Cold Spring Harbor Laboratory

The Resurgence of Reference Quality Genome

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Hayan Lee[1,2,5], James Gurtowski[1], Shinjae Yoo[3], Maria Nattestad[1], 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]DOE Joint Genome Institute, Walnut Creek, CA 94598

ABSTRACT

Several new 3rd generation long-range DNA sequencing and mapping technologies have recently become available that are starting to create a resurgence in genome sequence quality. Unlike their 2nd generation, short-read counterparts that can resolve a few hundred or a few thousand base-pairs, the new technologies can routinely sequence 10,000 bp reads or map across 100,000 bp molecules. The substantially greater lengths are being used to enhance a number of important problems in genomics and medicine, including de novo genome assembly, structural variation detection, and haplotype phasing.

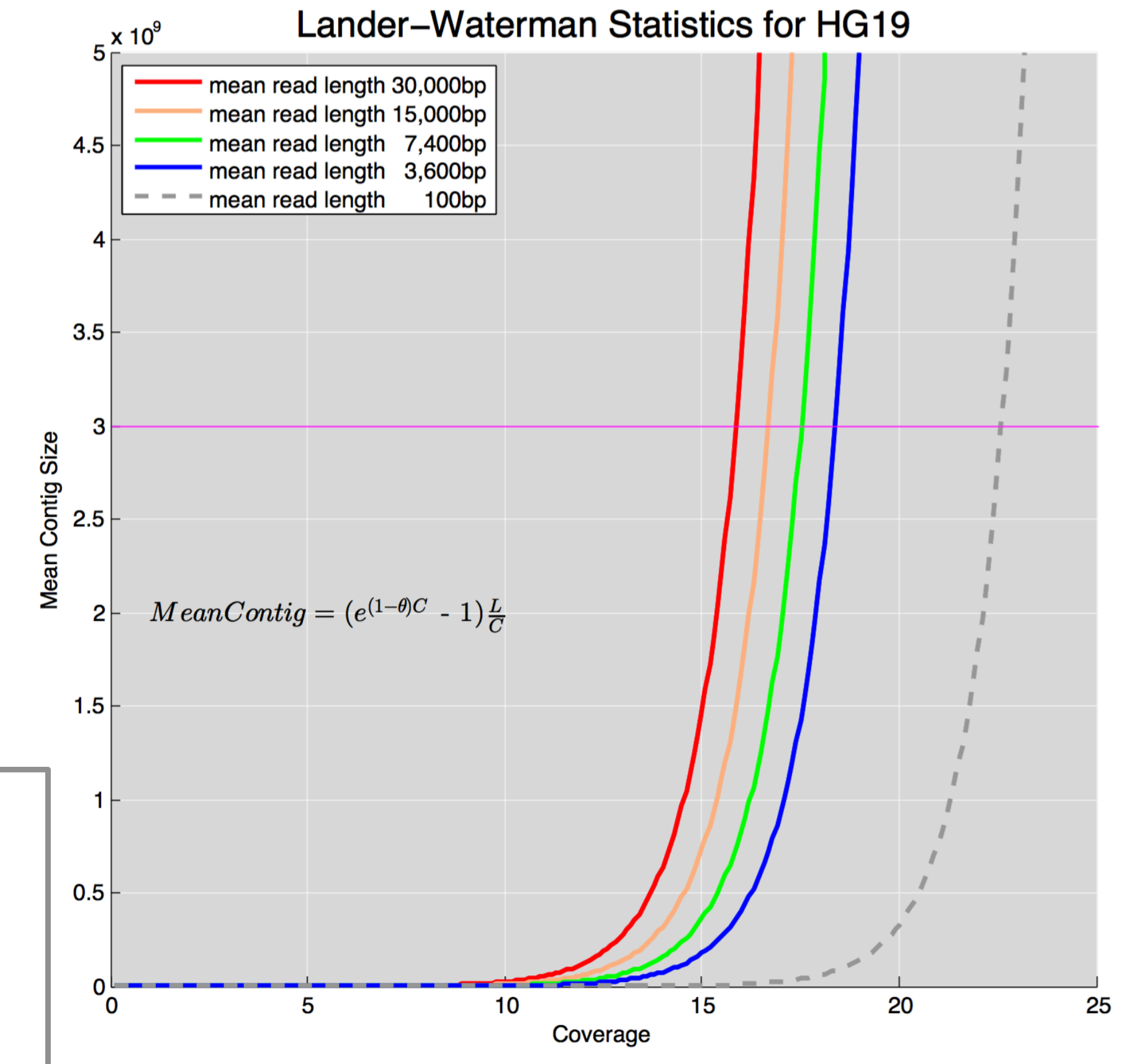
Here we discuss the capabilities of the latest echnologies, and show how they will improve the "3Cs of Genome Assembly": the contiguity, completeness, and correctness. We derive this analysis from (1) a meta-analysis of the currently available 3rd generation genome assemblies, (2) a retrospective analysis of the evolution of the reference human genome, and (3) extensive simulations with dozens of species across the tree of life.

We also propose a model using support vector regression (SVR) that predicts genome assembly performance using four features: read lengths(L) and coverage values(C) that can be used for evaluating potential technologies along with genome size(G) and repeats(R) that present species specific characteristics. The proposed model significantly improves genome assembly performance prediction by adopting data-driven approach and addressing limitations of the previous hypothesis-driven methodology.

