The Resurgence of Reference Quality Genome Sequence

Michael Schatz

Jan 12, 2016
PAG XXIV

@mike_schatz / #PAGXXIV

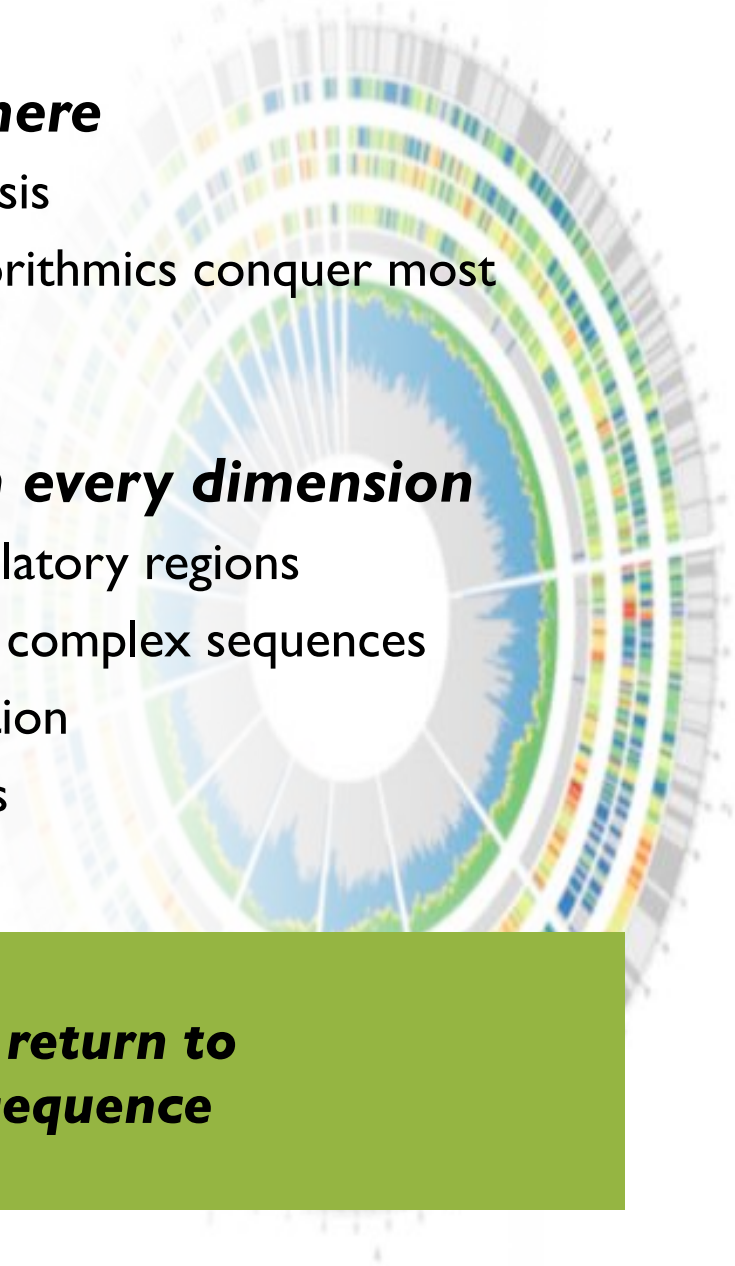
Summary & Recommendations

Reference quality genome assembly is here

- Use the longest possible reads for the analysis
- Don't fear the error rate, coverage and algorithmics conquer most problems

Megabase N50 improves the analysis in every dimension

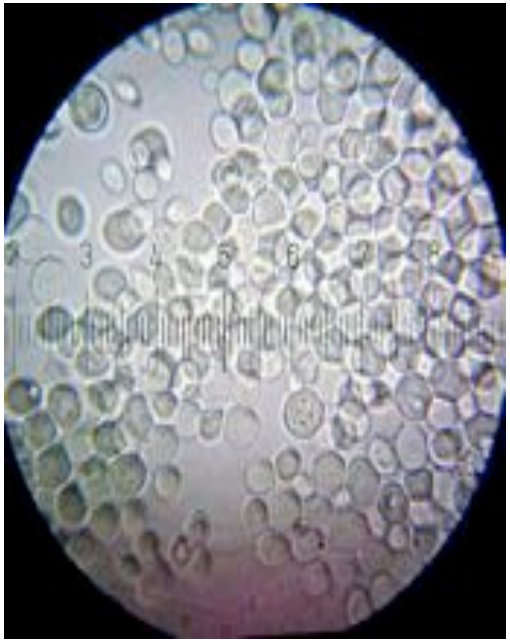
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis



The year 2015 will mark the return to reference quality genome sequence

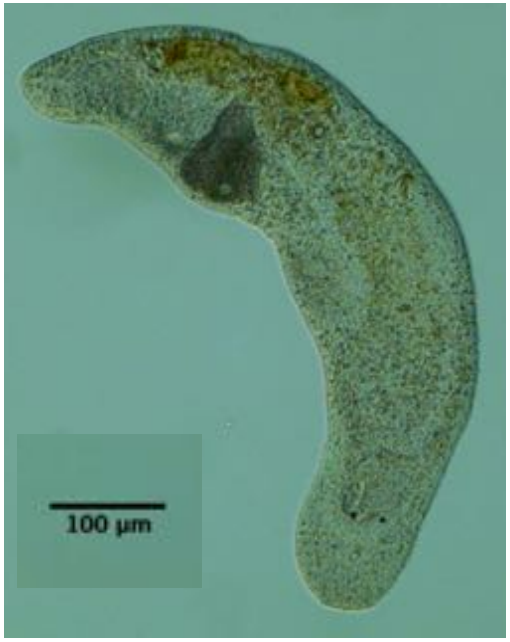
Selected Genomes from 2015

Saccharomyces cerevisiae
ONT + Illumina



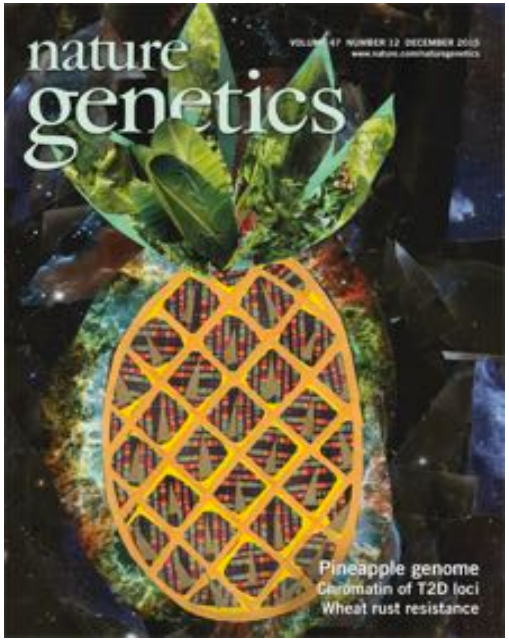
Goodwin *et al.* (2015)
Genome Research.
doi: 10.1101/gr.191395.115

Macrostomum lignano
PacBio



Wasik *et al.* (2015)
PNAS.
doi: 10.1073/pnas.1516718112

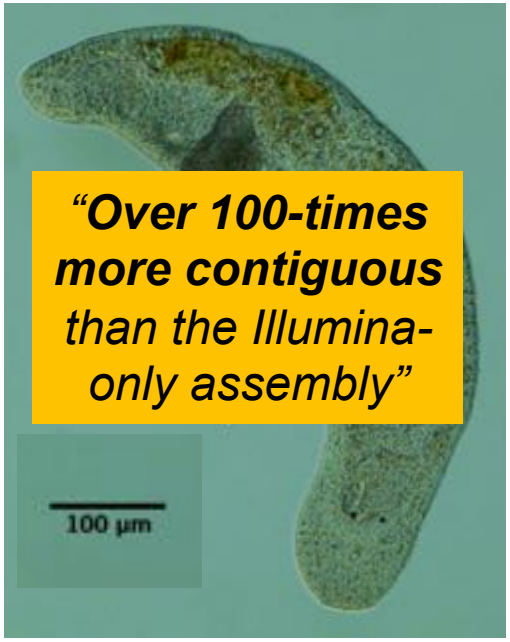
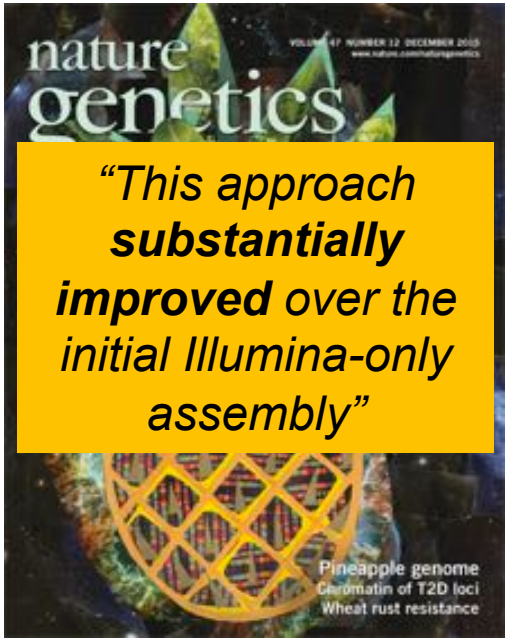
Ananas comosus
Illumina + Moleculo + PacBio



