

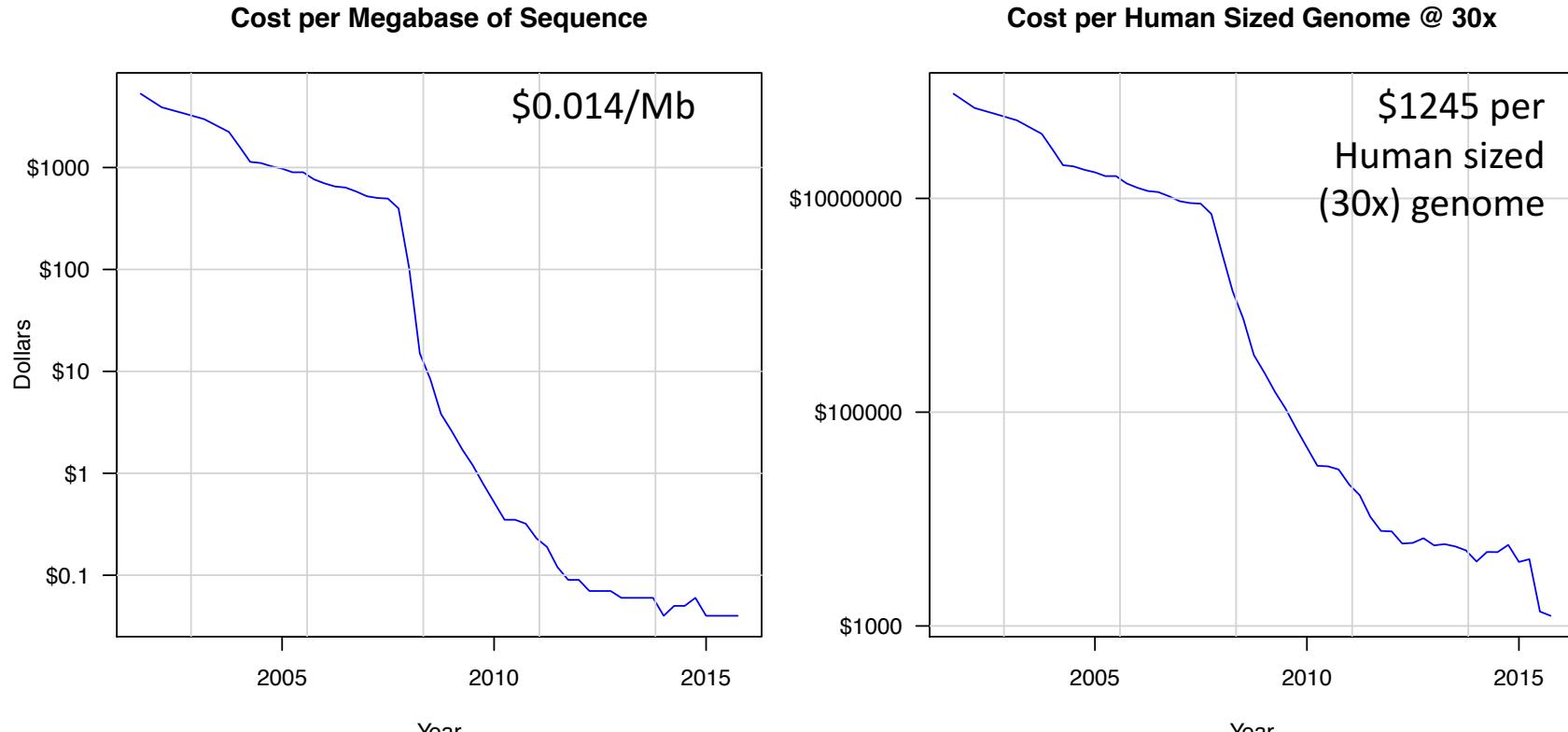
# Introduction to genomics / technologies

Dr. Matthew L. Settles

Genome Center  
University of California, Davis  
[settles@ucdavis.edu](mailto:settles@ucdavis.edu)