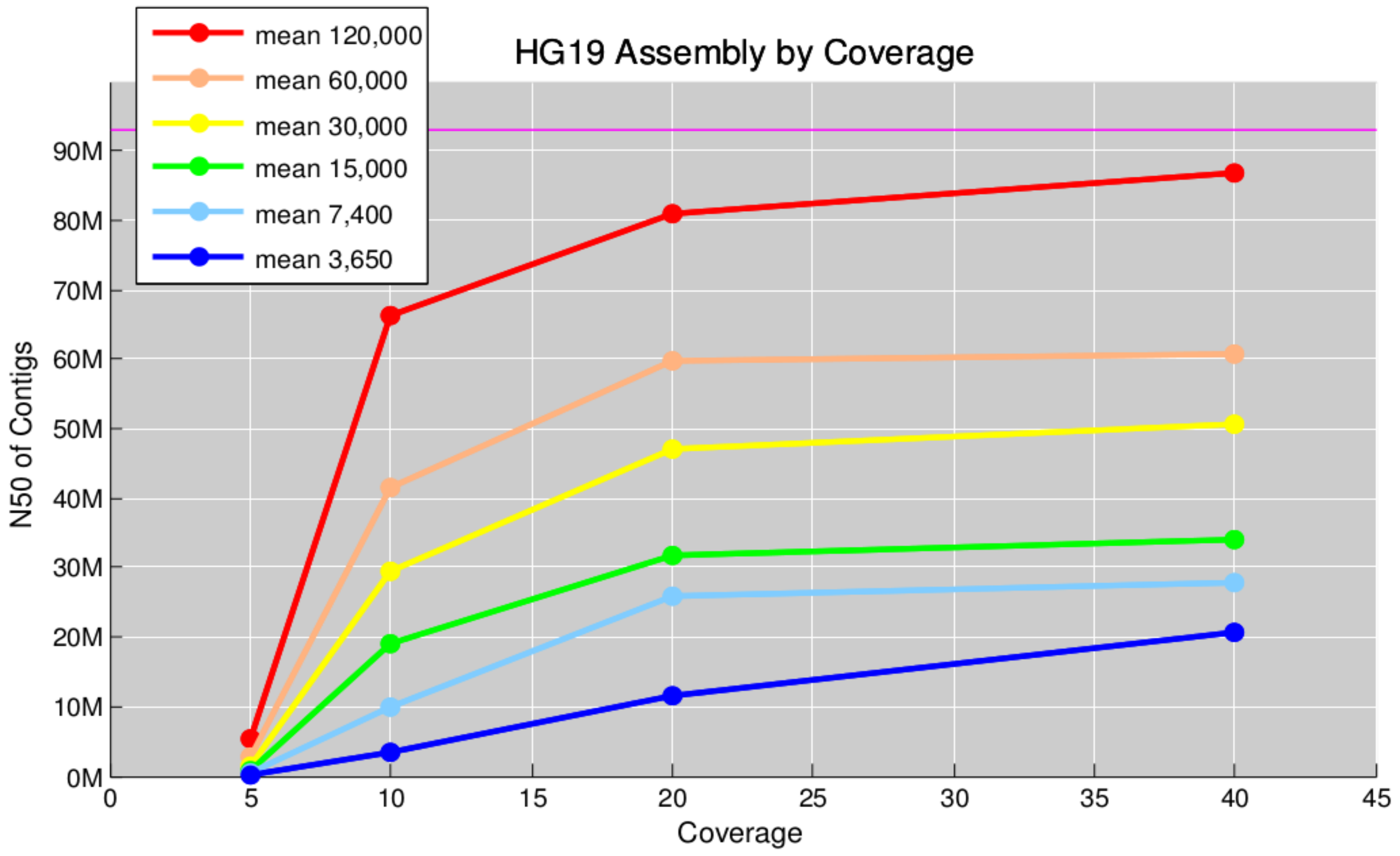
Overall, we anticipate these technologies unlock the genomic "dark matter", and provide many new insights into evolution, agriculture, and human diseases.

MOTIVATION

Two key observations
1. Contig over genome size
2. Read Length vs. Coverage
(Techonolgy vs. Money)



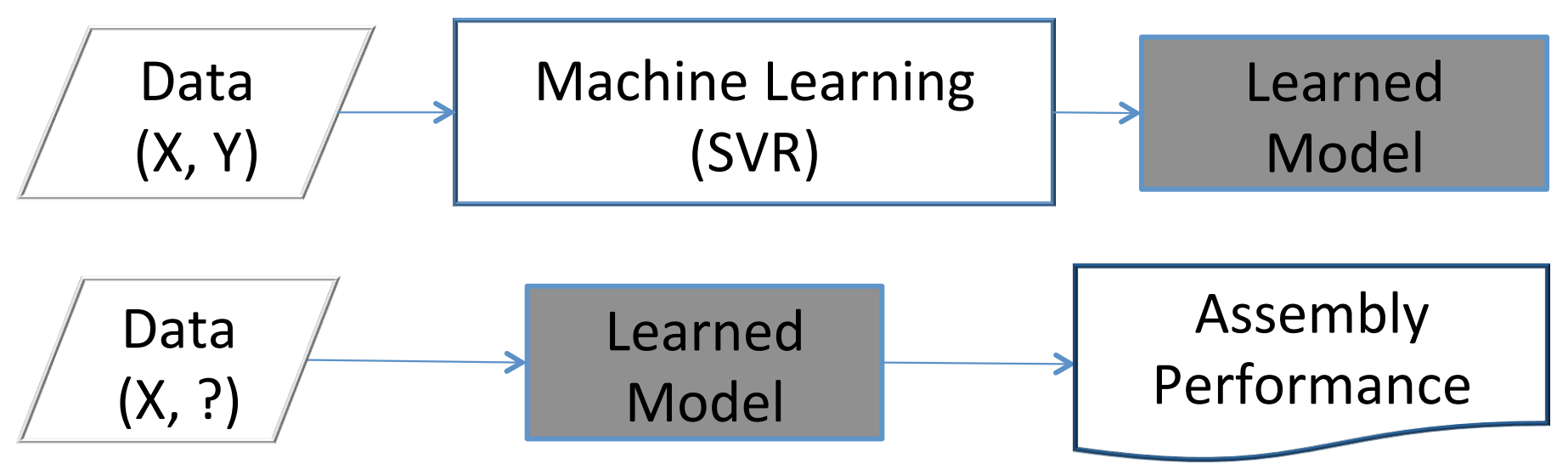
Lengths selected to represent idealized biochemistries:
mean32: ~Optical mapping
mean16: ~10x / Chromatin
mean8: ~10x / Chromatin
mean4: PacBio/ONT
mean2: PacBio/ONT
mean1: Moleculo
(log-normal with increasing means)