Ming *et al.* (2015)
Nature Genetics.
doi: doi:10.1038/ng.3435

#1MbpCtgClub

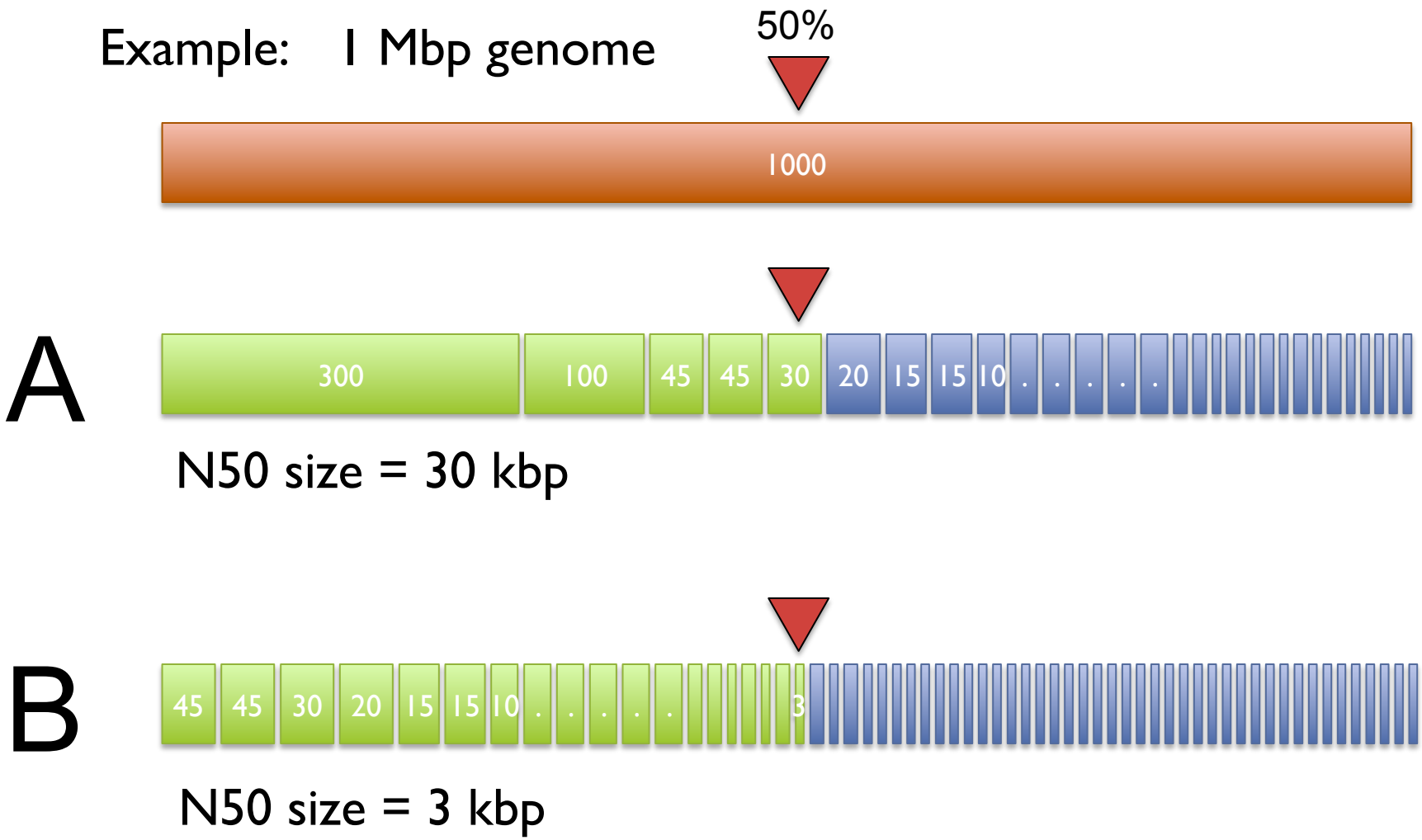
Selected Genomes from 2015

Saccharomyces cerevisiae *ONT + Illumina*	***Macrostomum lignano*** *PacBio*	***Ananas comosus*** *Illumina + Moleculo + PacBio*
		
Goodwin *et al.* (2015) *Genome Research.* doi: 10.1101/gr.191395.115	Wasik *et al.* (2015) *PNAS.* doi: 10.1073/pnas.1516718112	Ming *et al.* (2015) *Nature Genetics.* doi: doi:10.1038/ng.3435

#1MbpCtgClub

Contig N50

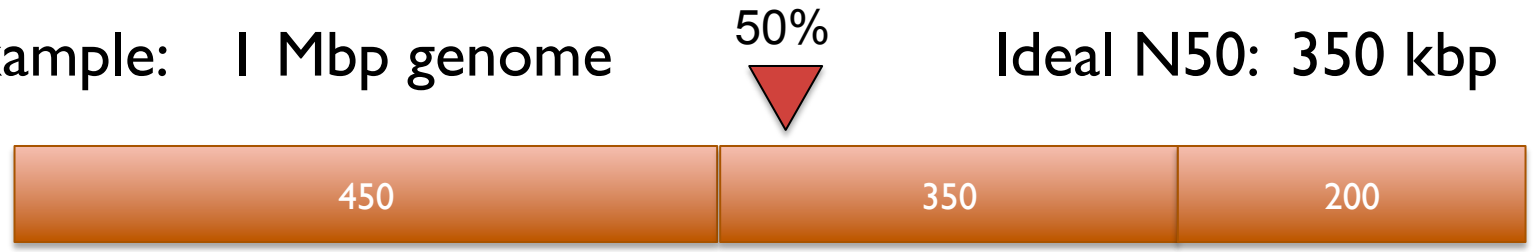
Def: 50% of the genome is in contigs as large as the N50 value



Assembly Performance

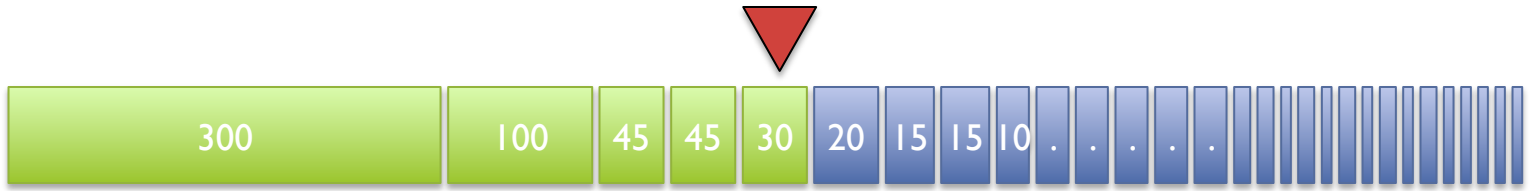
Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



Ideal N50: 350 kbp

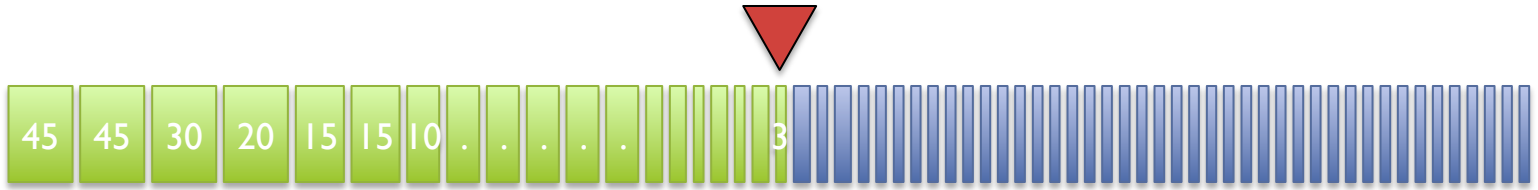
A



N50 size = 30 kbp

Assembly performance = 30 kbp / 350 kbp = 8.5%

B

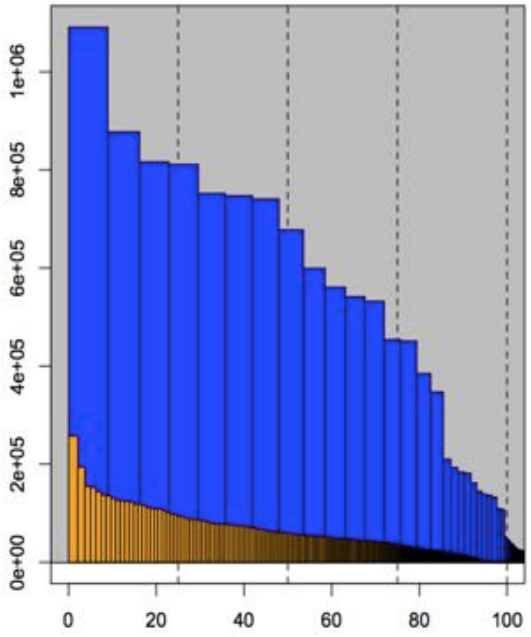


N50 size = 3 kbp

Assembly Performance = 3 kbp / 350 kbp = 0.85%

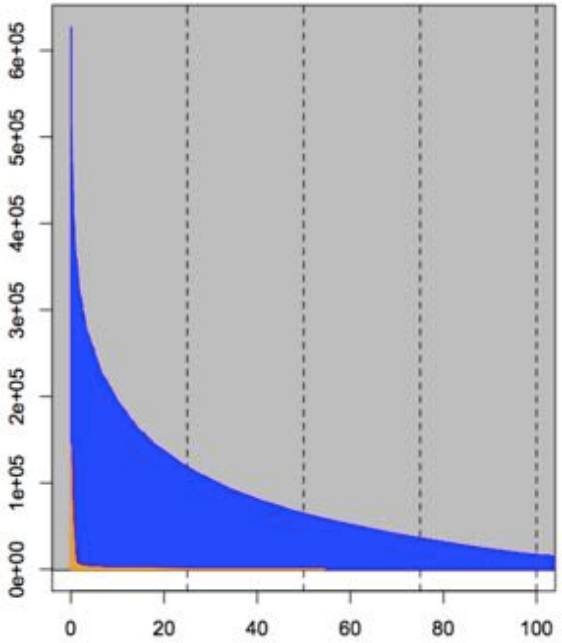
Selected Genomes from 2015

Saccharomyces cerevisiae
ONT + Illumina



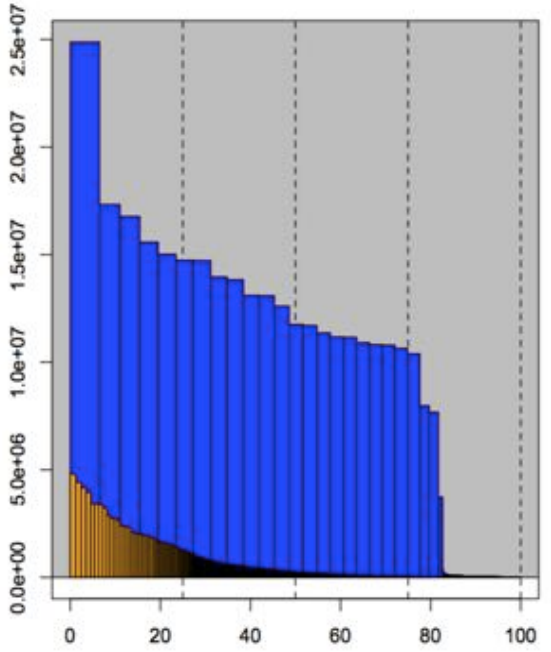
Goodwin *et al.* (2015)
Genome Research.
doi: 10.1101/gr.191395.115

Macrostomum lignano
PacBio



Wasik *et al.* (2015)
PNAS.
doi: 10.1073/pnas.1516718112

Ananas comosus
Illumina + Moleculo + PacBio



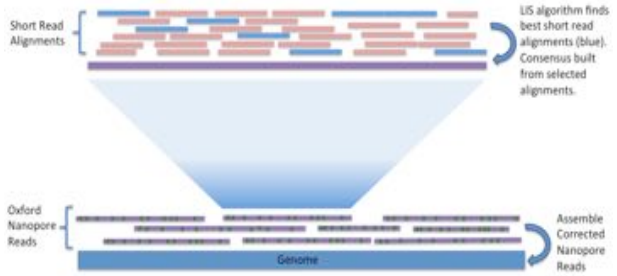
Ming *et al.* (2015)
Nature Genetics.
doi: doi:10.1038/ng.3435

