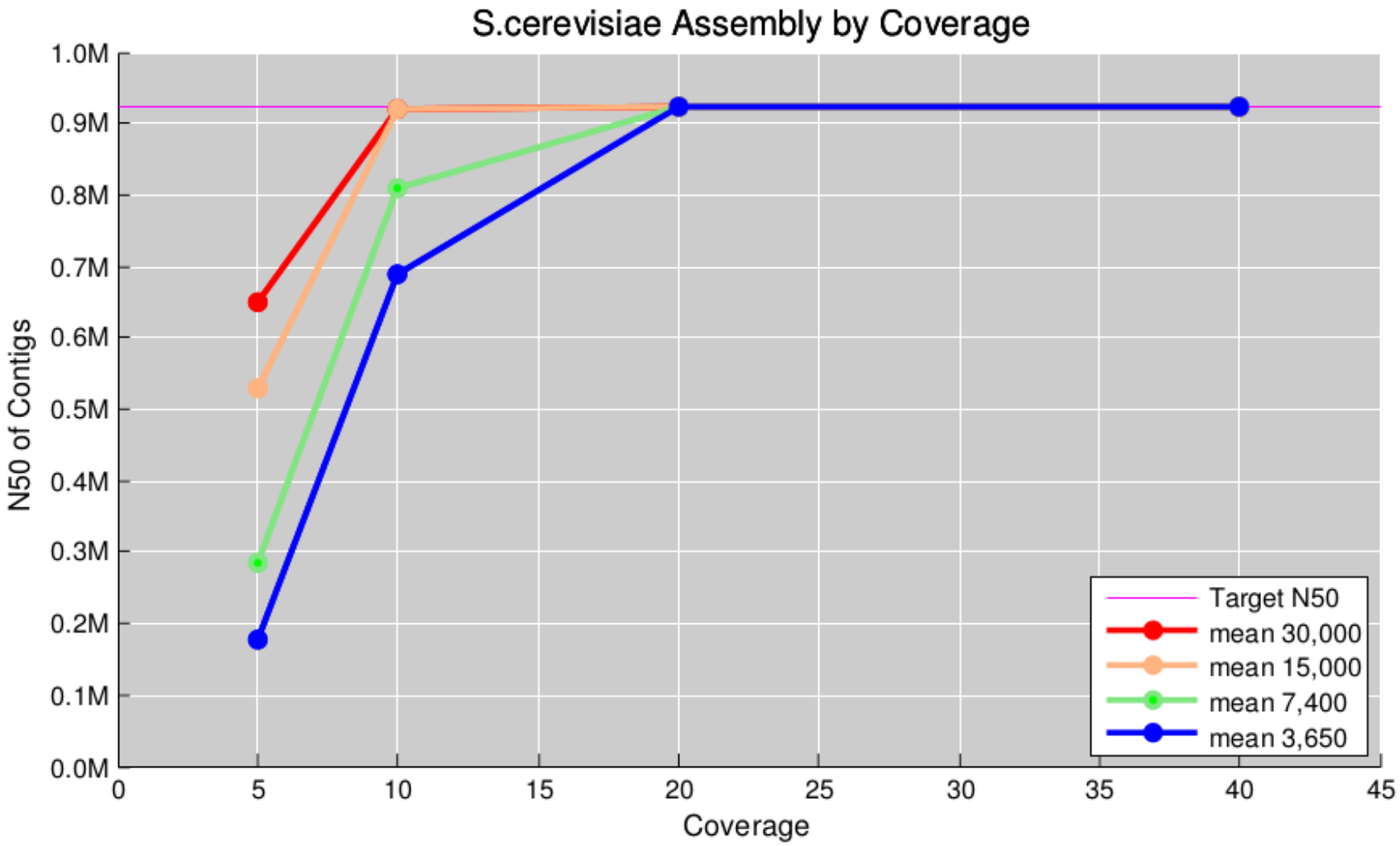



Longest Repeat Size and Genome Size
M.jannaschii (Euryarchaeota)
C.hydrogenoformans (Firmicutes)
E.coli(Eubacteria)
Y.pestis(Proteobacteria)
B.anthracis(Firmicutes)
A.mirum(Actinobacteria)
S.cerevisiae(Yeast)
Y.lipolytica(Fungus)
D.discoideum(Slime mold)
N.crassa (Red bread mold)
C.intestinalis(Sea squirt)
C.elegans(Roundworm)
C.reinhardtii(Green algae)
A.taliana(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)
Reference quality genomes
Finished genomes
Draft genomes
Longest Repeat Size(bp)
Genome Size(bp)
10^3
10^4
10^5
10^6
10^7
10^8
10^9
10^{10}

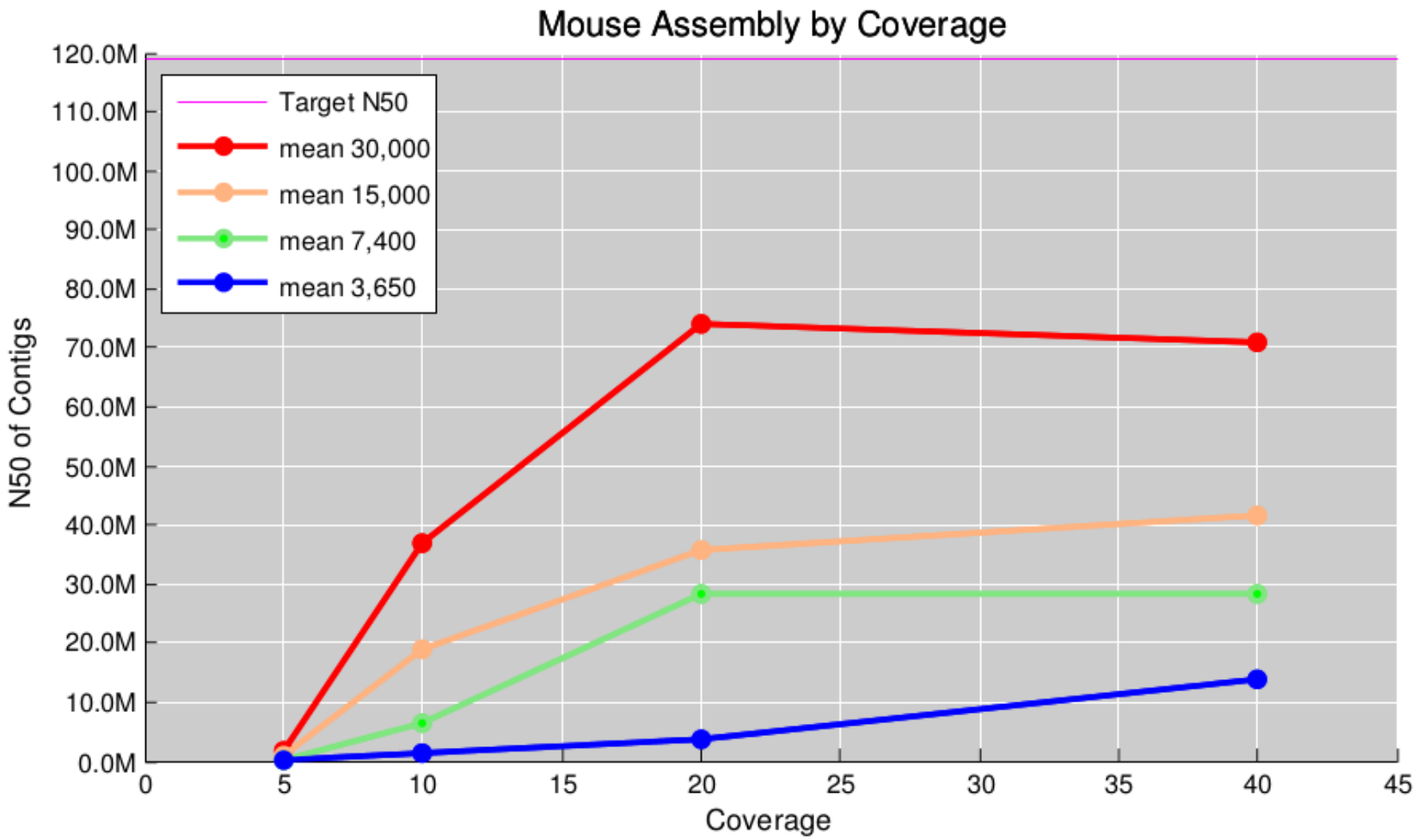
Genome Size

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

Genome Size

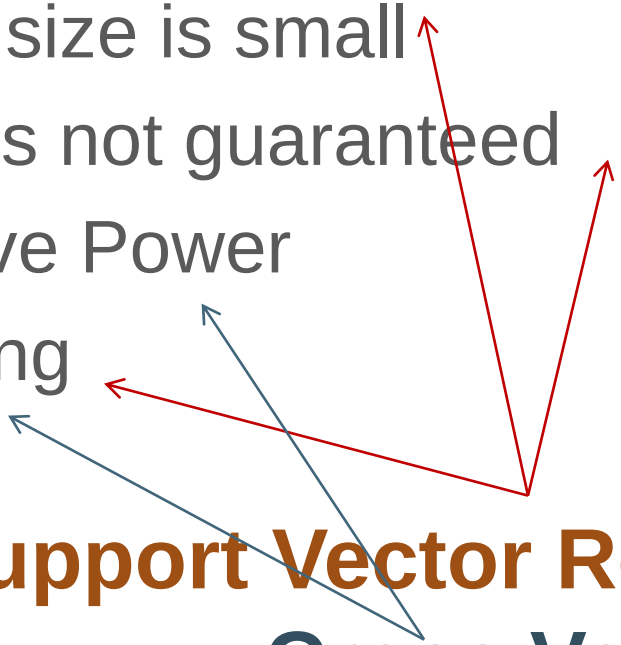

S.cerevisiae Assembly by Coverage
N50 of Contigs
1.0M
0.9M
0.8M
0.7M
0.6M
0.5M
0.4M
0.3M
0.2M
0.1M
0.0M
0
5
10
15
20
25
30
35
40
45
Coverage
Target N50
mean 30,000
mean 15,000
mean 7,400
mean 3,650

Genome Size


Mouse Assembly by Coverage
N50 of Contigs
Coverage
Target N50
mean 30,000
mean 15,000
mean 7,400
mean 3,650
120.0M
110.0M
100.0M
90.0M
80.0M
70.0M
60.0M
50.0M
40.0M
30.0M
20.0M
10.0M
0.0M
0
5
10
15
20
25
30
35
40
45

Challenges for Prediction

- Sample size is small
- Quality is not guaranteed
- Predictive Power
- Overfitting