METHODS

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

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

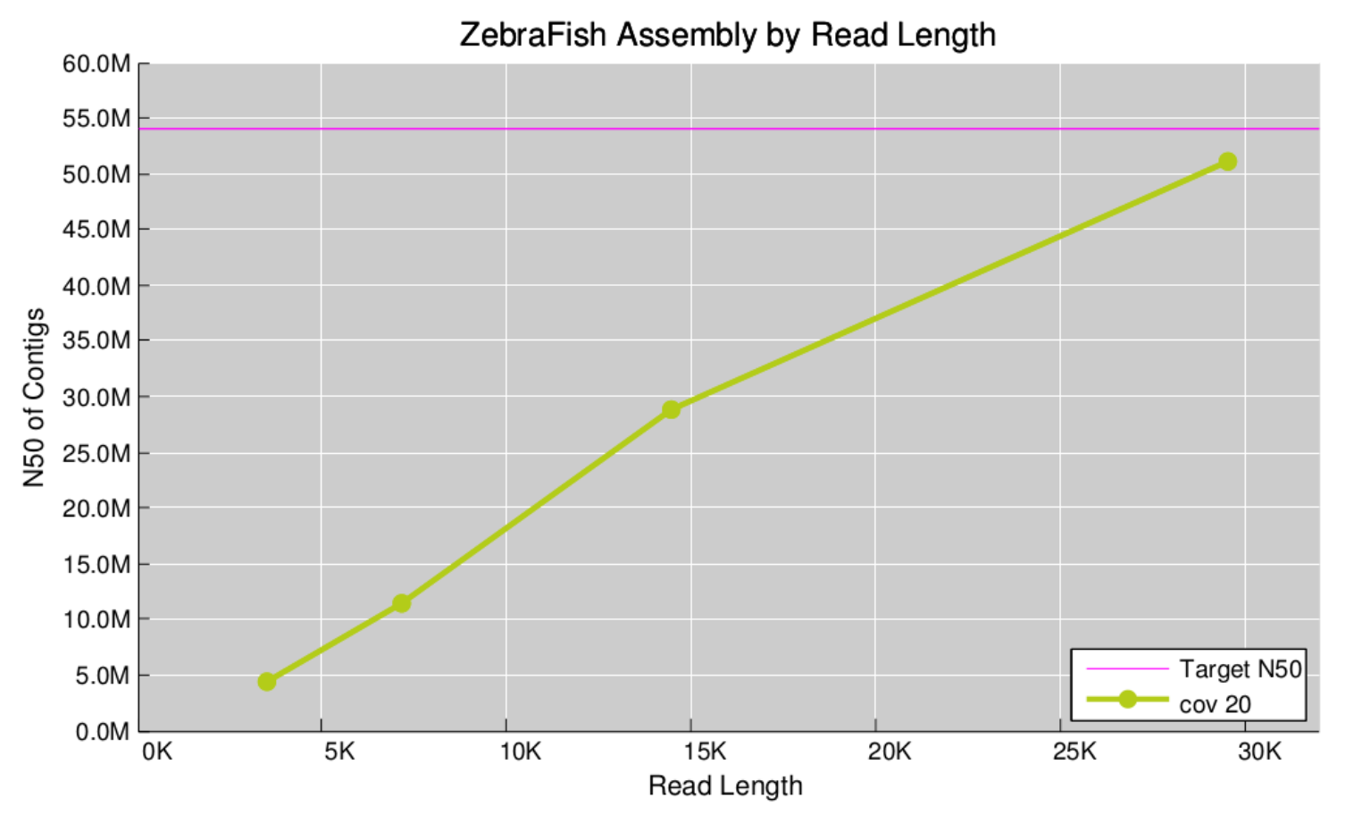

Data (X, Y)
Machine Learning (SVR)
Learned Model
Data (X, ?)
Learned Model
Assembly Performance

$$Performance(\%) \equiv \frac{N50\ fromAssembly}{N50\ fromChromosomeSegments} \times 100$$

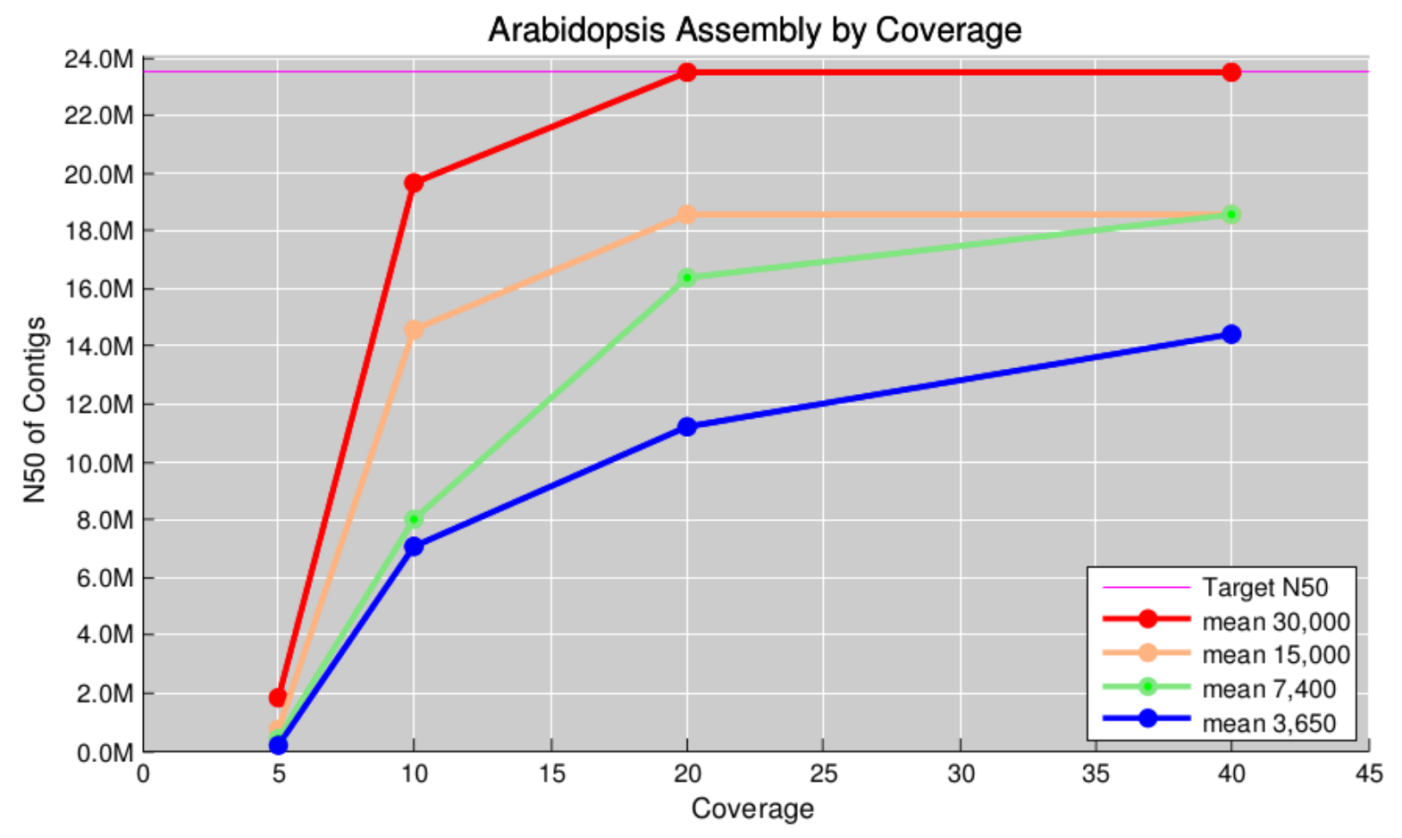
$$\approx f\left(\begin{array}{l} Read\ Length \\ Coverage \\ Genome\ Size \\ Repeat \end{array}\right)$$

We used four features; Read length(L), Coverage(C), Repeats(R), Genome size(G) to model de novo genome assembly contiguity after feature engineering.

