Long Read Sequencing Technology

- Algorithms and applications -

Hayan Lee@Schatz Lab

Aug 11, 2015

Dissertation

Outline

- **Background**
  - Long read sequencing technology
- **The limitations of short read mapping illustrated by Genome Mappability Score (GMS)**
  - Related works - Virmid
- **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)
  - Related works - MHAP
- **Sugarcane de novo genome assembly challenges**
  - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploid genome
  - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
  - Related works
    - Pineapple de novo genome assembly challenges - Heterozygous diploid genome
    - SK-BR-3 breast cancer study using SMRT reads - Benefits of long reads : From de novo assembly to structural variation detection
- **Contributions**

Background

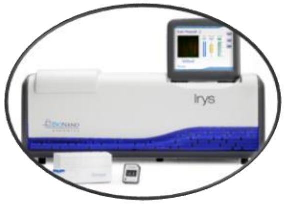
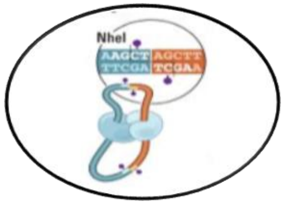
- **Sanger + BAC-by-BAC Era (1995 to 2007)**
  - Very high quality reference genomes for human, mouse, worm, fly, rice, Arabidopsis and a select few other high value species.
  - Contig sizes in the megabases, but costs in the 10s to 100s of millions of dollars
- **Next-Gen Era (2007 to current)**
  - 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 a *Resurgence of Reference Quality Genomes*
  - *De novo* assemblies of human and other large genomes with contig sizes over 1Mbp.

Third-Gen Sequencing 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)

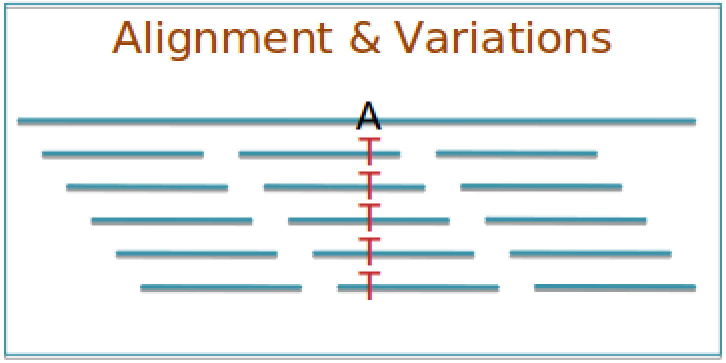
- **Long Span Sequencing: Chromosome Scaffolding, SV analysis, phasing**

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

Outline

- Background
  - Long read sequencing technology
- **The limitations of short read mapping illustrated by Genome Mappability Score (GMS)**
  - Related works - Virmid
- 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)
  - Related works - MHAP
- Sugarcane de novo genome assembly challenges
  - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploid genome
  - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
  - Related works
    - Pineapple de novo genome assembly challenges - Heterozygous diploid genome
    - SK-BR-3 breast cancer study using SMRT reads - Benefits of long reads : From de novo assembly to structural variation detection
- Contributions

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

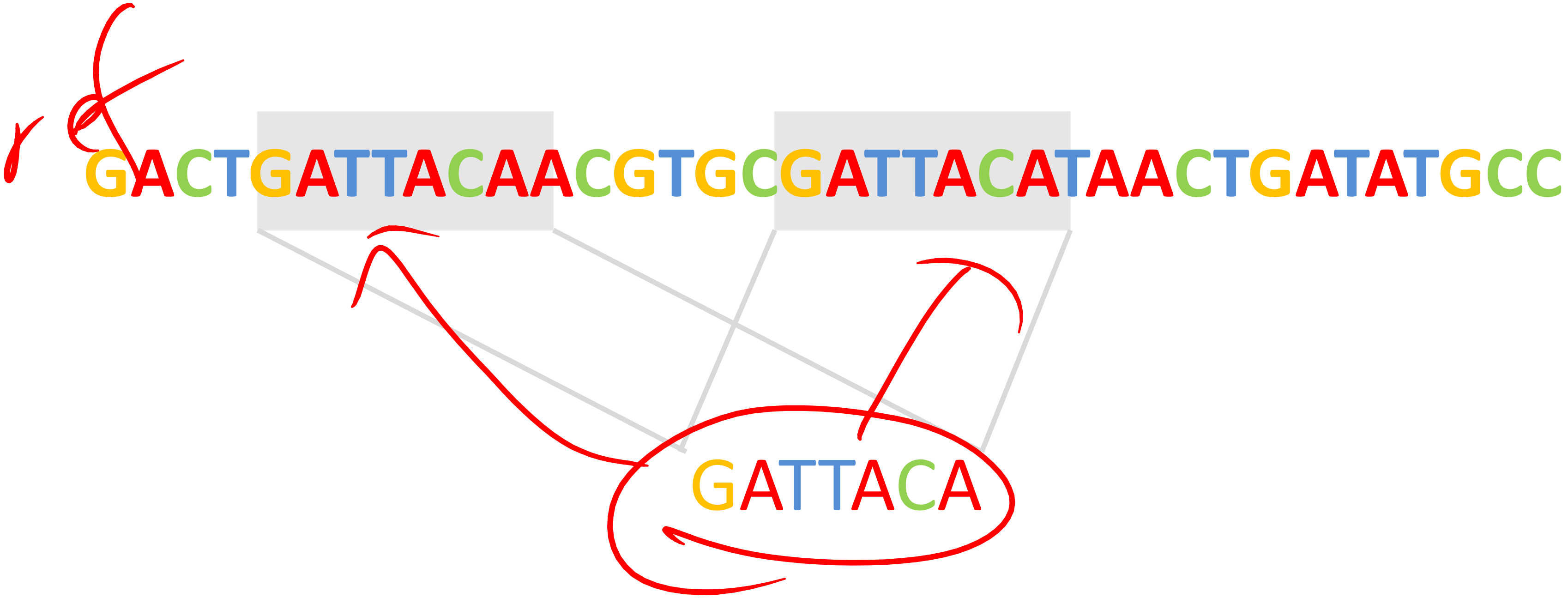
ENCODE

HMP

TCGA

NCI | NHGRI

Repeats



GACTGATTACAACGTGCGATTACATAACTGATATGCC

GATTACA

Read Quality Score - MAQ(1)

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

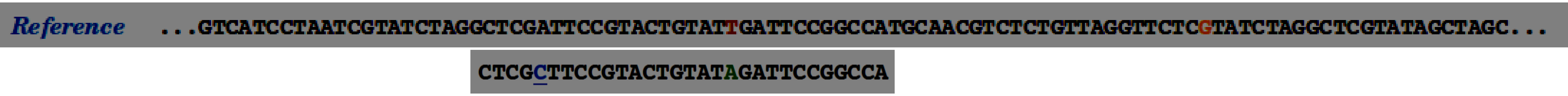
$$Q_s = -10\log_{10}[1 - p_s(u \mid x, z)]$$

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

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 error 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)



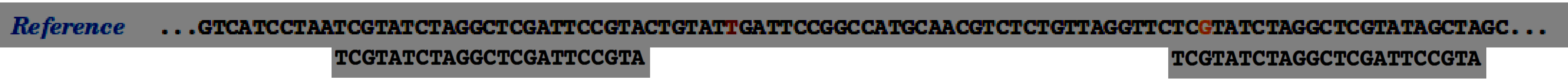
```
Reference   ...GTCATCCTAATCGTATCTAGGCTCGATTCCGTACTGTATTGATTCCGGCCATGCAACGTCTCTGTTAGGTTCTCGTATCTAGGCTCGTATAGCTAGC...
                                    CTCGCTTCCGTACTGTATAGATTCCGGCCA
```

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

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

- **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?**

Read Quality Score – MAQ (3)



```
Reference   ...GTCATCCTAATCGTATCTAGGCTCGATTCCGTACTGTATTGATTCCGGCCATGCAACGTCTCTGTTAGGTTCTCGTATCTAGGCTCGTATAGCTAGC...
                   TCGTATCTAGGCTCGATTCCGTA                                       TCGTATCTAGGCTCGATTCCGTA
```

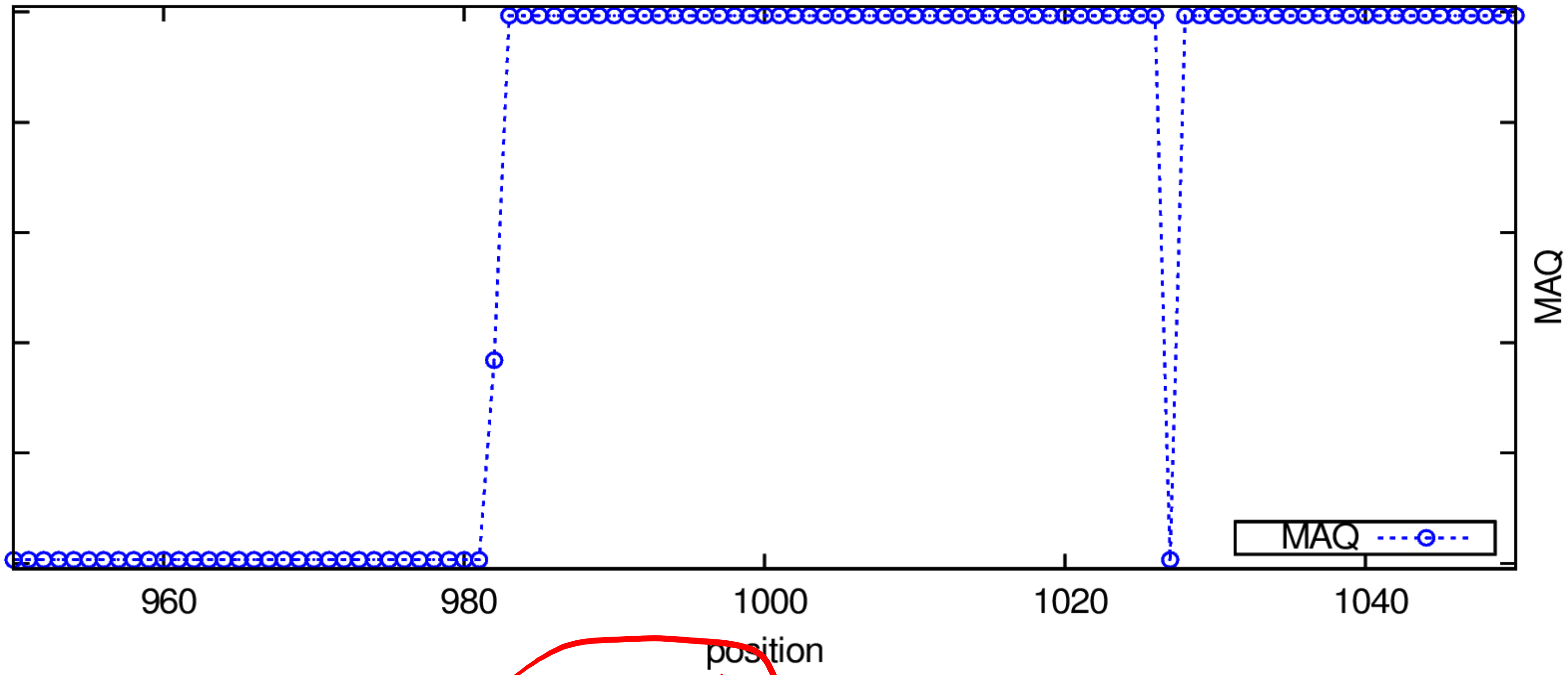
$$Q_s = -10\log_{10}[1 - p_s(u \mid x, z)]$$

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

- **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

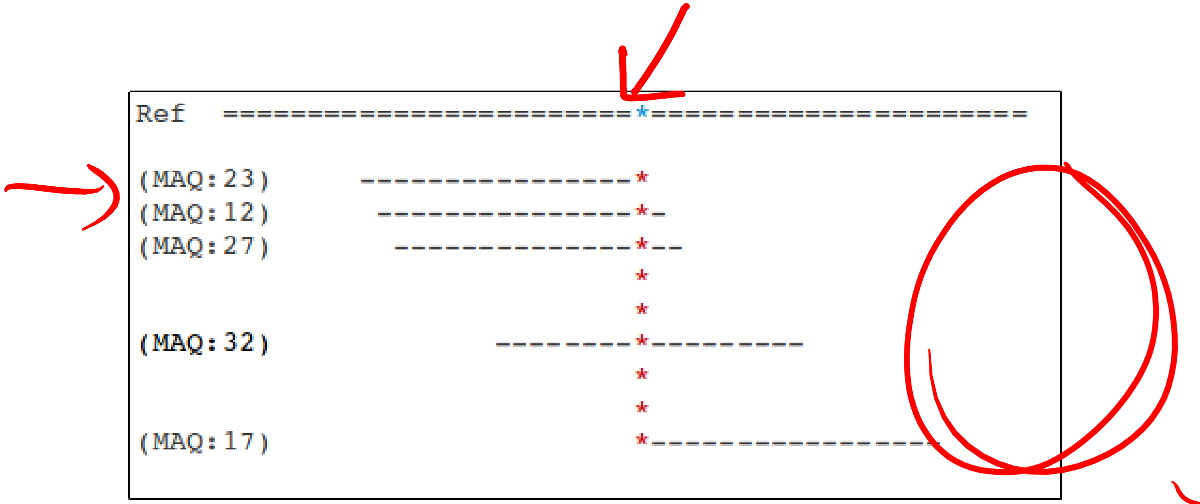


AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACTAGCATGCCCTAAGCCCGTAATGCAGTCATACTAGCTATCCTCGCCCTCTCCGTCAAGCTAC

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)



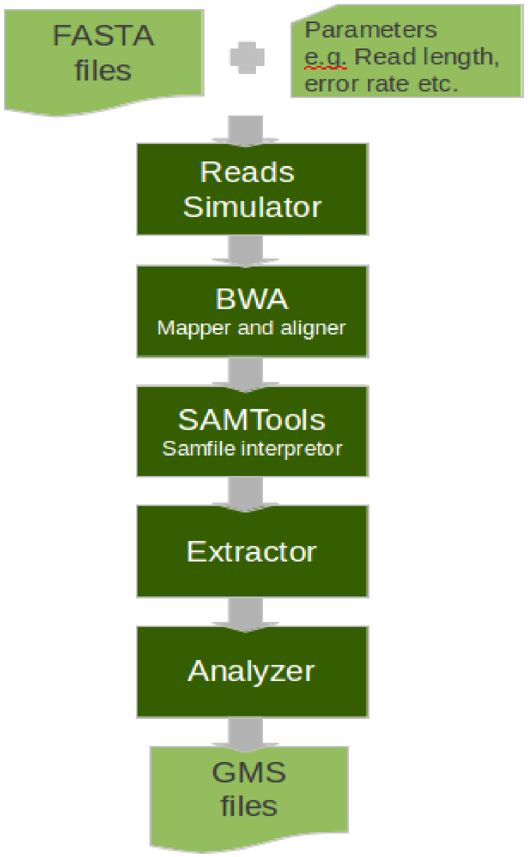
```
Ref    =======================*======================

(MAQ:23)       -----------------*
(MAQ:12)        ----------------*-
(MAQ:27)         ---------------*--
                                *
                                *
(MAQ:32)                --------*---------
                                *
                                *
(MAQ:17)                        *-----------------
```

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

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

Genome Mappability Analyzer (GMA)



FASTA files + Parameters e.g. Read length, error rate etc.

Reads Simulator

BWA
Mapper and aligner

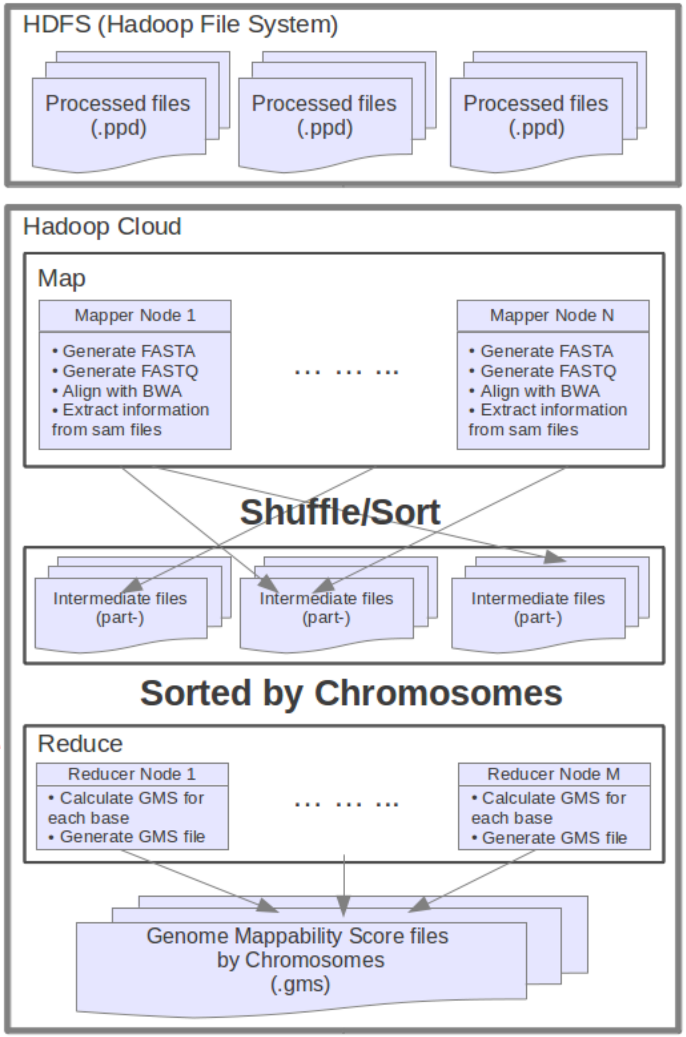
SAMTools
Samfile interpretor

Extractor

Analyzer

GMS files

lol



HDFS (Hadoop File System)

Processed files (.ppd) | Processed files (.ppd) | Processed files (.ppd)

Hadoop Cloud

Map

Mapper Node 1
- Generate FASTA
- Generate FASTQ
- Align with BWA
- Extract information from sam files

… … …

Mapper Node N
- Generate FASTA
- Generate FASTQ
- Align with BWA
- Extract information from sam files

Shuffle/Sort

Intermediate files (part-) | Intermediate files (part-) | Intermediate files (part-)

Sorted by Chromosomes

Reduce

Reducer Node 1
- Calculate GMS for each base
- Generate GMS file

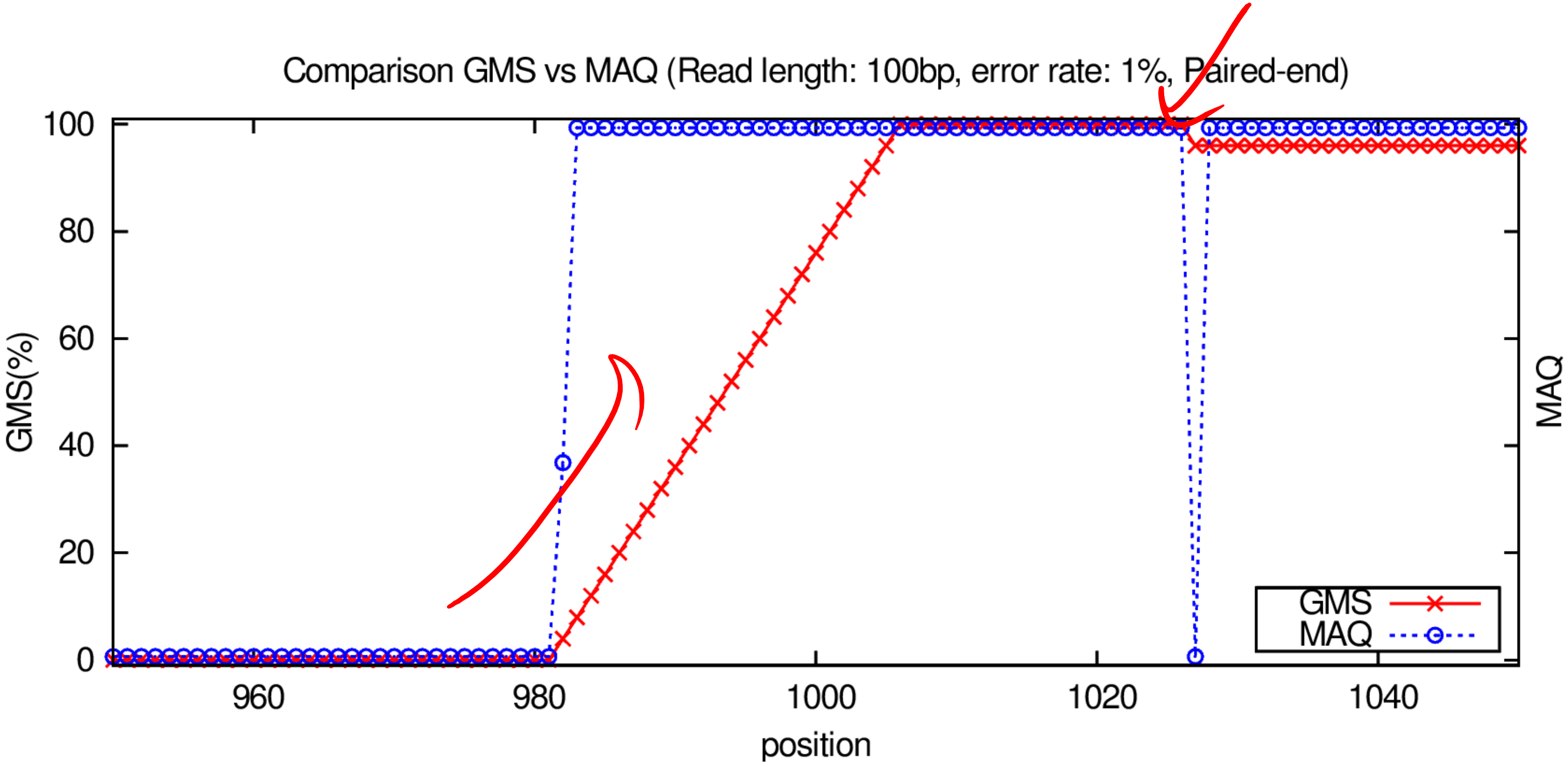
… … …

Reducer Node M
- Calculate GMS for each base
- Generate GMS file

Genome Mappability Score files by Chromosomes (.gms)

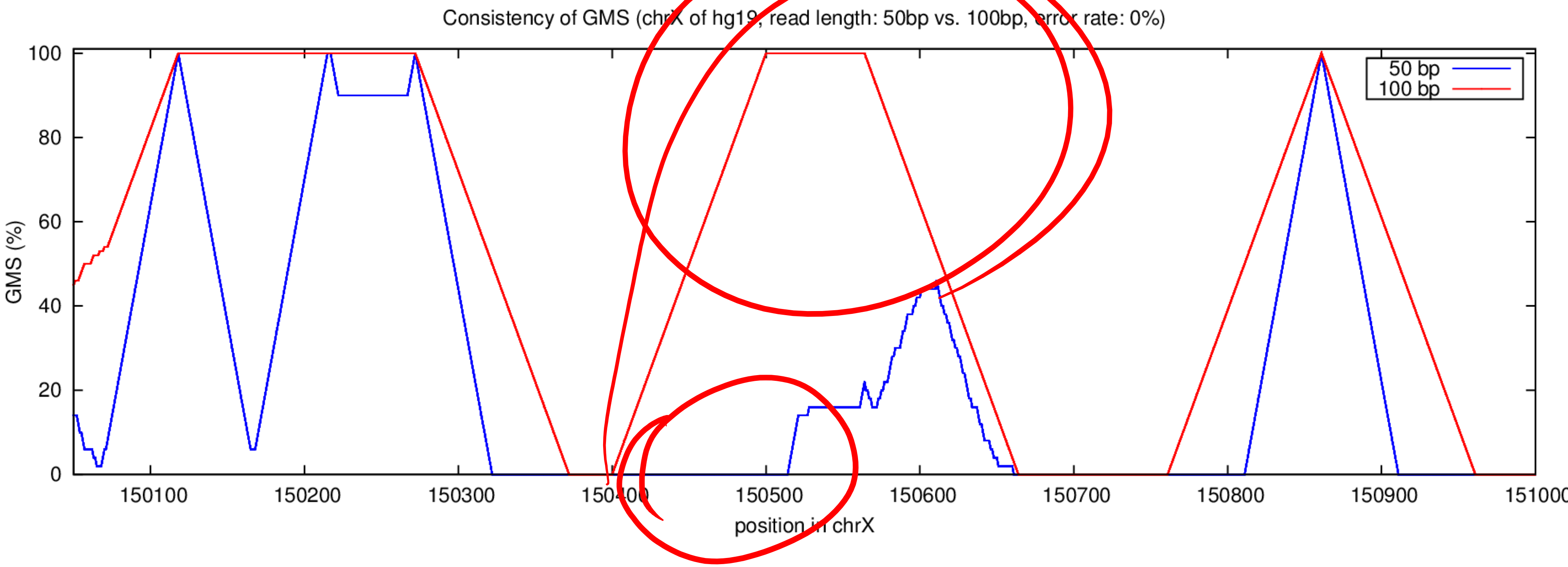
GMS vs. MAQ

Sensitivity of Read Mapping Score

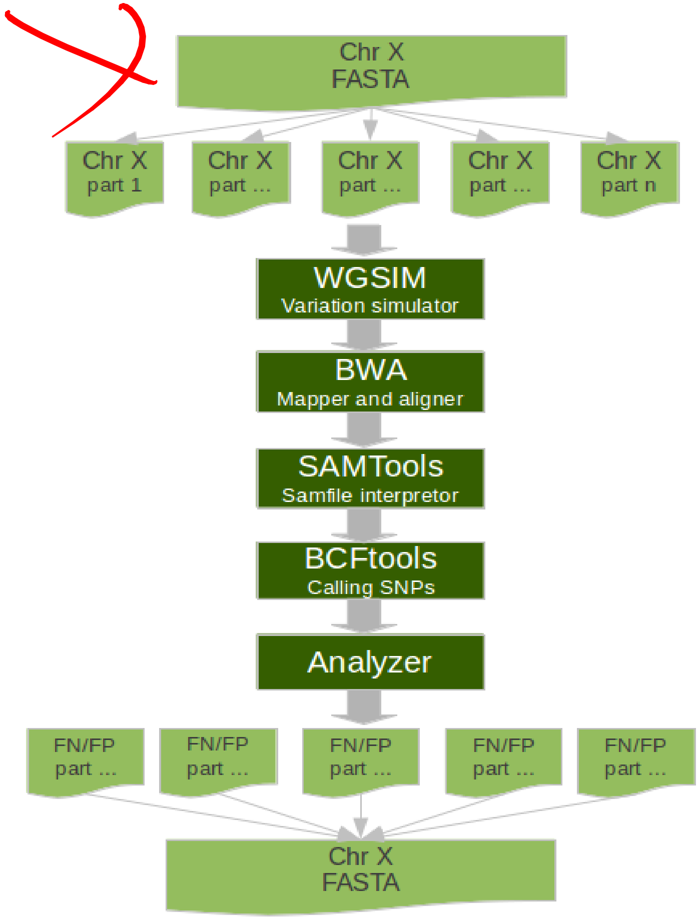


AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA CTAGCATGCCCTAAGCCCGTAATGCAGTCATACTAGCTATCCTCGCCCTCTCCGTCAAGCTAC

Map = f(Read Length)