Support Vector Regression (SVR)

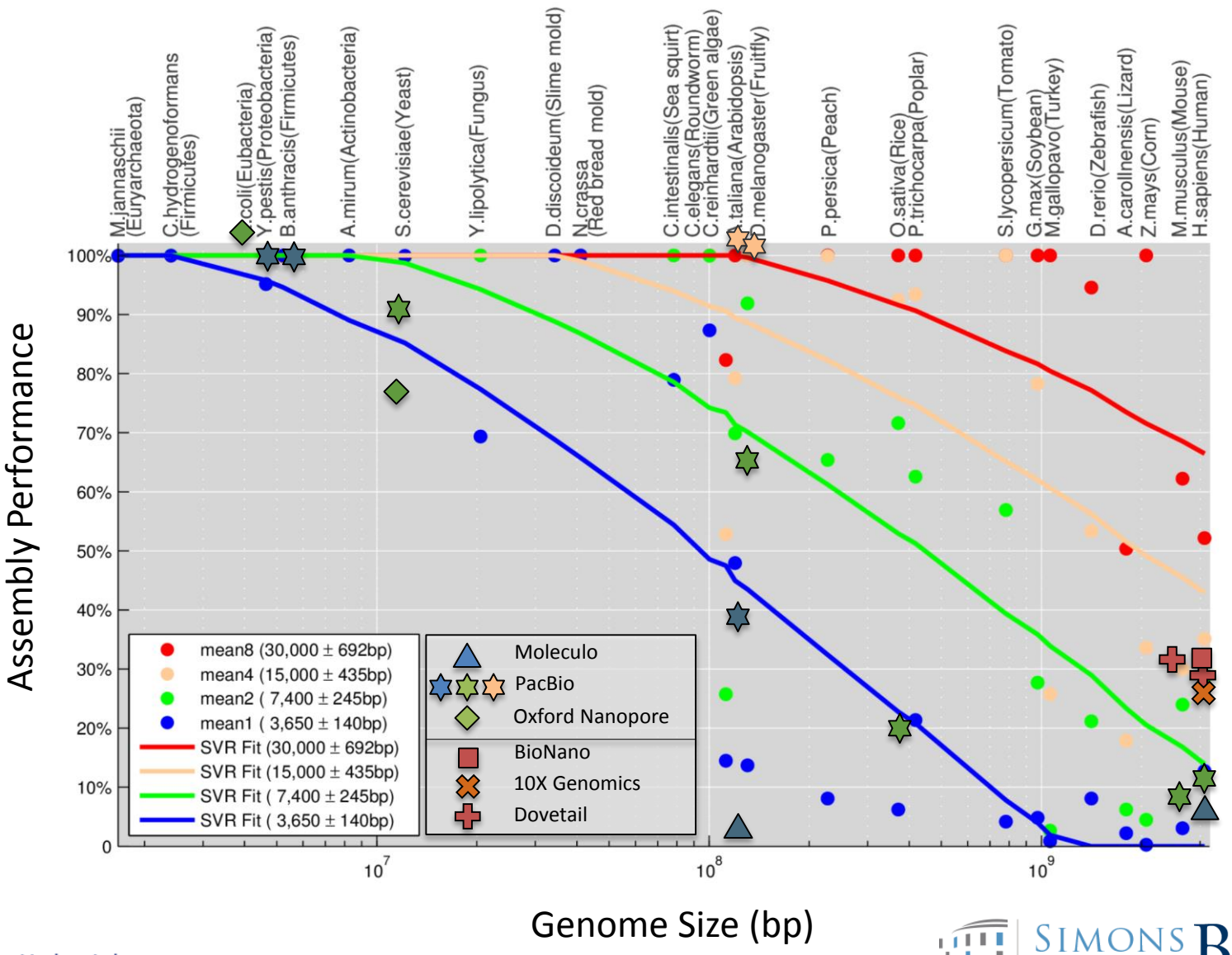
Cross Validation

Feature Engineering

- **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

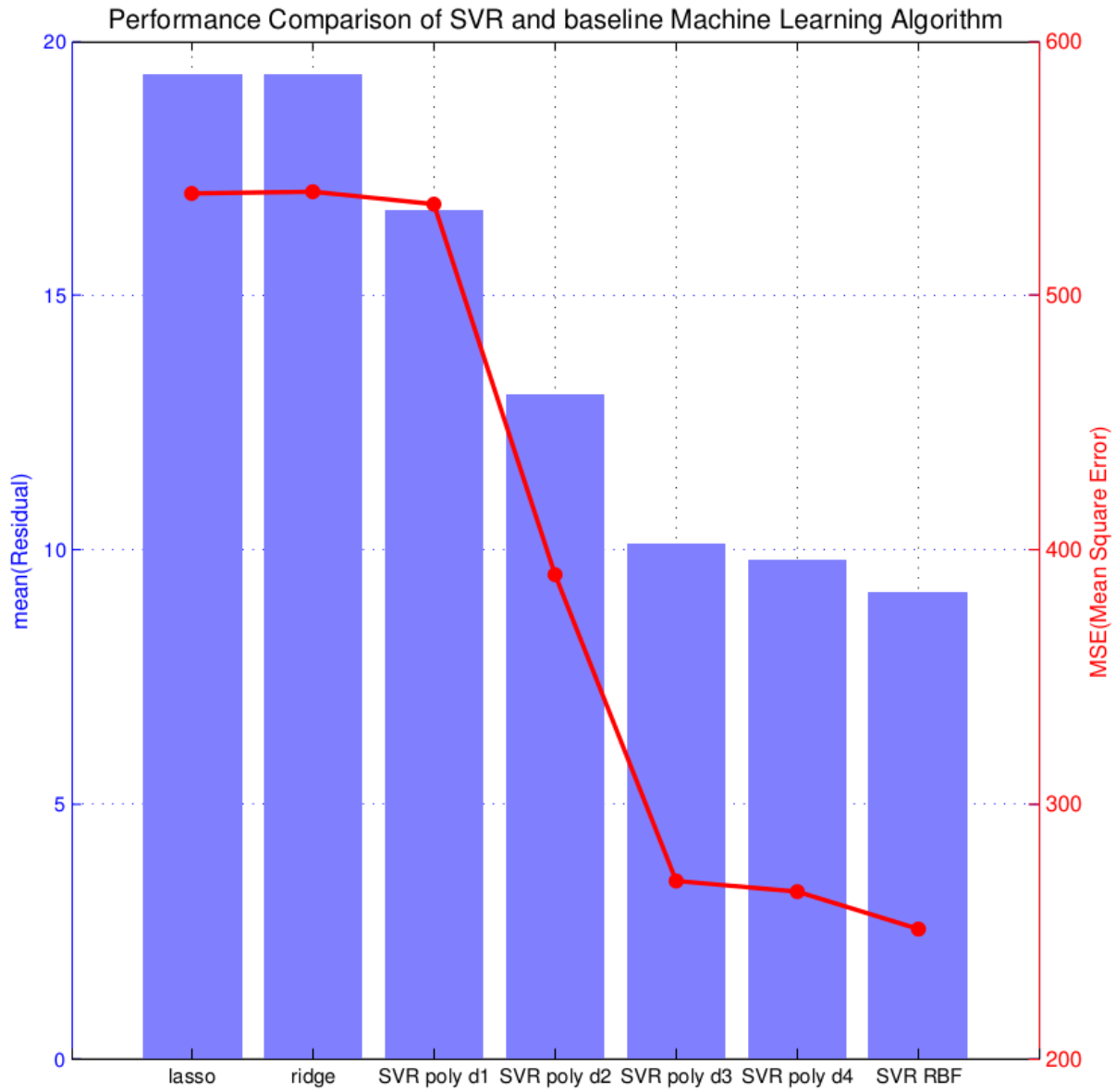
Reference Genome Quality



Cross Validation

- **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



Lee-Schatz Model

- Average of residual is 10%-15%
- We can predict the new genome assembly performance in 10%-15% of error residual boundary
- Read length, coverage and genome size used explicitly
- Repeats are included implicitly

	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

$$(e^{(1-q)C} - 1)\frac{L}{C}\mu\, L \times e^{C}$$

$$L \times e^{C}$$

$$\Downarrow$$

$$L \times C$$

$$\Downarrow$$

$$L \times \log C$$

$$\Downarrow$$

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

$$\Downarrow$$

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

$$\Downarrow$$

$$SVR(L, C, G, R)$$

Web Service for Contiguity Prediction

Genome Assembly Performance Prediction - Mozilla Firefox

Genome Assembly Perf...

qb.cshl.edu/asm_model/predict.html

research dic

Genome Assembly Performance Prediction

This is the Genome Assembly Performance Prediction Service. If you have any queries please email Hayan Lee(hlee@cshl.edu).

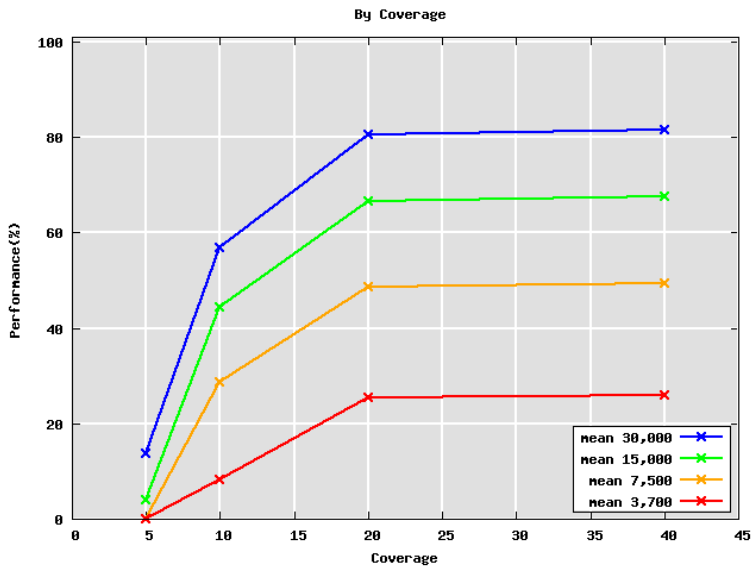
Http://qb.cshl.edu/asm_model/predict.html