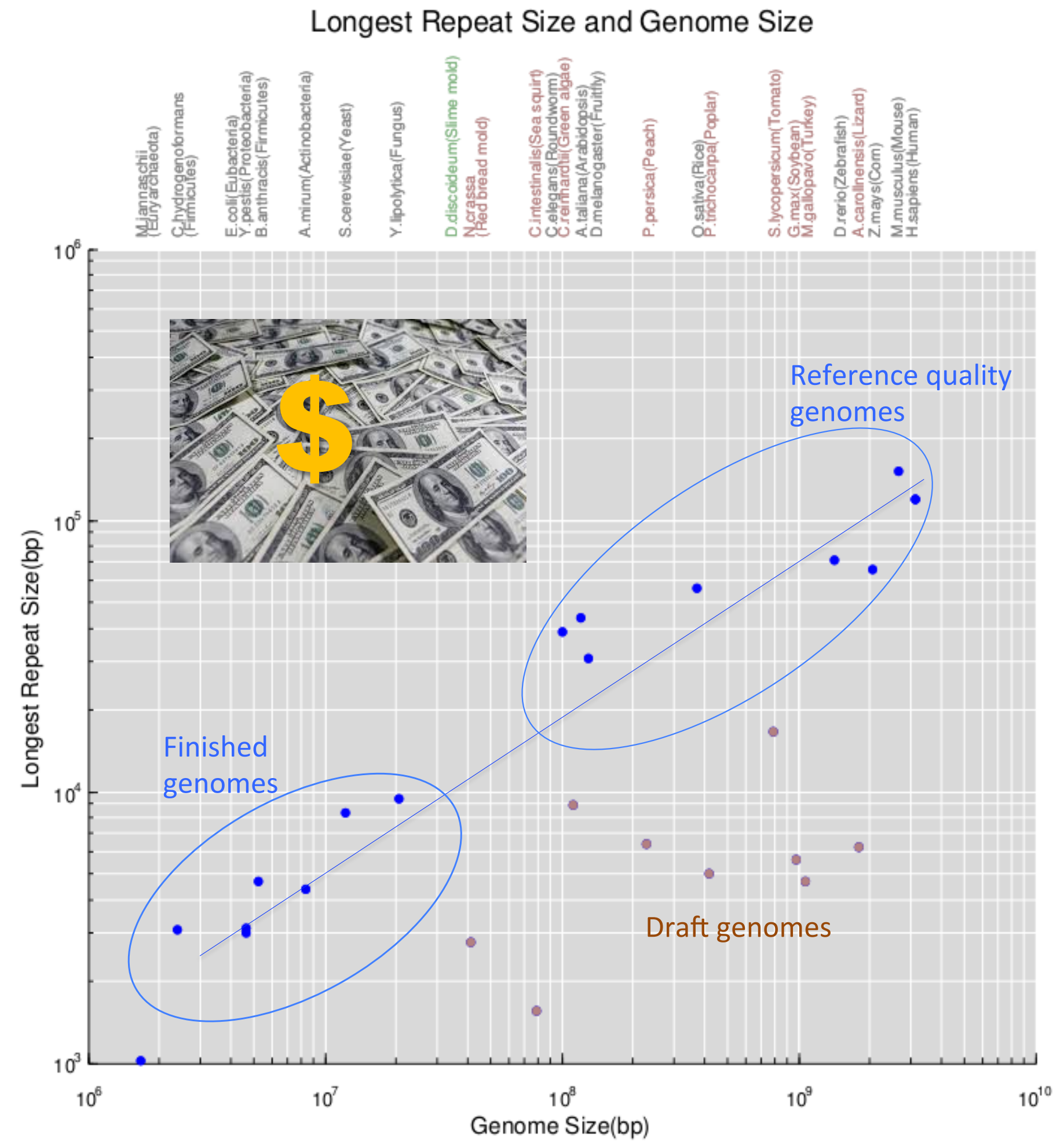
1. Read Length (L)



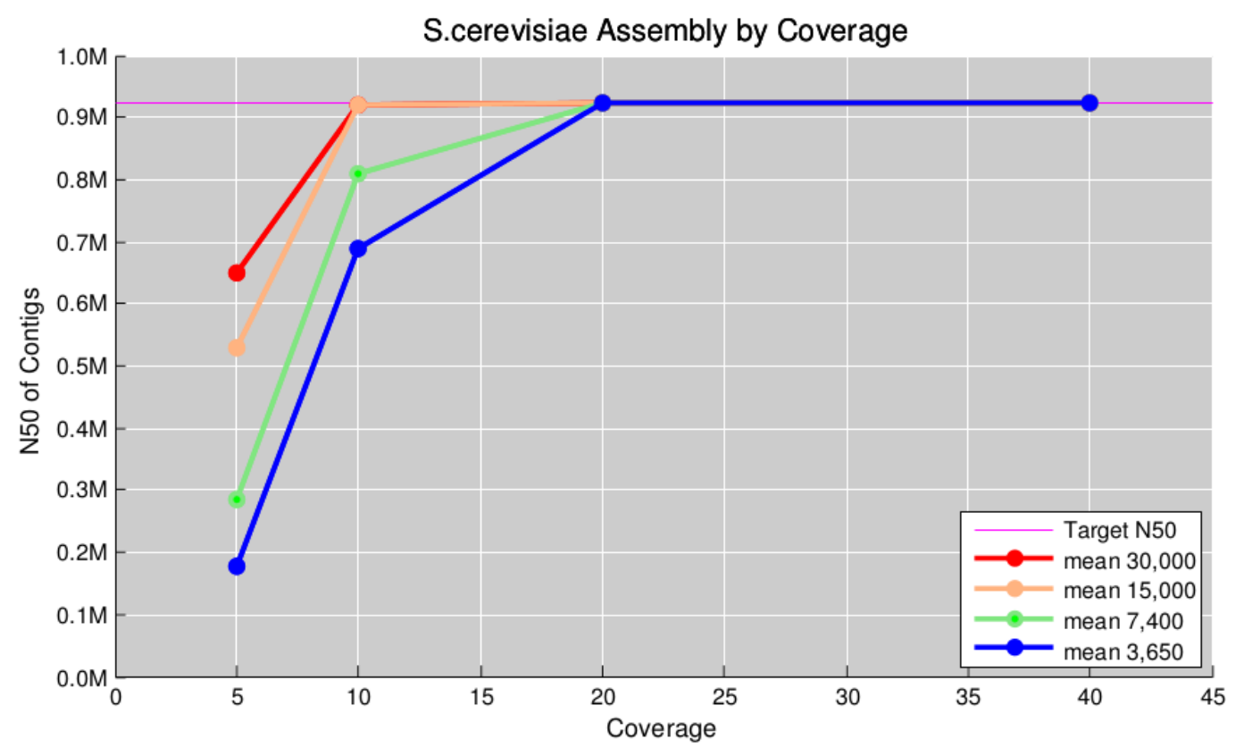
2. Coverage (C)