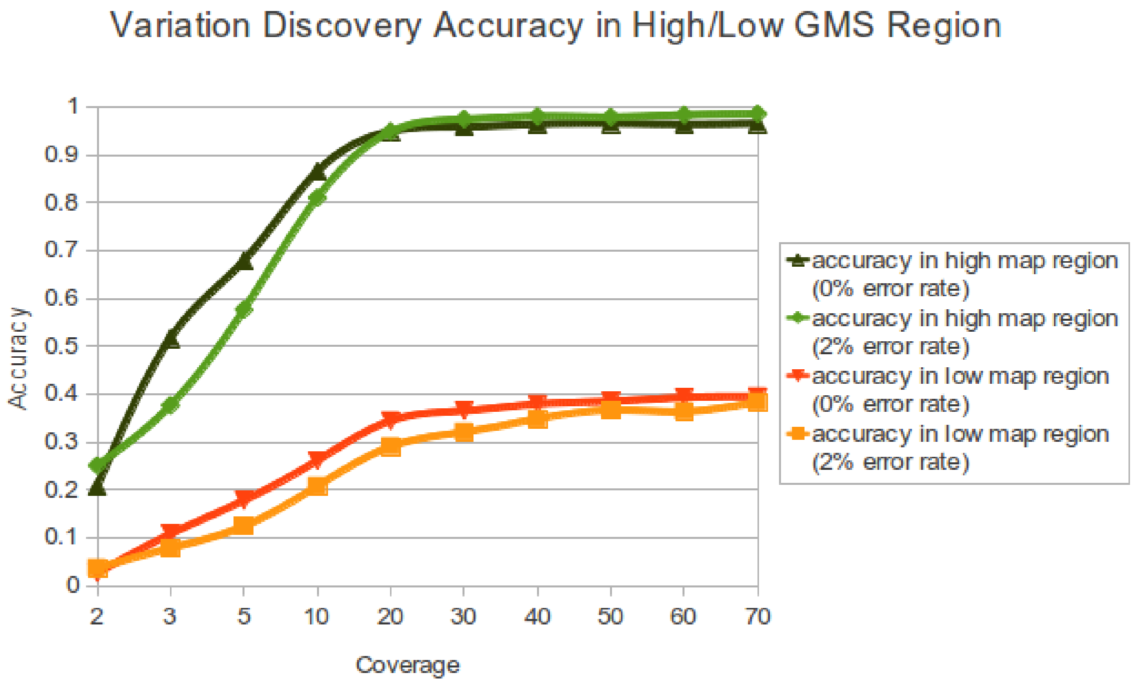

Consistency of GMS (chrX of hg19, read length: 50bp vs. 100bp, error rate: 0%)

Variation Accuracy Simulator (VAS)

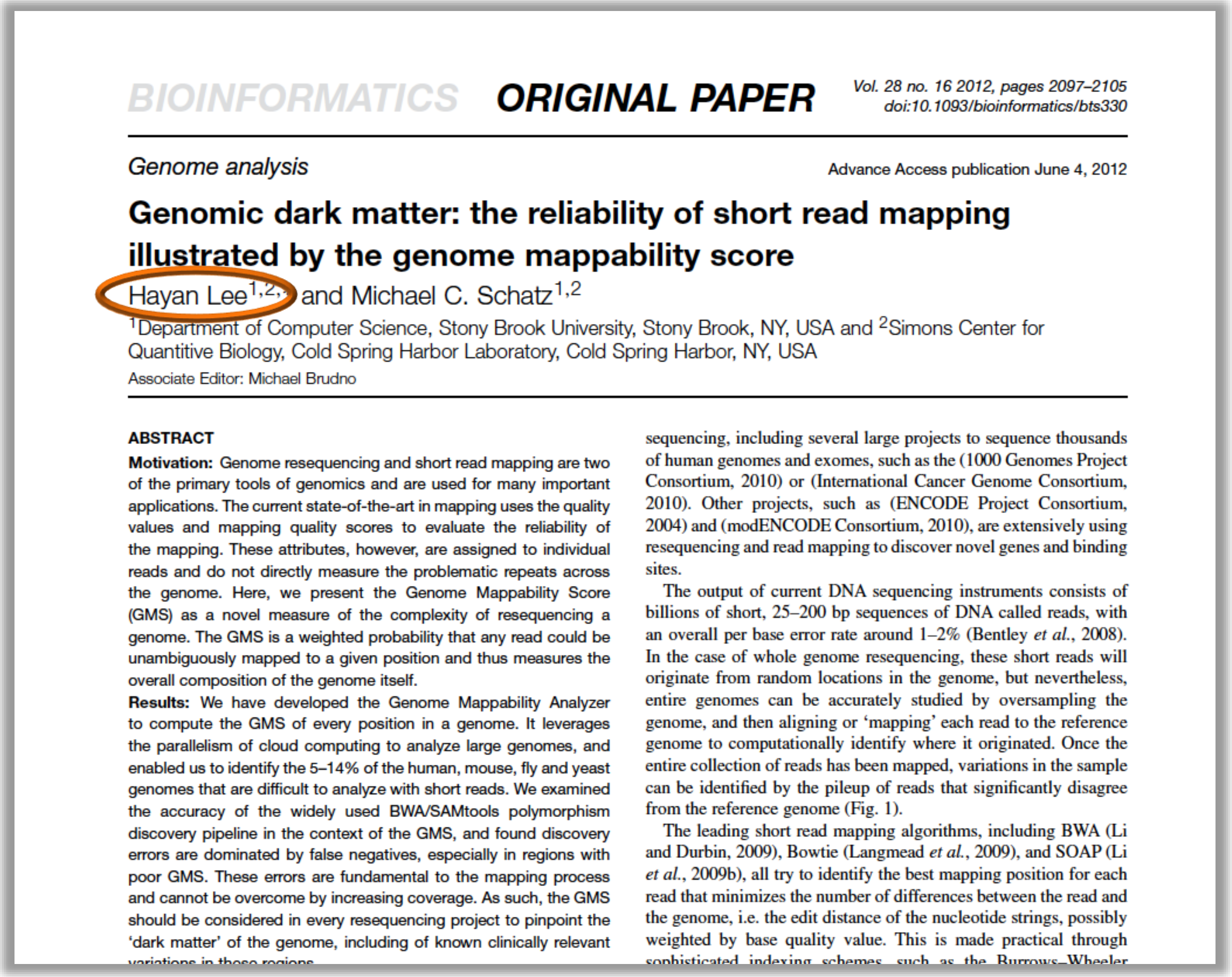


- Simulation of resequencing experiments to measure the accuracy of variation detection

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



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]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA and [2]Simons Center for Quantitive Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Genome resequencing and short read mapping are two of the primary tools of genomics and are used for many important applications. The current state-of-the-art in mapping uses the quality values and mapping quality scores to evaluate the reliability of the mapping. These attributes, however, are assigned to individual reads and do not directly measure the problematic repeats across the genome. Here, we present the Genome Mappability Score (GMS) as a novel measure of the complexity of resequencing a genome. The GMS is a weighted probability that any read could be unambiguously mapped to a given position and thus measures the overall composition of the genome itself.

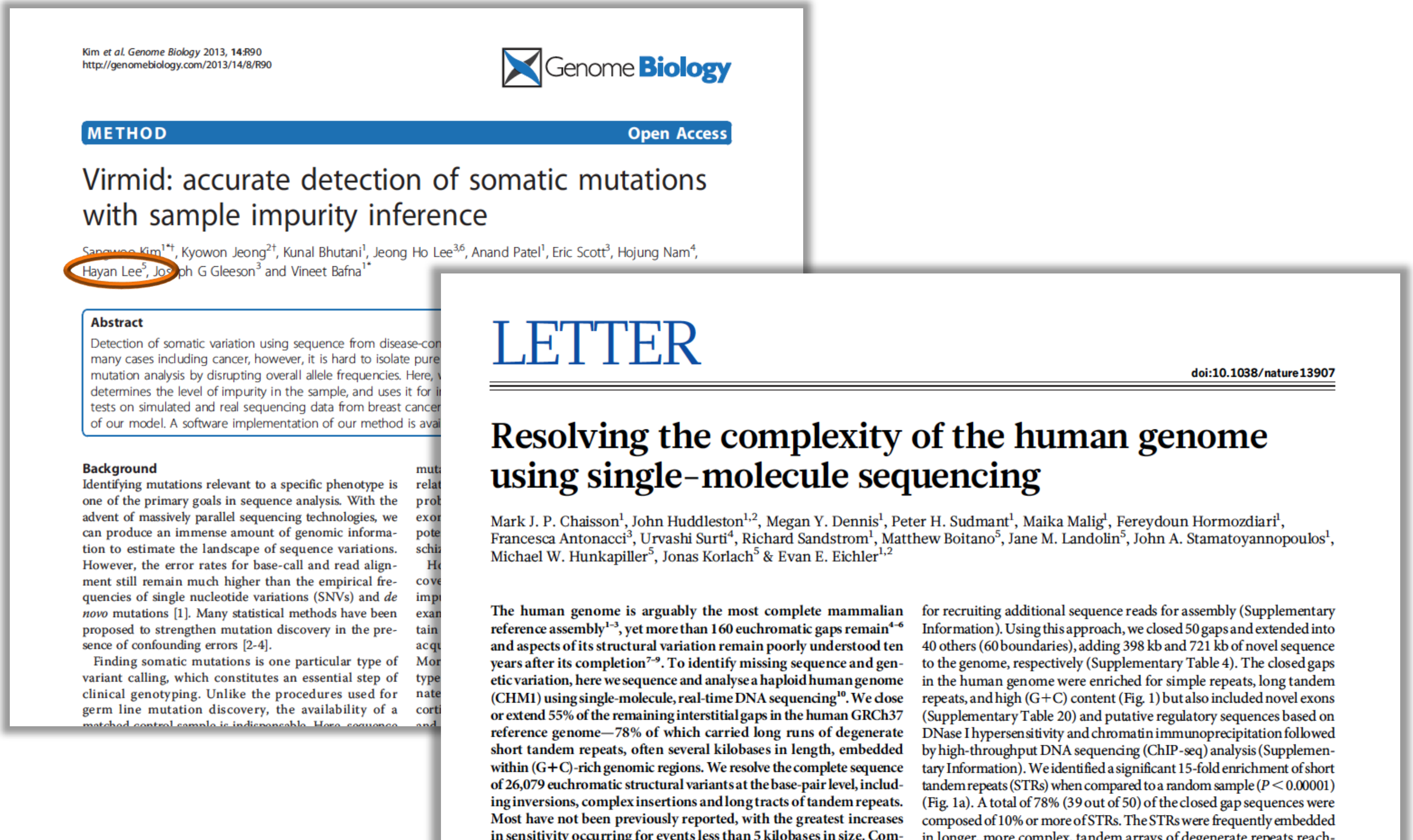
Results: We have developed the Genome Mappability Analyzer to compute the GMS of every position in a genome. It leverages the parallelism of cloud computing to analyze large genomes, and enabled us to identify the 5–14% of the human, mouse, fly and yeast genomes that are difficult to analyze with short reads. We examined the accuracy of the widely used BWA/SAMtools polymorphism discovery pipeline in the context of the GMS, and found discovery errors are dominated by false negatives, especially in regions with poor GMS. These errors are fundamental to the mapping process and cannot be overcome by increasing coverage. As such, the GMS should be considered in every resequencing project to pinpoint the 'dark matter' of the genome, including of known clinically relevant variations in these regions.

sequencing, including several large projects to sequence thousands of human genomes and exomes, such as the (1000 Genomes Project Consortium, 2010) or (International Cancer Genome Consortium, 2010). Other projects, such as (ENCODE Project Consortium, 2004) and (modENCODE Consortium, 2010), are extensively using resequencing and read mapping to discover novel genes and binding sites.

The output of current DNA sequencing instruments consists of billions of short, 25–200 bp sequences of DNA called reads, with an overall per base error rate around 1–2% (Bentley *et al.*, 2008). In the case of whole genome resequencing, these short reads will originate from random locations in the genome, but nevertheless, entire genomes can be accurately studied by oversampling the genome, and then aligning or 'mapping' each read to the reference genome to computationally identify where it originated. Once the entire collection of reads has been mapped, variations in the sample can be identified by the pileup of reads that significantly disagree from the reference genome (Fig. 1).

The leading short read mapping algorithms, including BWA (Li and Durbin, 2009), Bowtie (Langmead *et al.*, 2009), and SOAP (Li *et al.*, 2009b), all try to identify the best mapping position for each read that minimizes the number of differences between the read and the genome, i.e. the edit distance of the nucleotide strings, possibly weighted by base quality value. This is made practical through sophisticated indexing schemes, such as the Burrows–Wheeler

Cited by



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

Genome **Biology**

METHOD **Open Access**

Virmid: accurate detection of somatic mutations with sample impurity inference

Sangwoo Kim[1*†], Kyowon Jeong[2†], Kunal Bhutani[1], Jeong Ho Lee[3,6], Anand Patel[1], Eric Scott[3], Hojung Nam[4], Hayan Lee[5], Joseph G Gleeson[3] and Vineet Bafna[1*]

Abstract

Detection of somatic variation using sequence from disease-con
many cases including cancer, however, it is hard to isolate pure
mutation analysis by disrupting overall allele frequencies. Here, v
determines the level of impurity in the sample, and uses it for i
tests on simulated and real sequencing data from breast cancer
of our model. A software implementation of our method is avai

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 variant calling, which constitutes an essential step of clinical genotyping. Unlike the procedures used for germ line mutation discovery, the availability of a matched control sample is indispensable. Here sequence

mut
relat
prob
exon
pote
schi
Ho
cove
imp
exan
tain
acqu
Mor
type
nate
corti
and

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. Com-

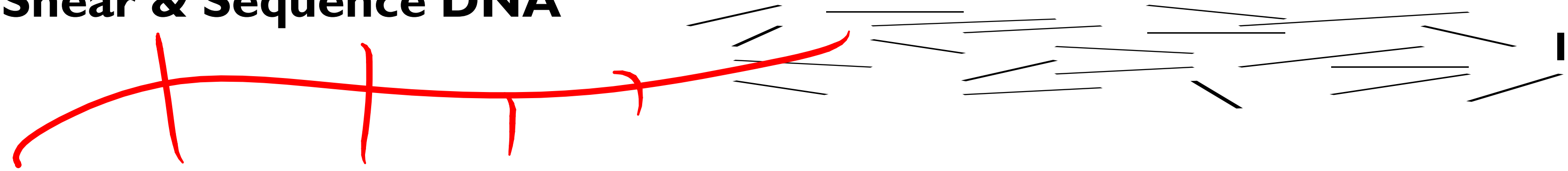
for recruiting additional sequence reads for assembly (Supplementary Information). Using this approach, we closed 50 gaps and extended into 40 others (60 boundaries), adding 398 kb and 721 kb of novel sequence to the genome, respectively (Supplementary Table 4). The closed gaps in the human genome were 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 reach-

Outline

- Background
  - Long read sequencing technology
- The limitations of short read mapping illustrated by Genome Mappability Score (GMS)
  - Related works - Virmid
- **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)
  - Related works - MHAP
- Sugarcane de novo genome assembly challenges
  - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploid genome
  - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
  - Related works
    - Pineapple de novo genome assembly challenges - Heterozygous diploid genome
    - SK-BR-3 breast cancer study using SMRT reads - Benefits of long reads : From de novo assembly to structural variation detection
- Contributions

De novo genome assembly

1. **Shear & Sequence DNA**



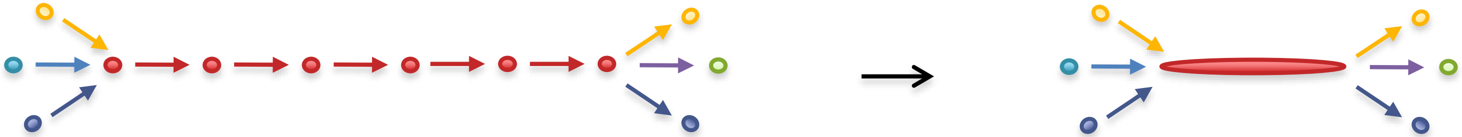
2. **Construct assembly graph from overlapping reads**

...AGCCTAGGGATGCGCGACACGT

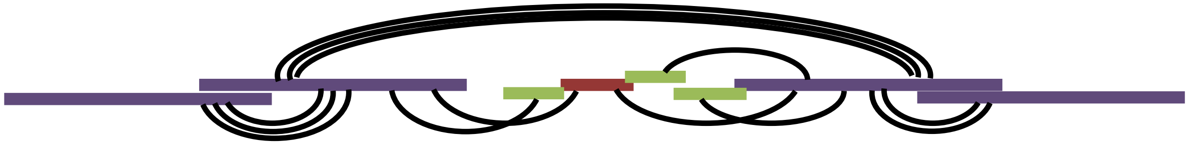
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

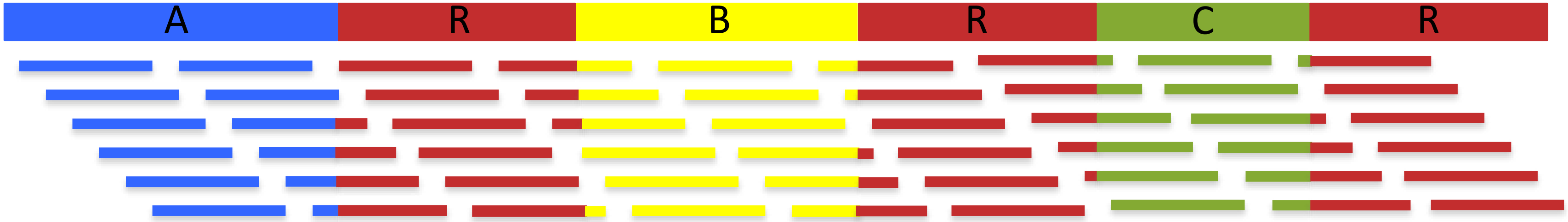
3. **Simplify assembly graph**

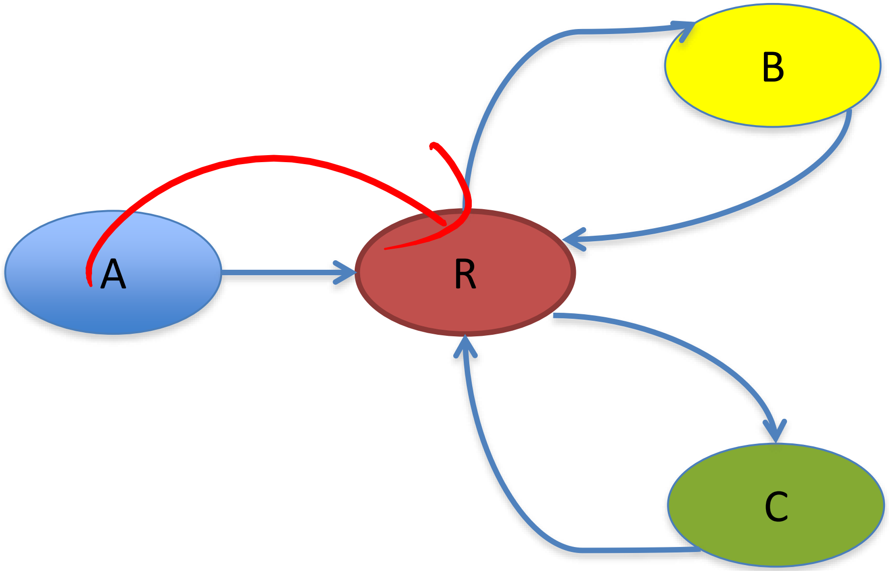


4. **Detangle graph with long reads, mates, and other links**



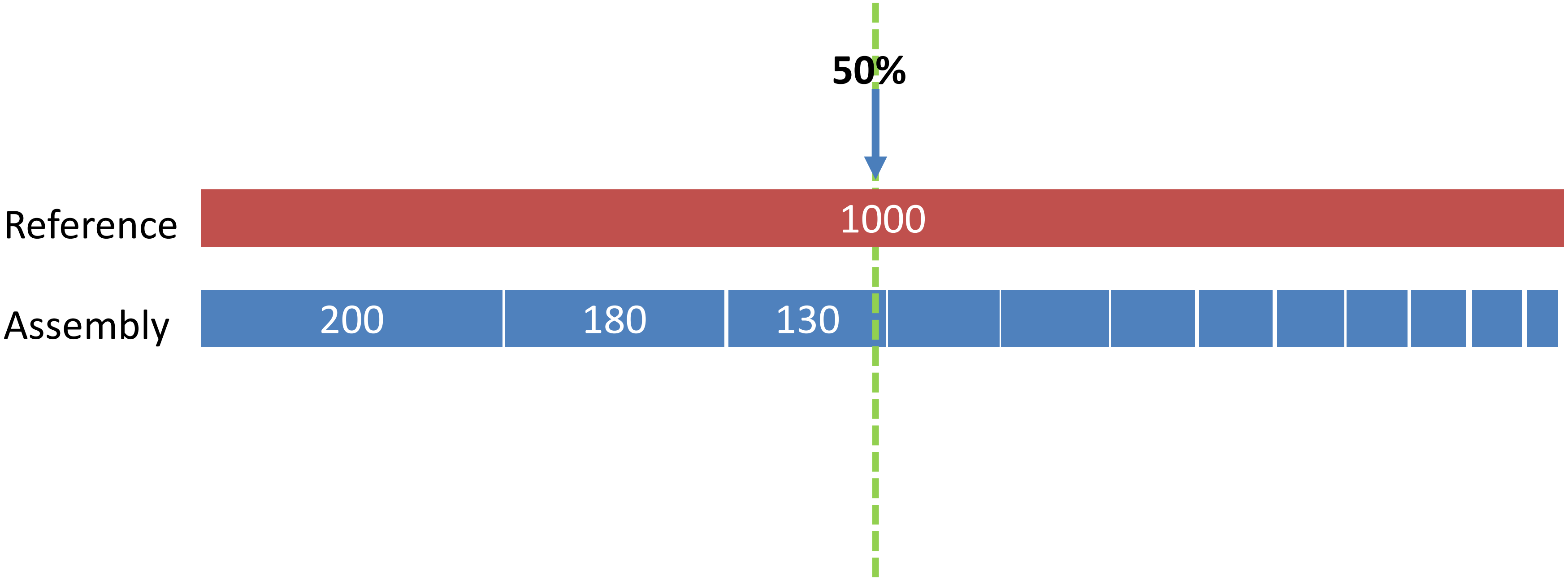
Assembly Complexity by Repeats





Long Reads is the solution!!!