Performance(%) = N50 of assembly / N50 of chromosome segments

Genome size : 1000123456

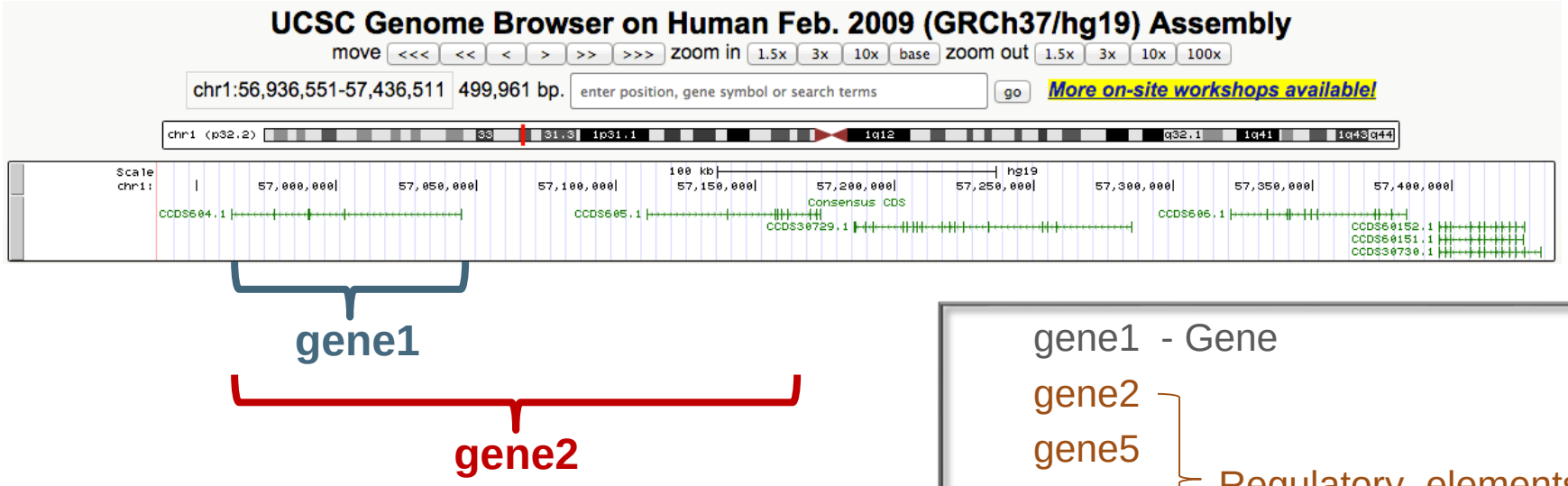
Submit

Assembly Prediction of Genome Size 1000123456



Completeness

Human Reference Genome Quality by gene block analysis



gene1

gene2

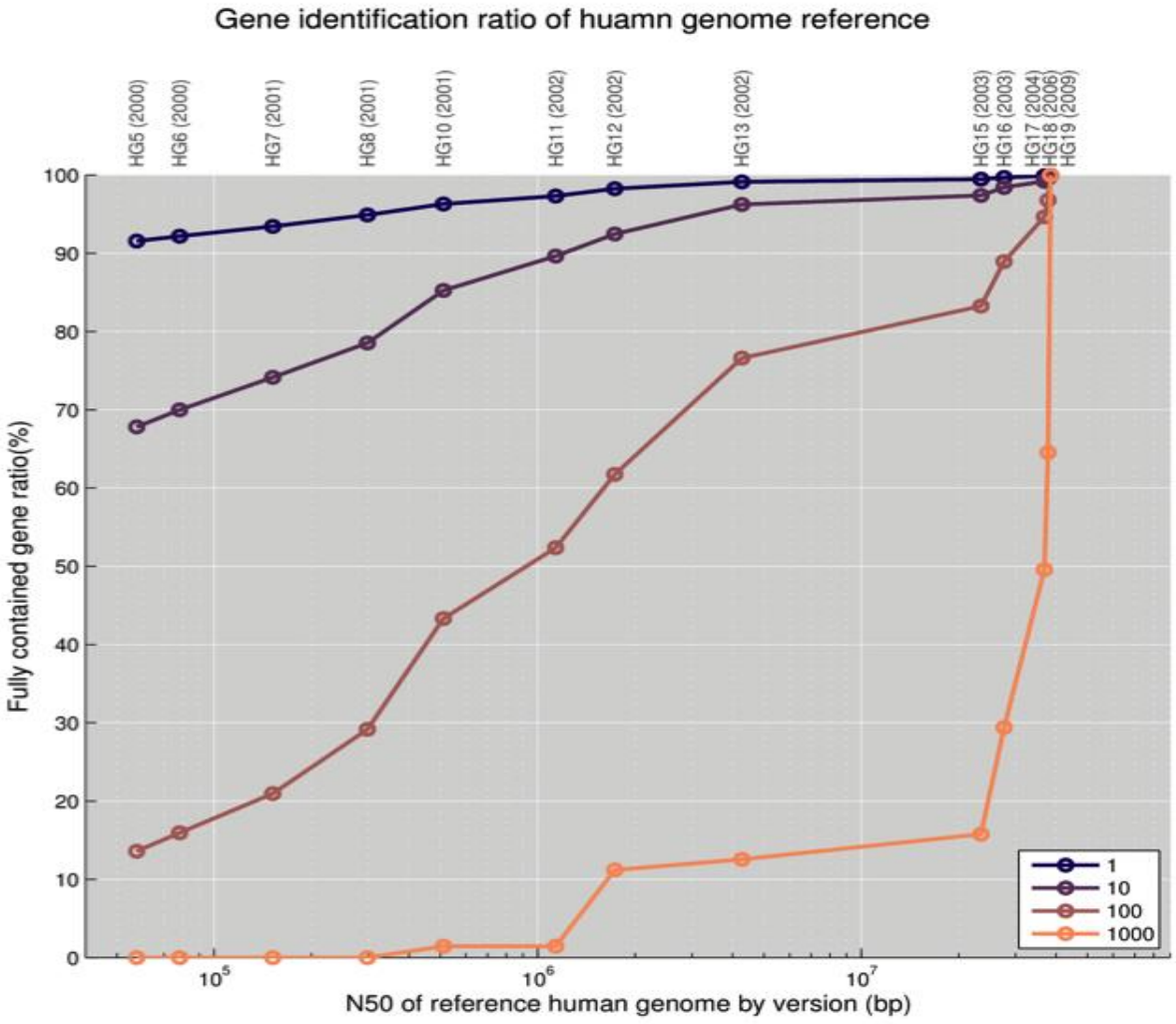
gene1 - Gene

gene2
gene5
gene10
gene20
Regulatory elements

gene50
gene100
gene200
gene500
Synteny blocks

gene1000 - **Chromosome structure**

Completeness
Human Reference Genome Quality by gene block analysis



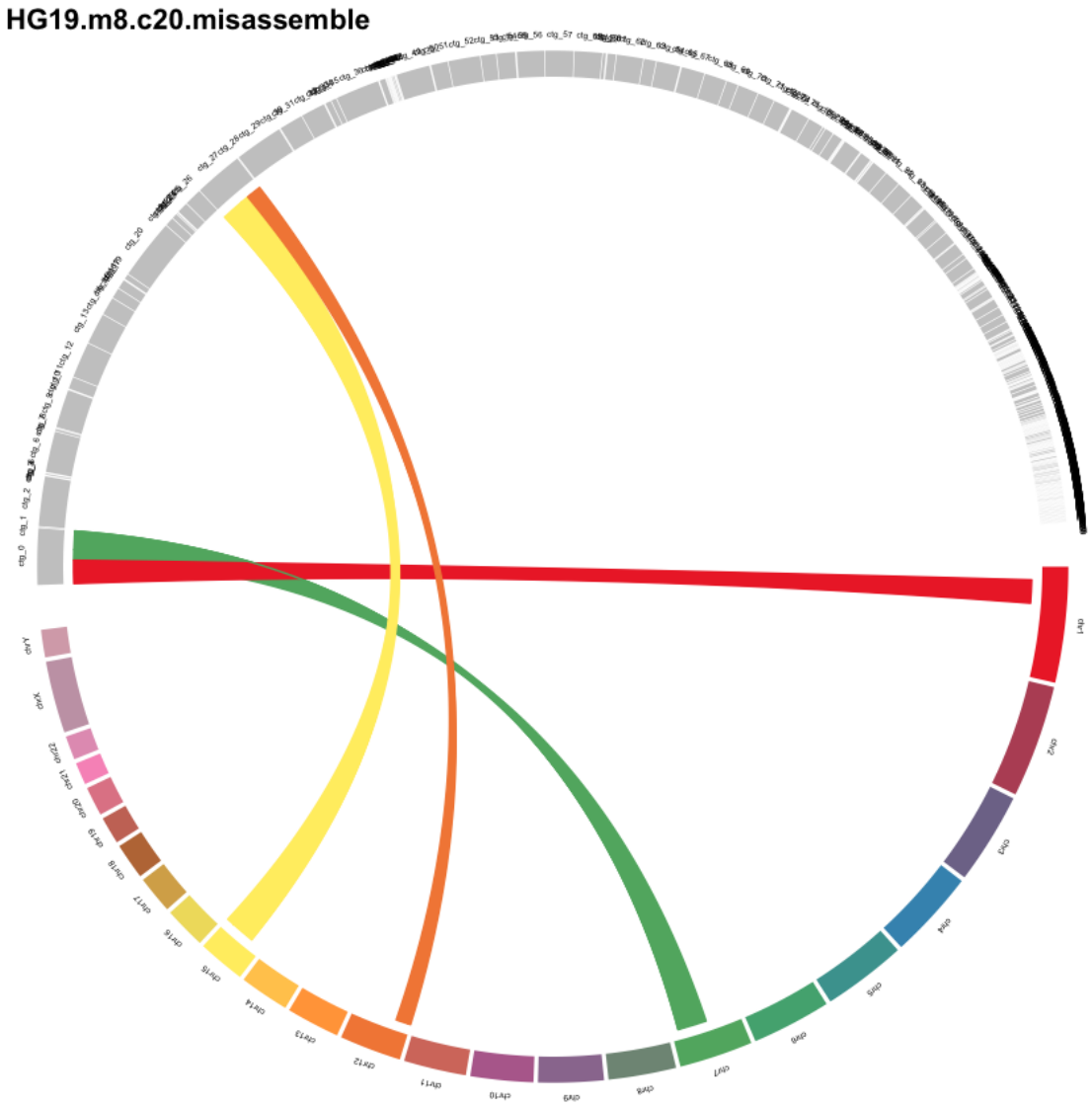
Larger contigs and scaffolds empowers analysis at every possible level.