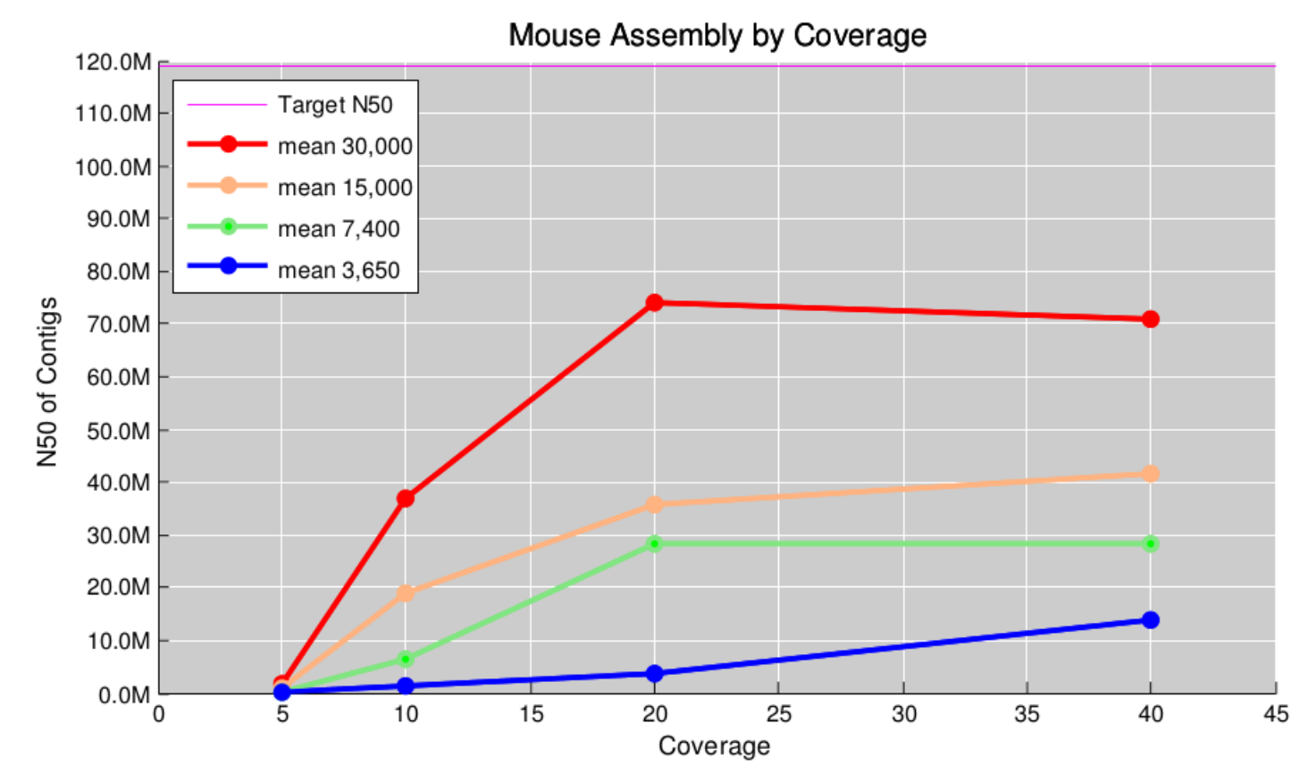


3. Repeats (R)



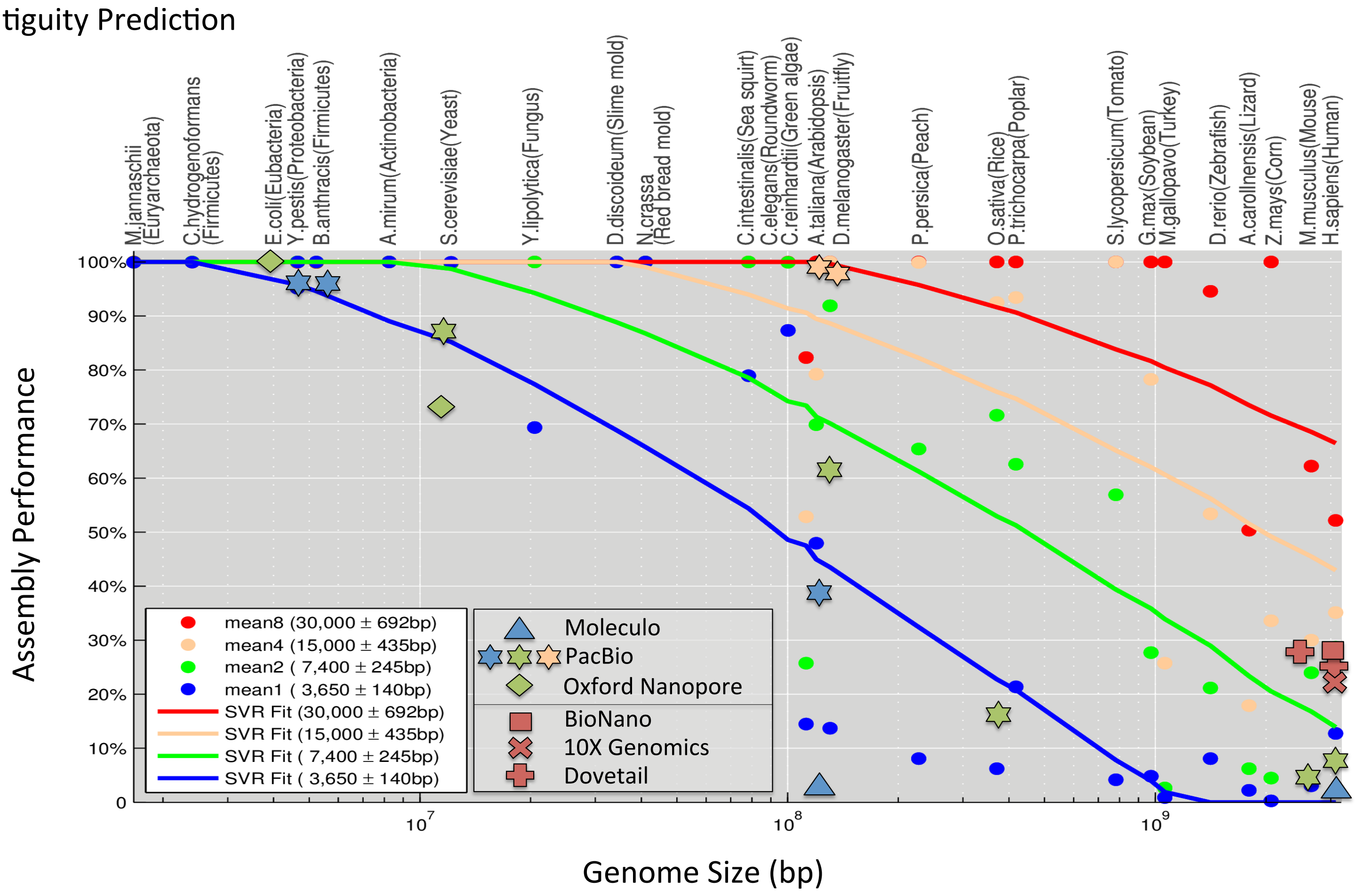
4. Genome Size (G)