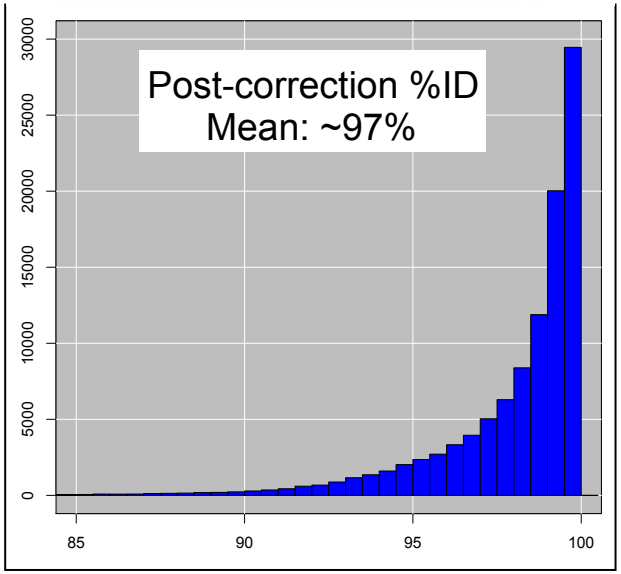
#1MbpCtgClub

NanoCorr: Nanopore-Illumina Hybrid Error Correction

http://schatzlab.cshl.edu/data/nanocorr/

1. BLAST Miseq reads to all raw Oxford Nanopore reads
2. Select non-repetitive alignments
   - First pass scans to remove "contained" alignments
   - Second pass uses Dynamic Programming (LIS) to select an optimal set of high-identity alignments
3. Compute consensus of each Oxford Nanopore read
   - State machine of most commonly observed base at each position in read

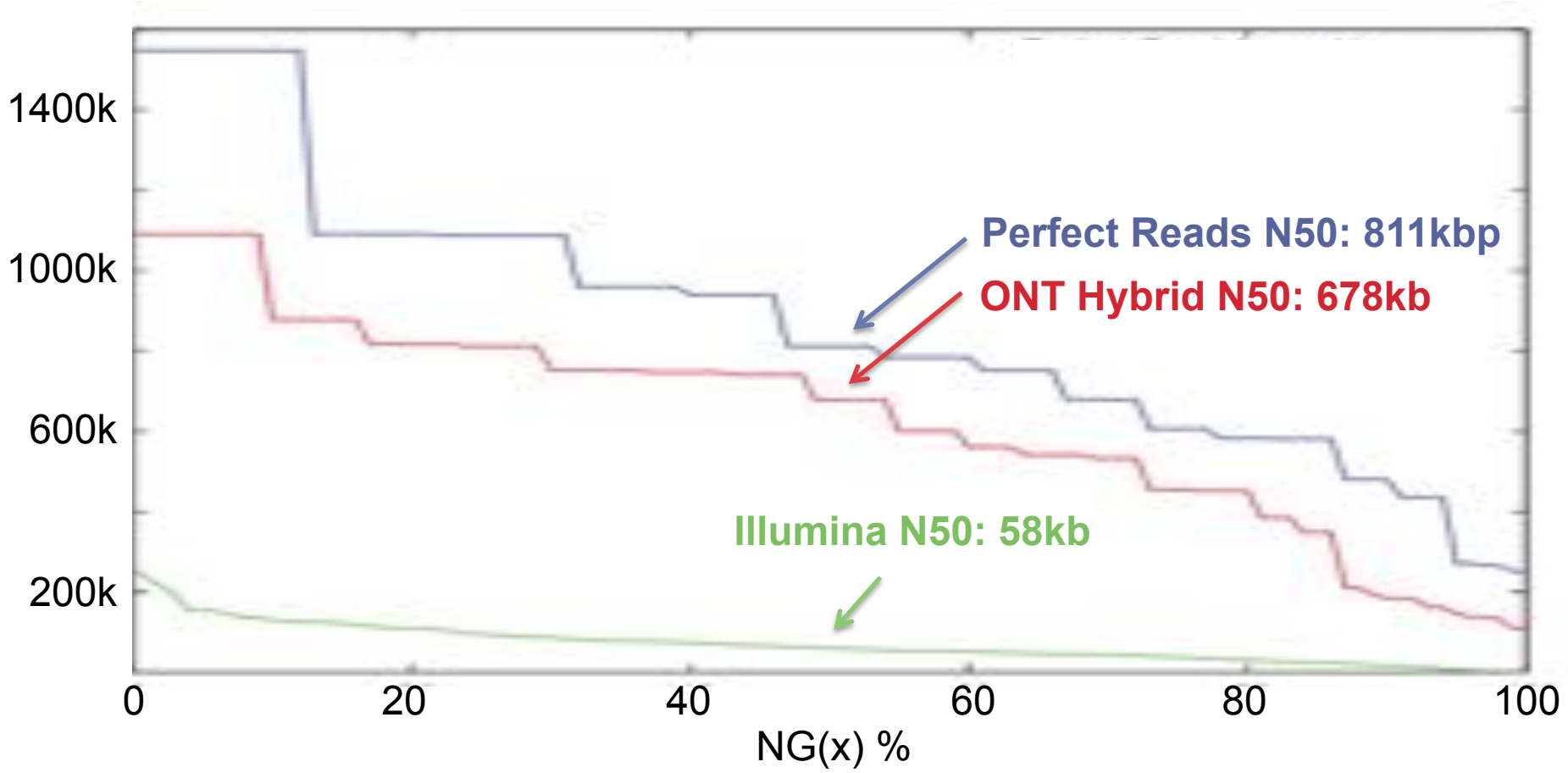




Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Goodwin, S *et al.* (2015) *Genome Research*. doi: 10.1101/gr.191395.115