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

Gene
Regulatory elements
Synteny blocks
Chromosome structure

Correctness

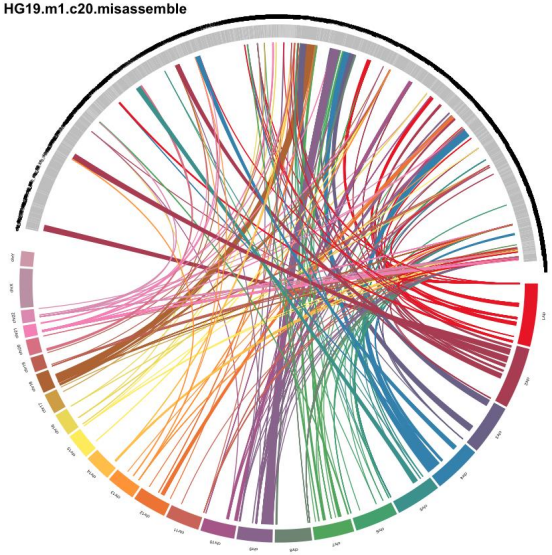
Misassembly - A critical error in de novo assembly



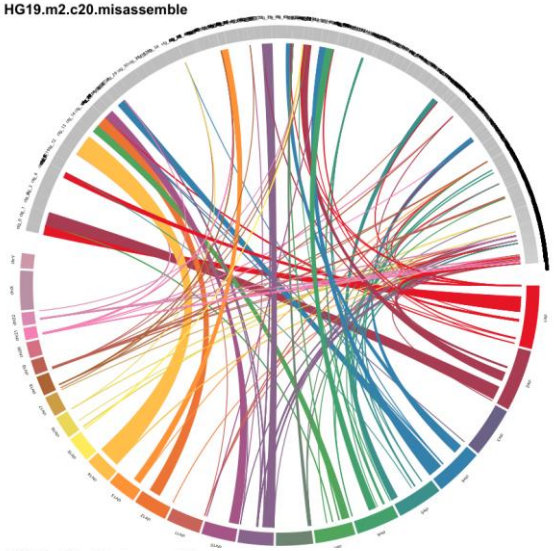
Correctness

Misassembly Analysis in HG19

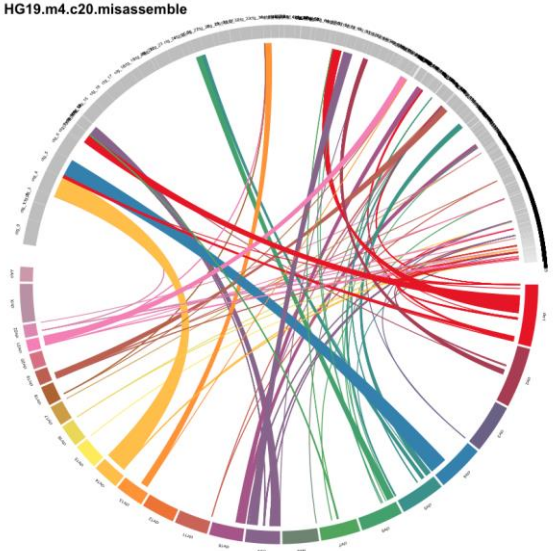
HG19.m1.c20.misassemble



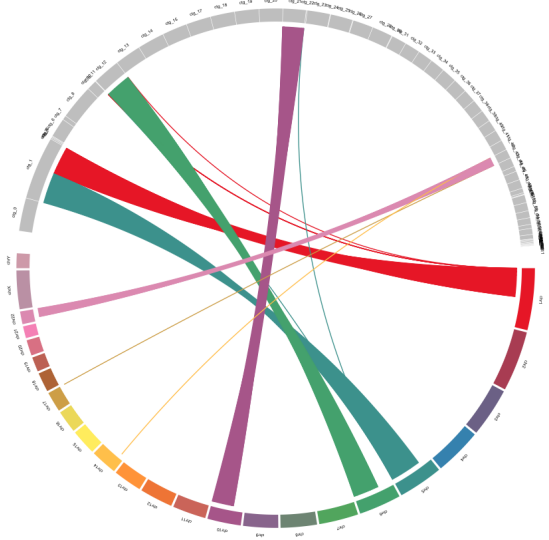
HG19.m2.c20.misassemble



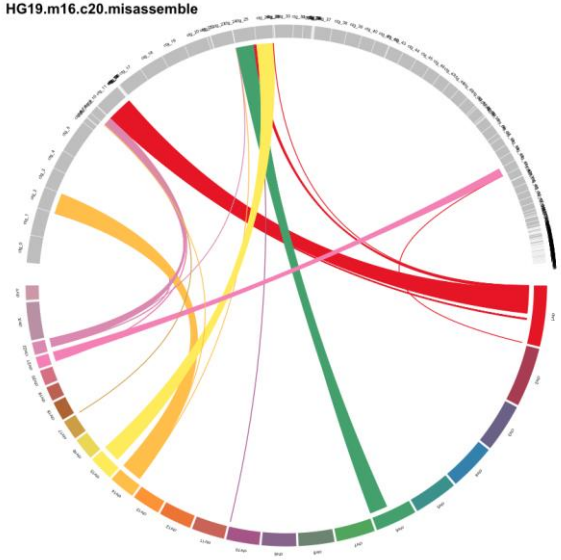
HG19.m4.c20.misassemble



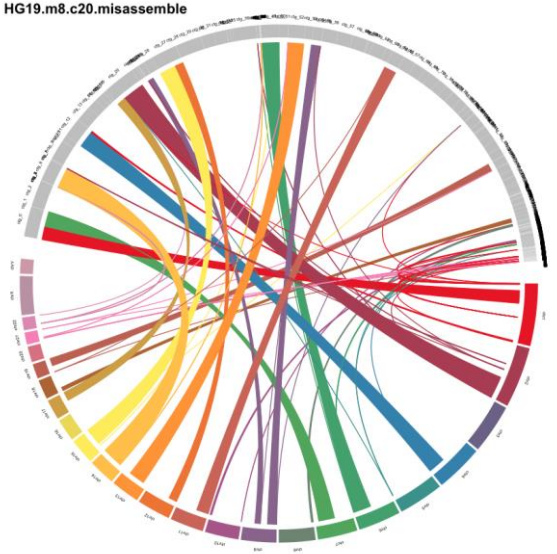
HG19.m32.c20.misassemble



HG19.m16.c20.misassemble

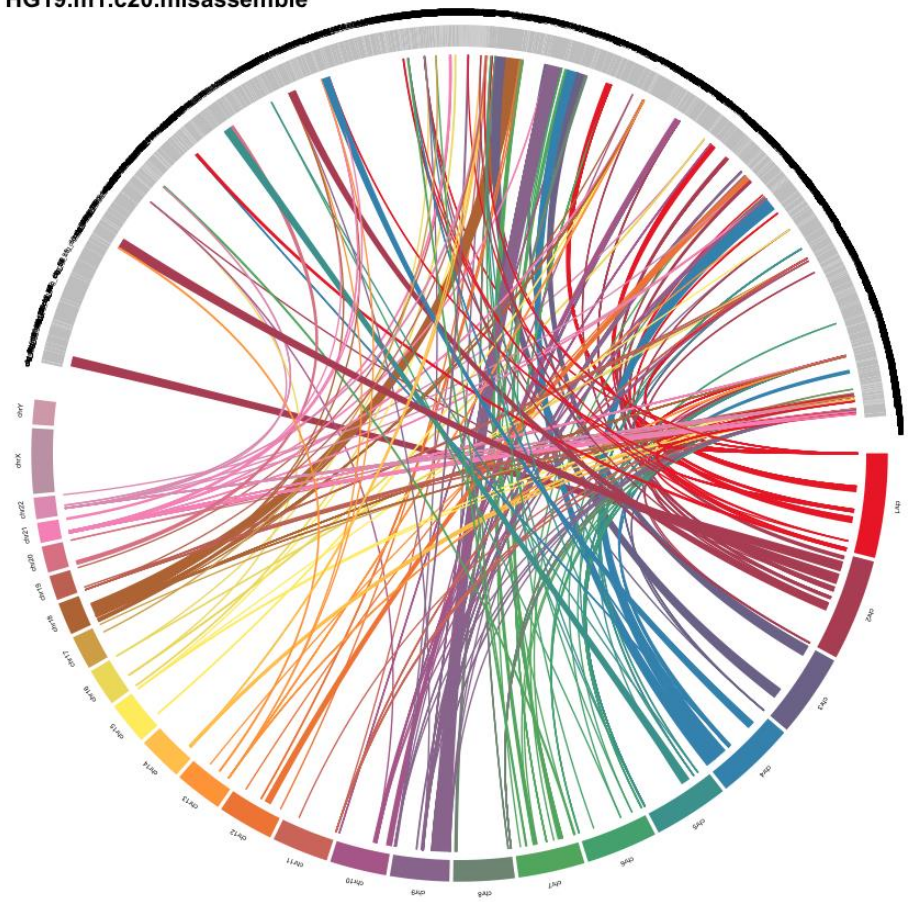


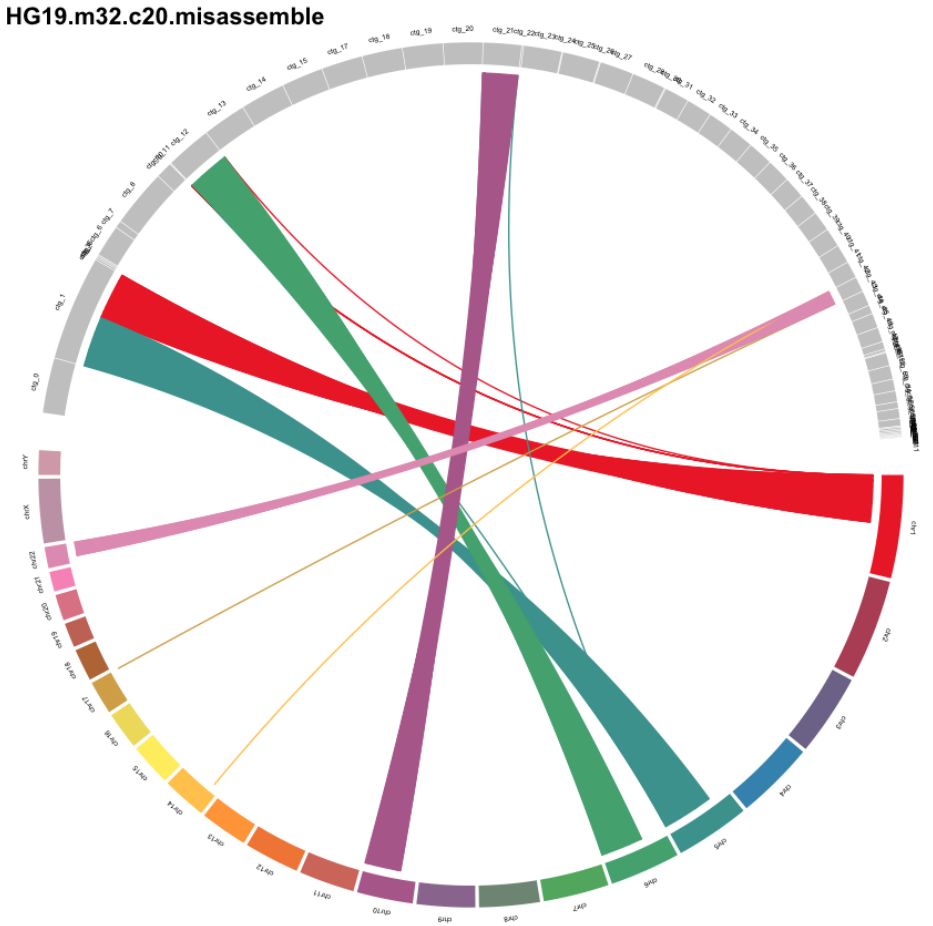
HG19.m8.c20.misassemble



Correctness

Misassembly Analysis in HG19



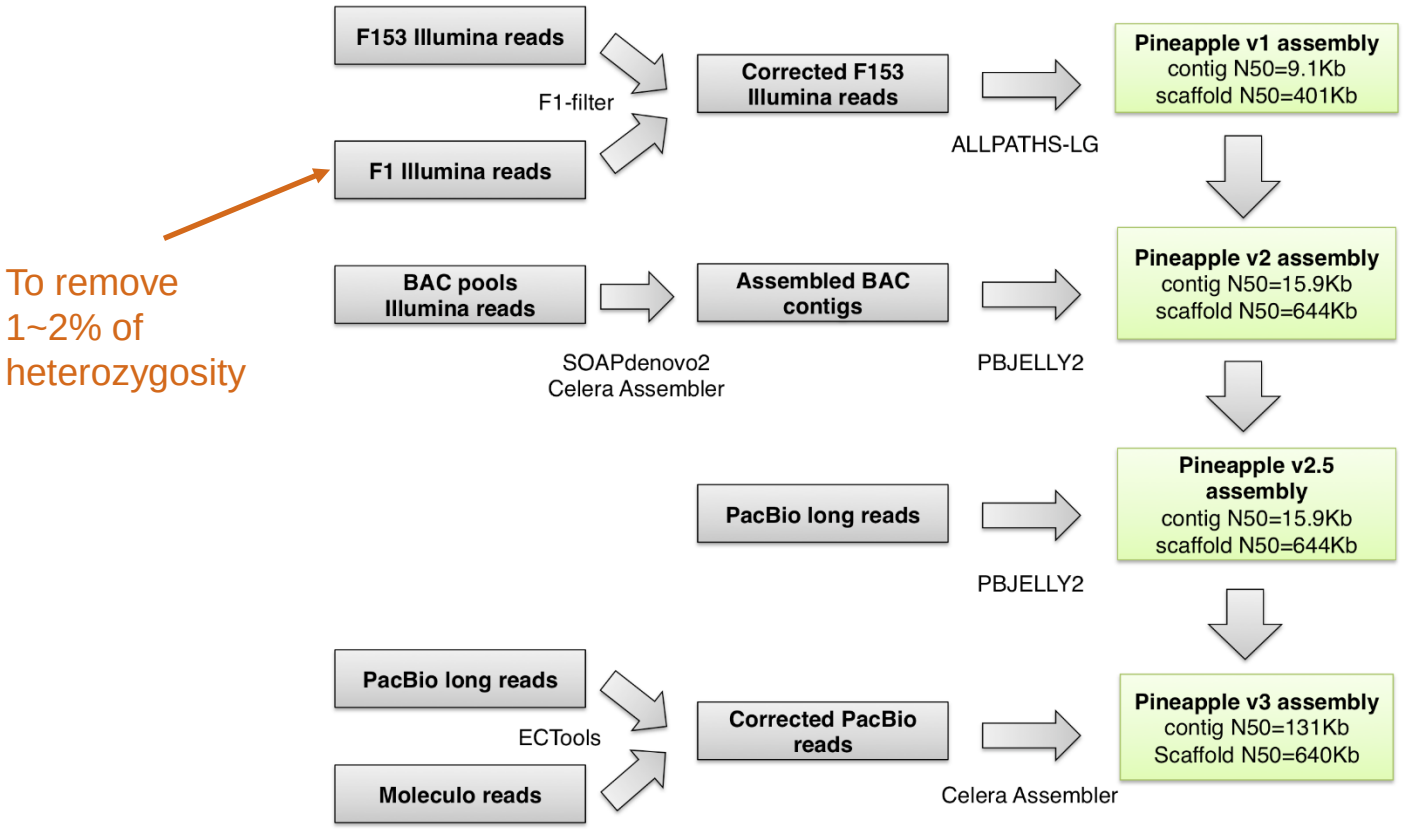


Long read sequencing technology helps to reduce both misassembly and breaks thus increase correctness of de novo genome assembly

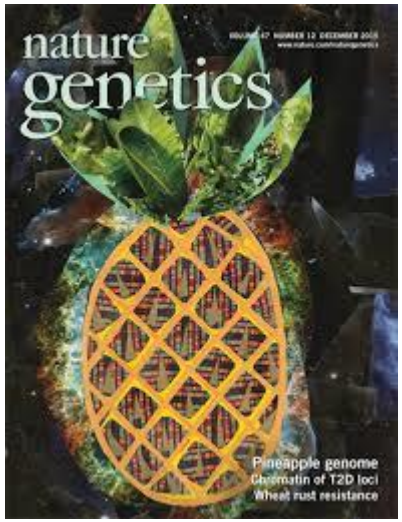
De novo assembly for highly polymorphic genomes

- Pineapple (highly heterozygous)
- Sugarcane (highly heterozygous and poly/aneuploidy)

Pineapple De Novo Assembly



Schematic workflow of the pineapple genome assembly and improvement



nature genetics

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journal ▾

home ▸ advance online publication ▸ full text

NATURE GENETICS | ARTICLE OPEN

The pineapple genome and the evolution of CAM photosynthesis