N50 : Contiguity Metric



N50 = 130

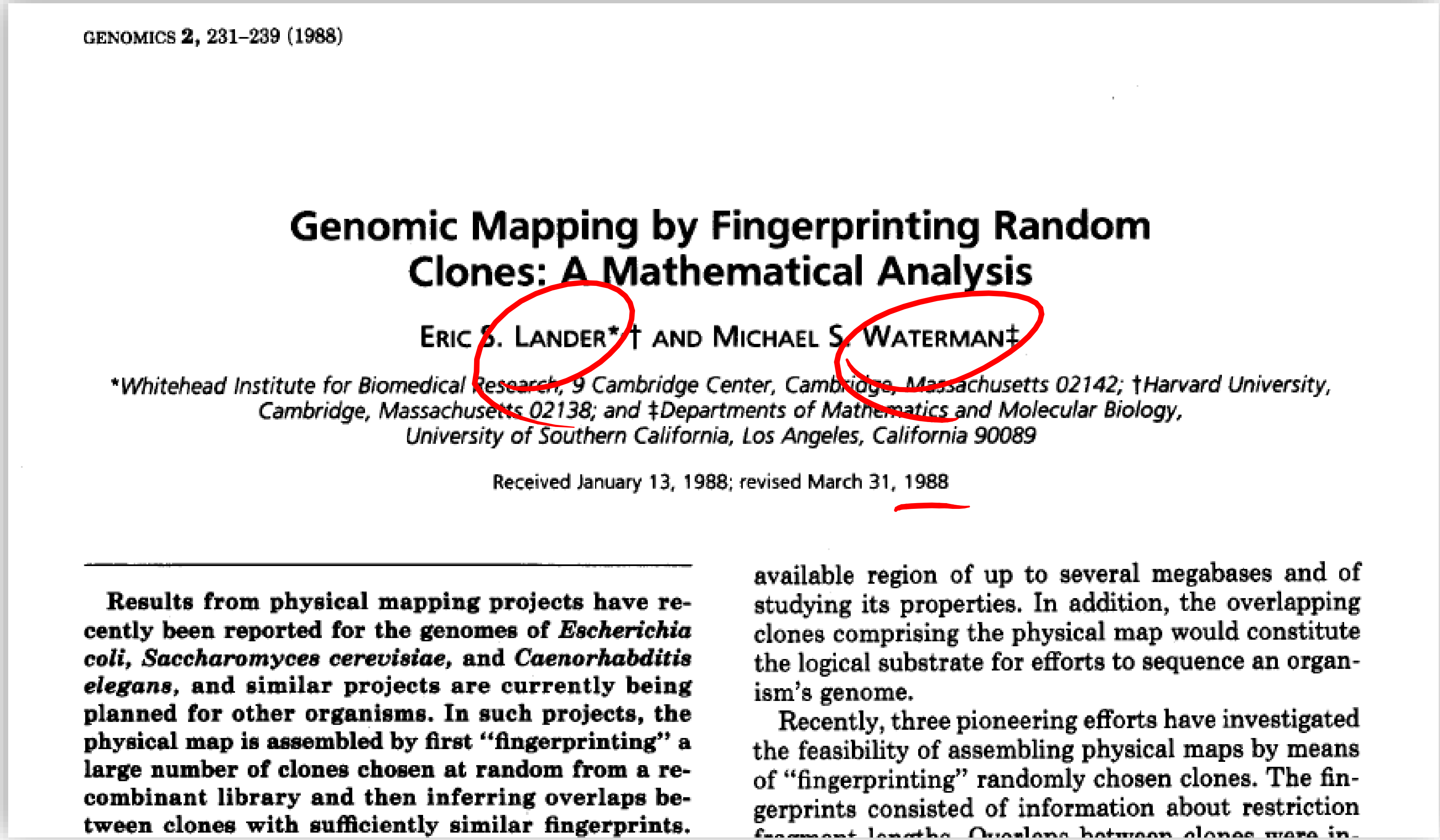
Many Genomes Are Sequenced… Many Questions Are Raised… But…

- 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

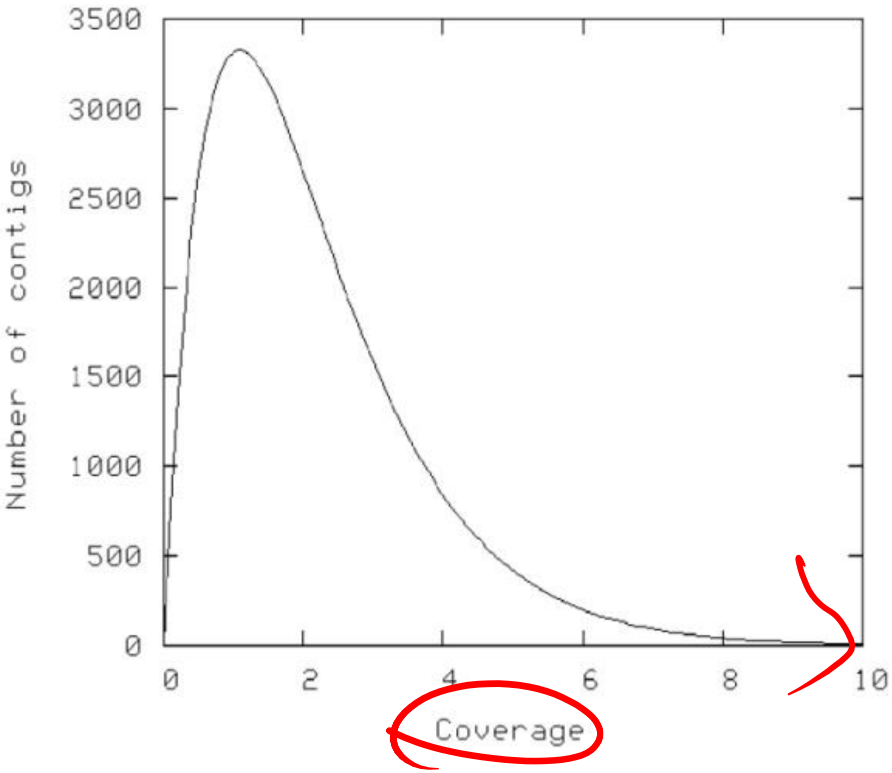
Received January 13, 1988; revised March 31, 1988

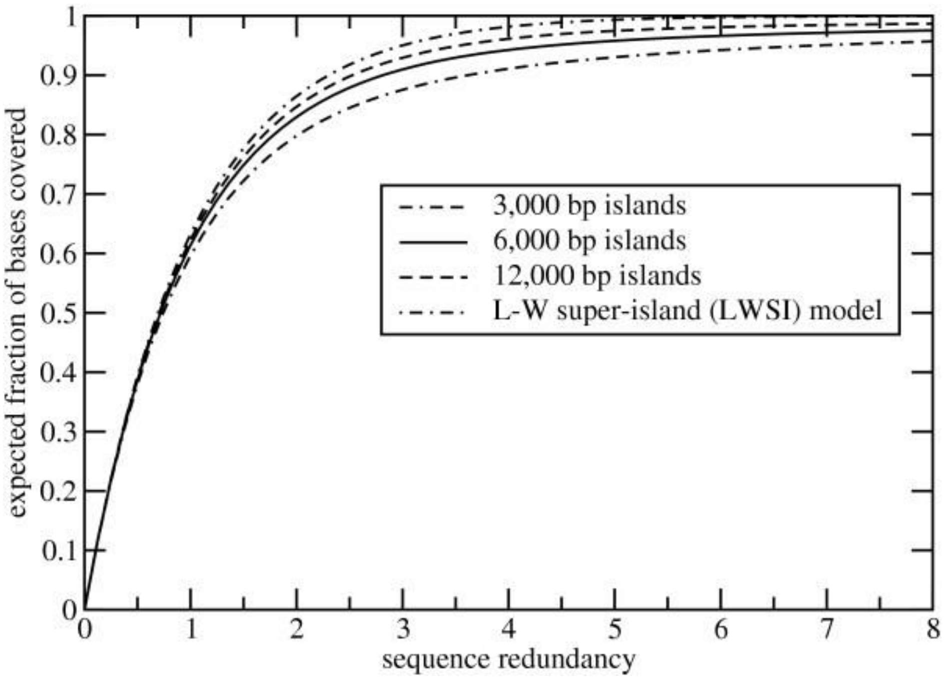
Results from physical mapping projects have recently been reported for the genomes of *Escherichia coli*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*, and similar projects are currently being planned for other organisms. In such projects, the physical map is assembled by first "fingerprinting" 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-

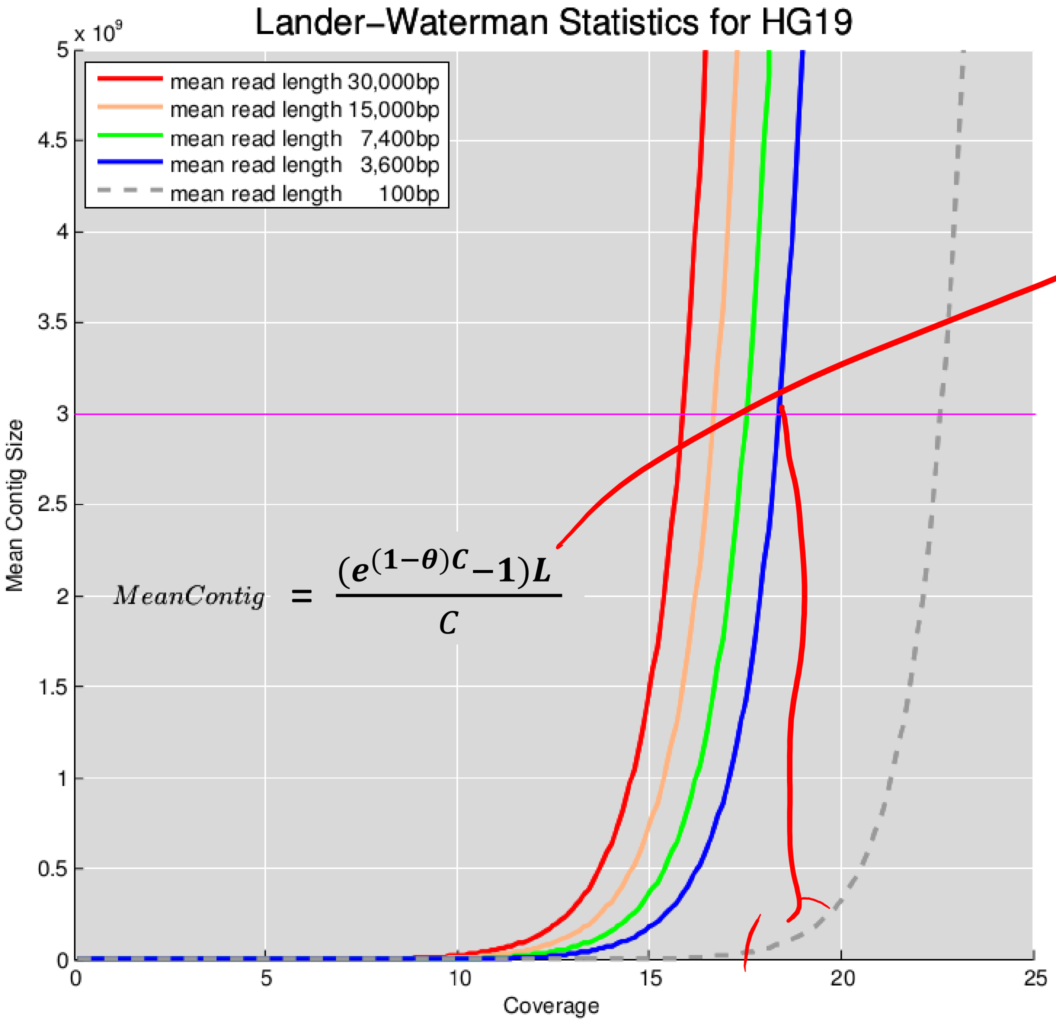
Lander-Waterman Statistics





In practice, it's useful only in low coverage (3-5x) but becomes nonsensical in high coverage.

HG19 Genome Assembly Performance by Lander-Waterman Statistics



$$MeanContig = \frac{(e^{(1-\theta)C}-1)L}{C}$$

Two key observations

1. Contig over genome size

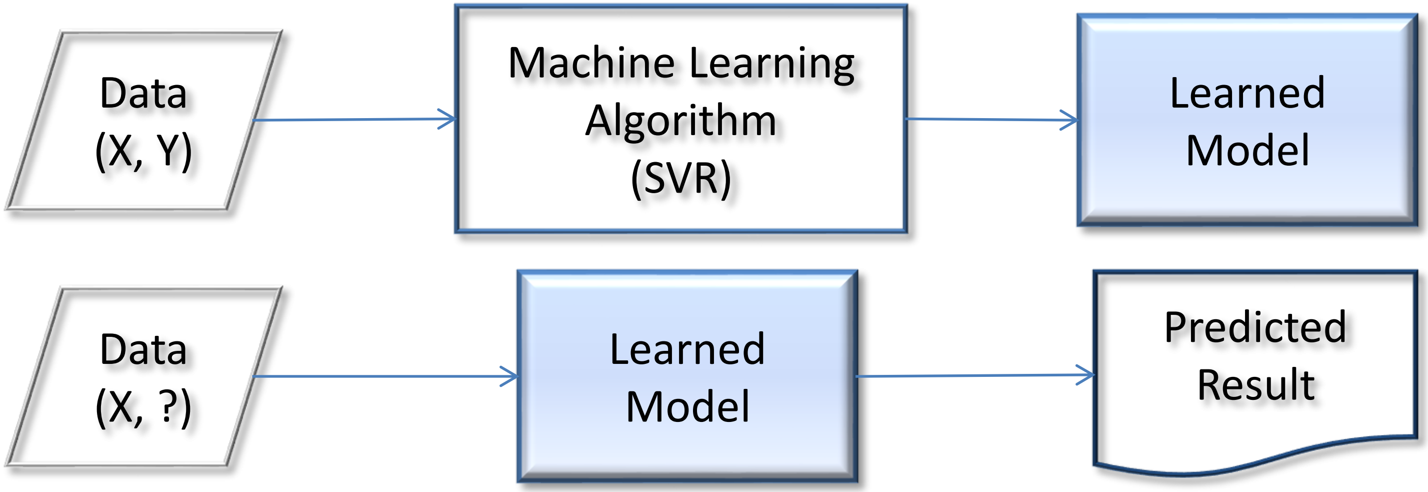
2. Read Length vs. Coverage

lin

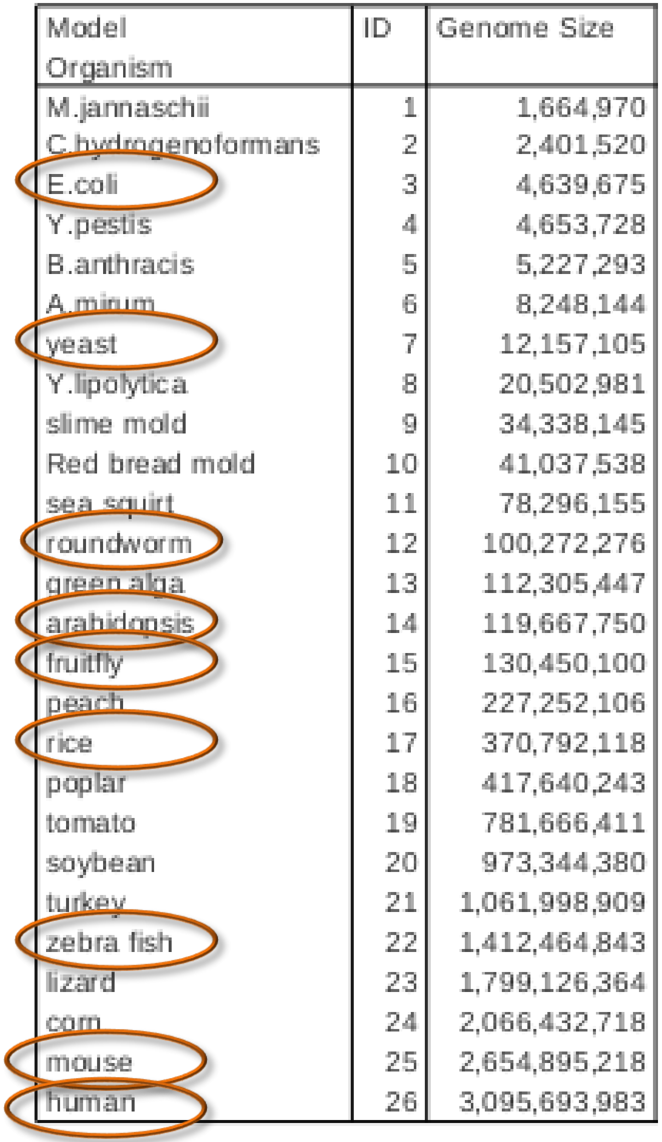
Technology vs. Money

Empirical Data-driven Approach

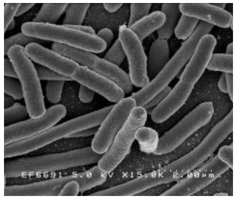
- **We selected 26 species across tree of life and exhaustively analyzed their assemblies using simulated reads for 4 different length (6 for HG19) and 4 different coverage per species**
- **For the extra long reads, we fixed the Celera Assembler(CA) to support reads up to 0.5Mbp**

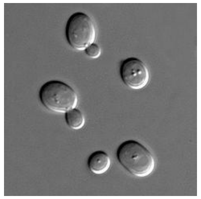


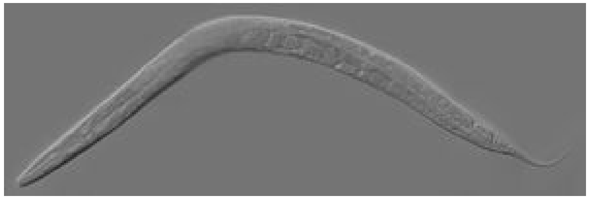
26 Species Across Tree of Life



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.minum	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







HG19 Genome Assembly Performance

by

Lengths selected to represent idealized biotechnologies:

mean32: ~Optical mapping

mean16: ~10x / Chromatin

mean8: ~10x / Chromatin

mean4: PacBio/ONT

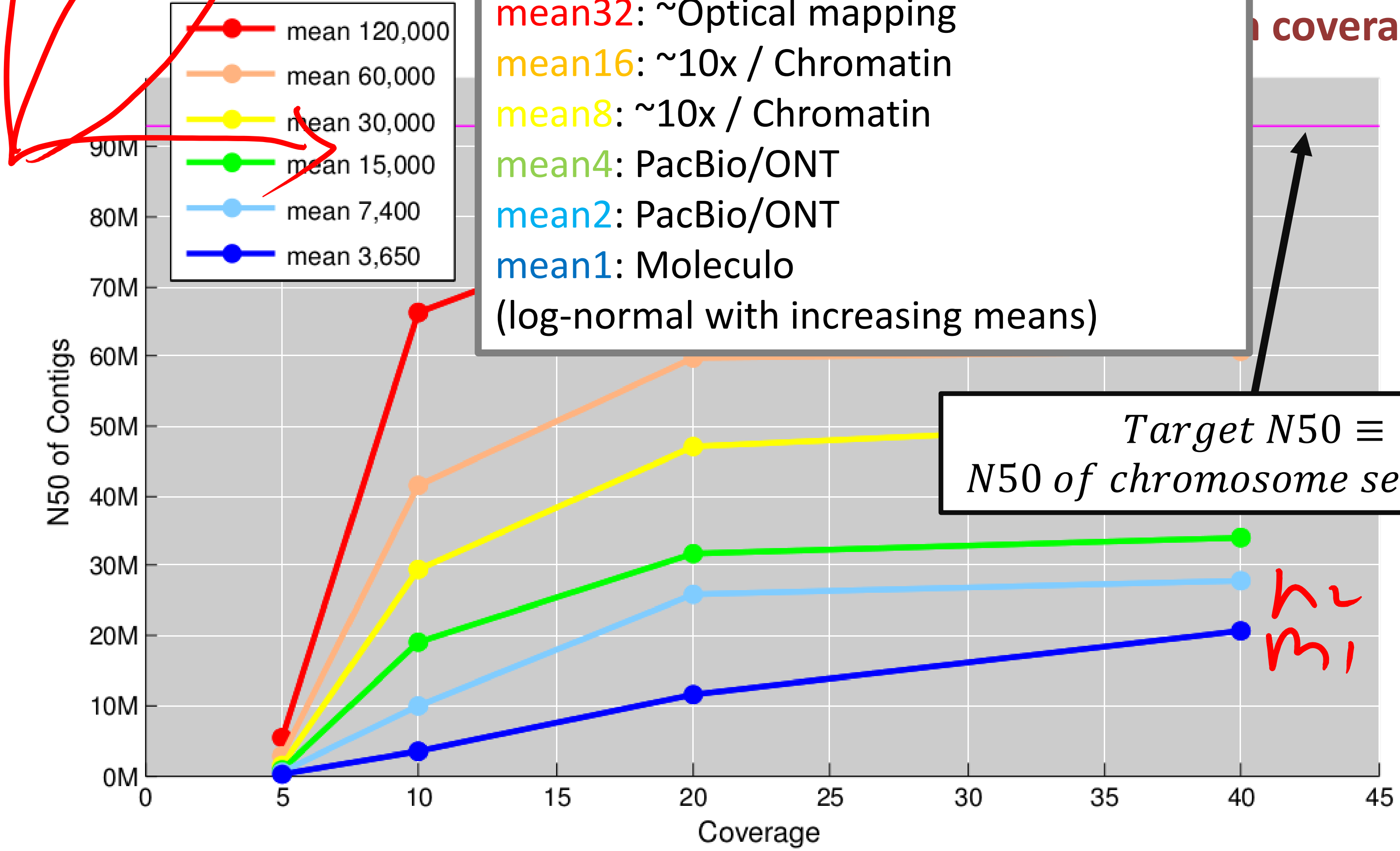
mean2: PacBio/ONT

mean1: Moleculo

(log-normal with increasing means)

n has stronger

n coverage



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

Why?

Lander-Waterman Statistics

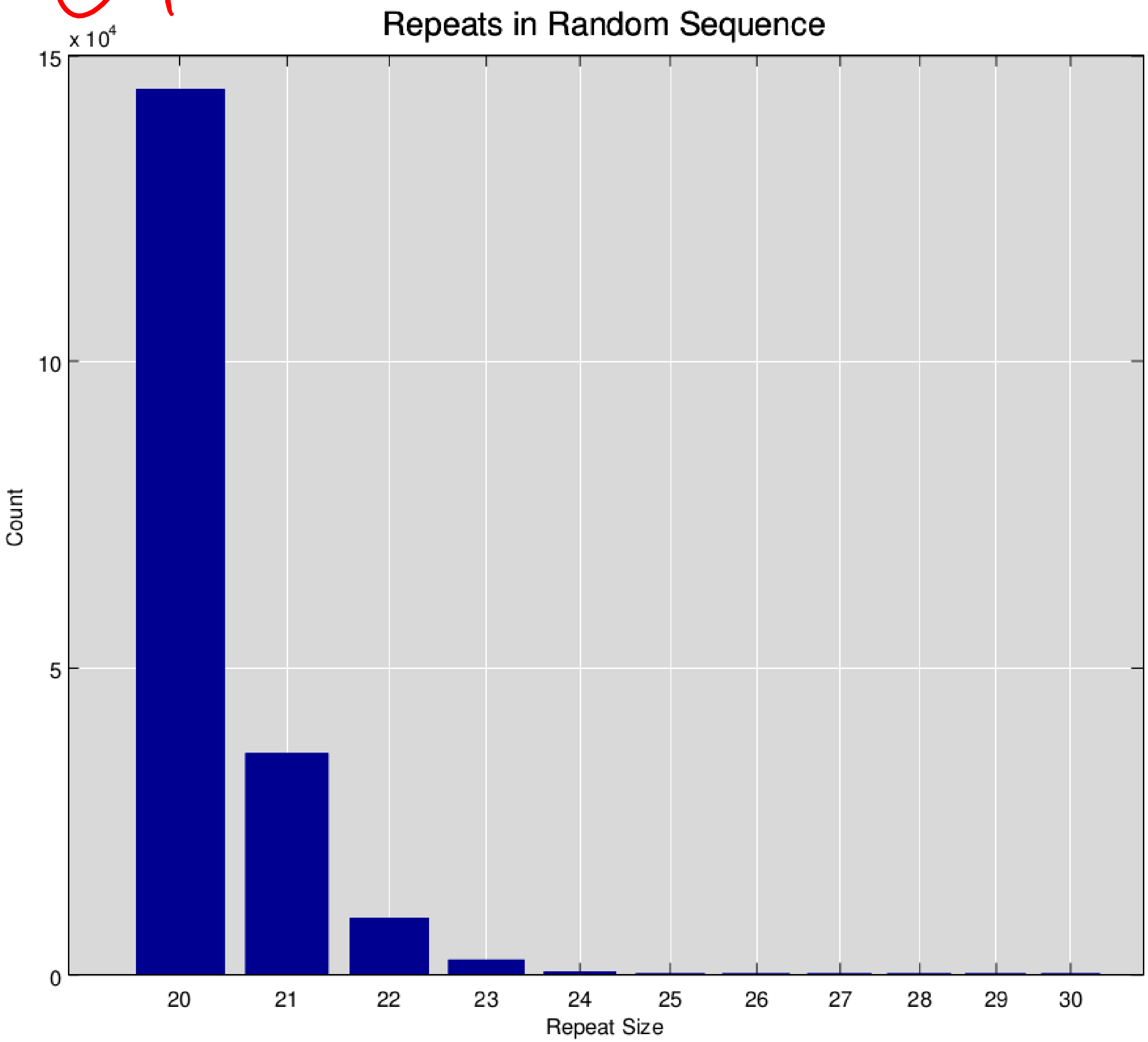
- **Assumptions!!!**
- **If genome is a random sequence, it will work**
- **It works only in low coverage 3-5x**
- **It works for small genomes (< yeast)**