NanoCorr Yeast Assembly

Contiguity: Idealized and Realized Contig Length

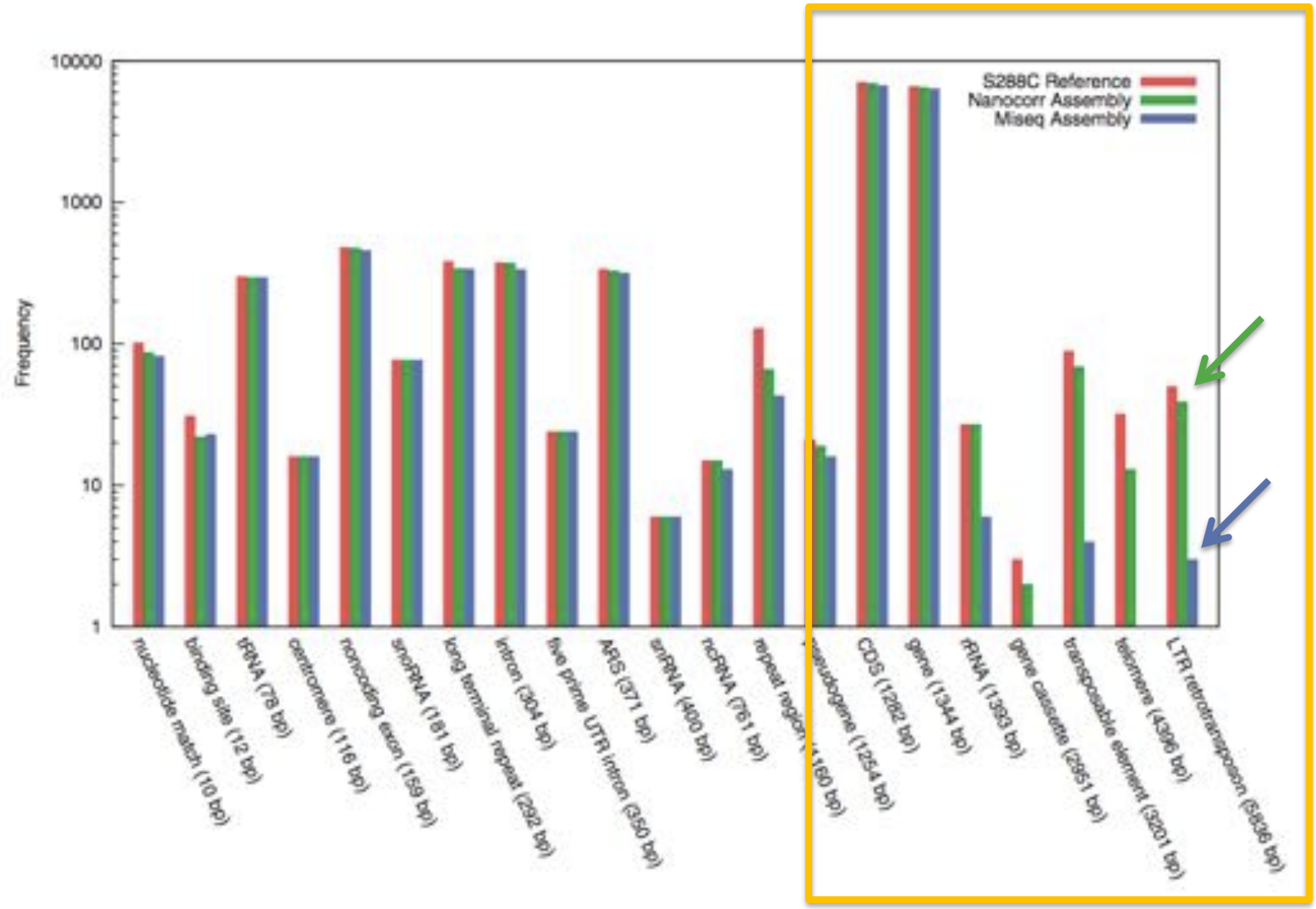


Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome

Goodwin, S *et al.* (2015) *Genome Research*. doi: 10.1101/gr.191395.115

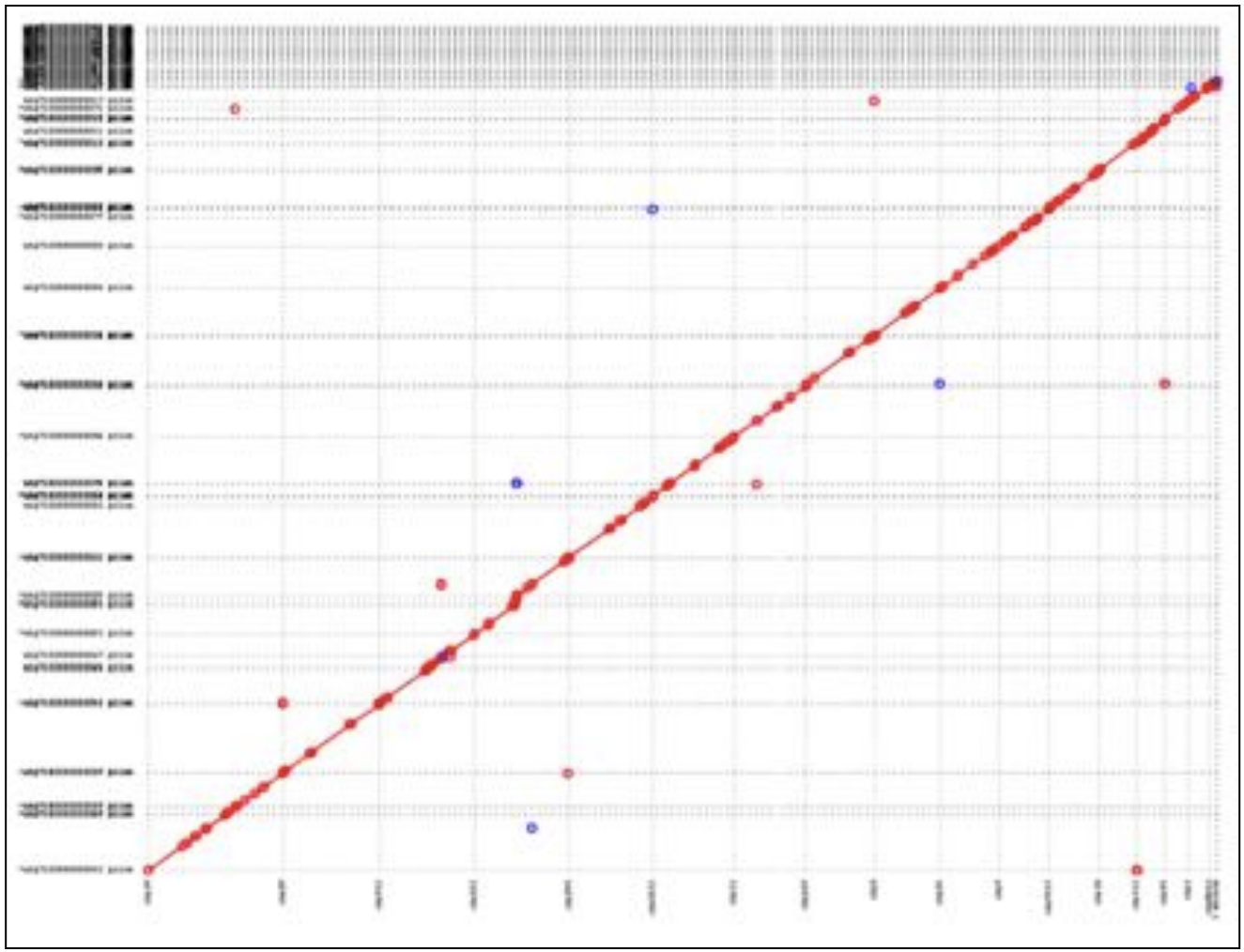
NanoCorr Yeast Assembly

Completeness: Genomic Feature Analysis



NanoCorr Yeast Assembly

Correctness: Structural errors + Sequence fidelity



Structural Analysis:
Most structural differences genuine biological variants between S228C and W303.

Sequence Fidelity:
Raw accuracy: 99.78%
Pilon polishing: 99.88%
Gene accuracy: 99.90%

Most residual errors present in homopolymer sequences

What should we expect from an assembly?

The Three C's of Genome Quality

1. Contiguity

How does read length and sequence coverage impact contig lengths?

2. Completeness

How successful will we be reconstructing genes and other features?

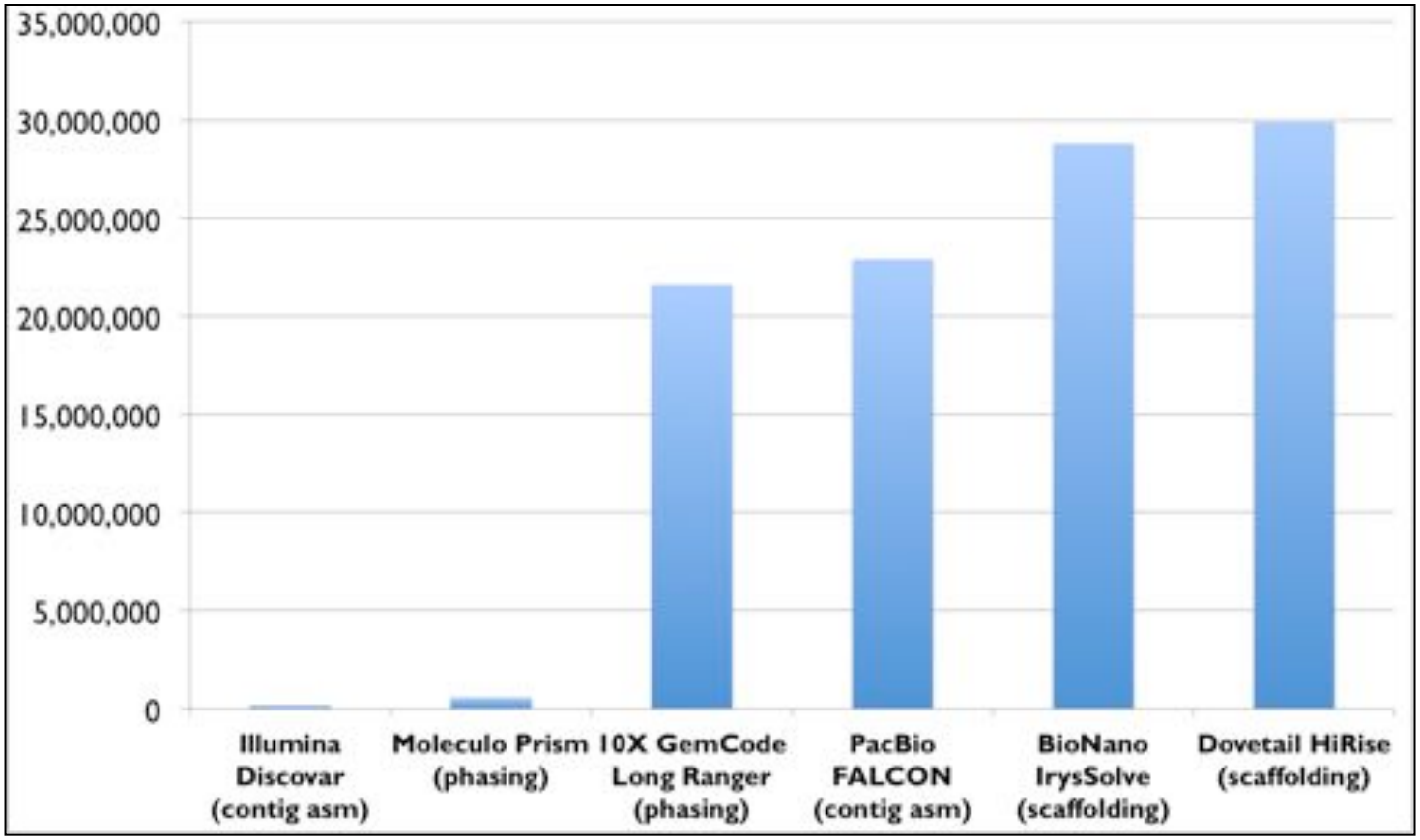
3. Correctness

Does the assembled sequence faithfully represent the genome?

Data Sources:

- *Meta-analysis of available 2nd and 3rd generation assemblies*
- *Historical analysis to the improvements to the human genome*
- *De novo assemblies of idealized sequencing data*

Human Analysis N50s*

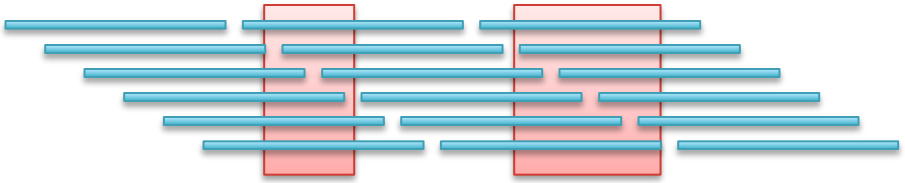


Technology	Application	N50	Sample	Citation
Illumina Discovar	contig asm	178,000	NA12877	Putnam *et al.* (2015) arXiv:1502.05331
Moleculo Prism	phasing	563,801	NA12878	Kuleshov *et al.* (2014) Nature BioTech. doi:10.1038/nbt.2833
10X GemCode Long Ranger	phasing	21,600,000	GIAB	Zook *et al.* (2015) bioRxiv. doi: http://dx.doi.org/10.1101/026468
PacBio FALCON	contig asm	22,900,000	JCV-1	Jason Chin, PAG2016
BioNano IrysSolve	scaffold	28,800,000	NA12878	Pendleton *et al.* (2015) Nature Methods. doi:10.1038/nmeth.3454
Dovetail HiRise	scaffold	29,900,000	NA12878	Putnam *et al.* (2015) arXiv:1502.05331

*Cross analysis of different applications

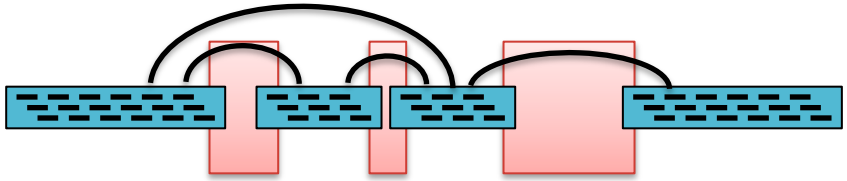
3rd Generation Sequencing Applications

a) De novo Contig Assembly



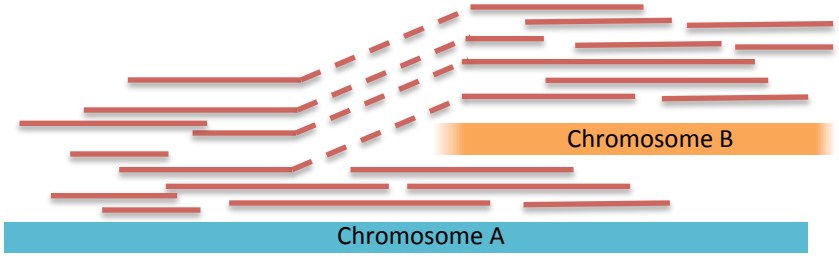
Reconstruct the genome sequence directly from the sequenced reads (blue). Longer reads will span more repetitive elements (red), and produce longer contigs.

b) Chromosome Scaffolding



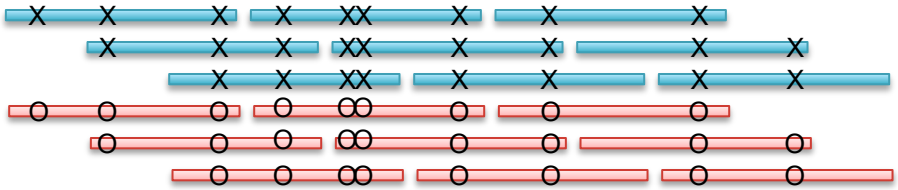
Order and orient contigs (blue) assembled from overlapping reads (black) into longer pseudo-molecules. Longer spans are more likely to connect distantly spaced contigs, especially those separated by long repeats (red).