Ray Ming, Robert VanBuren, Ching Man Wai, Haibao Tang, Michael C Schatz, John E Bowers, Eric Lyons, Ming-Li Wang, Jung Chen, Eric Biggers, Jisen Zhang, Lixian Huang, Lingmao Zhang, Wenjing Miao, Jian Zhang, Zhangyao Ye, Chenyong Miao, Zhicong Lin, Hao Wang, Hongye Zhou, Won C Yim, Henry D Priest, Chunfang Zheng, Margaret Woodhouse, Patrick P Edger, Romain Guyot, Hao-Bo Guo, Hong Guo, Guangyong Zheng, Ratnesh Singh, Anupma Sharma, Xiangjia Min, Yun Zheng, Hayan Lee, James Gurtowski, Fritz J Sedlazeck, Alex Harkess, Michael R McKain, Zhenyang Liao, Jingping Fang, Juan Liu, Xiaodan Zhang, Qing Zhang, Weichang Hu, Yuan Qin, Kai Wang, Li-Yu Chen, Neil Shirley, Yann-Rong Lin, Li-Yu Liu, Alvaro G Hernandez, Chris L Wright, Vincent Bulone, Gerald A Tuskan, Katy Heath, Francis Zee, Paul H Moore, Ramanjulu Sunkar, James H Leebens-Mack, Todd Mockler, Jeffrey L Bennetzen, Michael Freeling, David Sankoff, Andrew H Paterson, Xinguang Zhu, Xiaohan Yang, J Andrew C Smith, John C Cushman, Robert E Paull & Qingyi Yu – Show fewer authors

Affiliations | Contributions | Corresponding authors

Nature Genetics (2015) | doi:10.1038/ng.3435

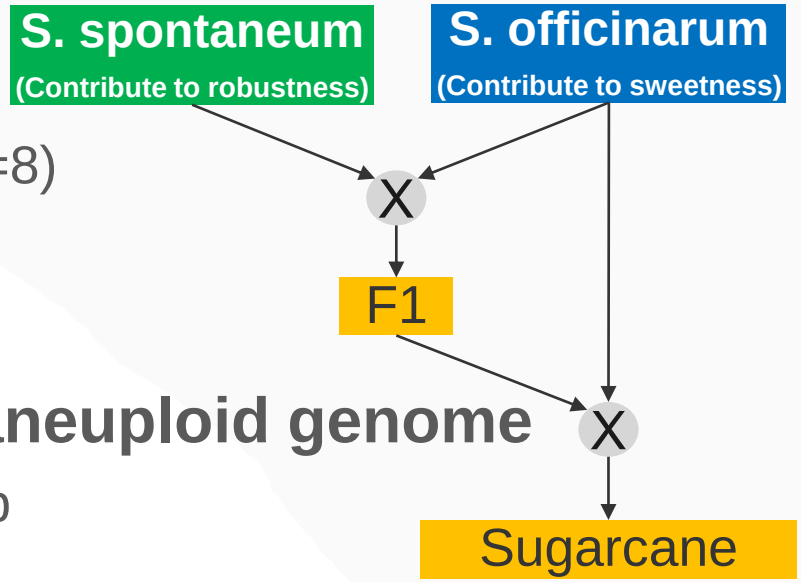
Received 09 July 2015 | Accepted 05 October 2015 | Published online 02 November 2015

Sugarcane for food and biofuel

- **Food**
  - By 2050, the world's population will grow by 50%, thus another 2.5 billion people will need to eat!
  - Rapidly rising oil prices, adverse weather conditions, speculation in agricultural markets are causing more demand
- **Biofuel**
  - By 2050, global energy needs will double as will carbon dioxide emission
  - Low-carbon solution
  - Sugarcane ethanol is a clean, renewable fuel that produces on average 90 percent less carbon dioxide emission than oil and can be an important tool in the fight against climate change.