Our Approach

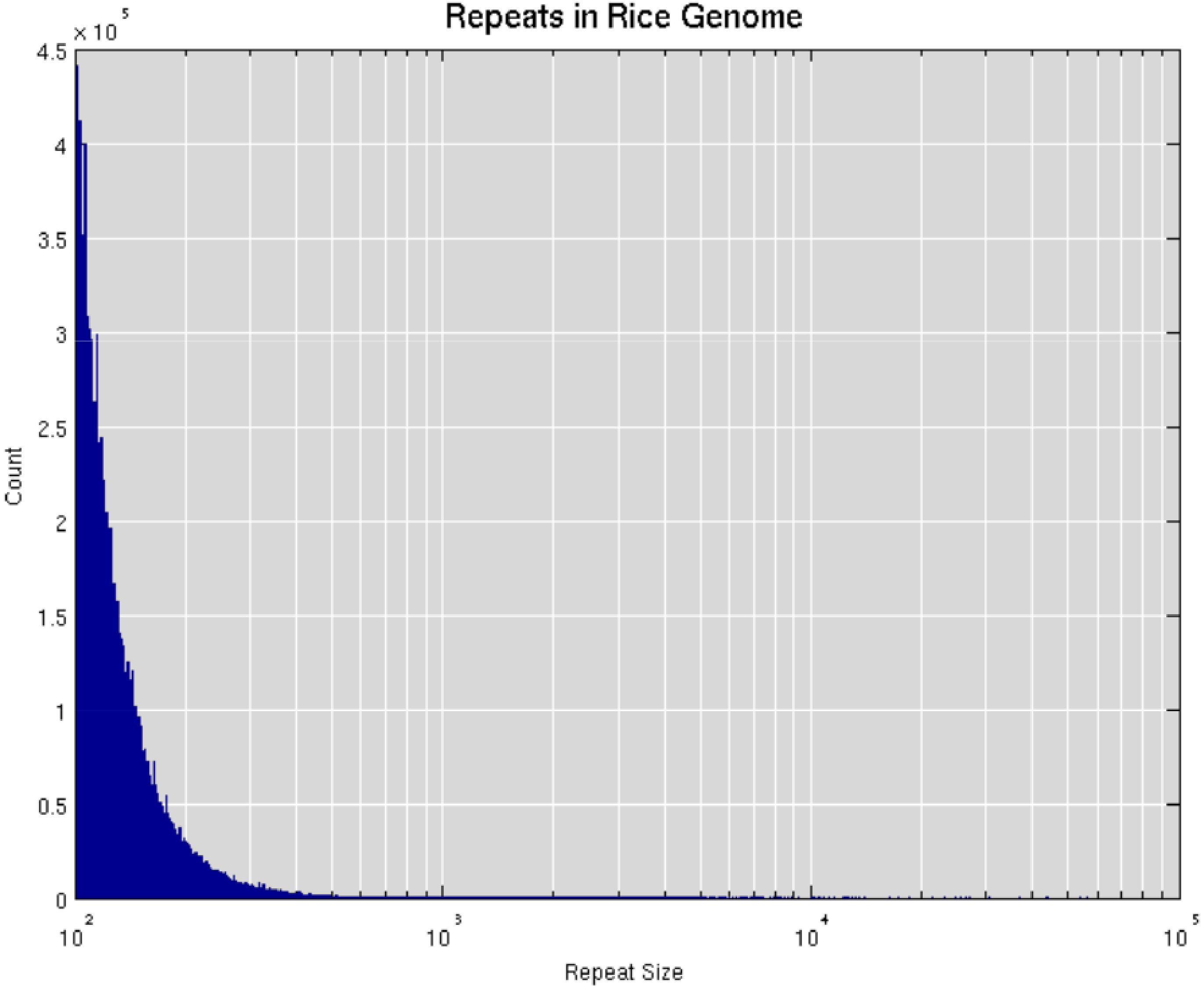
- Stop assuming what we cannot guarantee!!!
- We tried to assume as little as possible.
- Instead of building on top of assumptions, we let the model learn from the data
- Empirical data-driven approach

400Ms

Repeats



Repeats in Rice



Our Goal

- To **predict** genome assembly **contiguity**

$$Performance(\%) \equiv \frac{N50\ from\ assembly}{N50\ of\ chromosome\ segments} \times 100$$

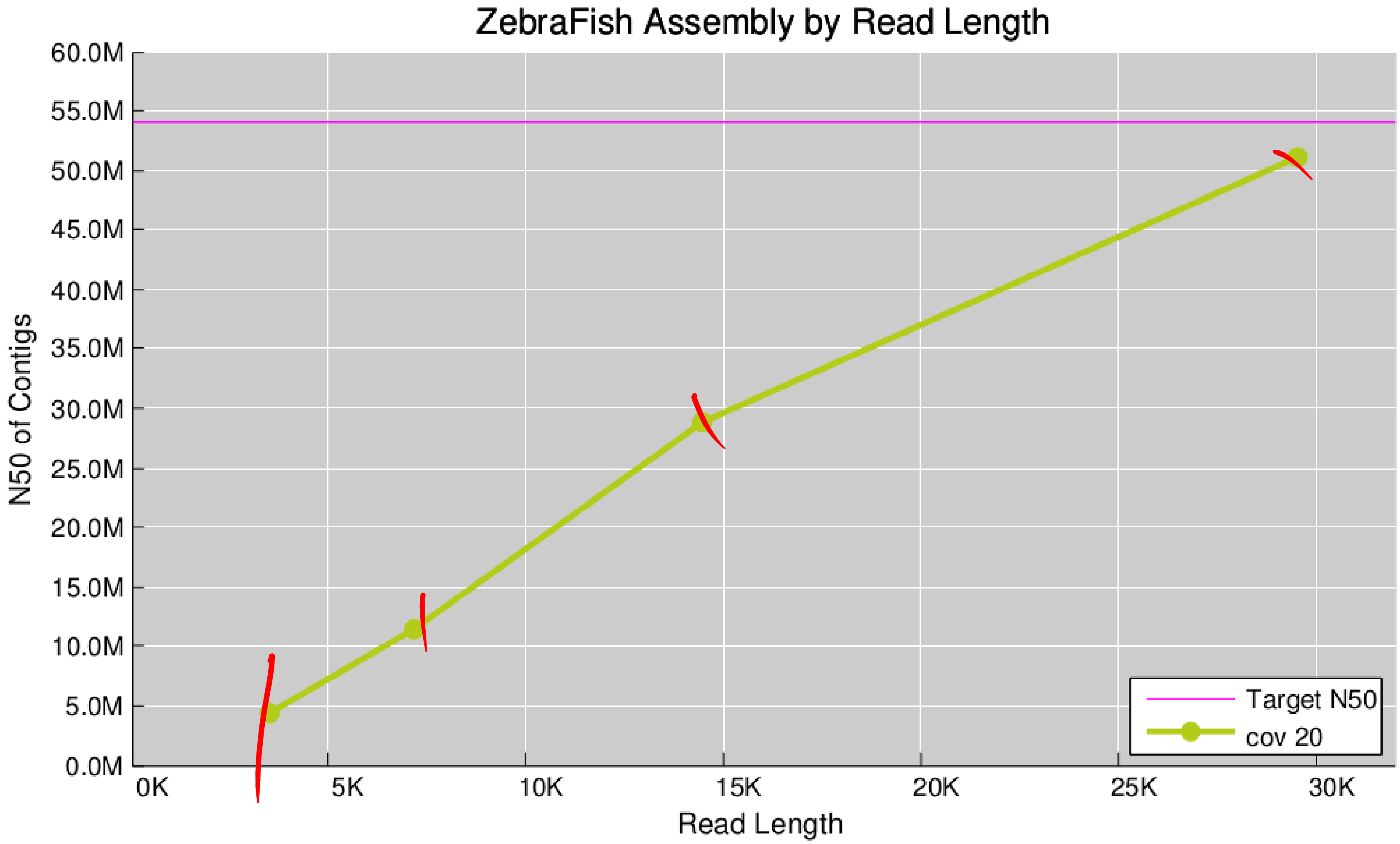
$$\approx f\begin{pmatrix} Read\ Length \\ Coverage \\ Repeats \\ Genome\ Size \end{pmatrix}$$

Assembly Challenge (1)
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%

- Assembly Challenge (1) -

Read Length



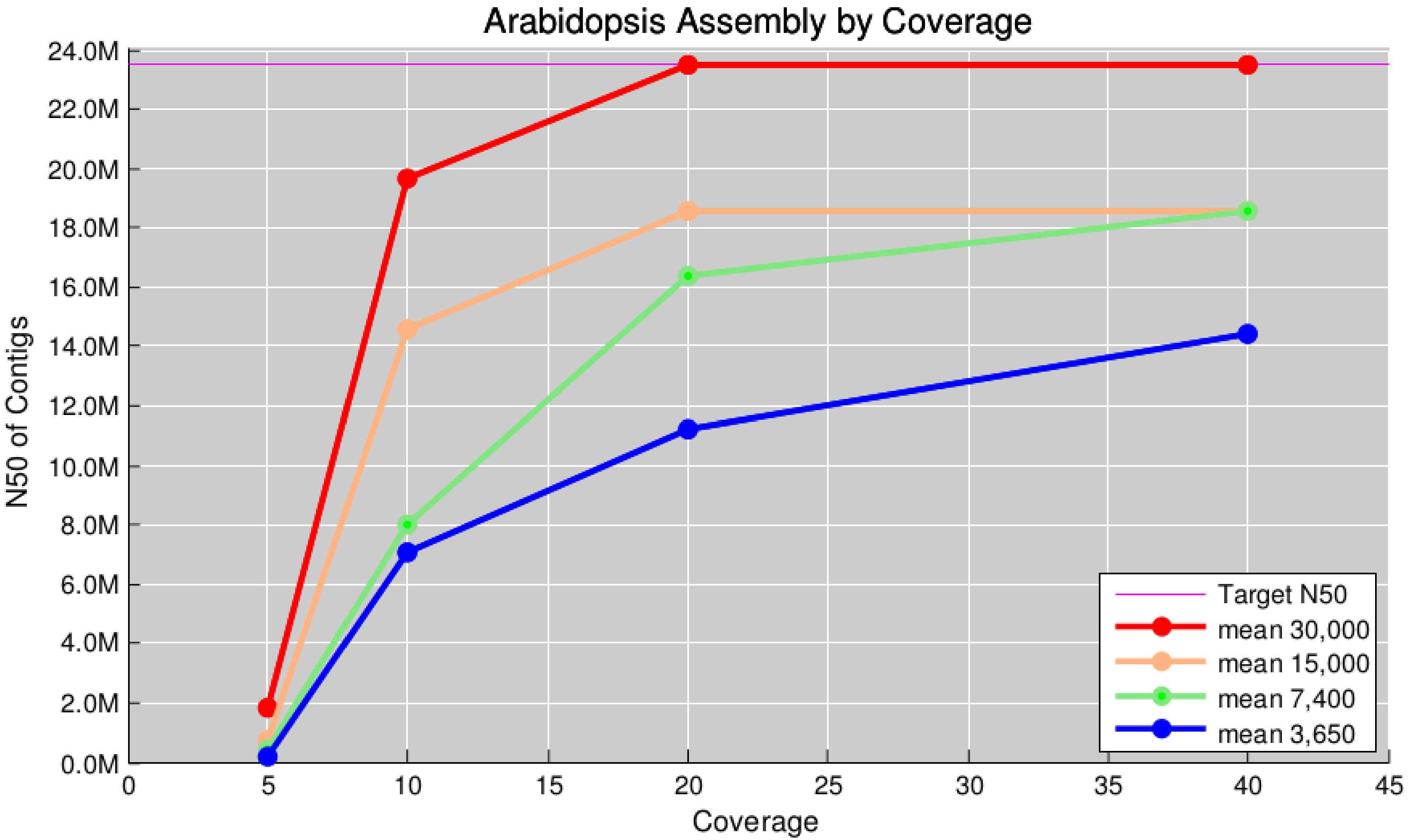
Assembly Challenge (2)

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)

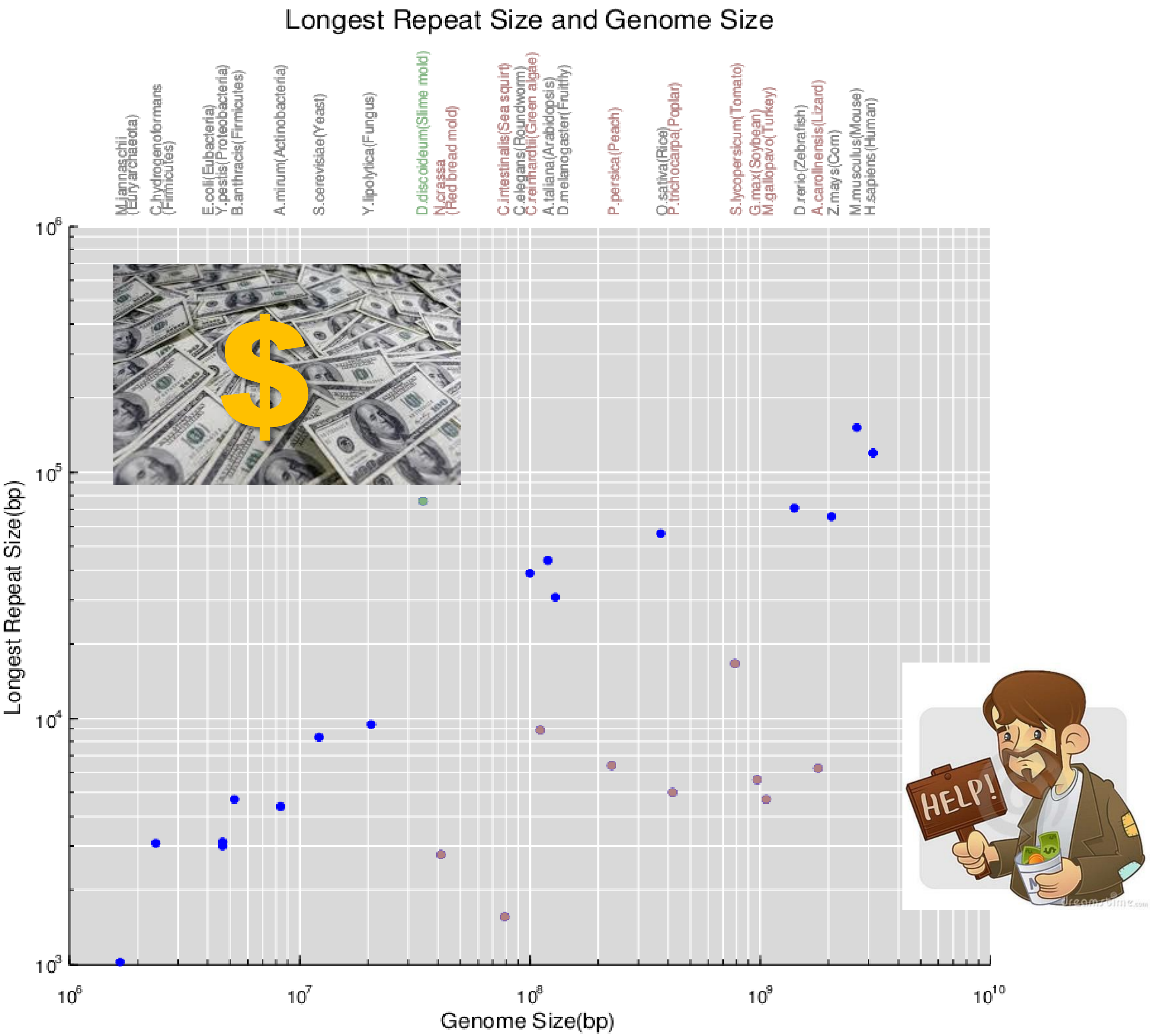
Assembly Challenge (2)

Coverage

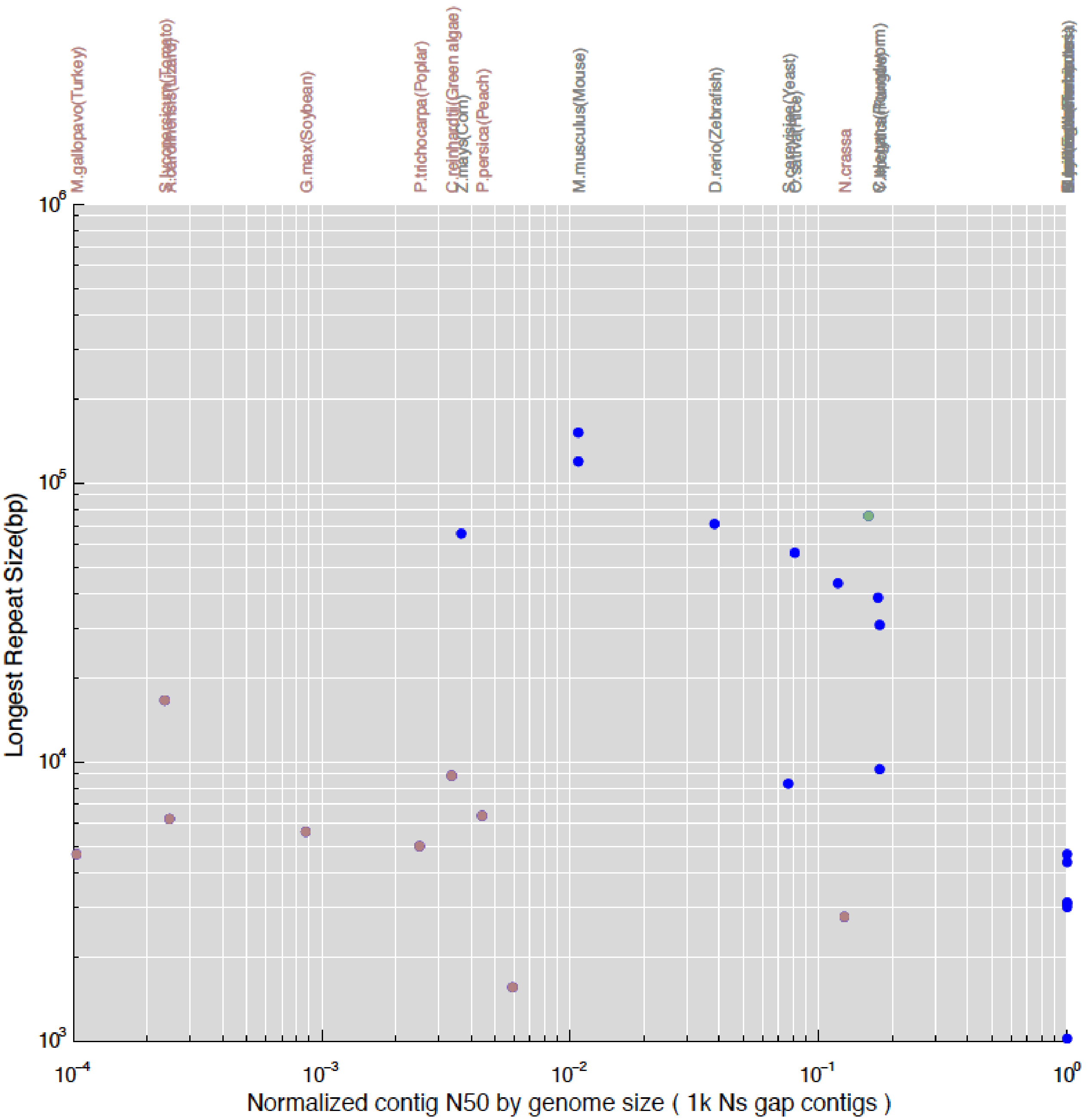


Assembly Challenge (3)
Repeats

- **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


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)
Longest Repeat Size(bp)
10^6
10^5
10^4
10^3
Genome Size(bp)
10^6
10^7
10^8
10^9
10^{10}
$
HELP!

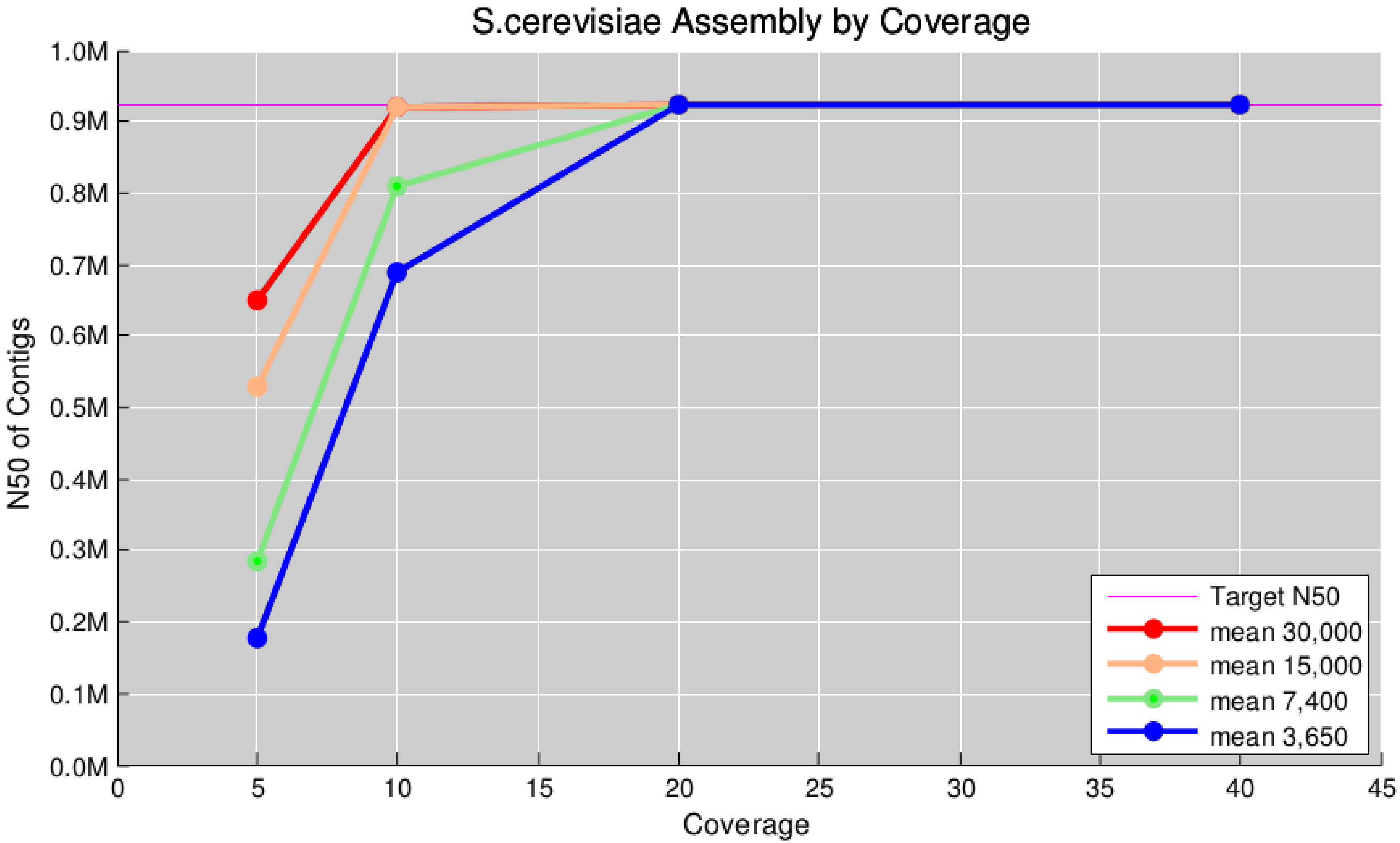
Longest Repeat Size and Normalized reference contig N50



Assembly Challenge (4)
Genome Size

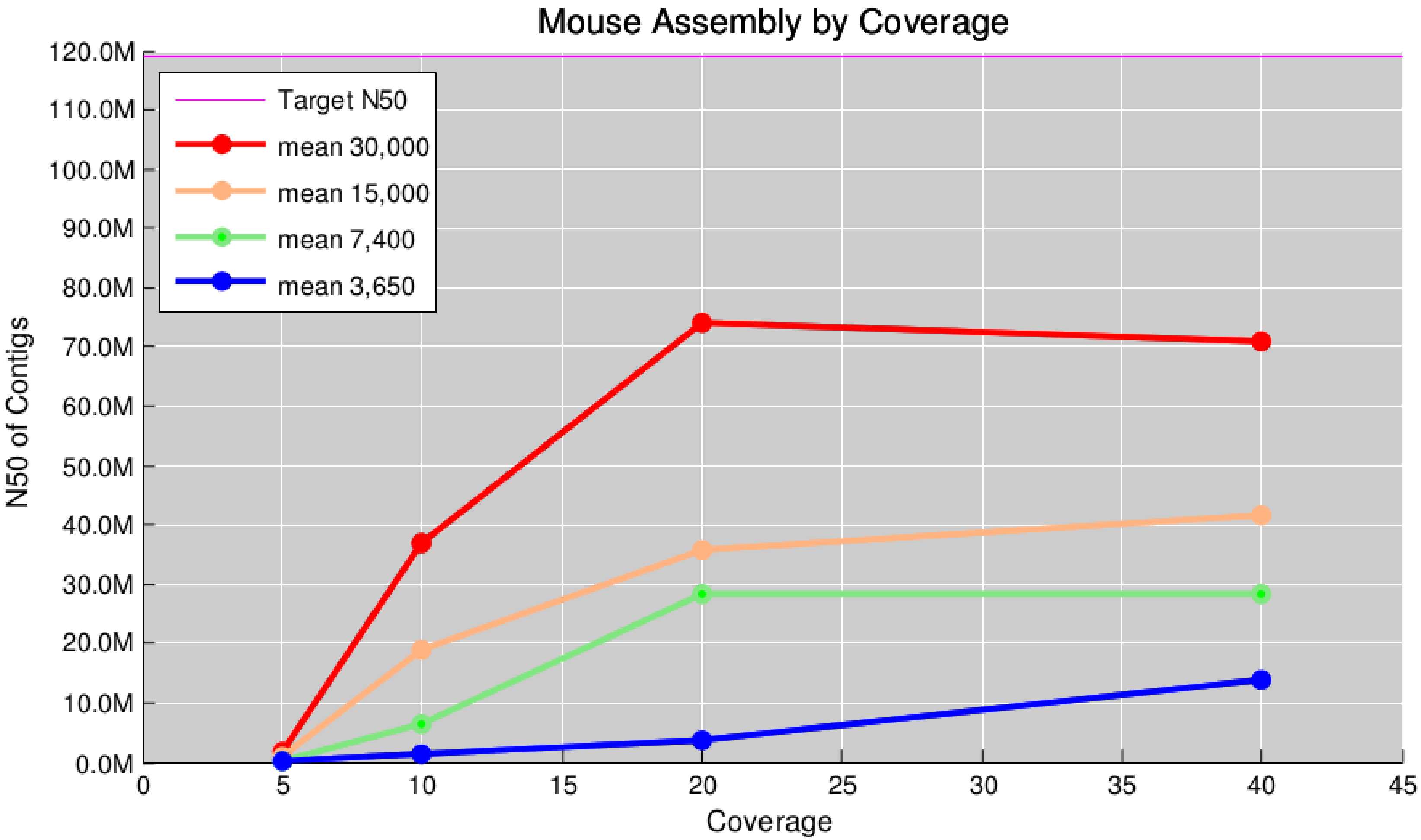
- **Increase the assembly complexity**
- **Make a hard problem harder.**