c) Structural Variation Analysis



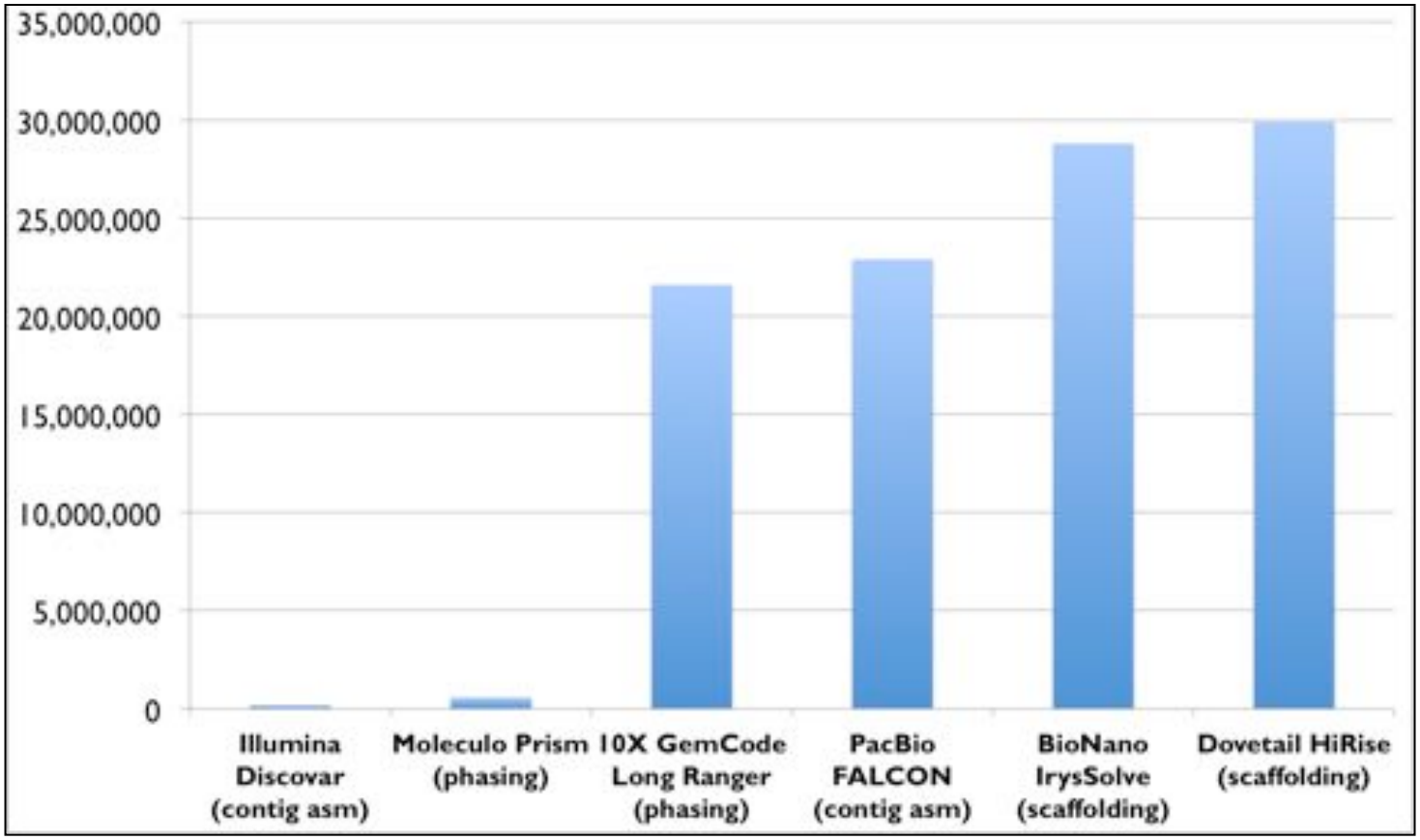
Identify reads/spans (red) that map to different chromosomes or discordantly within one. The longer the read/span, the more likely to capture the SV, and will have improved mappability to resolve SVs in repetitive element.

d) Haplotype Phasing



Link heterozygous variants (X/O) into phased sequences representing the original maternal (red) and paternal (blue) chromosomes. Longer reads and longer spans will be able to connect more distantly spaced variants.

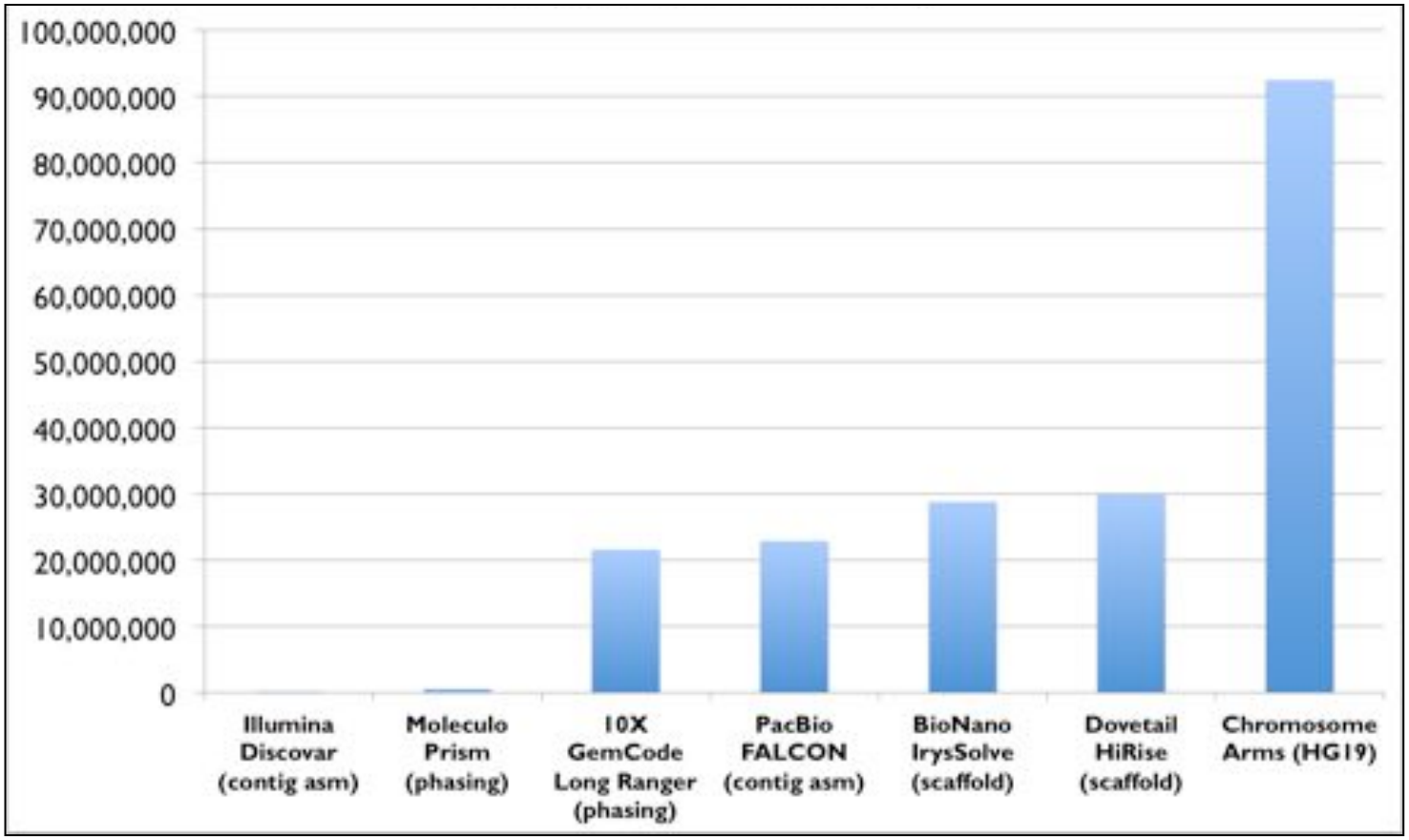
Human Analysis N50s*



Technology	Application	N50	Sample	Citation
Illumina Discovar	contig asm	178,000	NA12877	Putnam *et al.* (2015) arXiv:1502.05331
Moleculo Prism	phasing	563,801	NA12878	Kuleshov *et al.* (2014) Nature BioTech. doi:10.1038/nbt.2833
10X GemCode Long Ranger	phasing	21,600,000	GIAB	Zook *et al.* (2015) bioRxiv. doi: http://dx.doi.org/10.1101/026468
PacBio FALCON	contig asm	22,900,000	JCV-1	Jason Chin, PAG2016
BioNano IrysSolve	scaffold	28,800,000	NA12878	Pendleton *et al.* (2015) Nature Methods. doi:10.1038/nmeth.3454
Dovetail HiRise	scaffold	29,900,000	NA12878	Putnam *et al.* (2015) arXiv:1502.05331