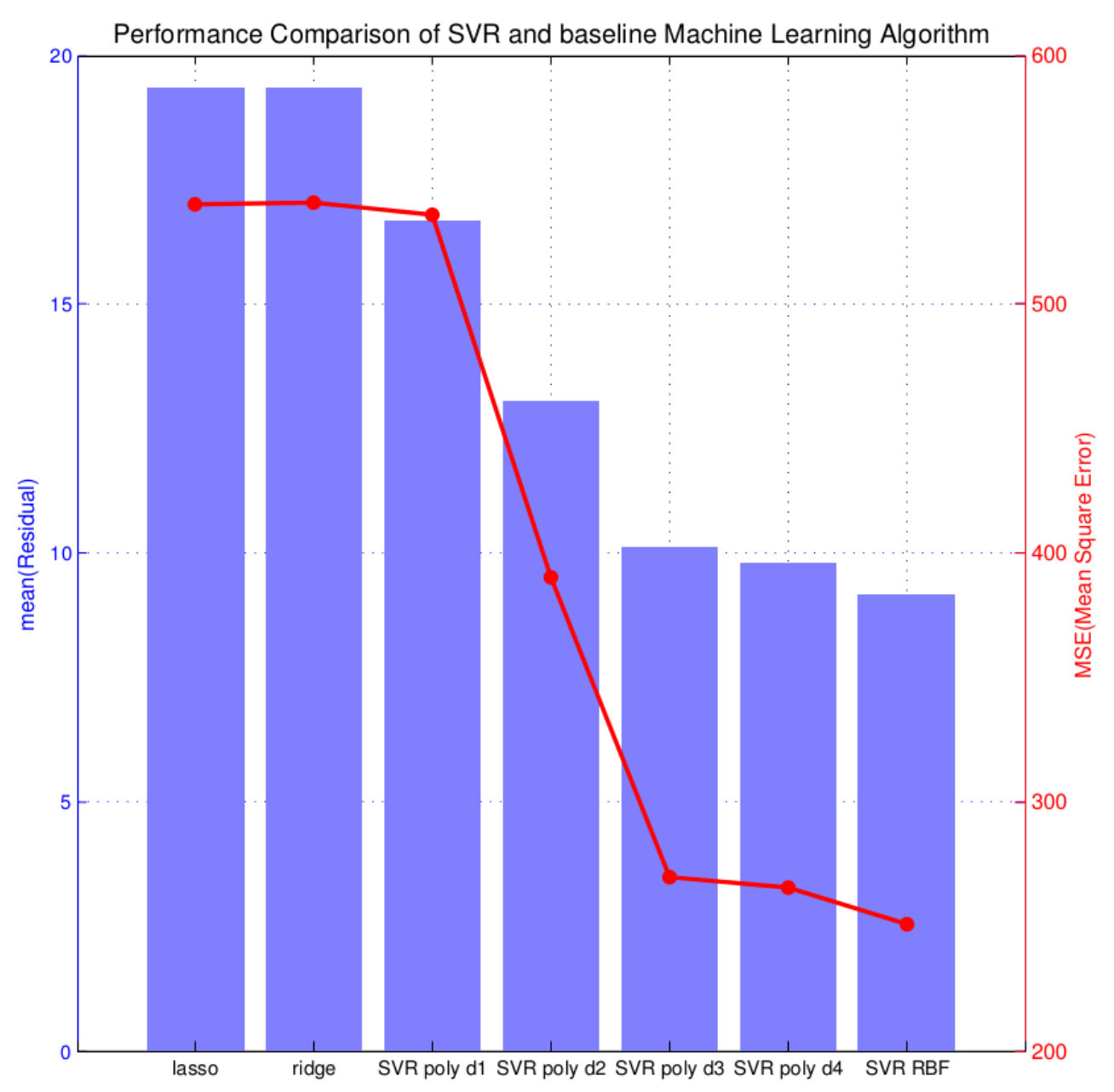


RESULTS (in terms of 3Cs)