Assembly Challenge (4)
Genome Size

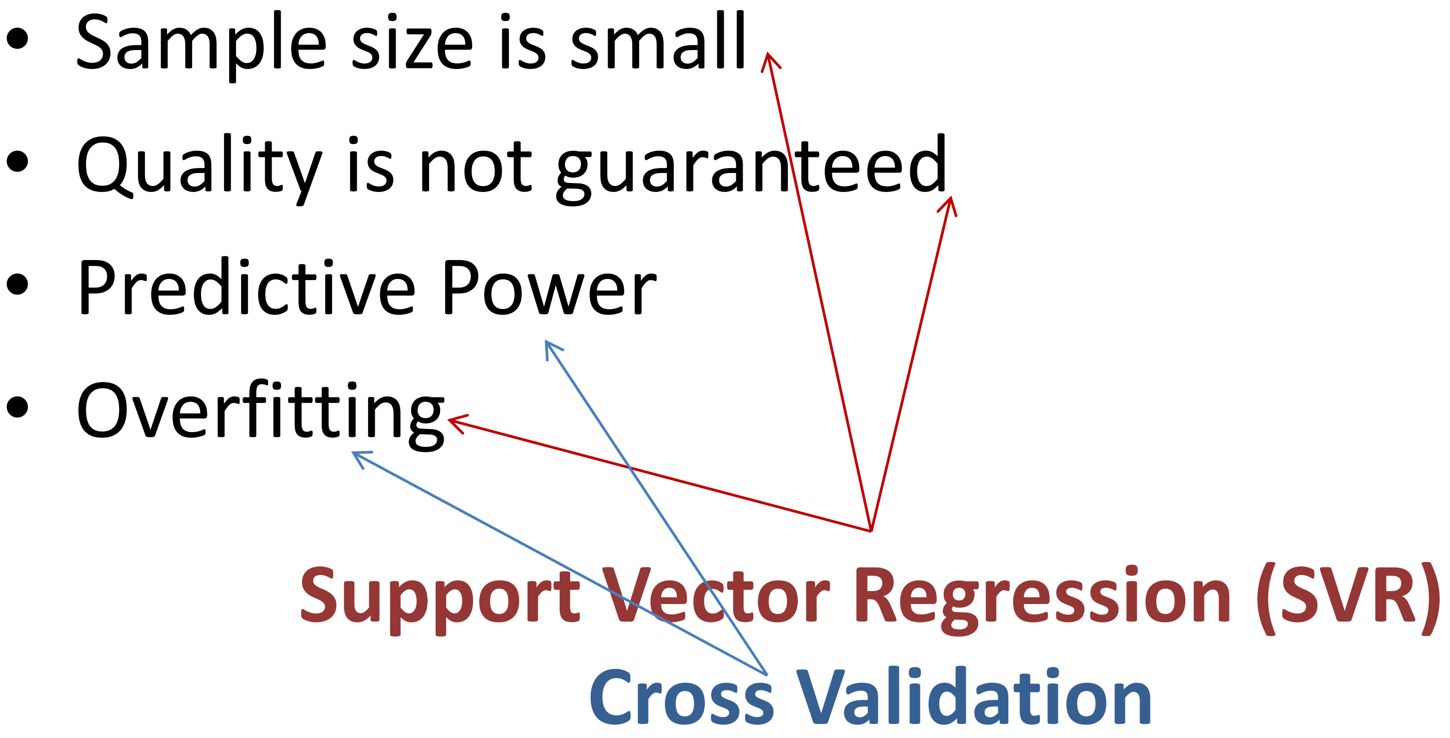


Assembly Challenge (4)

Genome Size



Challenges for Prediction



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

Support Vector Regression (SVR)

Cross Validation

Feature Engineering (1)

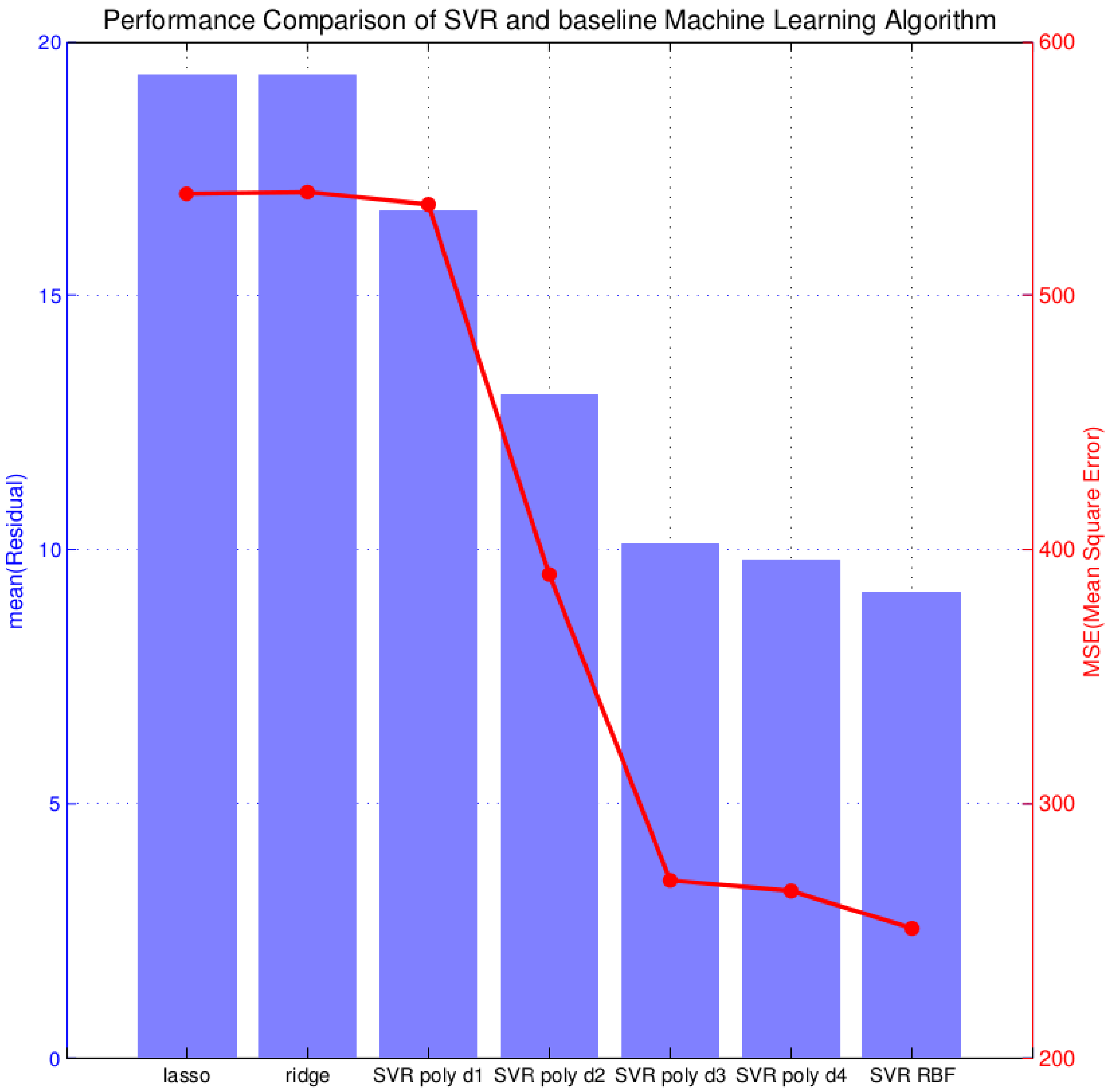
- **Correlation Coefficient**
  - Performance vs. genome size
    - R = -0.38
  - Performance vs. Read Length
    - R = 0.2

Feature Engineering (2)

- **Correlation Coefficient**
  - Performance and *log* (genome size)
    - R = -0.49
  - Performance and *log* (read length)
    - R = 0.32
  - 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

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**



The resurgence of reference genome qaultiy

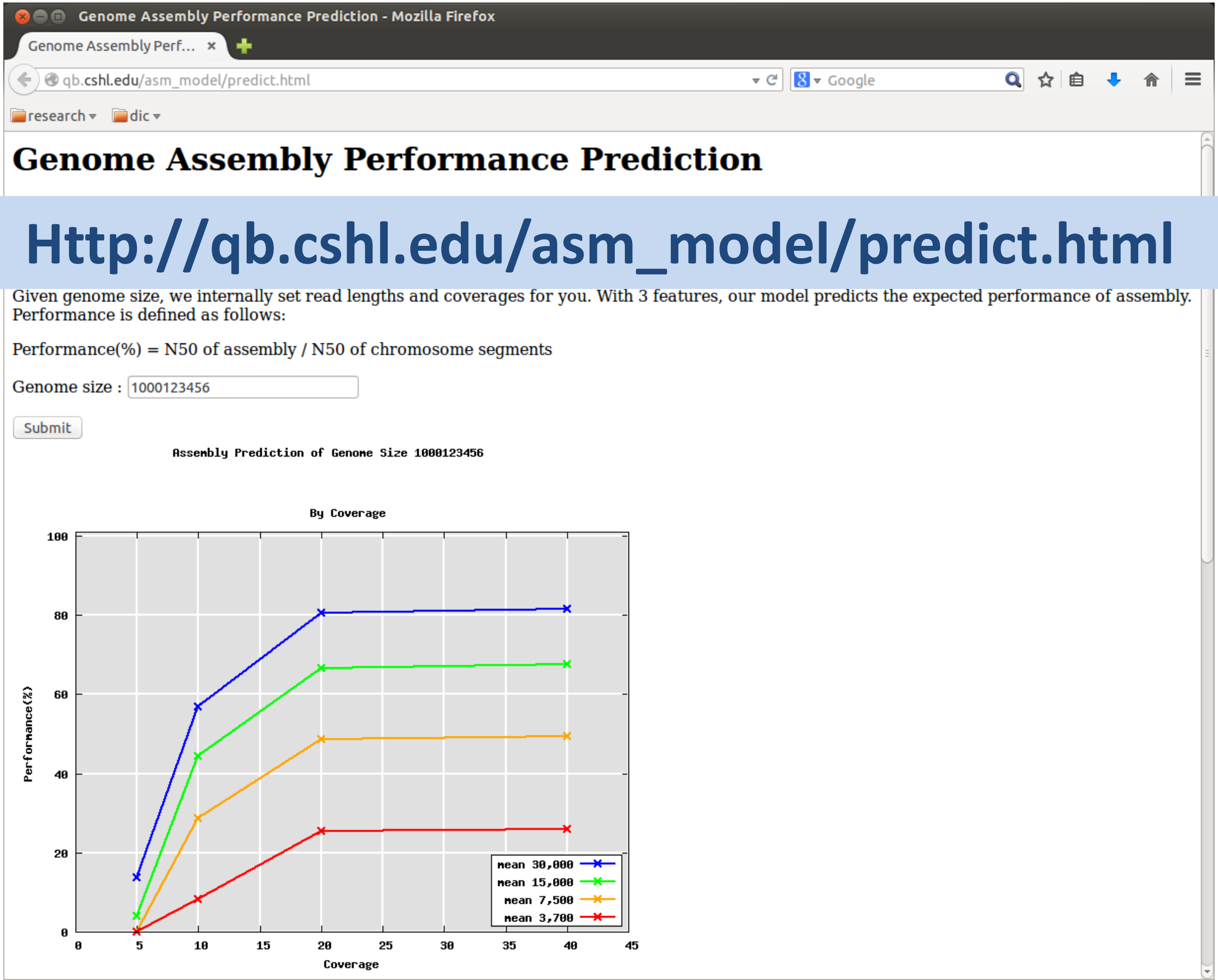
Lee, H, Gurtowski, J, Yoo, S, Nattestad, M, Marcus, S, Goodwin, S, McCombie, WR, Schatz MC *et al.* (2015) *Under review*

Predictive Power

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

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

Web Service for Contiguity Prediction



Genome Assembly Performance Prediction - Mozilla Firefox

Genome Assembly Perf... ×

qb.cshl.edu/asm_model/predict.html

Google

research ▾ dic ▾

Genome Assembly Performance Prediction

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

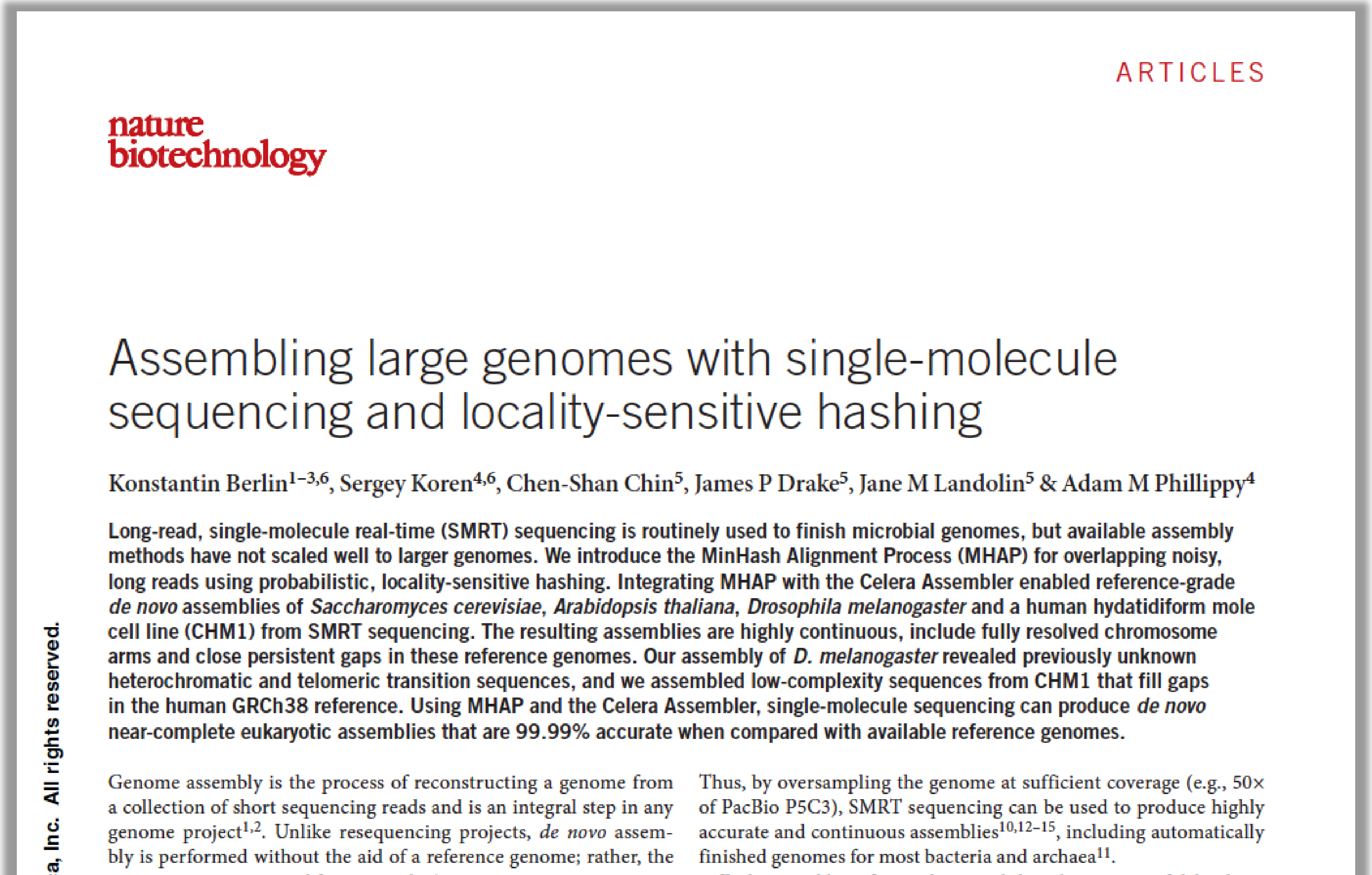
Given genome size, we internally set read lengths and coverages for you. With 3 features, our model predicts the expected performance of assembly. Performance is defined as follows:

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

Genome size : 1000123456

Submit

Validated by MHAP



ARTICLES

nature biotechnology

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing

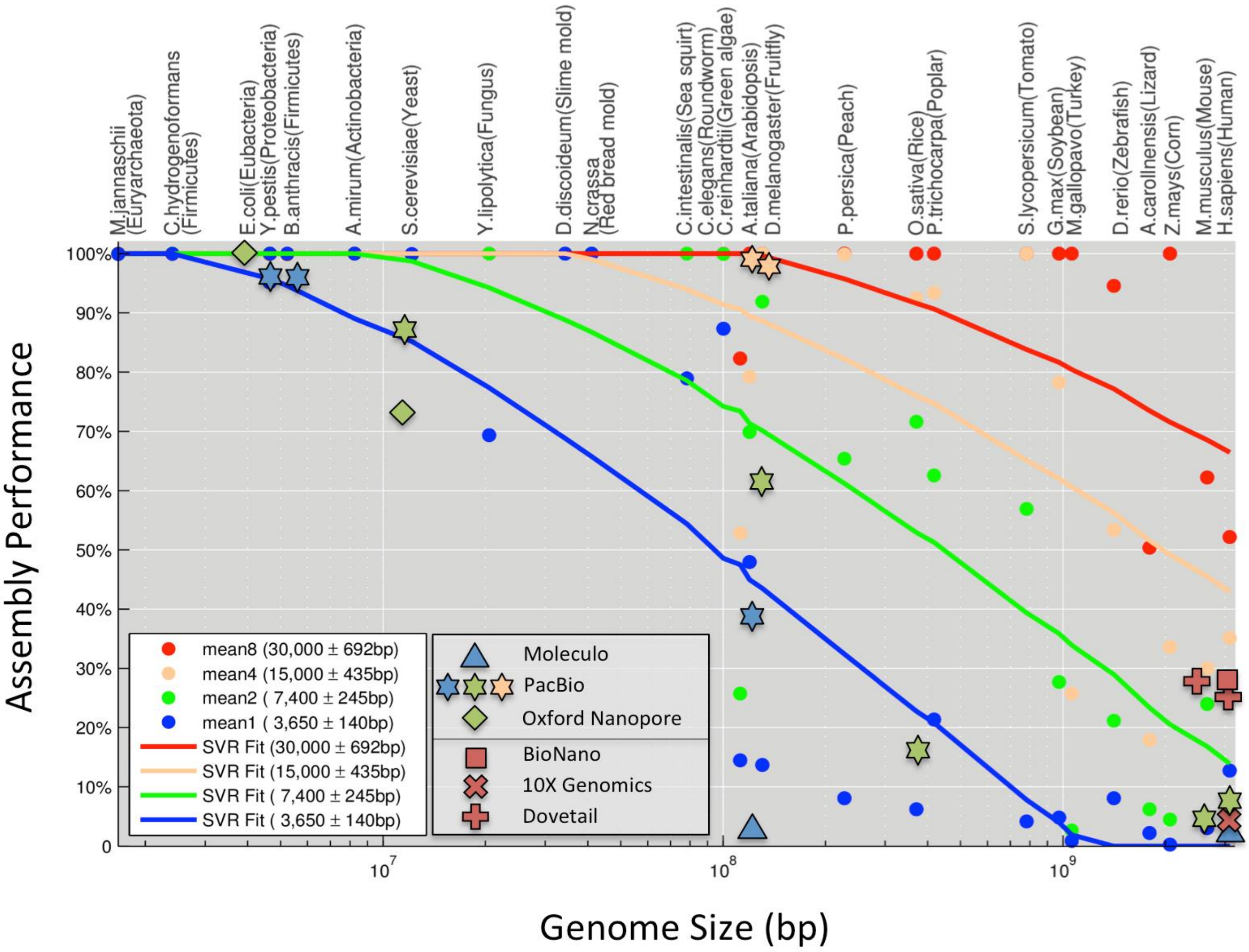
Konstantin Berlin[1–3,6], Sergey Koren[4,6], Chen-Shan Chin[5], James P Drake[5], Jane M Landolin[5] & Adam M Phillippy[4]

Long-read, single-molecule real-time (SMRT) sequencing is routinely used to finish microbial genomes, but available assembly methods have not scaled well to larger genomes. We introduce the MinHash Alignment Process (MHAP) for overlapping noisy, long reads using probabilistic, locality-sensitive hashing. Integrating MHAP with the Celera Assembler enabled reference-grade *de novo* assemblies of *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster* and a human hydatidiform mole cell line (CHM1) from SMRT sequencing. The resulting assemblies are highly continuous, include fully resolved chromosome arms and close persistent gaps in these reference genomes. Our assembly of *D. melanogaster* revealed previously unknown heterochromatic and telomeric transition sequences, and we assembled low-complexity sequences from CHM1 that fill gaps in the human GRCh38 reference. Using MHAP and the Celera Assembler, single-molecule sequencing can produce *de novo* near-complete eukaryotic assemblies that are 99.99% accurate when compared with available reference genomes.

Genome assembly is the process of reconstructing a genome from a collection of short sequencing reads and is an integral step in any genome project[1,2]. Unlike resequencing projects, *de novo* assembly is performed without the aid of a reference genome; rather, the

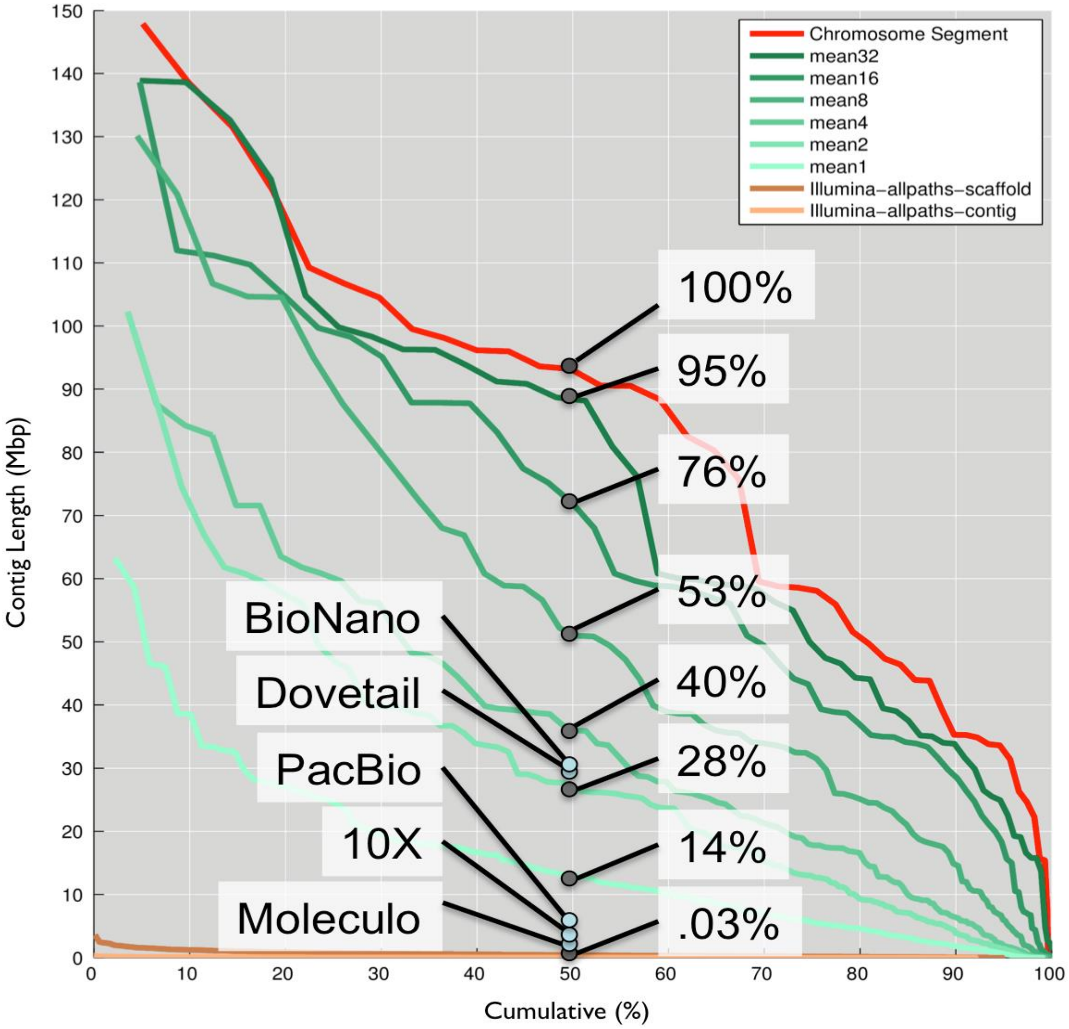
Thus, by oversampling the genome at sufficient coverage (e.g., 50× of PacBio P5C3), SMRT sequencing can be used to produce highly accurate and continuous assemblies[10,12–15], including automatically finished genomes for most bacteria and archaea[11].