*Cross analysis of different applications

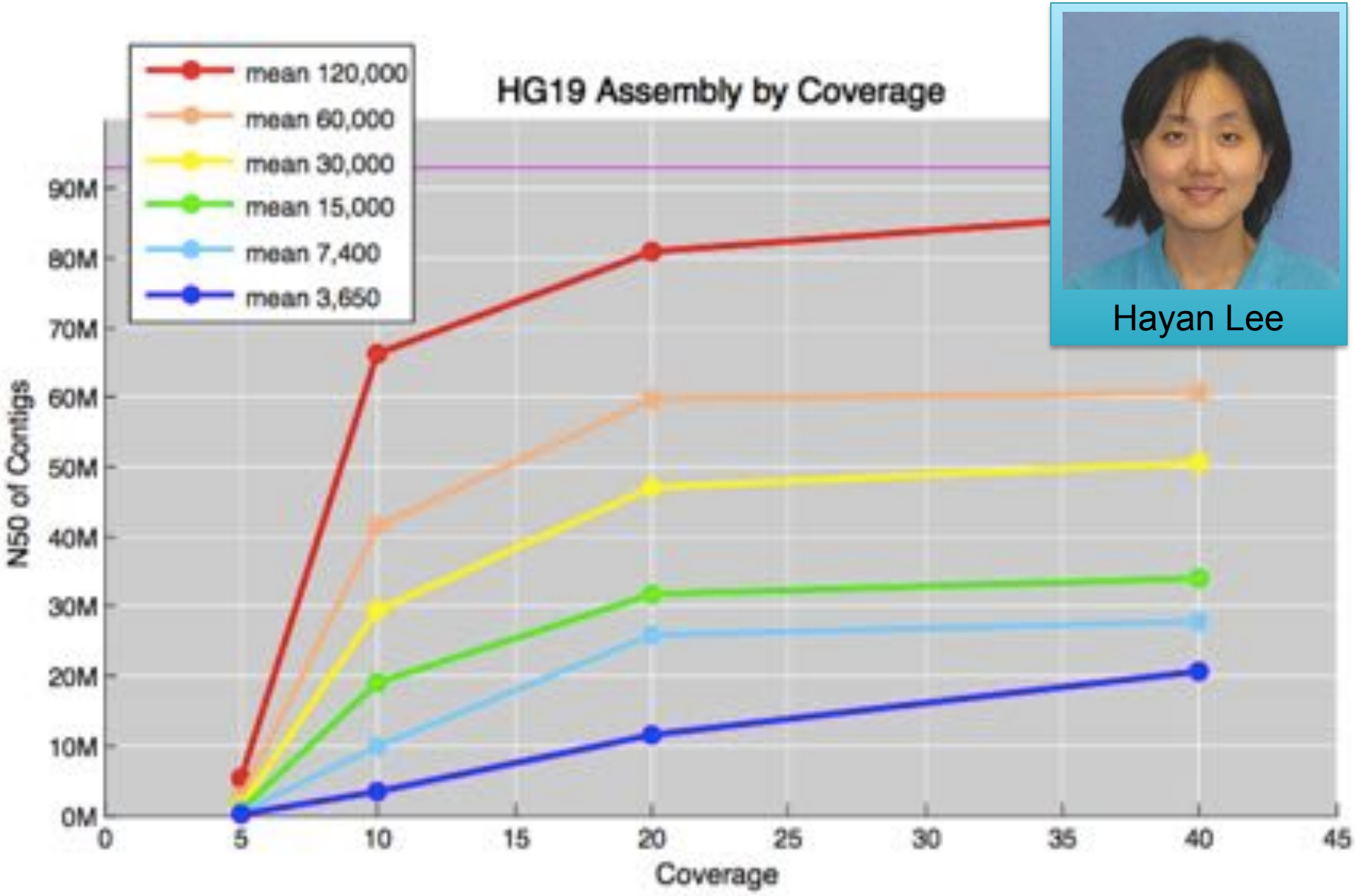
Human Analysis N50s*



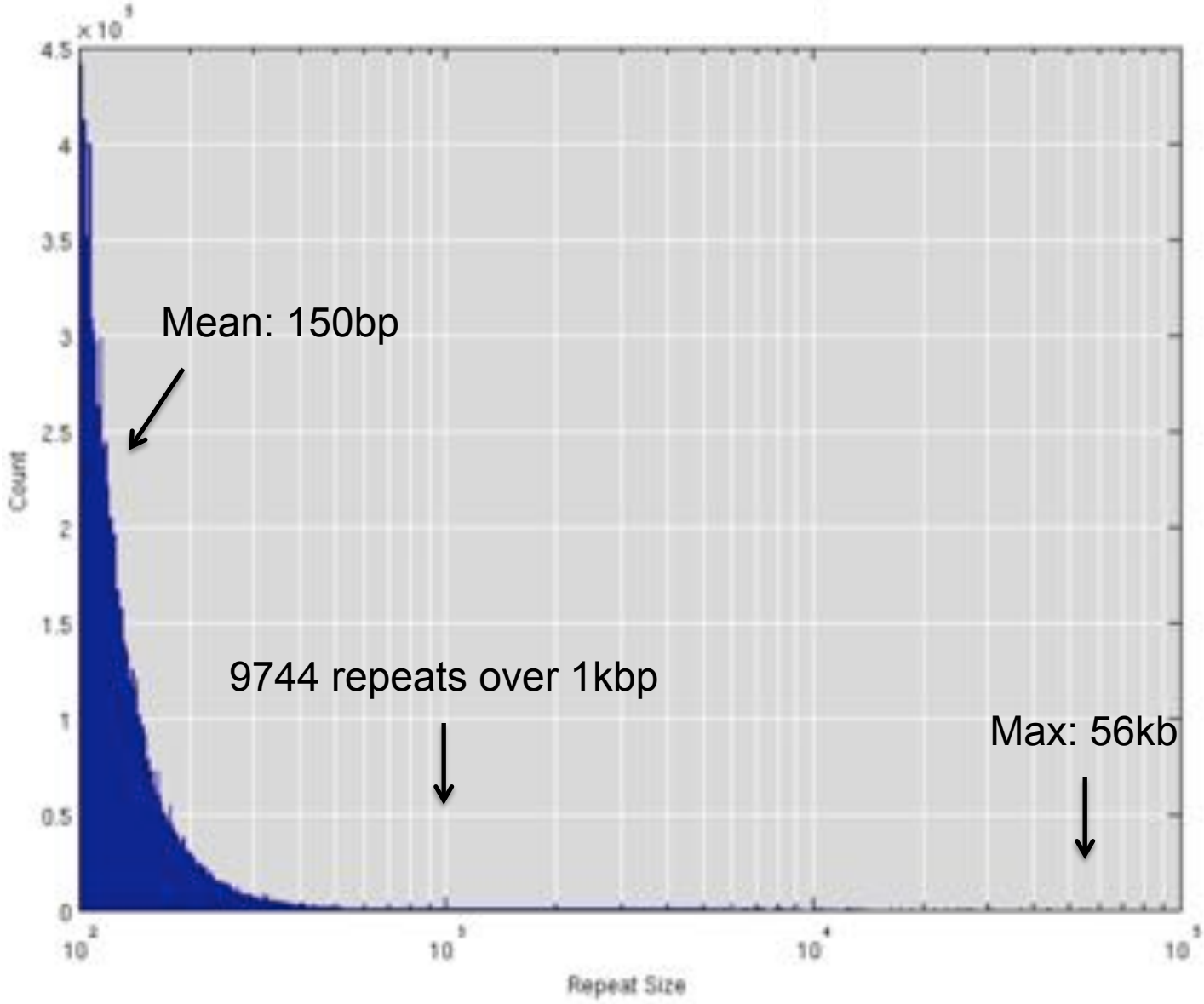
Technology	Application	N50	Sample	Citation
Illumina Discovar	contig asm	178,000	NA12877	Putnam *et al.* (2015) arXiv:1502.05331
Moleculo Prism	phasing	563,801	NA12878	Kuleshov *et al.* (2014) Nature BioTech. doi:10.1038/nbt.2833
10X GemCode Long Ranger	phasing	21,600,000	GIAB	Zook *et al.* (2015) bioRxiv. doi: http://dx.doi.org/10.1101/026468
PacBio FALCON	contig asm	22,900,000	JCV-1	Jason Chin, PAG2016
BioNano IrysSolve	scaffold	28,800,000	NA12878	Pendleton *et al.* (2015) Nature Methods. doi:10.1038/nmeth.3454
Dovetail HiRise	scaffold	29,900,000	NA12878	Putnam *et al.* (2015) arXiv:1502.05331