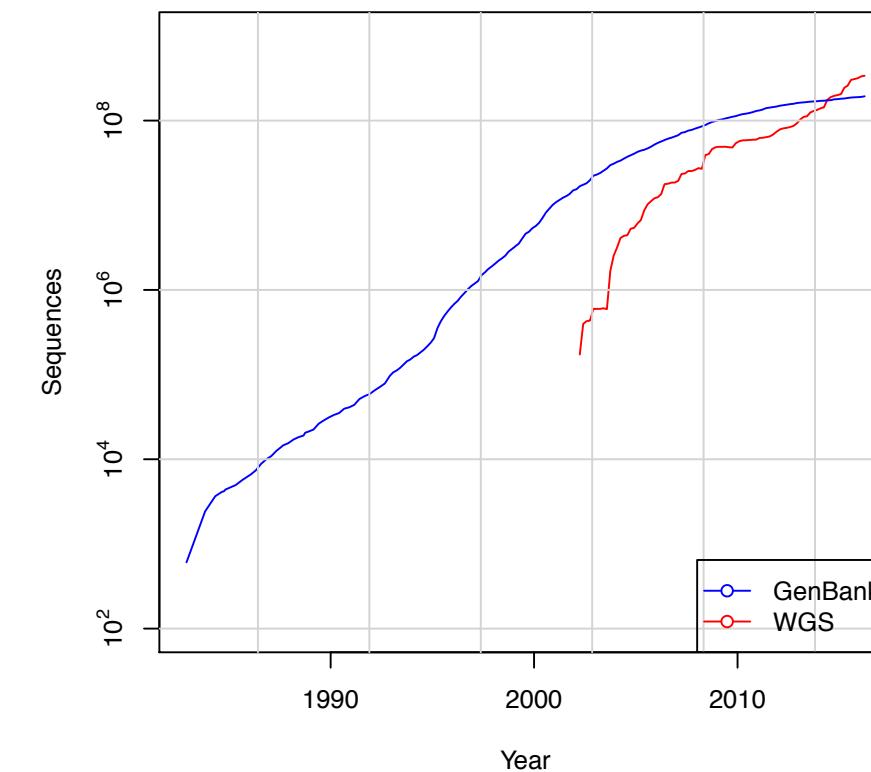
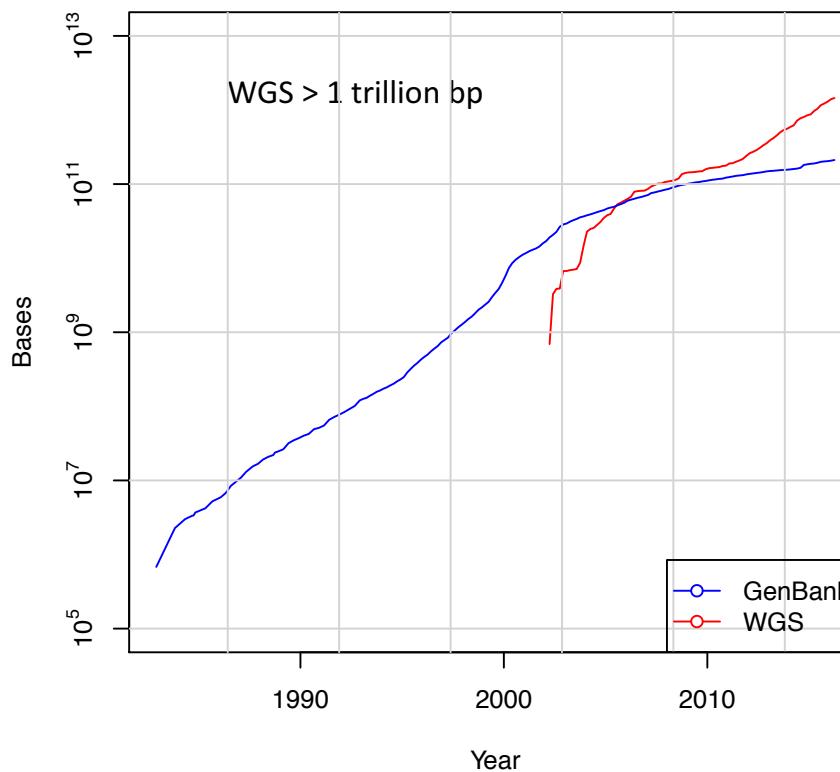
# Sequencing Costs