Reference Genome Quality



Contiguity

de novo human genome assembly



What happens as we sequence the human genome with longer reads?

- Red: Sizes of the chromosome arms of HG19 from largest to shortest
- Green: Results of our assemblies using progressively longer and longer reads
- Orange: Results of Allpaths/Illumina assemblies

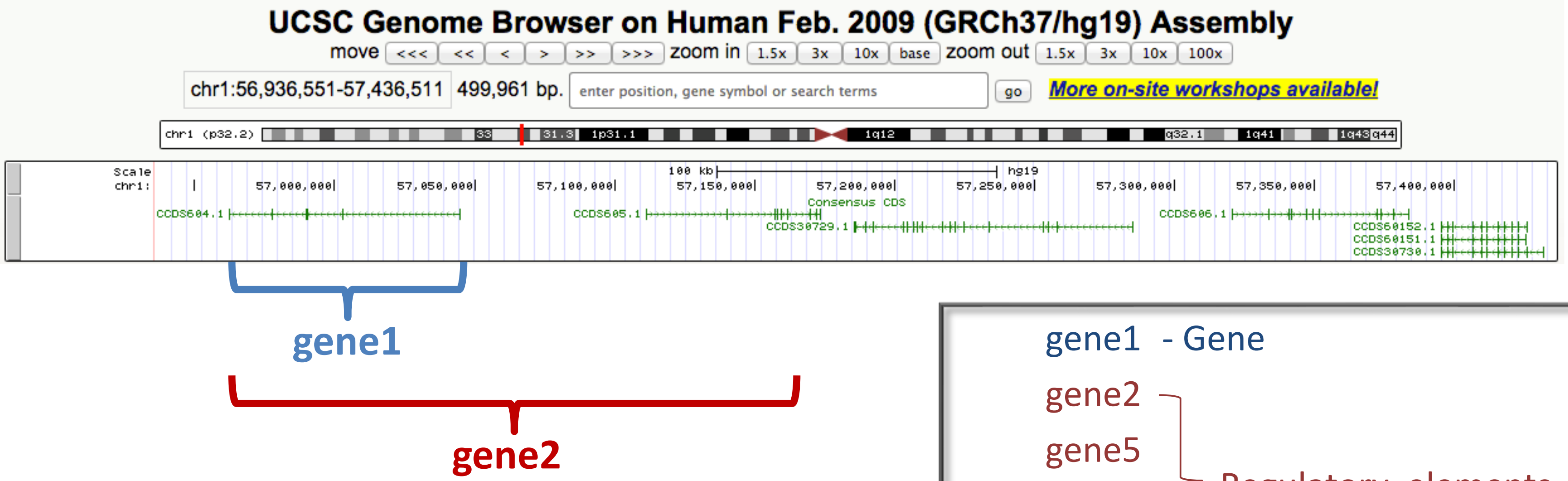
Lengths selected to represent the biotechnologies:

- mean1: ~Moleculo
- mean2: ~PacBio/ONT
- mean16: ~10x / Chromatin
- mean32: ~Optical mapping

(log-normal with increasing means)

Completeness

Human Reference Genome Quality by gene block analysis

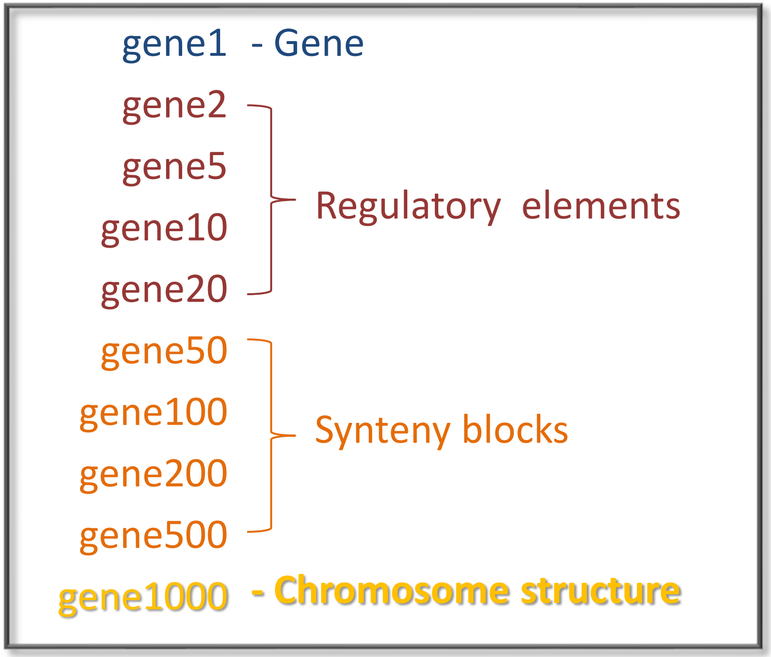


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr1:56,936,551-57,436,511 499,961 bp.

enter position, gene symbol or search terms

More on-site workshops available!



gene1 - Gene

gene2

gene5

gene10

gene20

Regulatory elements

gene50

gene100

gene200

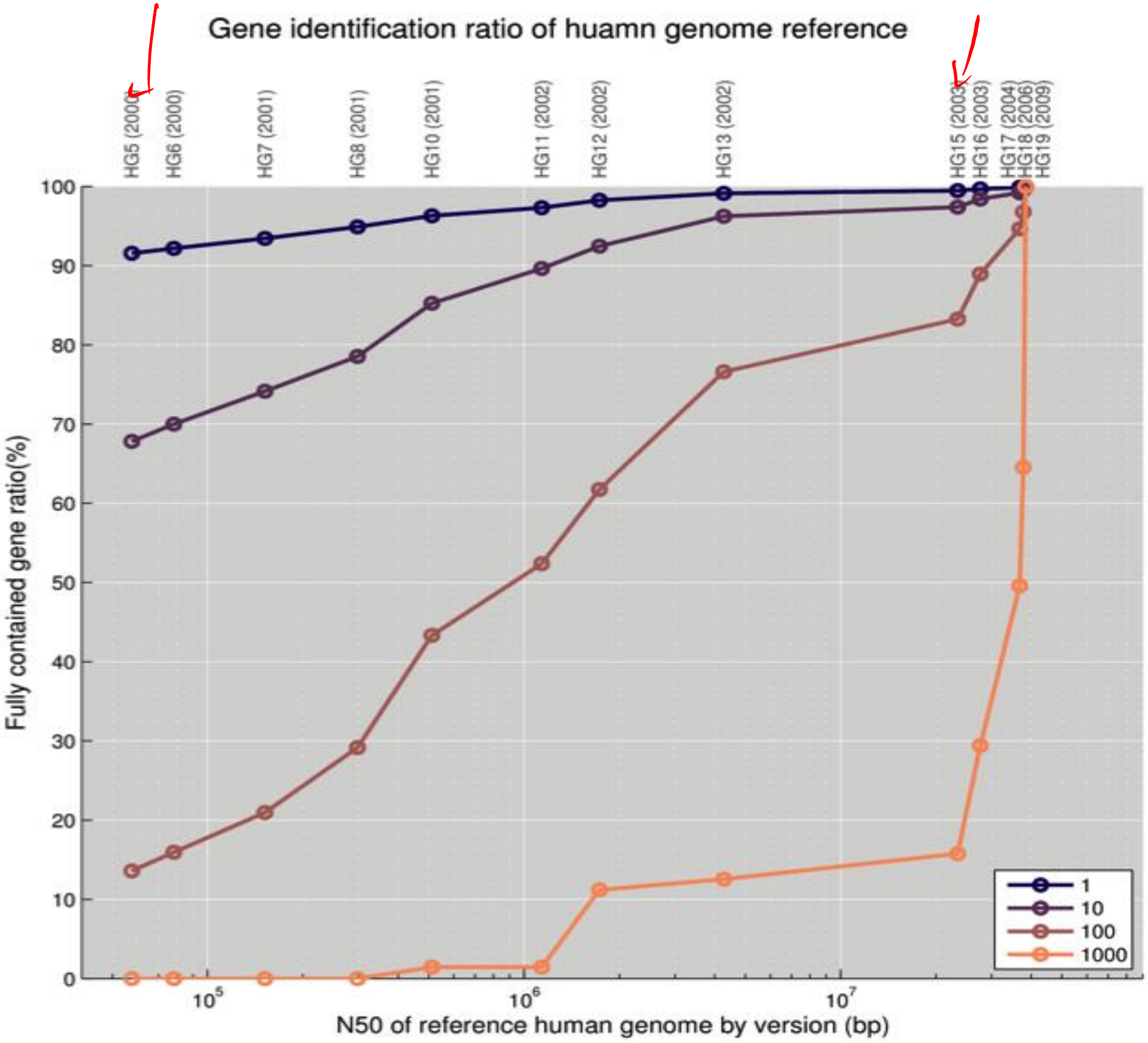
gene500

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

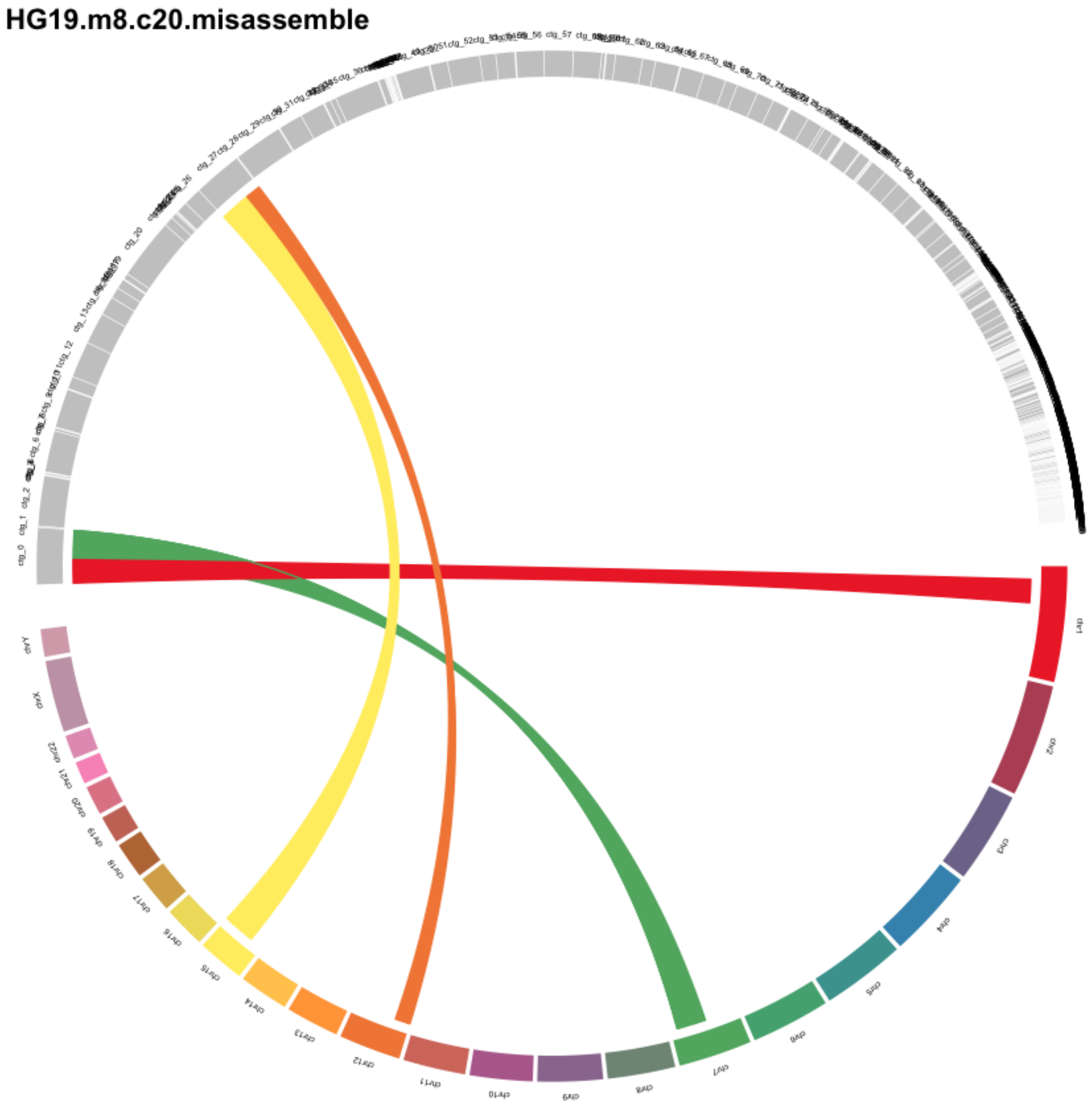
Correctness Summary in HG19

N50 misleading

HG19	(major) misassembly	(major) breaks
	False Positive	False Negative
	Increase N50 (falsely lengthen contiguity)	Decrease N50 (shorten contiguity)
	Mislead us in biological meaning	Negatively impact on downstream research
Mean1	209	4069
Mean2	70	462
Mean4	49	296
Mean8	33	197
Mean16	9	42
Mean32	7	5

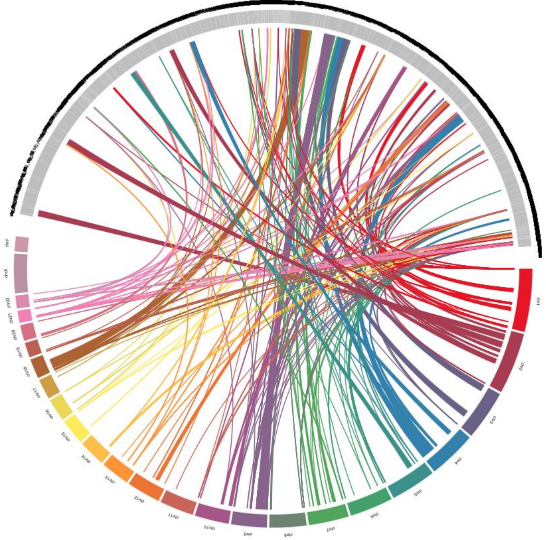
Misassembly

A critical error in de novo assembly

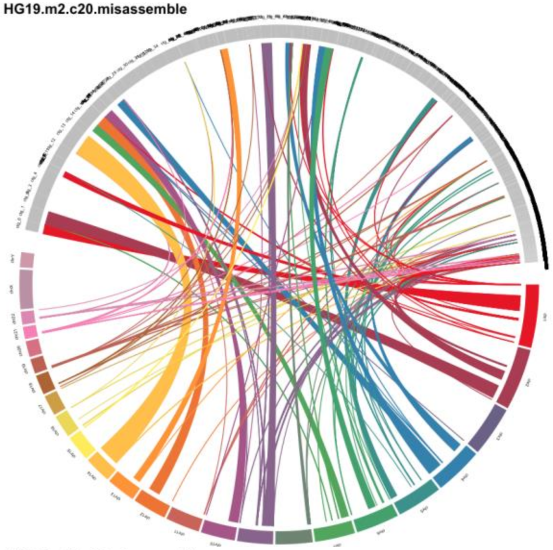


Misassembly Analysis in HG19

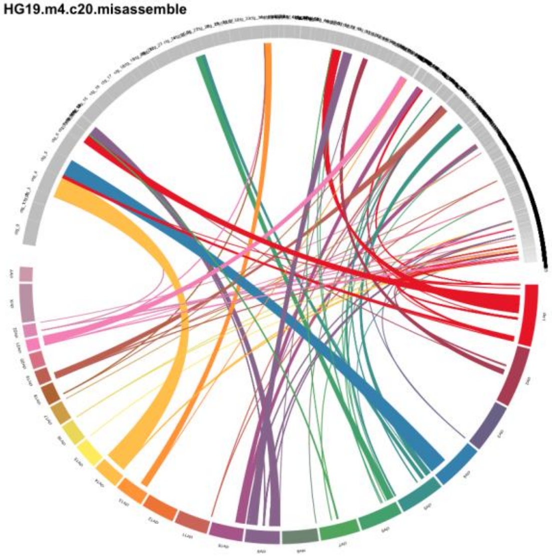
HG19.m1.c20.misassemble



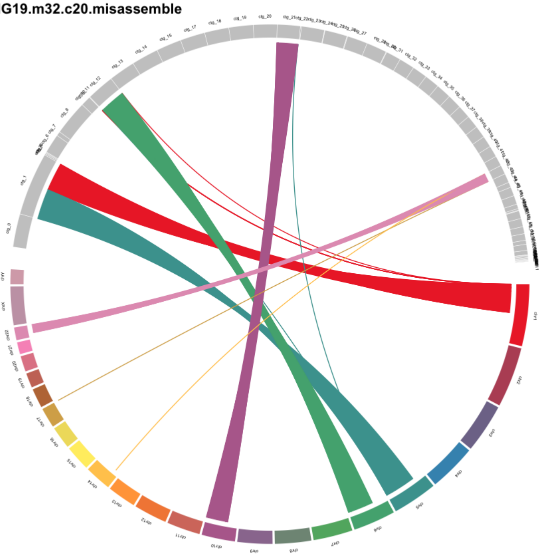
HG19.m2.c20.misassemble



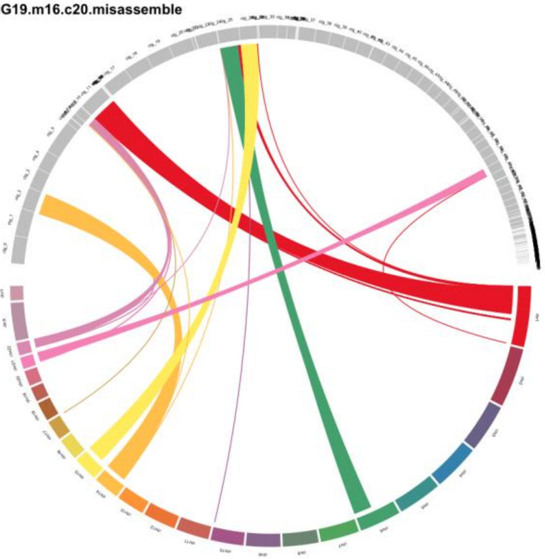
HG19.m4.c20.misassemble



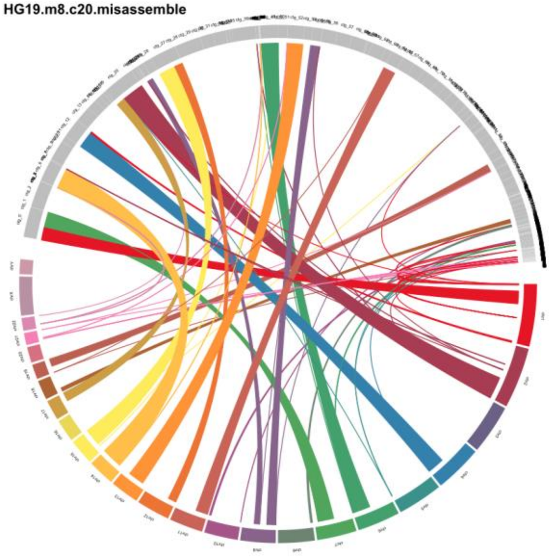
HG19.m32.c20.misassemble



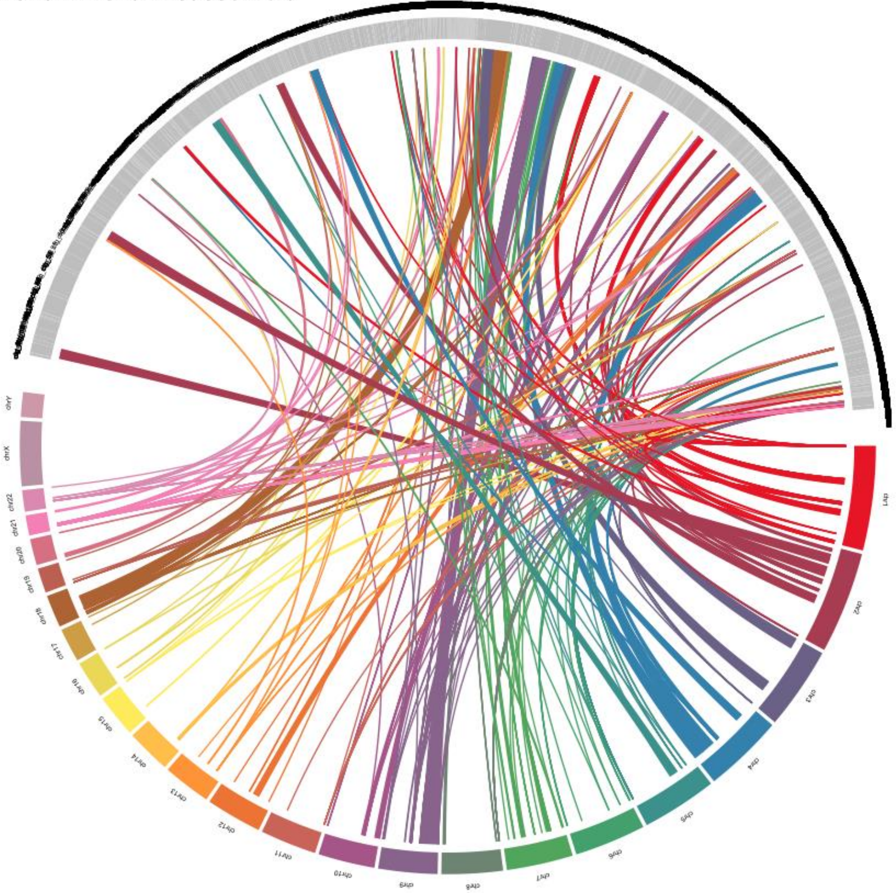
HG19.m16.c20.misassemble

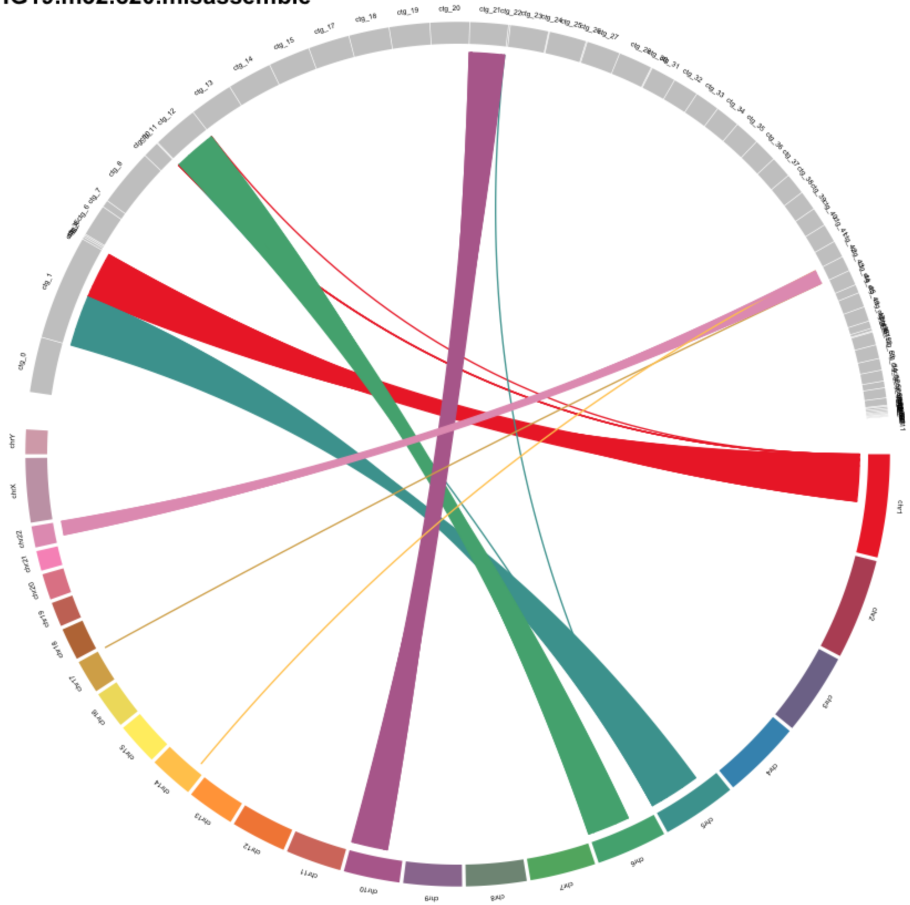


HG19.m8.c20.misassemble



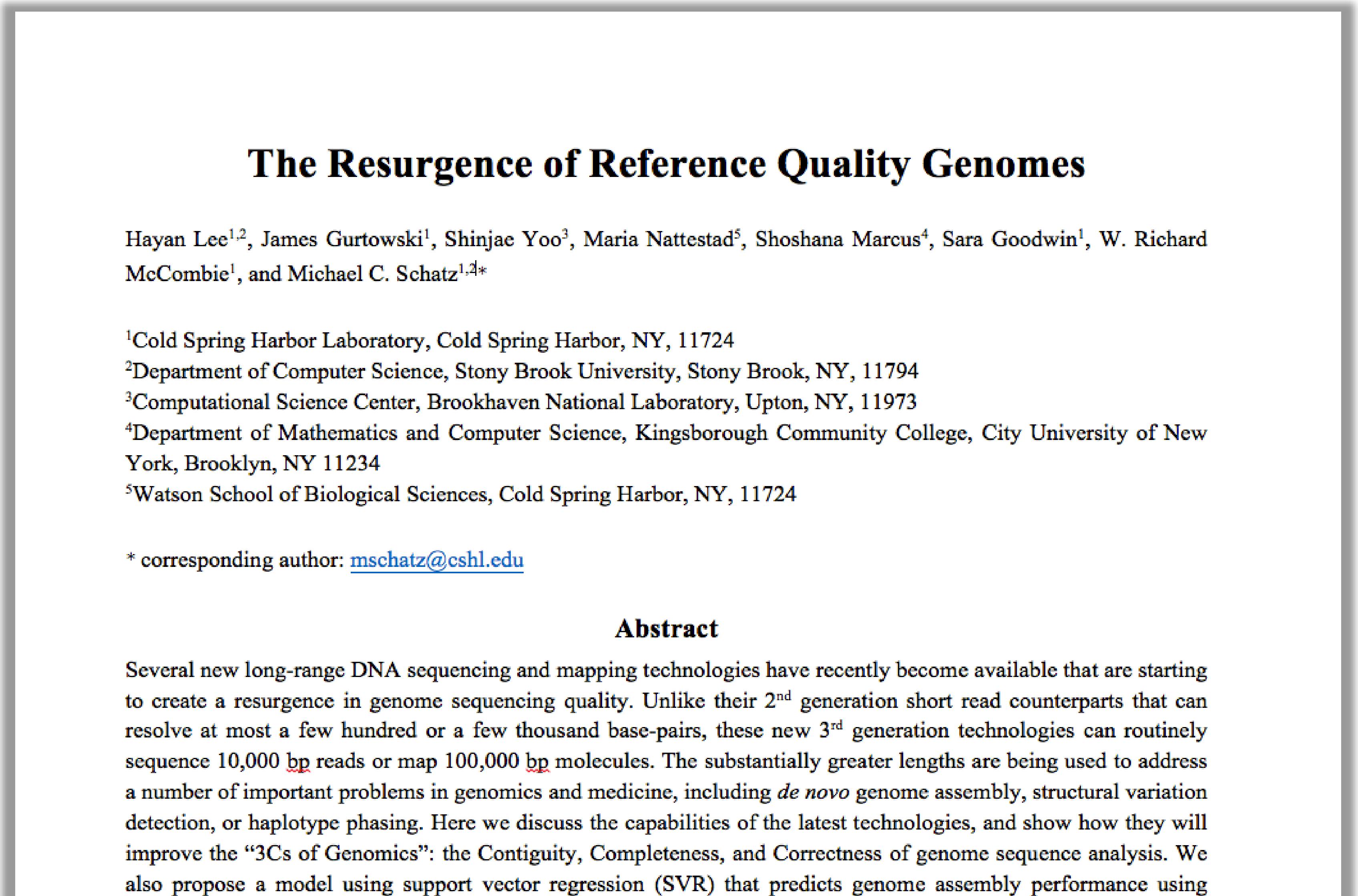
Misassembly Analysis in HG19


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

New Preprint



The Resurgence of Reference Quality Genomes

Hayan Lee[1,2], James Gurtowski[1], Shinjae Yoo[3], Maria Nattestad[5], Shoshana Marcus[4], Sara Goodwin[1], W. Richard McCombie[1], and Michael C. Schatz[1,2]*