A hybrid sugarcane cultivar SP80-3280

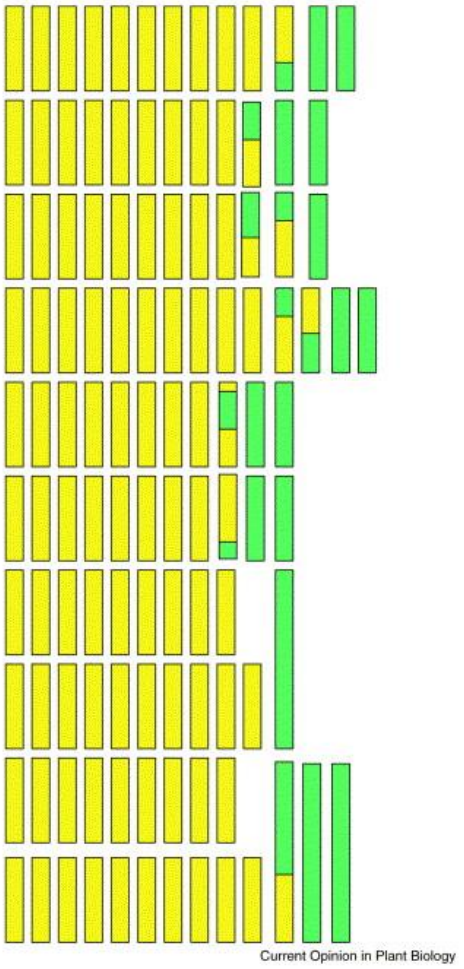
- **S.spontaneum x S.officinarum**
- **A century ago….**
- **Saccharum genus**
  - S. spontaneum (2n=40-128, x=8)
  - S. officinarum (2n=8x=80)



- **Big, highly polyploid and aneuploid genome**
  - Haploid genome is about 1Gbp
  - 8-12 copies per chromosome
  - In total, 100-130 chromosomes
  - Total size is about 10Gbp

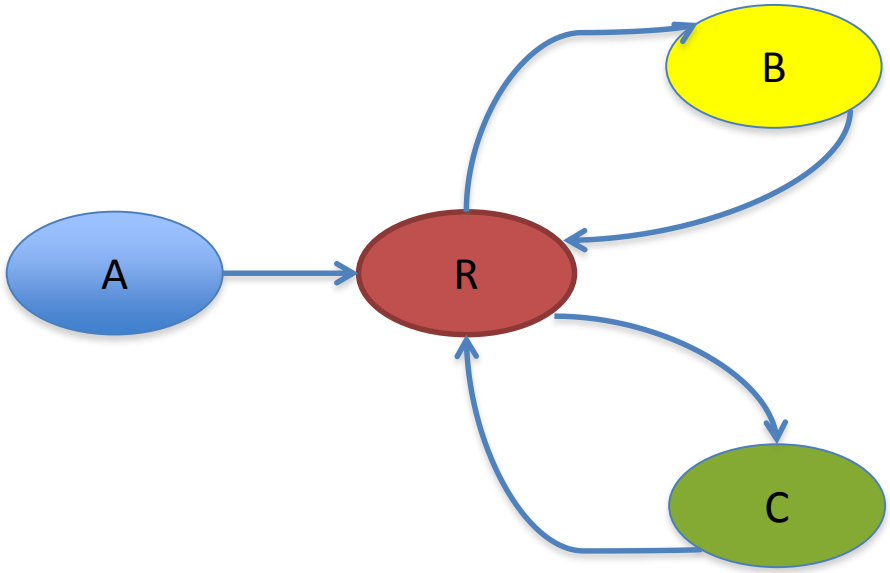
Why is sugarcane assembly harder?

- **Polyploidy/Aneuploidy**
  - 10% of the chromosomes are inherited in their entirety from *S. spontaneum*, 80% are inherited entirely from *S. officinarum*
- **Large scale recombination**
  - 10% is the result of recombination between chromosomes from the two ancestral species, a few being double recombinants



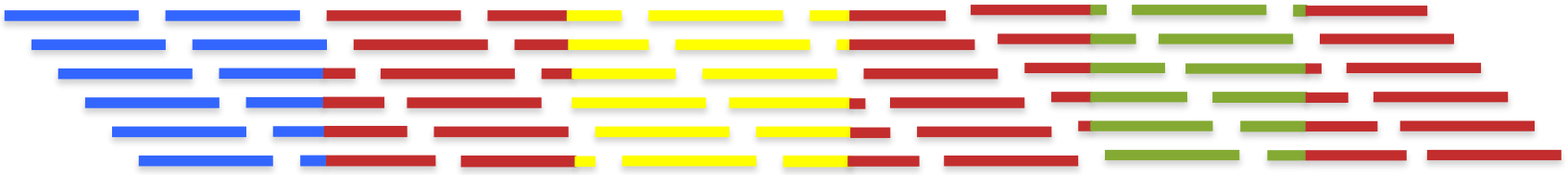
(source) http://ars.els-cdn.com/content/image/1-s2.0-S1369526602002340-gr1.jpg

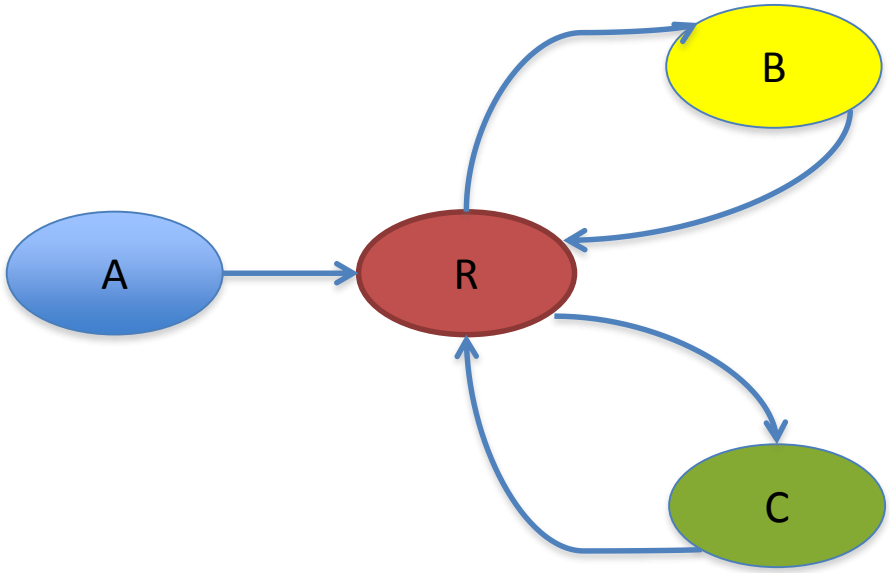
Assembly Complexity by Repeats



Long Reads is the solution!!!

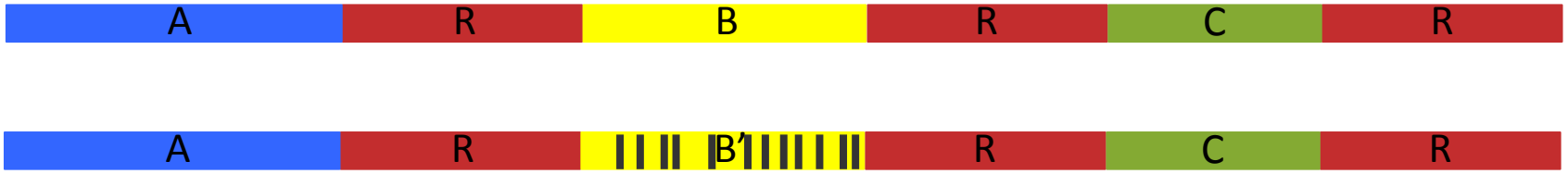
Assembly Complexity by Repeats

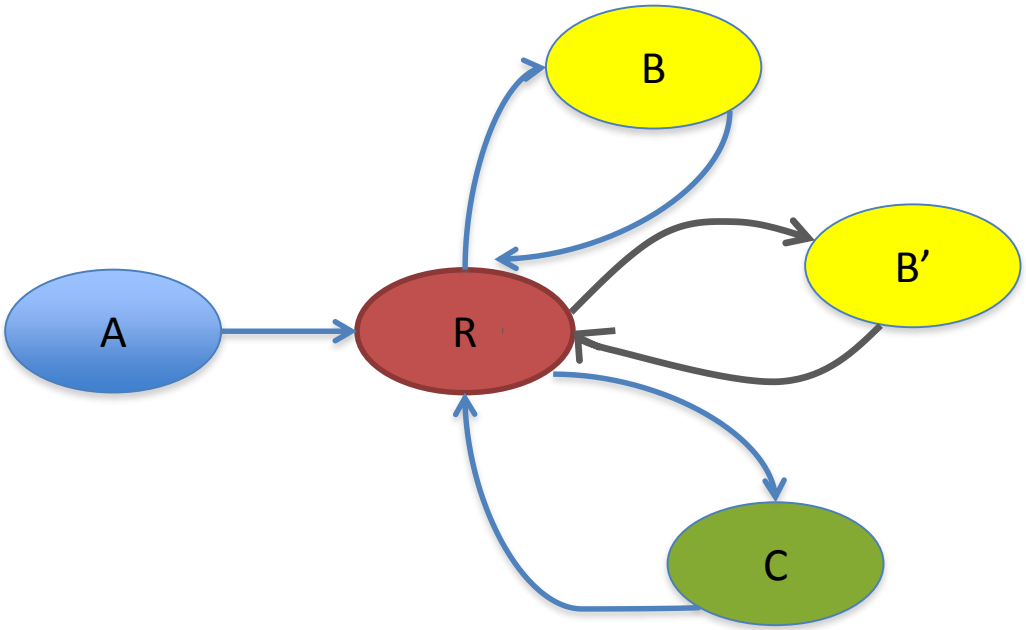




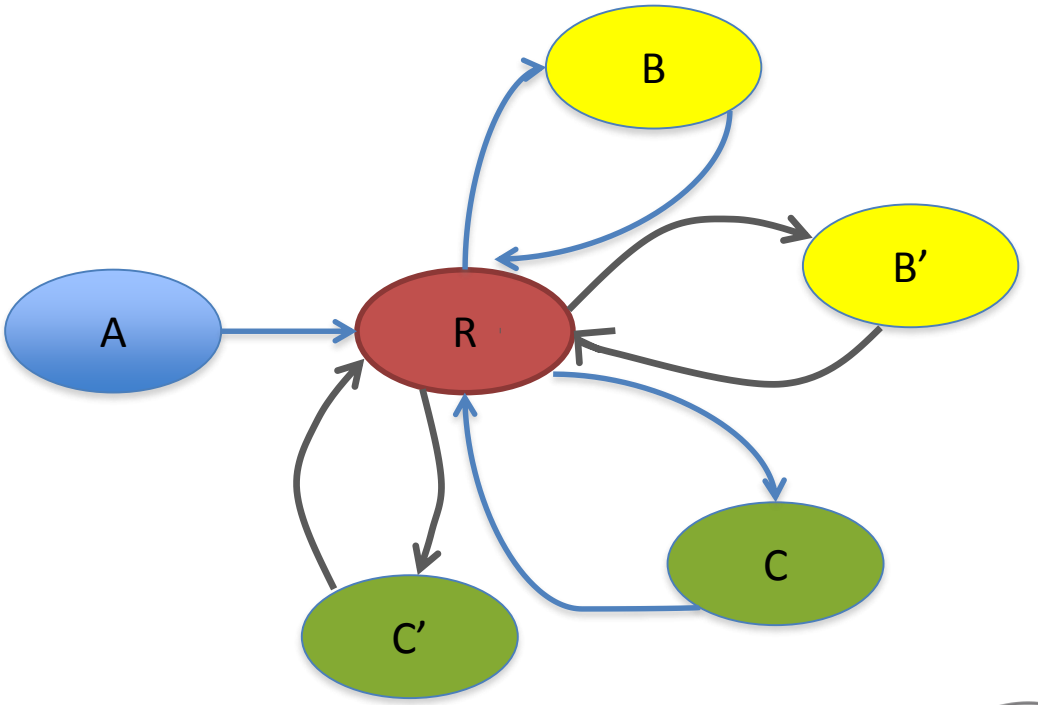
Long Reads is the solution!!!