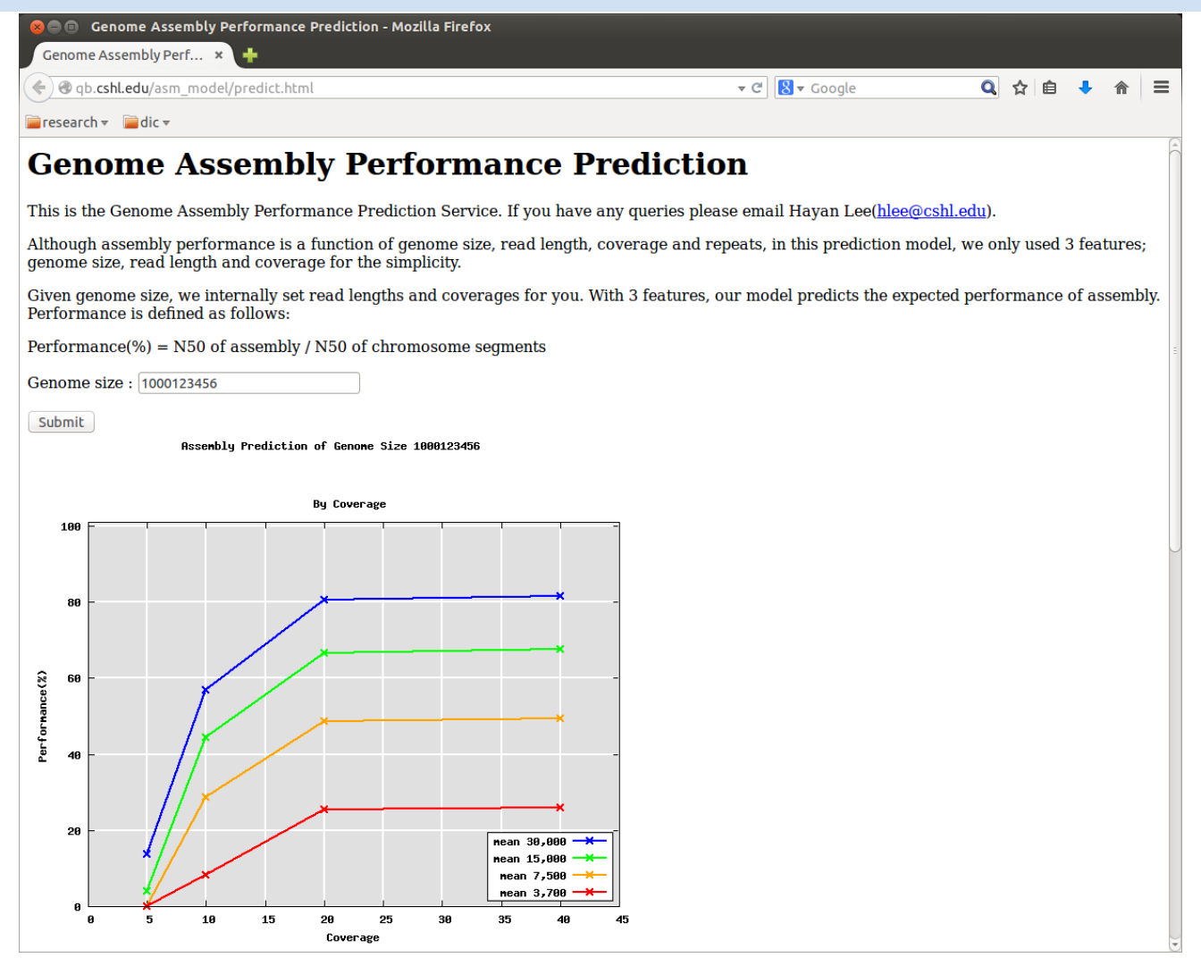
1. Contiguity Prediction



We started our web service for contiguity prediction.



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


Genome Assembly Performance Prediction - Mozilla Firefox
Genome Assembly Perf...
qb.cshl.edu/asm_model/predict.html
Google
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).

Although assembly performance is a function of genome size, read length, coverage and repeats, in this prediction model, we only used 3 features; genome size, read length and coverage for the simplicity.