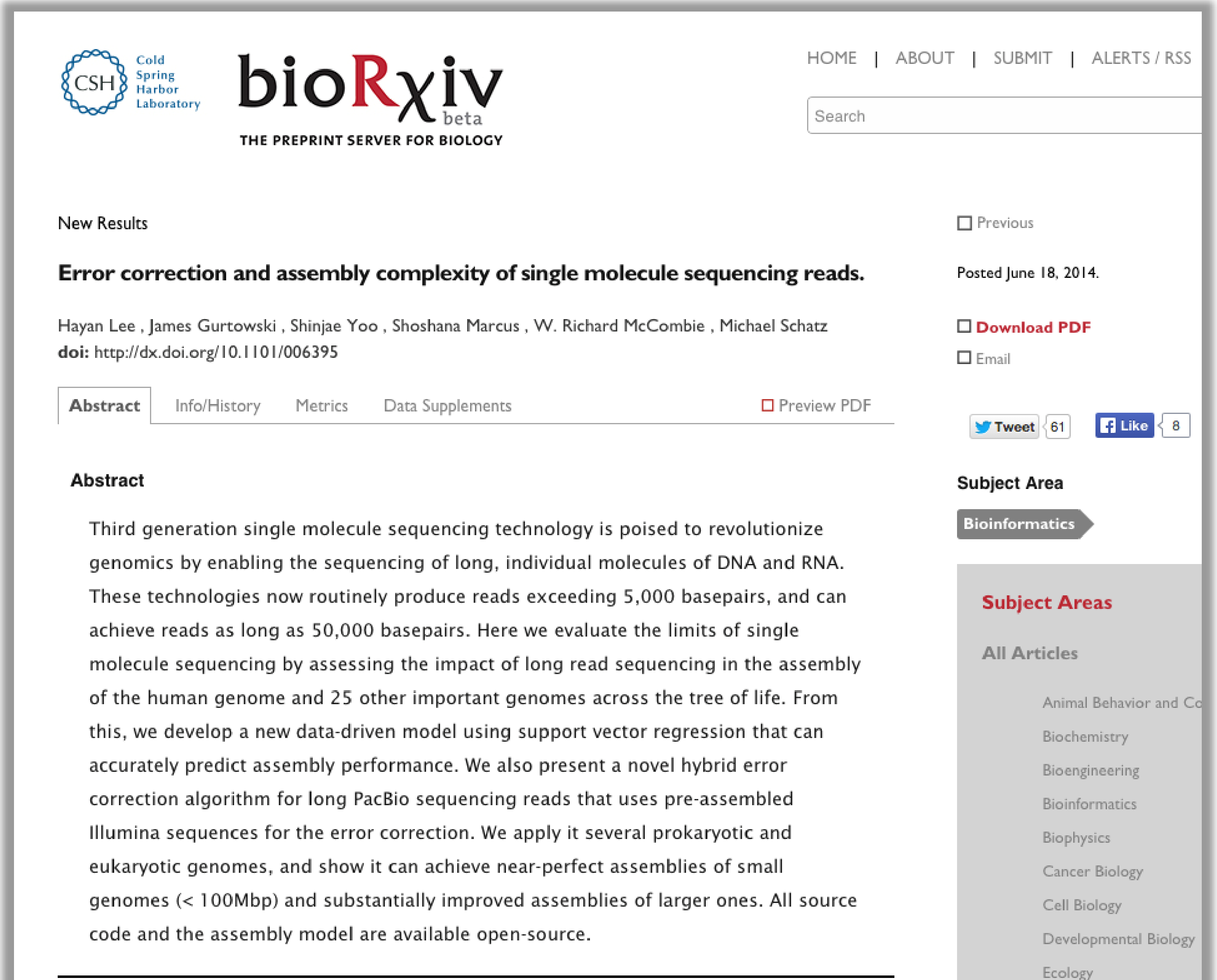
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724
[2]Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794
[3]Computational Science Center, Brookhaven National Laboratory, Upton, NY, 11973
[4]Department of Mathematics and Computer Science, Kingsborough Community College, City University of New York, Brooklyn, NY 11234
[5]Watson School of Biological Sciences, Cold Spring Harbor, NY, 11724

* corresponding author: mschatz@cshl.edu

Abstract

Several new long-range DNA sequencing and mapping technologies have recently become available that are starting to create a resurgence in genome sequencing quality. Unlike their 2nd generation short read counterparts that can resolve at most a few hundred or a few thousand base-pairs, these new 3rd generation technologies can routinely sequence 10,000 bp reads or map 100,000 bp molecules. The substantially greater lengths are being used to address a number of important problems in genomics and medicine, including *de novo* genome assembly, structural variation detection, or haplotype phasing. Here we discuss the capabilities of the latest technologies, and show how they will improve the "3Cs of Genomics": the Contiguity, Completeness, and Correctness of genome sequence analysis. We also propose a model using support vector regression (SVR) that predicts genome assembly performance using

Old Preprint



CSH Cold Spring Harbor Laboratory
bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | ALERTS / RSS
Search

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

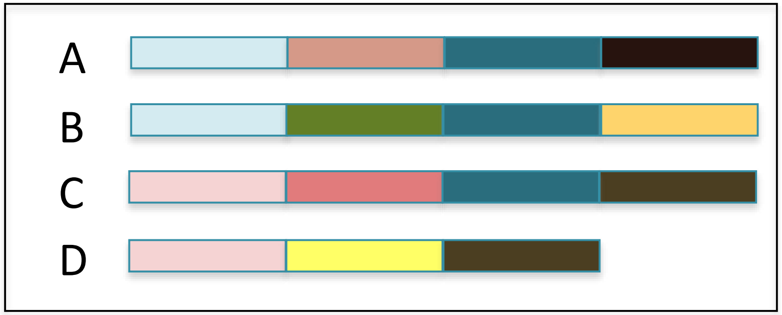
doi: http://dx.doi.org/10.1101/006395

Abstract Info/History Metrics Data Supplements □ Preview PDF
□ Previous
Posted June 18, 2014.
□ Download PDF
□ Email
Tweet 61 Like 8

Abstract

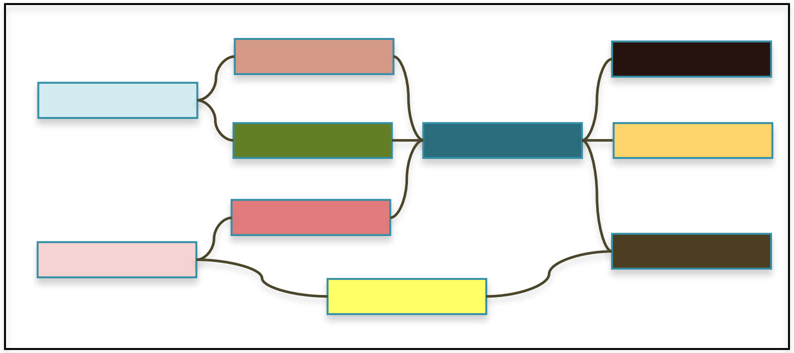
Third generation single molecule sequencing technology is poised to revolutionize genomics by enabling the sequencing of long, individual molecules of DNA and RNA. These technologies now routinely produce reads exceeding 5,000 basepairs, and can achieve reads as long as 50,000 basepairs. Here we evaluate the limits of single molecule sequencing by assessing the impact of long read sequencing in the assembly of the human genome and 25 other important genomes across the tree of life. From this, we develop a new data-driven model using support vector regression that can accurately predict assembly performance. We also present a novel hybrid error correction algorithm for long PacBio sequencing reads that uses pre-assembled Illumina sequences for the error correction. We apply it several prokaryotic and eukaryotic genomes, and show it can achieve near-perfect assemblies of small genomes (< 100Mbp) and substantially improved assemblies of larger ones. All source code and the assembly model are available open-source.

Pan-Genome Alignment & Assembly



Time to start considering problems for which N complete genomes is the input to study the "pan-genome"

•Available today for many microbial species, near future for higher eukaryotes



Pan-genome colored de Bruijn graph

- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips
Marcus, S, Lee, H, Schatz MC (2014) *Bioinformatics.* doi: 10.1093/bioinformatics/btu756

Outline

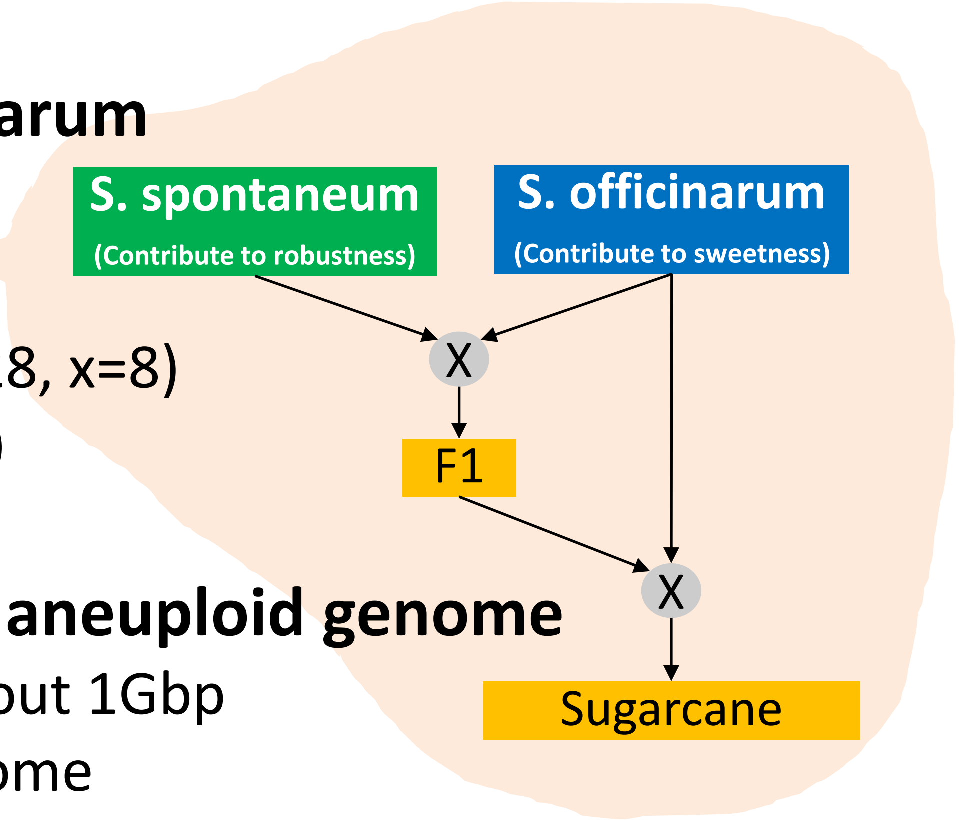
- **Background**
  - Long read sequencing technology
- **The limitations of short read mapping illustrated by Genome Mappability Score (GMS)**
  - Related works - Virmid
- **The Resurgence of reference quality genome (3Cs)**
  - The next version of Lander-Waterman Statistics (Contiguity)
  - Historical human genome quality by gene block analysis (Completeness)
  - The effectiveness of long reads in de novo assembly (Correctness)
  - Related works - MHAP
- **Sugarcane de novo genome assembly challenges**
  - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploid genome
  - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
  - Related works
    - Pineapple de novo genome assembly challenges - Heterozygous diploid genome
    - SK-BR-3 breast cancer study using SMRT reads - Benefits of long reads : From de novo assembly to structural variation detection
- **Contributions**

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

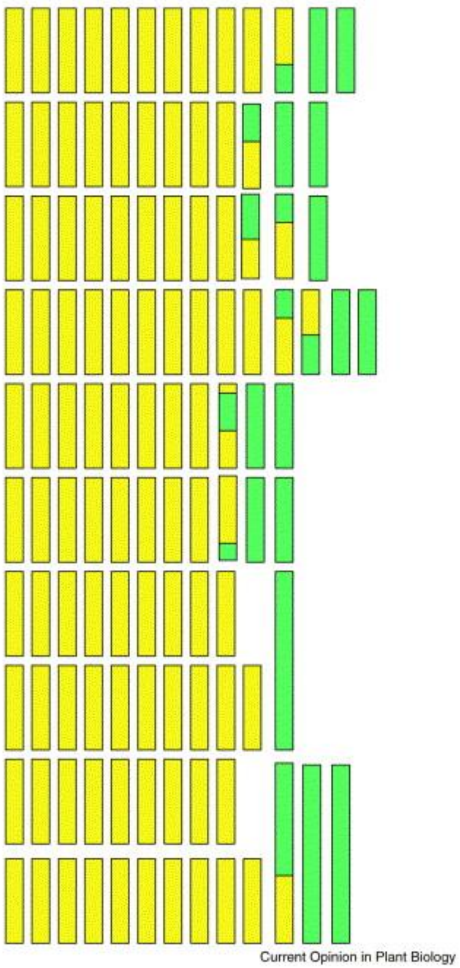
- **S.spontaneum x S.officinarum**
- **A century ago….**
- **Saccharum genus**
  - S. spontaneum (2n=40-128, x=8)
  - S. officinarum (2n=8x=80)



- **Big, highly polyploid and aneuploid genome**
  - Monoploid 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



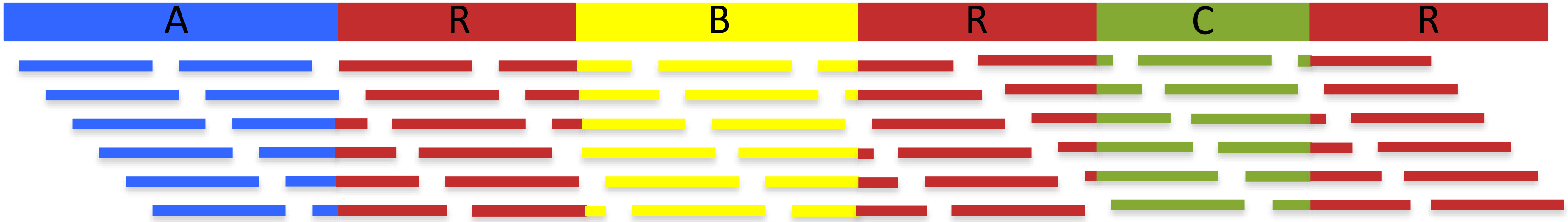
(source) http://ars.els-cdn.com/content/image/1-s2.0-S1369526602002340-gr1.jpg

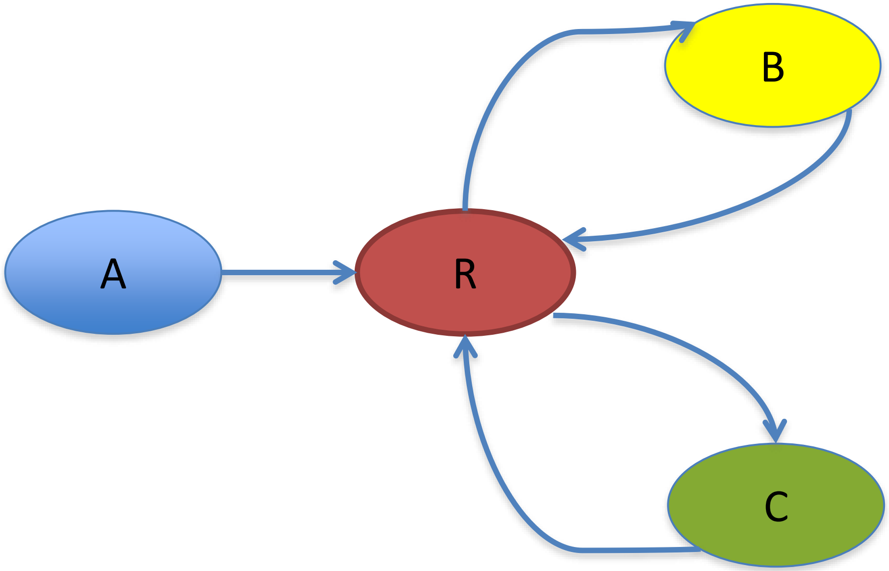
Four Important Questions in Sugarcane

- **Scaffold polyploidy/aneuploidy genome**
  - How do we connect contigs/cluster contigs per chromosome/fill gaps among contigs?
- **Phasing haplotypes**
  - Not solved in diploid genome yet
- **Heterozygosity**
  - How do we measure heterozygosity in polyploidy/aneuploidy genome?
  - How do we quantify alleles and get ratio?
- **Inference of polyploidy/aneuploidy estimation**
  - How do we infer the number of copies per chromosome in aneuploidy genome, especially in the large scale of recombination?

Margarido GRA, Heckerman D (2015) ConPADE: Genome Assembly Ploidy Estimation from Next-Generation Sequencing Data. PLoS Comput Biol 11(4): e1004229. doi: 10.1371/journal.pcbi.1004229

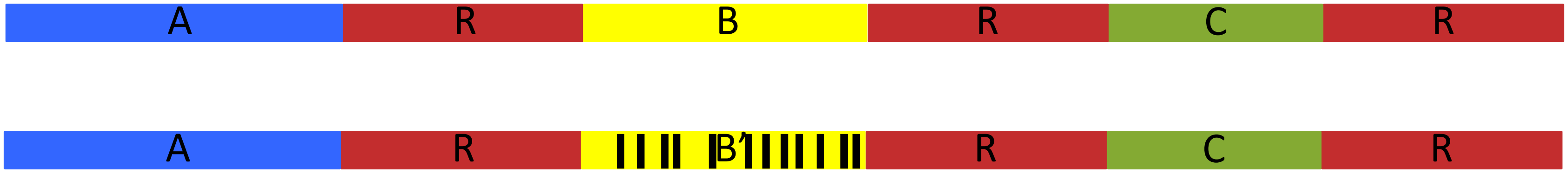
Assembly Complexity by Repeats

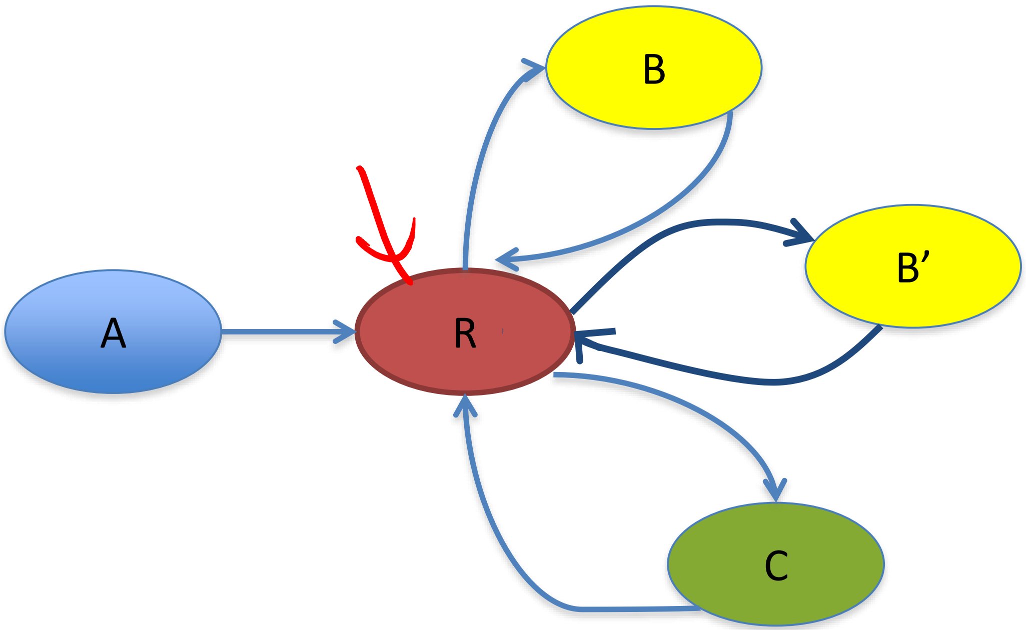




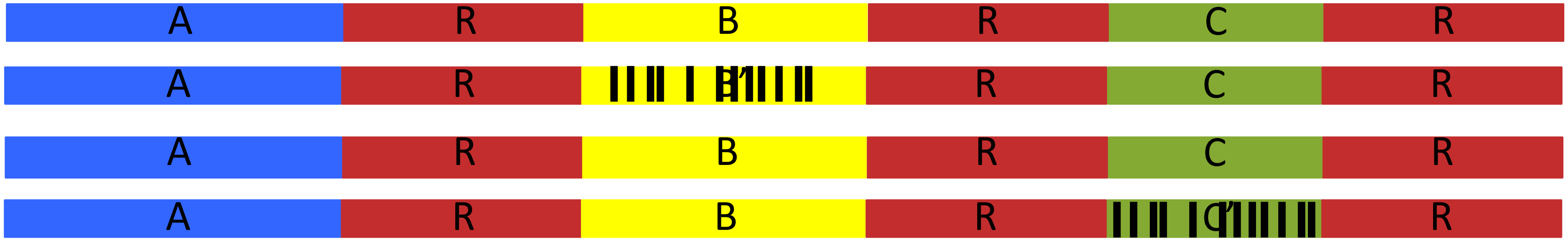
Long Reads is the solution!!!