Assembly Complexity by Heterozygosity





Assembly Complexity by Polyploidy



Long Reads is the solution!!!

Heterozygosity

Choose the right data and the right method

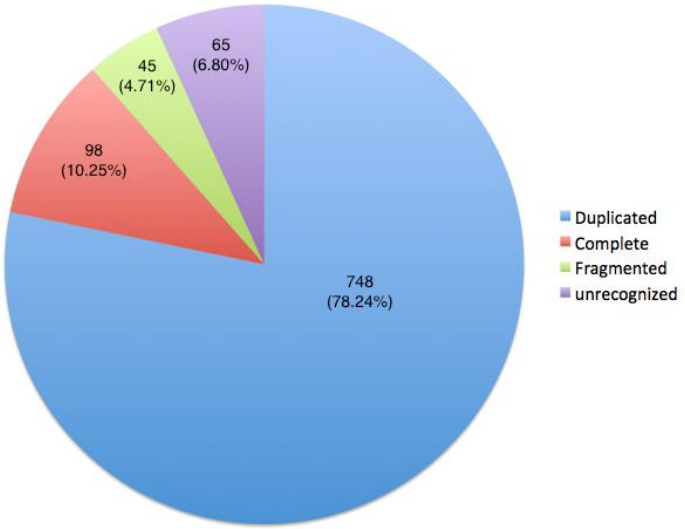
DATA	Hiseq 2000 PE (2x100bp) - 575Gbp - **600x** of haploid genome Roche454 - 9x of haploid genome - [min=20 max=1,168] - Mean=332bp	Moleculo - 19Gbp - **19x** of haploid genome - [min=1,500 max=22,904] - Mean = 4,930bp
Algorithm	SOAPdenovo (De Bruijn Graph)	Celera Assembler (Overlap Graph)
RESULT	Max contig = **21,564** bp NG50=**823** bp Coverage=**0.86x**	Max contig = **467,567** bp NG50=**41,394** bp Coverage=**3.59x** # of contigs = **450K**

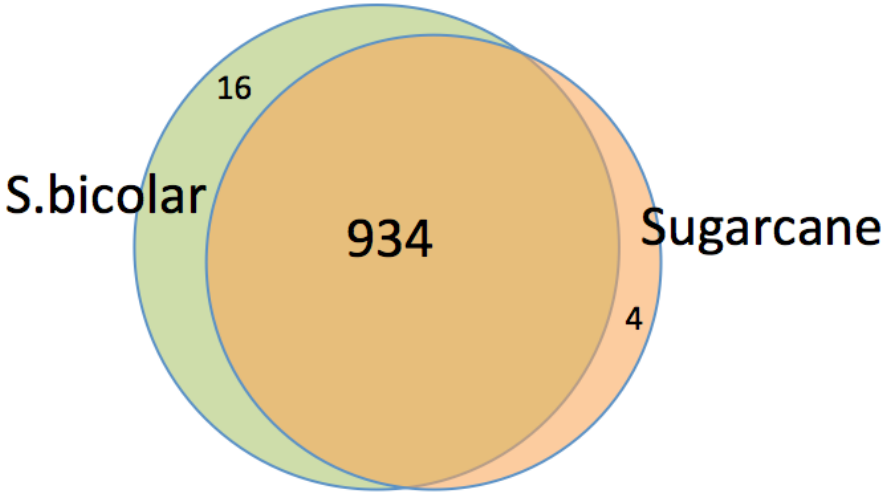
Validation by CEGMA/BUSCO

- **CEGs**
  - Korf Lab in UC. Davis selected 248 core eukaryotic genes
- **Statistics of the completeness**

	Prots	%Completeness	Total	Average	%Ortho
Complete	219	88.31	827	3.78	89.04
Partial	242	97.58	1083	4.48	95.45

- **BUSCO (Benchmarking Universal Single-Copy Orthologs)**





Detectability

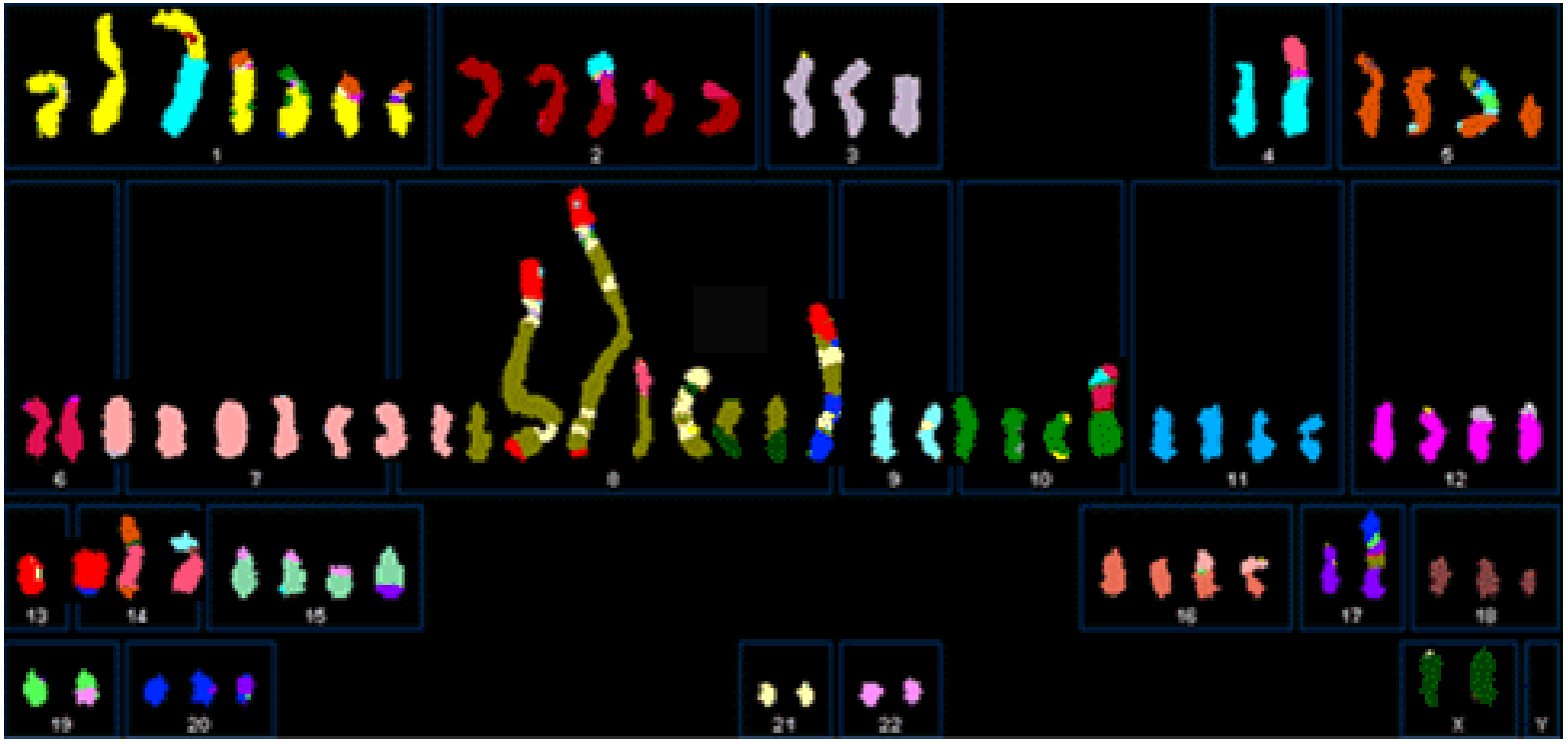
There is an event in your target genome. What is the probability that you can actually detect the event?

Long Reads vs. Short Reads

- **Mappability**
- **Assemblability**
- **Detectability**
  - Insertion
  - Deletion
  - Long range structural variation
  - Translocation
    - Inter-chromosomal
    - Intra-chromosomal
  - Duplication
    - Interspersed duplication
    - Tandem duplication

SKBR-3

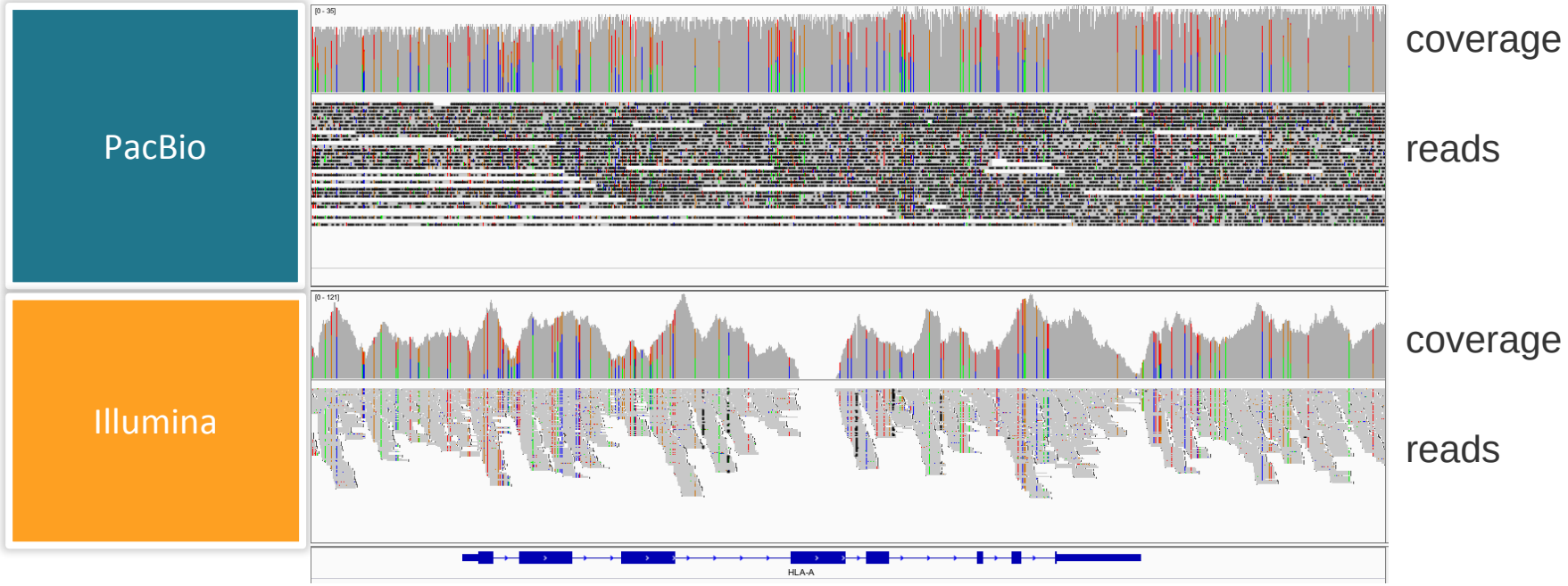
- **Most commonly used Her2-amplified breast cancer cell line**