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

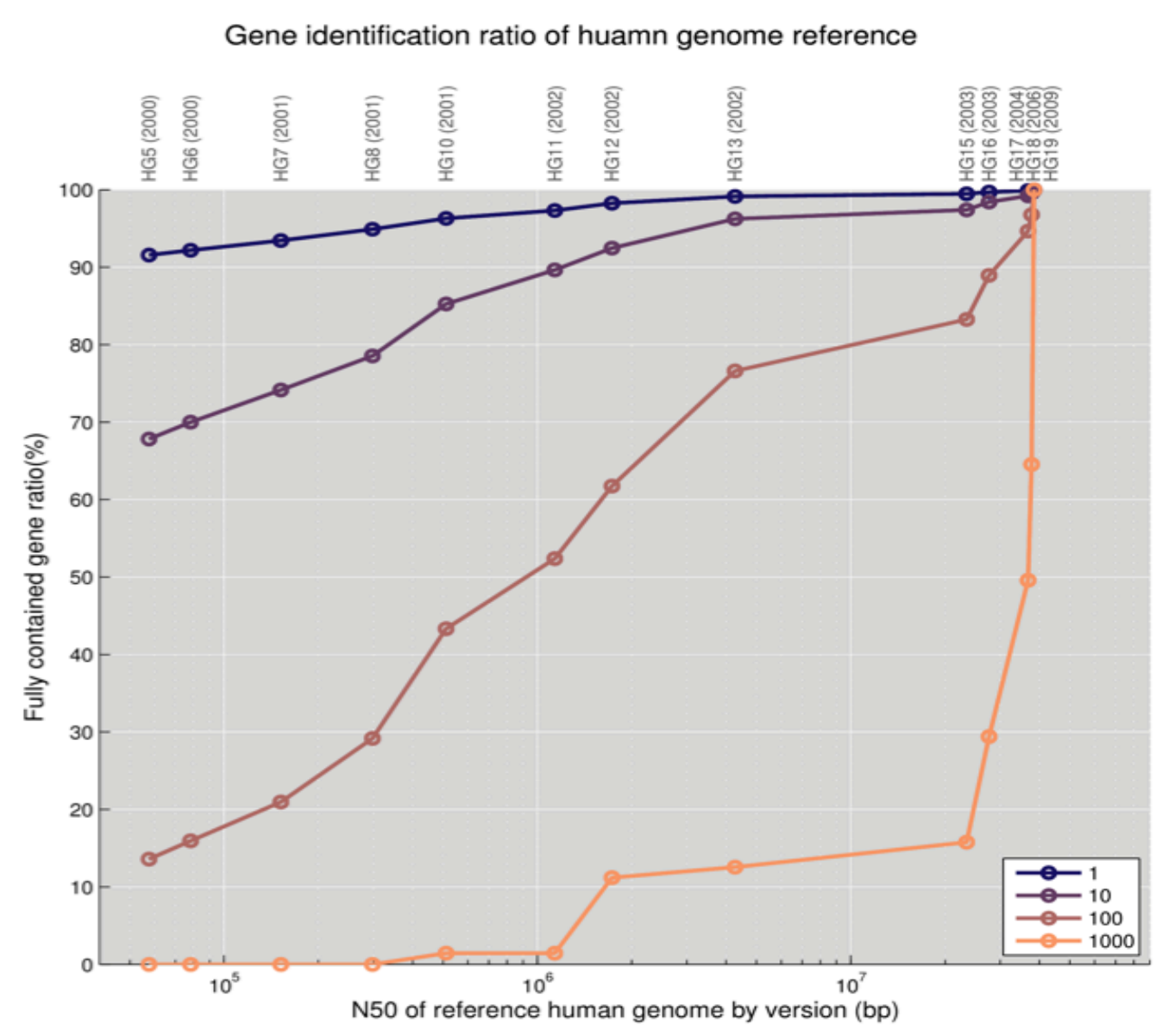
Genome size : 1000123456

Submit

Assembly Prediction of Genome Size 1000123456

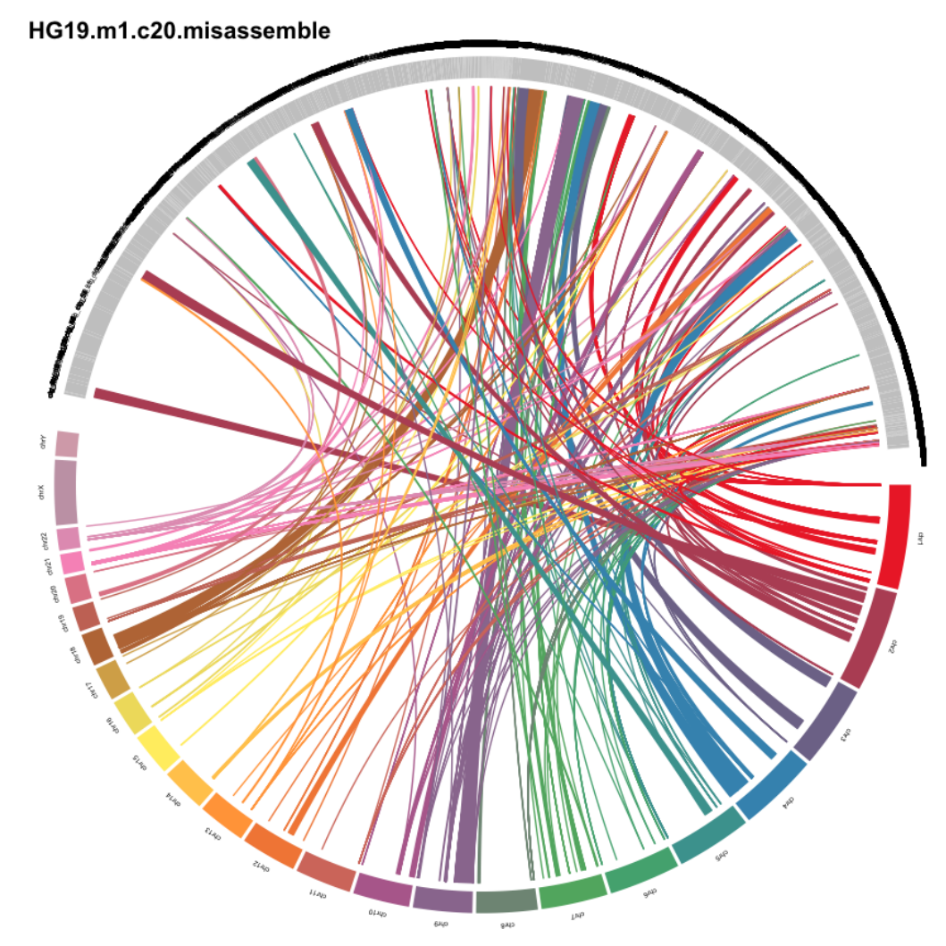
2. Completeness

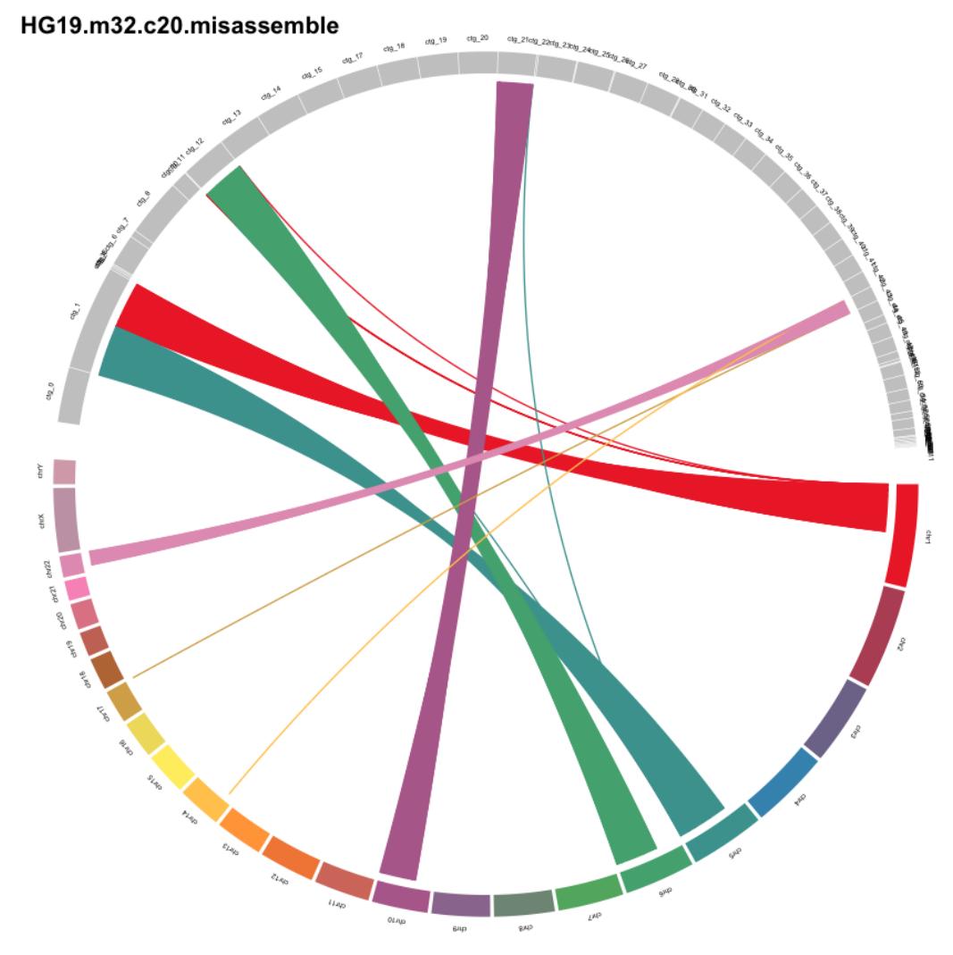
Gene1 : A single gene
Gene10 : 10 genes in a serial order
Regulatory elements
Gene100 : 100 genes in a serial order
Synteny blocks
Gene1000 : 1000 genes in a serial order
Chromosomal structure



3. Correctness

Misassemblies are one of the most severe problems of de novo assemblies, including producing contigs that falsely merge between two different chromosomes. It is a critical problem because (1) it can mislead us to incorrect biological conclusions, and (2) it can falsely increase the N50 length. We can reduce the number of misassemblies by using longer reads. Shown here is a plot of the major misassemblies when using reads averaging 3600bp (m1) versus those made when using 120Kbp (m32).


HG19.m1.c20.misassemble


HG19.m32.c20.misassemble