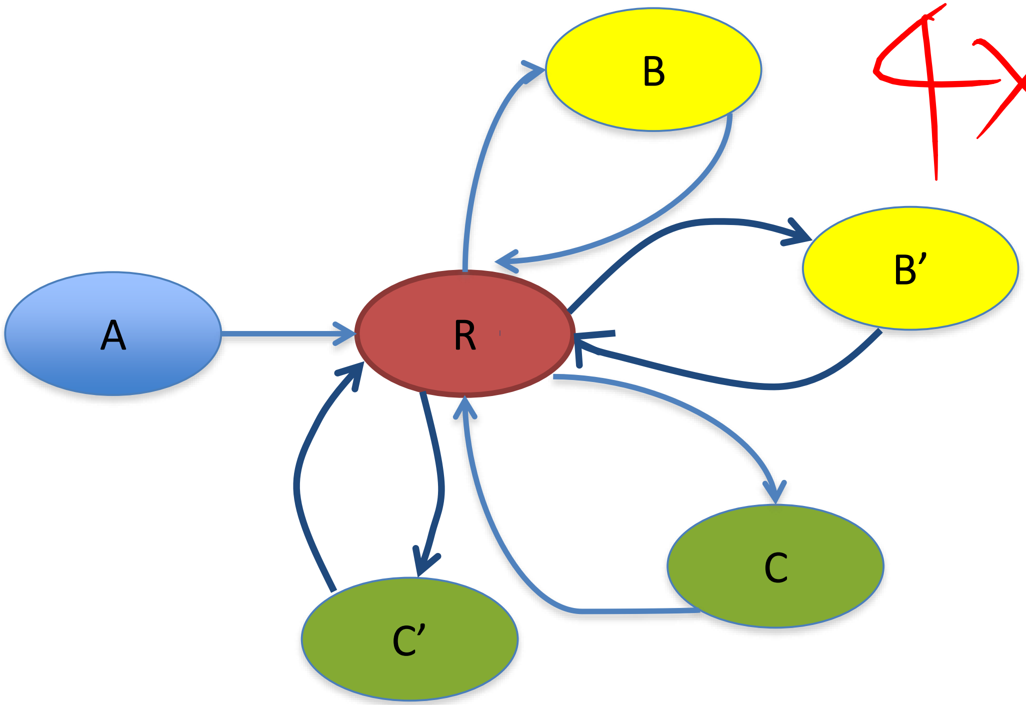
Assembly Complexity by Heterozygosity





Assembly Complexity by Polyploidy





4x3

Long Reads is the solution!!!

Moleculo Reads

(1) The DNA is sheared into fragments of about 10Kbp

(2) Sheared fragments are then diluted

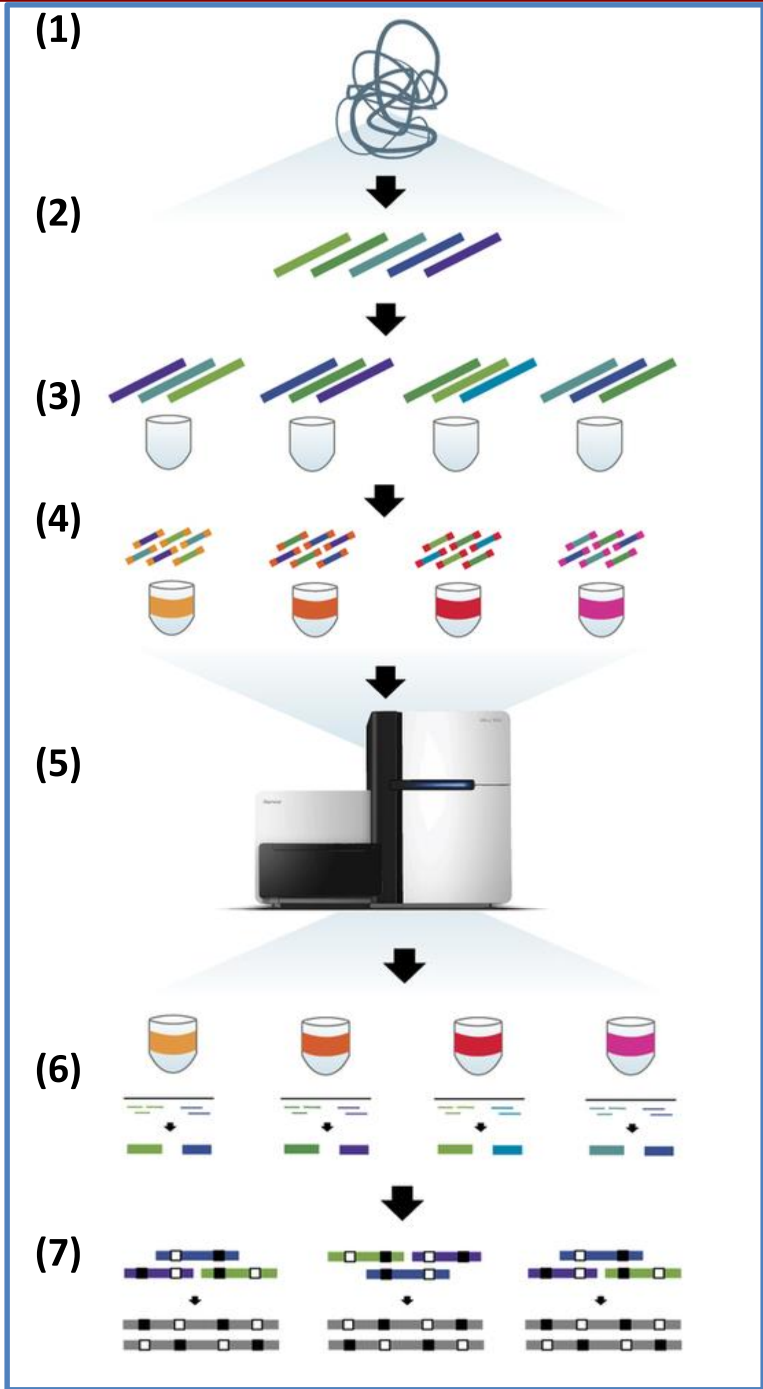
(3) and placed into 384 wells, at about 3,000 fragments per well.

(4) Within each well, fragments are amplified through long-range PCR, cut into short fragments and barcoded

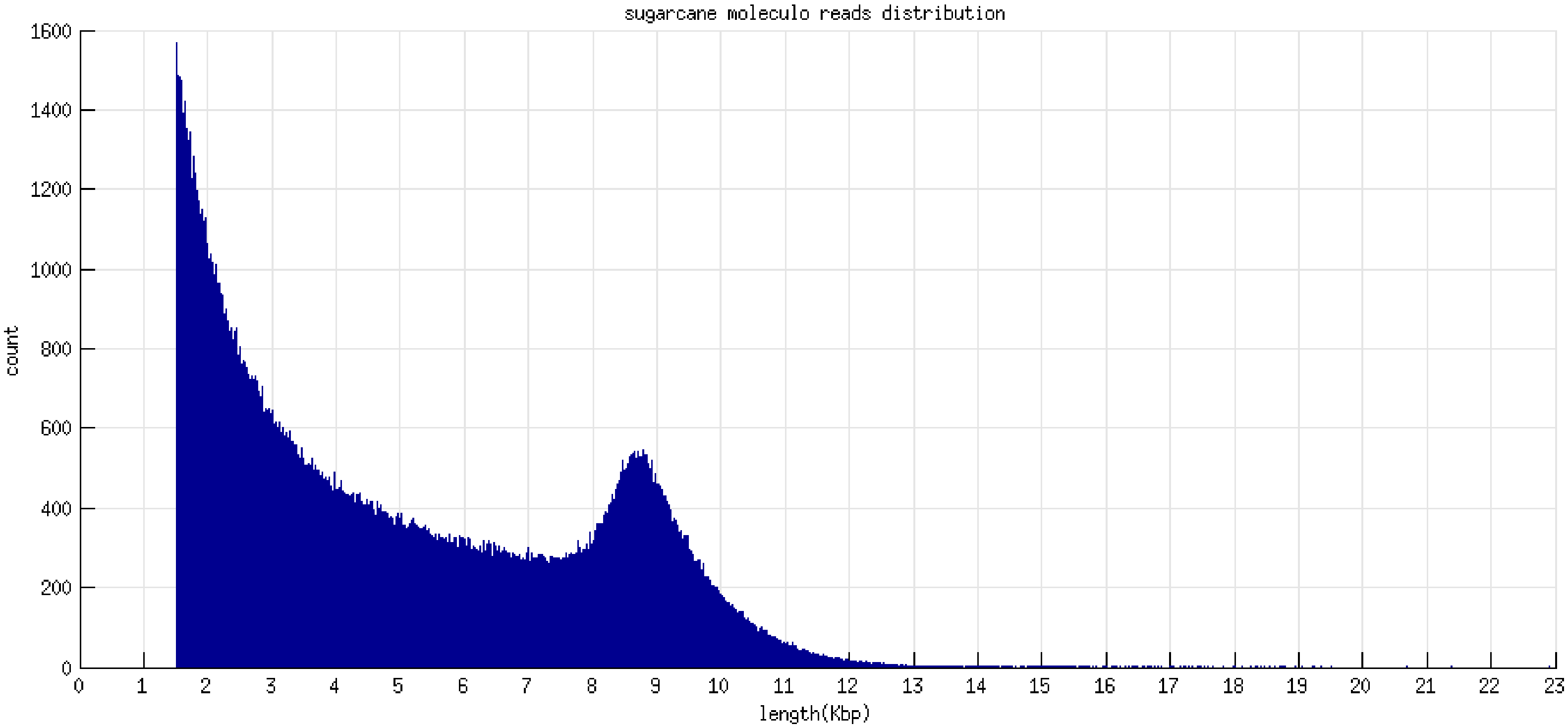
(5) before finally being pooled together and sequenced.

(6) Sequenced short reads are aligned and mapped back to their original well using the barcode adapters.

(7) Within each well, reads are grouped into fragments, which are assembled to long reads.



Read length distribution in Moleculo



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**

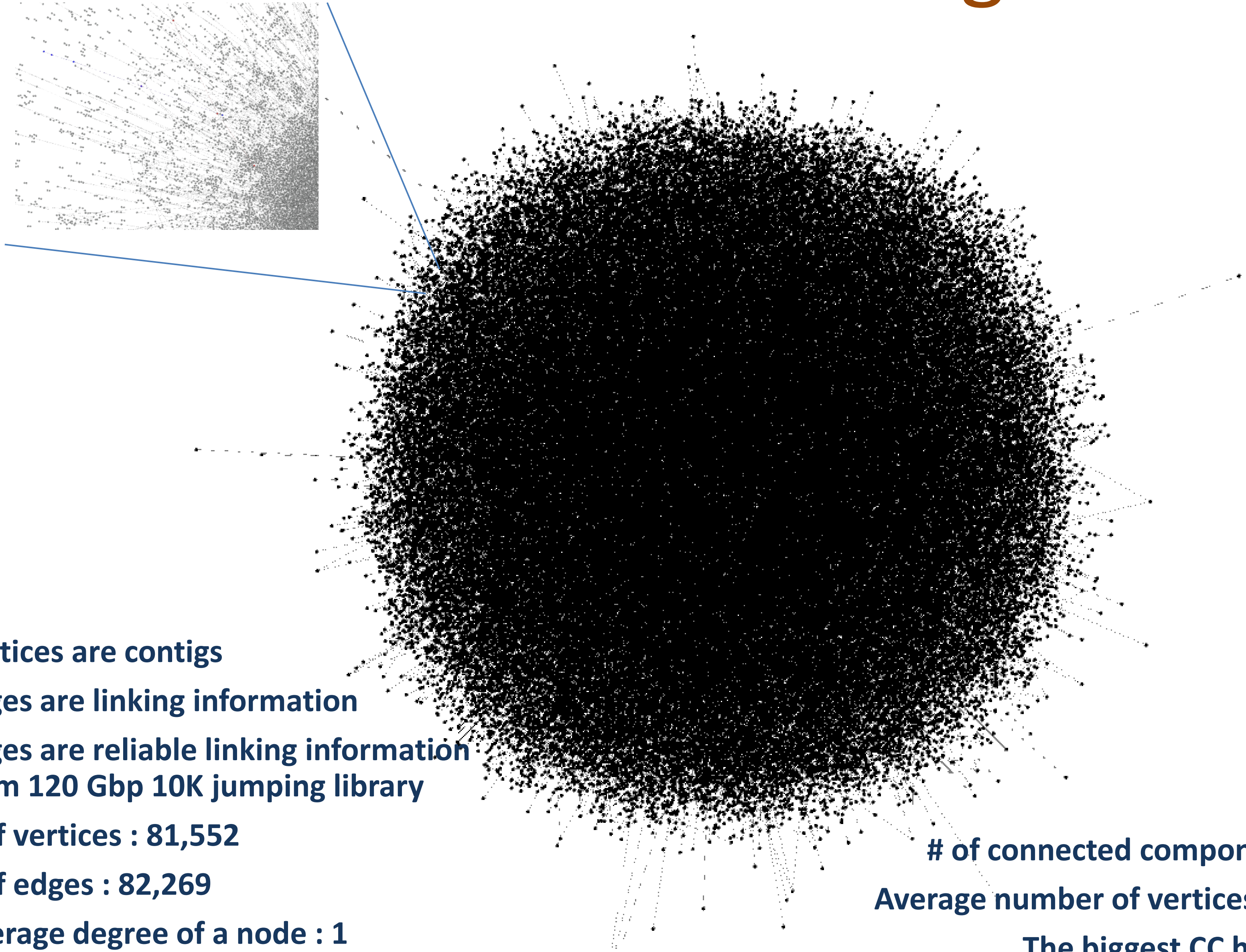
CEGMA

- **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

- **Gene prediction aided by sorghum gene model**
  - In progess…
  - 39k sorghum genes were found in sugarcane contigs at least partially

NP-Hard Hairball of Sugarcane



Vertices are contigs

Edges are linking information

Edges are reliable linking information from 120 Gbp 10K jumping library

# of vertices : 81,552

# of edges : 82,269

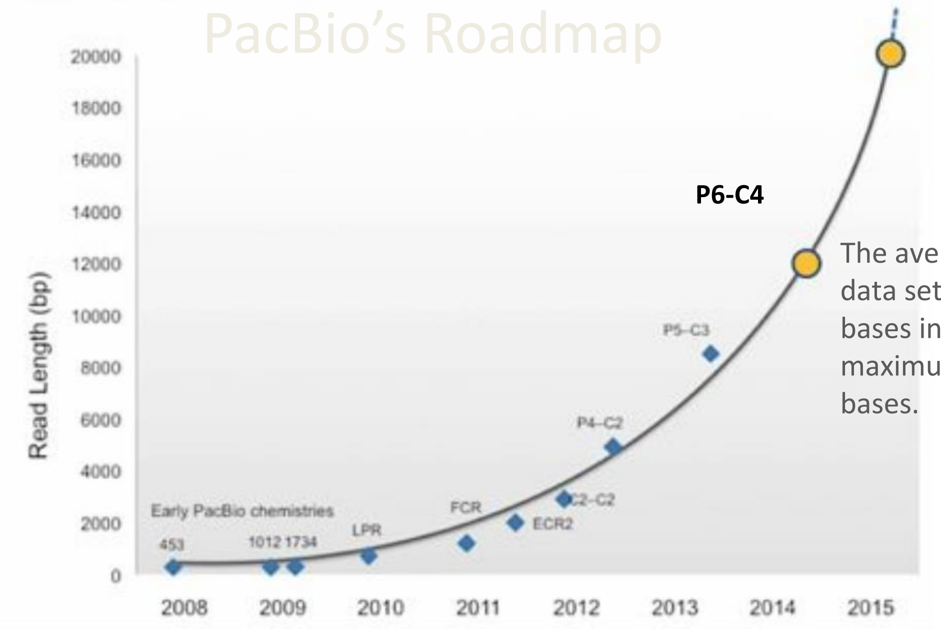
Average degree of a node : 1

# of connected components = 17,919

Average number of vertices per CC= 2.54

The biggest CC has 25 vertices

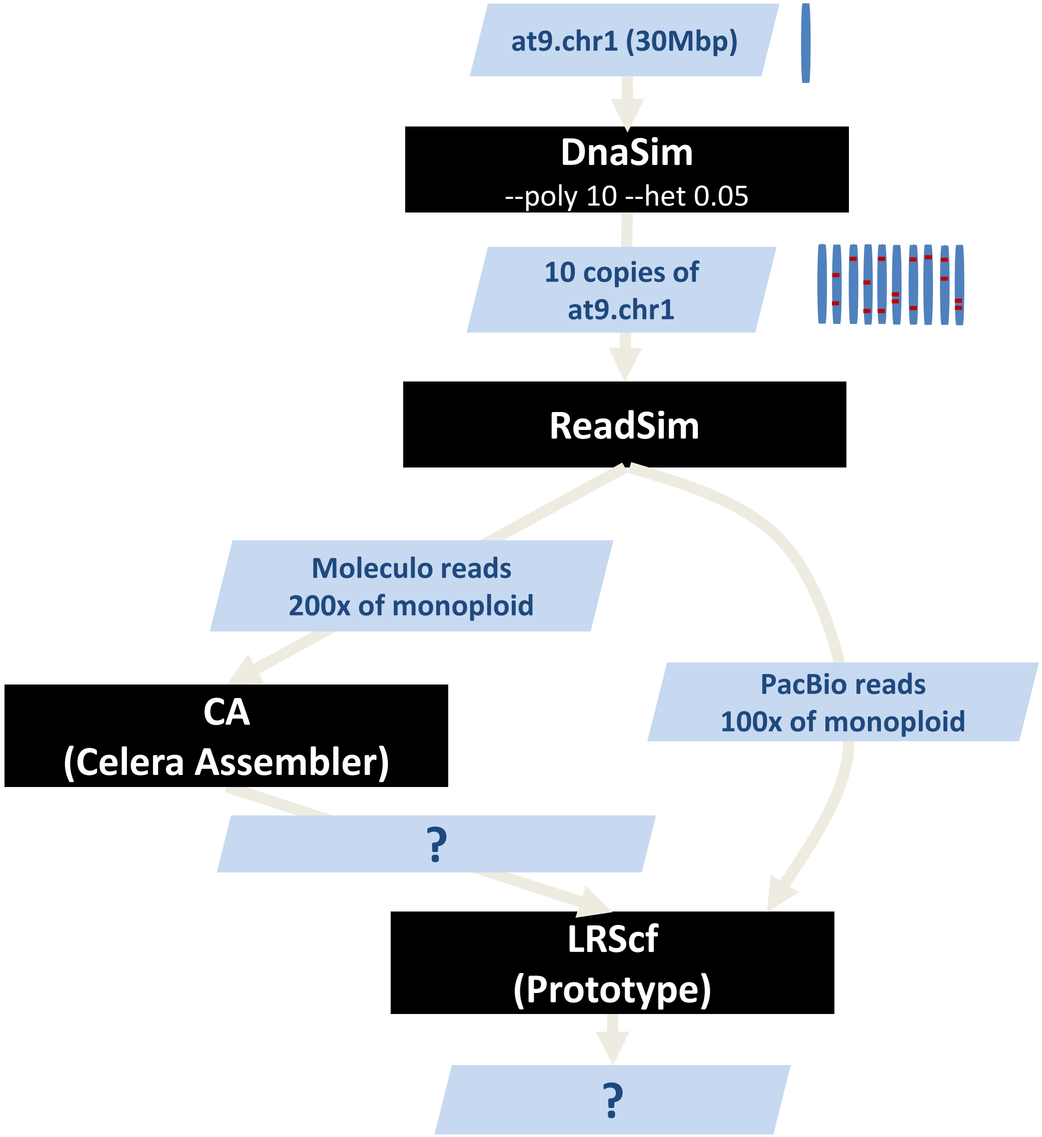
Benefits of Long Read Scaffolding



The average read length of the raw data set is >14 kb, with half of the bases in reads > 21 kb and the maximum read length of 64,500 bases.

- **Read Length is increasing, the cost is decreasing**
- **Very informative whether it has high error rate or not**
- **More repeats resolved**
- **Better scaffolding solution than long jumping library**
- **We don't have to approximate insert size by MLE or so.**
- **It's much better to fill gaps with some base information rather than just NNNNNN.**

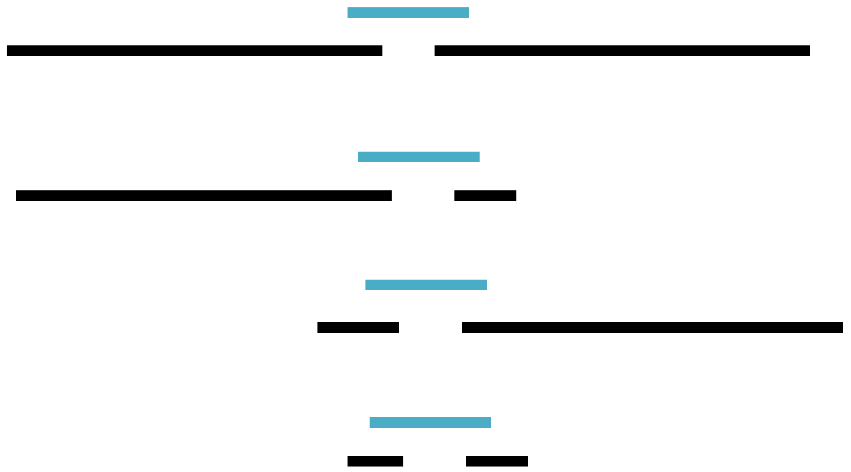
Prototype for scaffolding



1. **Simulate heterozygous polyploidy genome**
   - 10 copies with 5% of difference from original chromosome
2. **Simulate Moleculo reads from polyploidy genome**
   - Read length distribution follows exactly real molecule read distribution
3. **Simulate PacBio reads from polyploidy genome**
   - Simulate P6-C4, the lastest PacBio chemistry
4. **Run Celera Assembler(CA) to assemble contigs with Moleculo reads**
5. **Run LRScf to scaffold the contigs with PacBio reads**

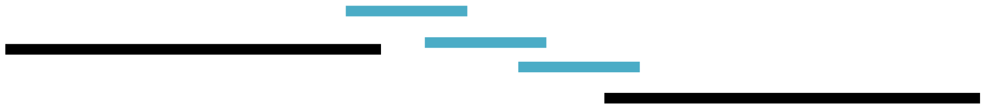
Preliminary Results

- **Moleculo-based contigs from CA**
  - Around 700 contigs
- **Long Read Scaffolding**
  - Align PacBio reads to all contigs
  - Find PacBio reads that link between two contigs
  - Around 1600 alignments out of 40K PacBio Reads

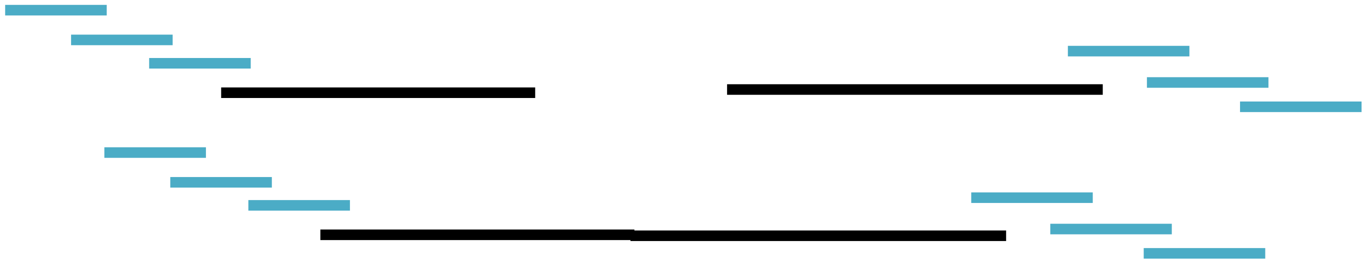


Sugarcane Scaffolding Challenges

- **How to represent aneuploidy genome?**
- **How to screen out false positive link information?**
  - # Weakly connected components 5
  - # Strongly connected components 61
  - True value 5 < 10 < 61
- **How to assemble PacBio reads across gaps?**



- **How to extend contigs with PacBio reads?**



Outline

- **Background**
  - Long read sequencing technology
- **The limitations of short read mapping illustrated by Genome Mappability Score (GMS)**
  - Related works - Virmid
- **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)
  - Related works - MHAP
- **Sugarcane de novo genome assembly challenges**
  - The effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploid genome
  - Pure long read de novo assembly, combine with accurate long reads and erroneous long reads
  - Related works
    - Pineapple de novo genome assembly challenges - Heterozygous diploid genome
    - SK-BR-3 breast cancer study using SMRT reads - Benefits of long reads : From de novo assembly to structural variation detection
- **Contributions**

Ming et al., (Under Review)

Illumina Contig N50 :2kbp
Moleculo+PacBio Contig N50 : 131kbp

Nattestad et al. (In preparation)

Illumina Contig N50 : 3.3kbp
PacBio Contig N50 : 2.17Mbp

Long Reads vs. Short Reads

- **Assembly**
- **Coverage analysis**
- **Structural Variant Discovery**
  - Insertion
  - Deletion
  - Translocation
    - Inter-chromosomal
    - Intra-chromosomal
  - Duplication
    - Interspersed duplication
    - Tandem duplication

Contributions

- **The limitations of short read mapping illustrated by the Genome Mappability Score (GMS)**
  - A new metric that measure reliability per position of a genome
  - Cloud computing pipeline for efficient computation for big genomes
  - Analysis of biological importance in variation discover low/high GMS region
- **The Resurgence of reference genome quality (3Cs)**
  - Provide the data-driven model, a.k.a. the next version of Lander-Waterman Statistics to predict contiguity of de novo genome assembly project
  - Analysis of completeness and correctness in historical human genome assembly
- **Sugarcane de novo genome assembly challenge**
  - Showed the effectiveness of accurate long reads in de novo assembly especially for highly heterozygous aneuploidy genome
    - NG50 contig length improved 50 times
    - The longest contig extended 25 times to half million bp
  - Pure long read de novo assembly for both contigs and scaffolding

Committee

Steven Skiena

Rob Patro

Michael Schatz

Adam Siepel

David Heckerman

Stony Brook University | The State University of New York

CSH

Microsoft Research

Acknowledgements

CSH

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

BROOKHAVEN NATIONAL LABORATORY

Shinjae Yoo

Microsoft Research

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

Stony Brook University | The State University of New York

Thank You

Q & A