*Cross analysis of different applications

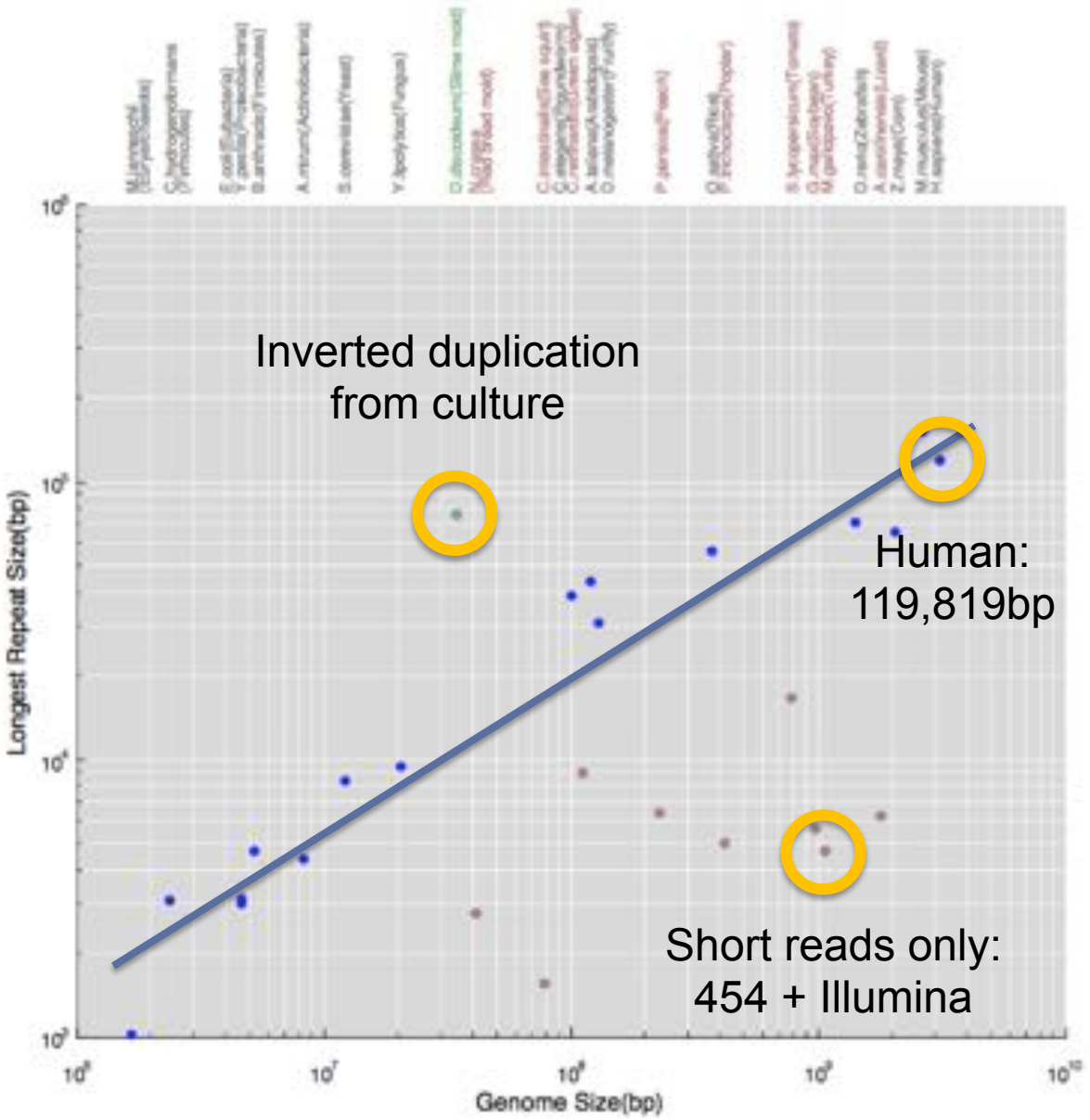
Idealized Human Assemblies



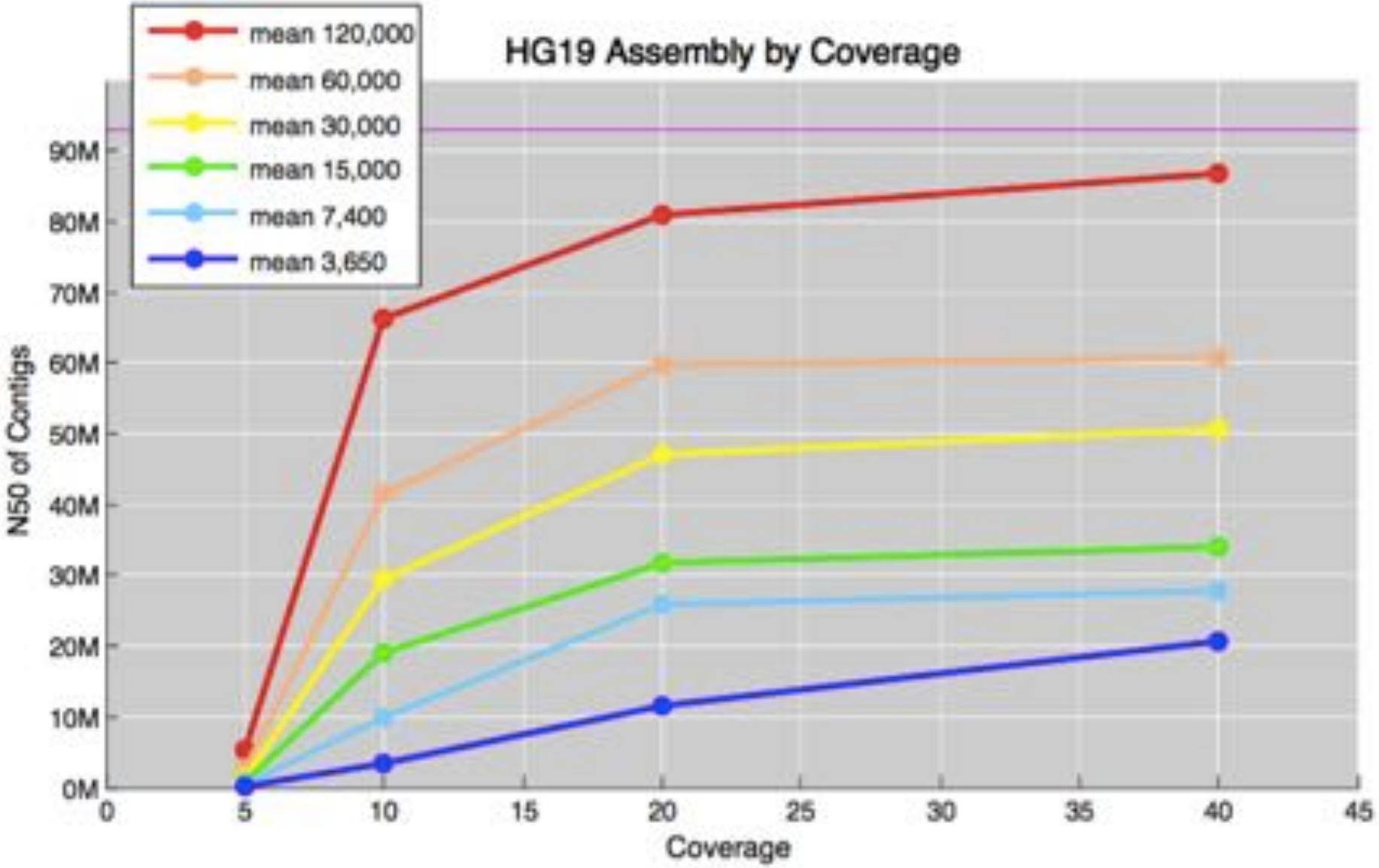
Perfect Repeats in the Rice Genome



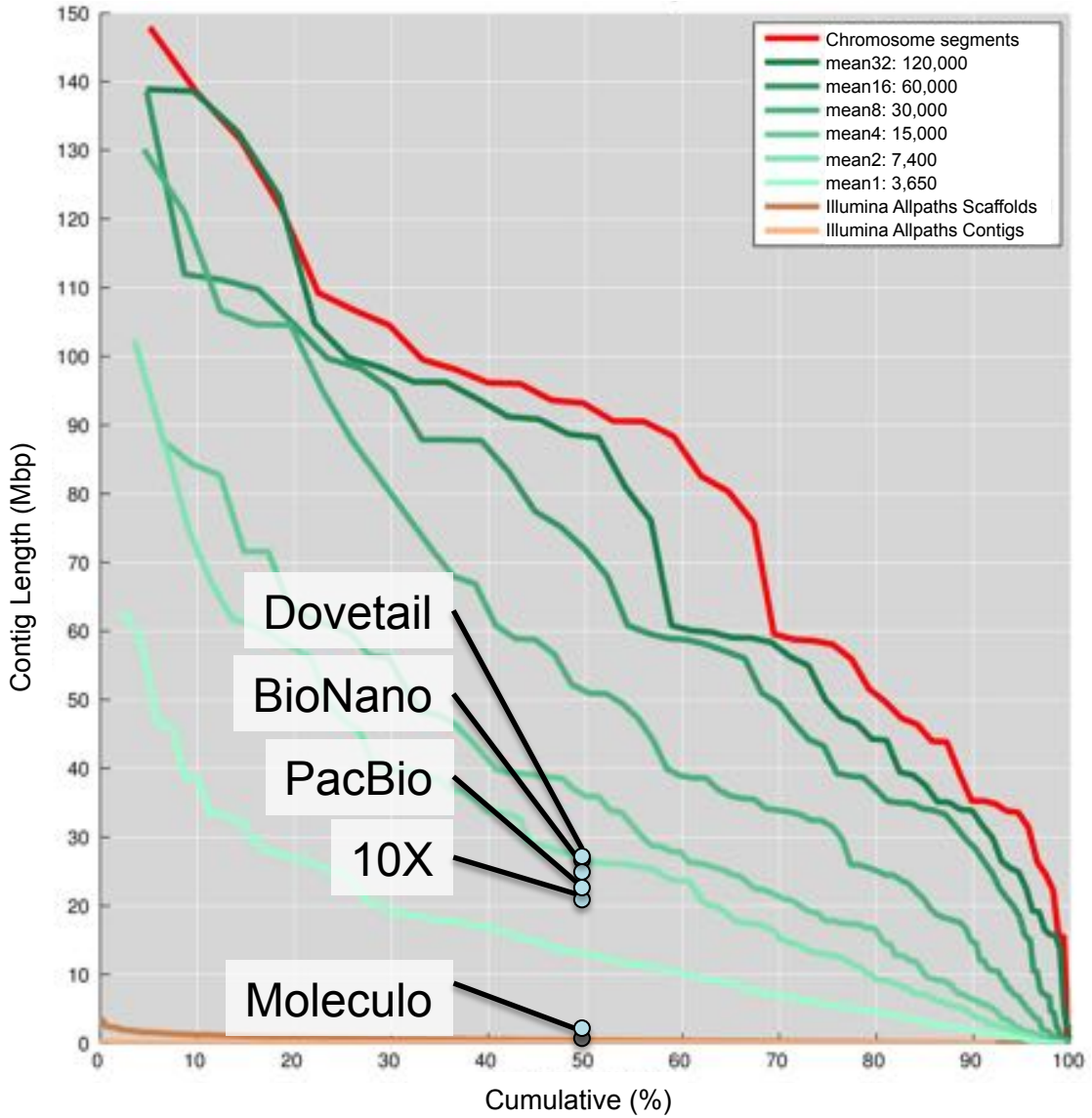
Perfect Repeats Across the Tree of Life



Idealized Human Assemblies



De novo human assemblies



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 simulated reads
- Orange: Results of Illumina/ ALLPATHS assemblies

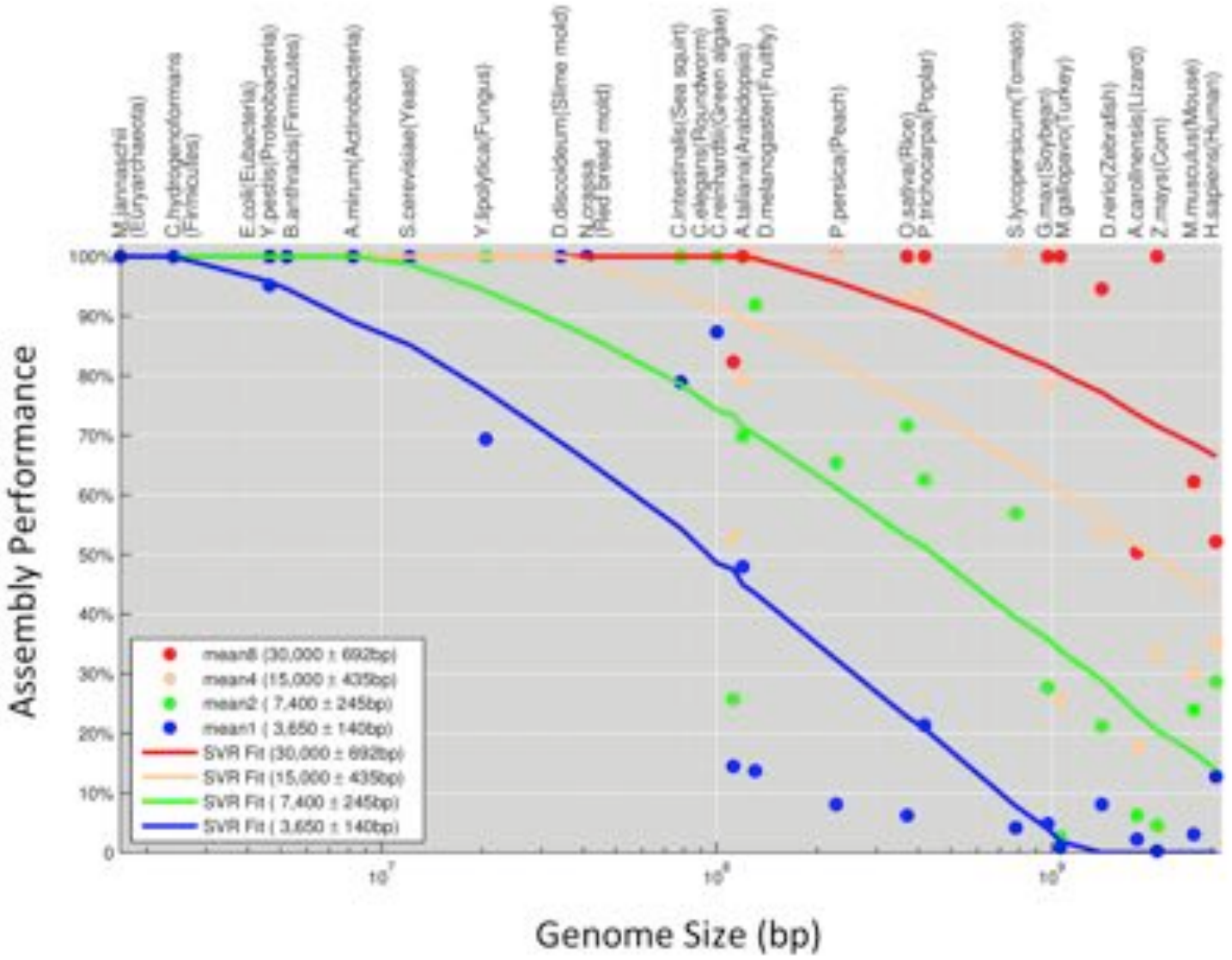
Lengths selected to represent idealized biotechnologies:

- mean1-2: Moleculo/PacBio/ONT
- mean2-4: ~10x / Chromatin
- mean16-32: ~Optical mapping

(log-normal with increasing means)