Source : http://old-www.path.cam.ac.uk/~pawefish/BreastCellLineDescriptions/sk-br-3.htm
(Davidson et al, 2000)

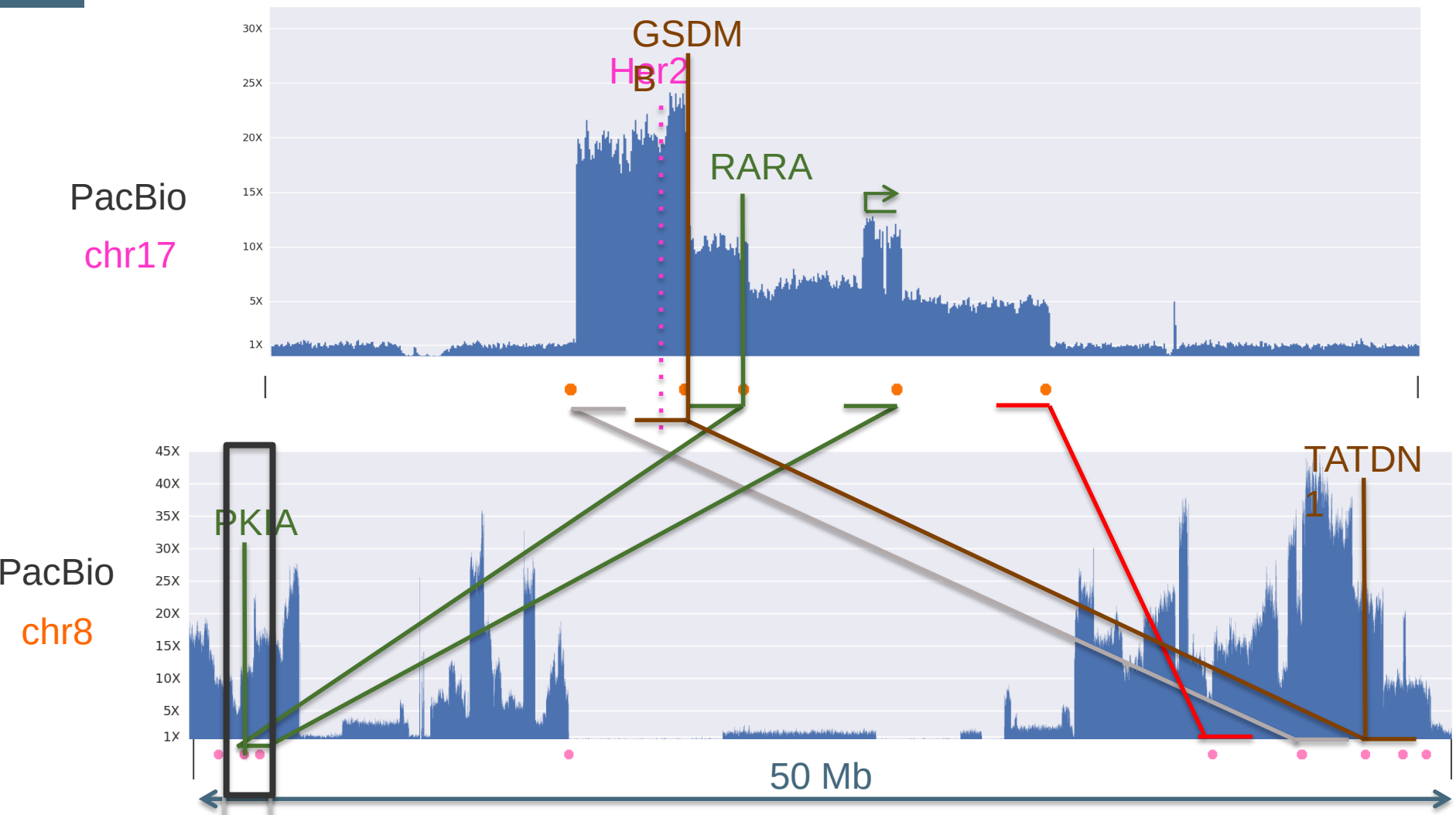
Benefits of Long Reads

- **PacBio coverage is more stable than Illumina coverage in repetitive regions**



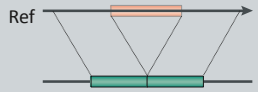
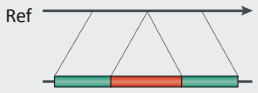
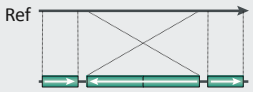
HLA-A gene

Translocation of HER2 in Chr17



Confirmed both known gene fusions in this region

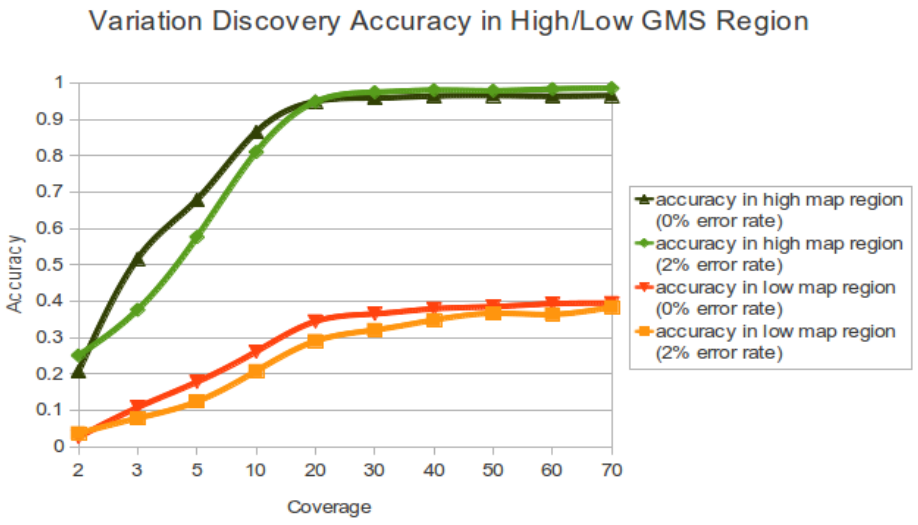
Structural Variation Discovery

Structural variation type		Short reads	Long reads
Deletion		Easy	Easier
Insertion		Hard	Easy
Duplication		Moderate	Easier
Inversion		Moderate	Easier
Translocation		Hard	Easy

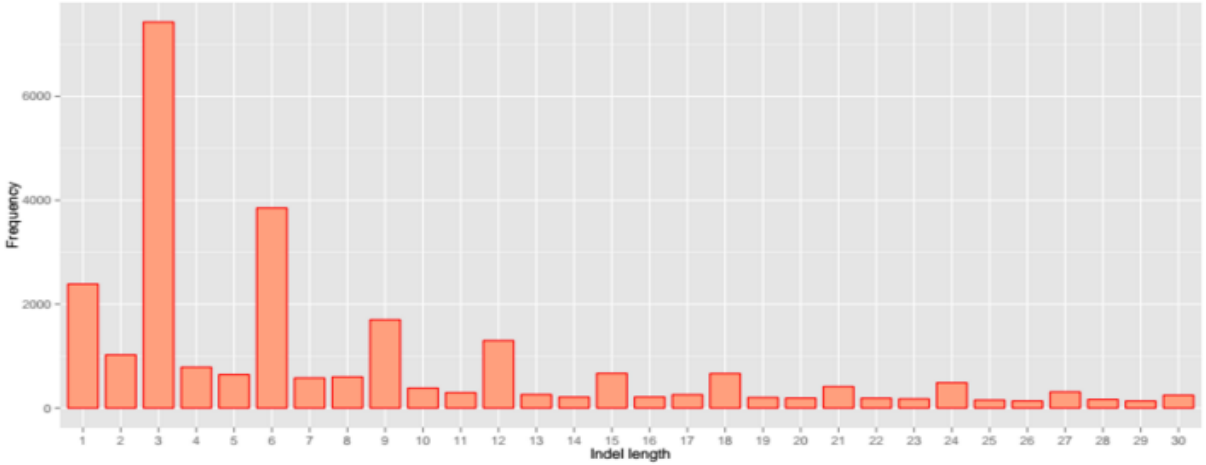
The illusion of deep coverage short read sequencing

- **Deep coverage short read mapping works**
  - If you work on small genomes such as virus or bacteria because it has no significant mappability issues.
  - If you precisely select the regions that you are interested

The illusion of deep coverage short read sequencing



There is no assembler or variant caller to fit highly heterozygous polyploid/aneuploide genome such as sugarcane or cancer genome



Cold Spring Harbor Laboratory

Simons Institute for the Theory of Computing

Berkeley UNIVERSITY OF CALIFORNIA

Acknowledgements

Schatz Lab
Michael Schatz
Fritz Sedlazeck
James Gurtowski
Sri Ramakrishnan
Han fang
Maria Nattestad
Rob Aboukhalil
Tyler Garvin
Mohammad Amin
Shoshana Marcus

Shinjae Yoo

Ravi Pandya
Bob Davidson
David Heckerman

University of São Paulo
Gabriel Rodrigues Alves Margarido
Jonas W. Gaiarsa
Carolina G. Lembke
Marie-Anne Van Sluys
Glaucia M. Souza

McCombie Lab
Dick McCombie
Sara Goodwin

JGI
JOINT GENOME INSTITUTE

Thank You

Q & A