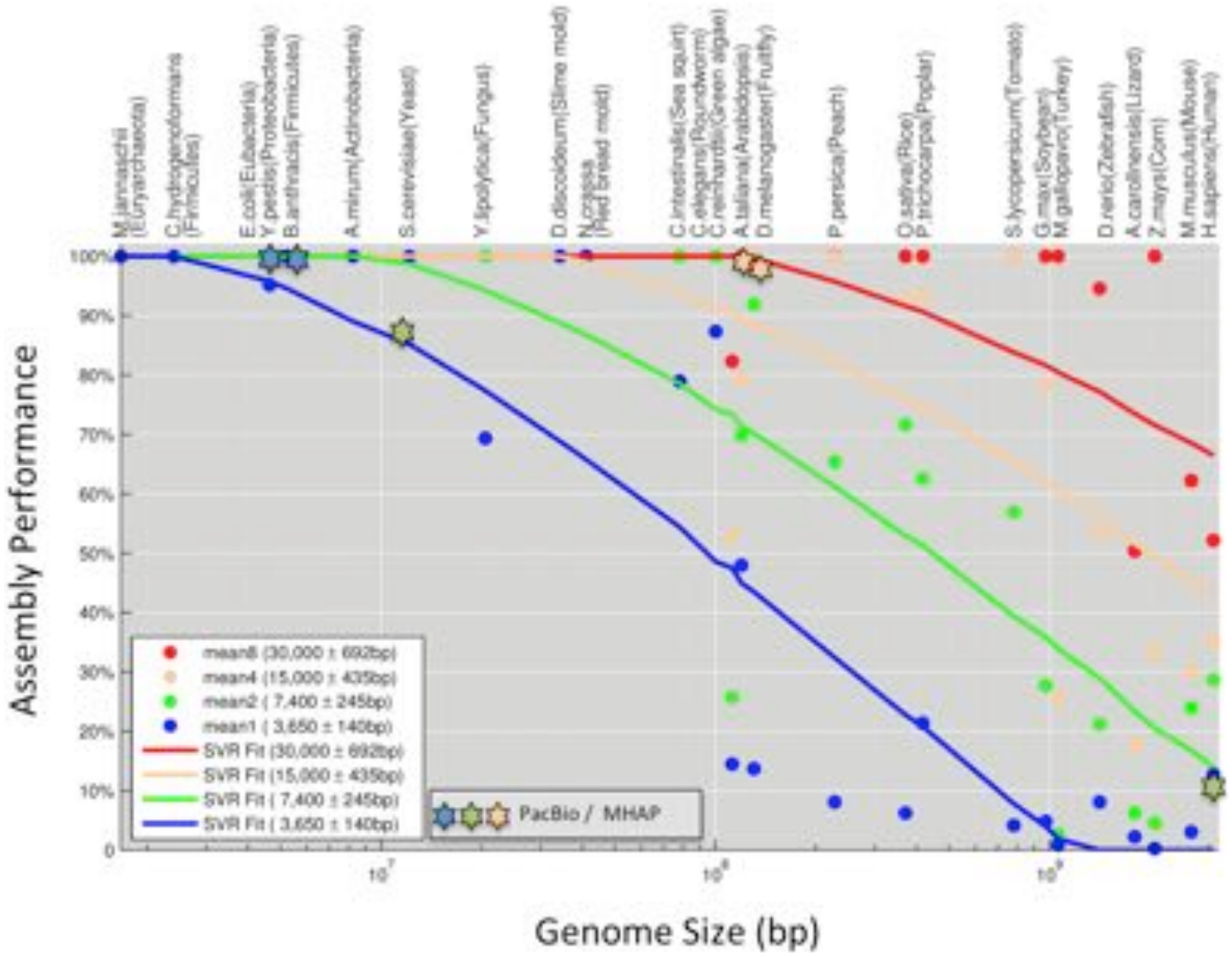
Assembly Contiguity

How long will the contigs be using reads/spans of different lengths?



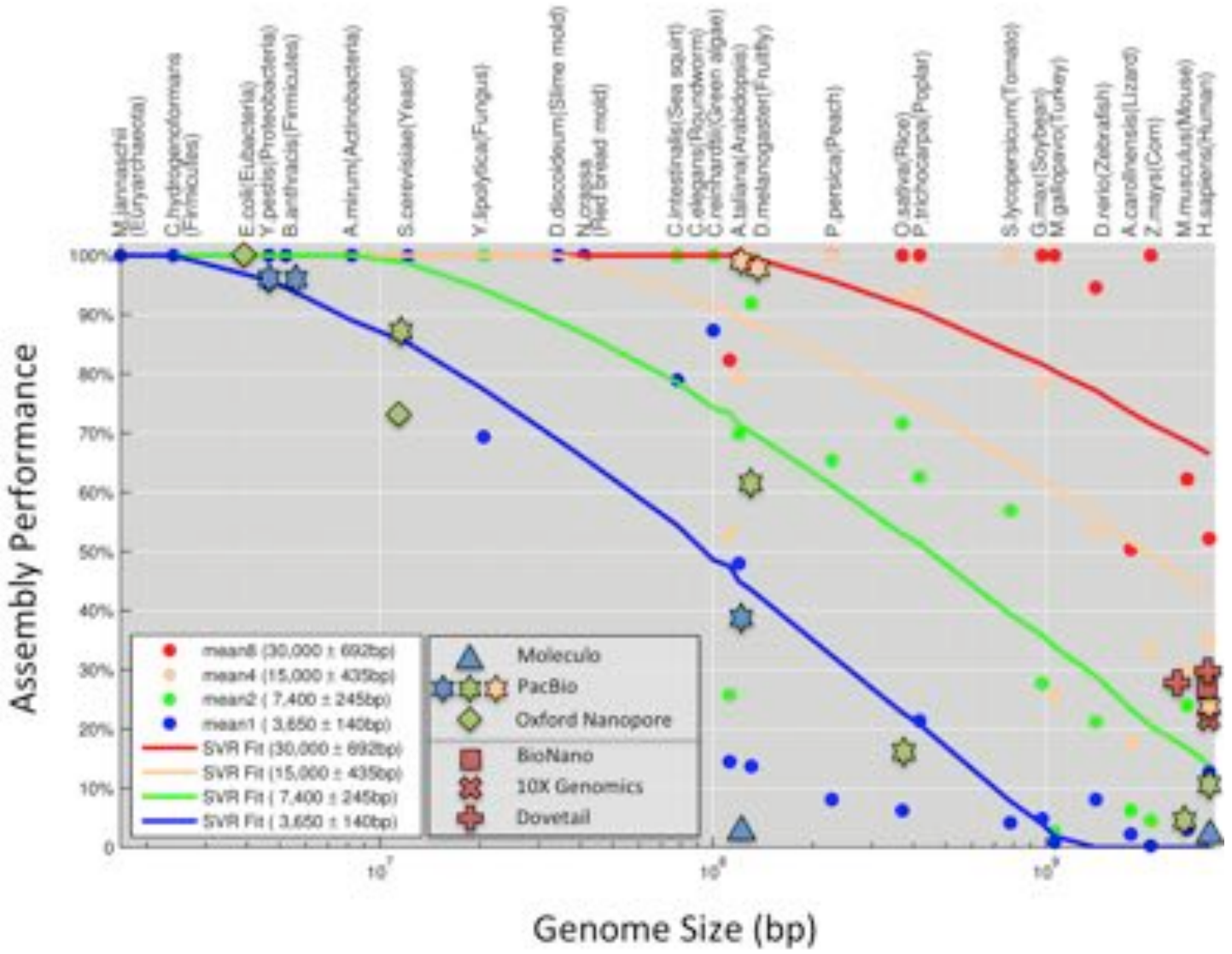
Assembly Contiguity

How long will the contigs be using reads/spans of different lengths?



Assembly Contiguity

How long will the contigs be using reads/spans of different lengths?



Assembly Contiguity

How

Assembly Performance

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

* corresponding author: mschatz@cshl.edu

Abstract

Several new 3rd generation long-range DNA sequencing and mapping technologies have recently become available that are creating a resurgence in high quality genome sequencing. Unlike their 2nd generation, short-read counterparts that can resolve a few hundred base-pairs, the new technologies routinely sequence 10,000 bp reads or map 100,000 bp molecules. The greater lengths are being used to enhance a number of important problems in genomics and medicine, including *de novo* genome assembly, structural variation analysis, and haplotype phasing. Here we discuss the capabilities of the technologies, and show how they will improve the "3Cs of Genomics": the contiguity, completeness, and correctness of genome sequencing. We also propose a model using support vector regression that predicts assembly performance using different read lengths or coverage that can be used for evaluating technologies. Overall, we anticipate these will unlock the genomic "dark matter", and provide many new insights into evolution.

Summary & Predictions

The Three C's of Genome Quality

1. Contiguity

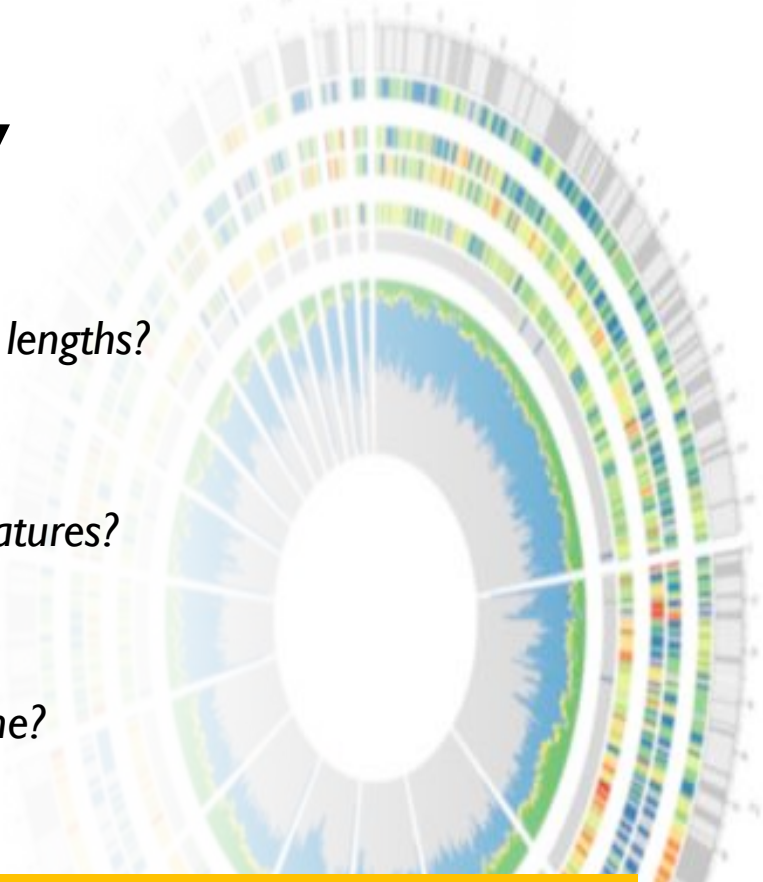
How does read length and sequence coverage impact contig lengths?

2. Completeness

How successful will we be reconstructing genes and other features?

3. Correctness

Does the assembled sequence faithfully represent the genome?



Predictions for 2016

- First 100 genomes will join the #1MbpCtgClub
- Enter the era of complete chromosome-level scaffolding
- First glimpses of the true complexity of chromosome evolution

Acknowledgements

Schatz Lab
Rahul Amin
Han Fang
Tyler Gavin
James Gurtowski
Hayan Lee
Zak Lemmon
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

SBU
Skiena Lab
Patro Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NYU
Jane Carlton
Elodie Ghedin

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

ALFRED P. SLOAN FOUNDATION

Thank you

http://schatzlab.cshl.edu

@mike_schatz / PAGXXIV