October 2016



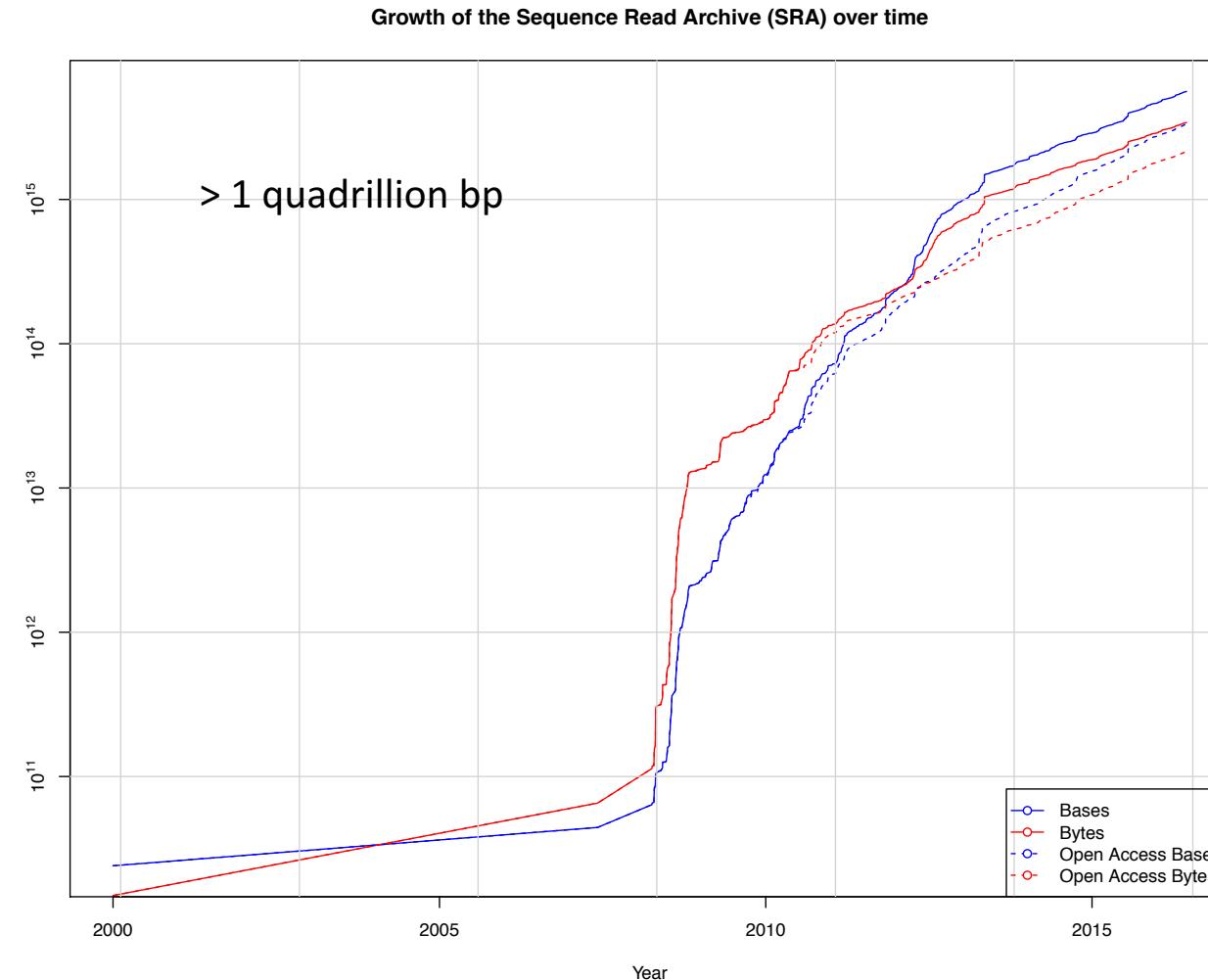
- Includes: labor, administration, management, utilities, reagents, consumables, instruments (amortized over 3 years), informatics related to sequence productions, submission, indirect costs.
- <http://www.genome.gov/sequencingcosts/>

# Growth in Public Sequence Database

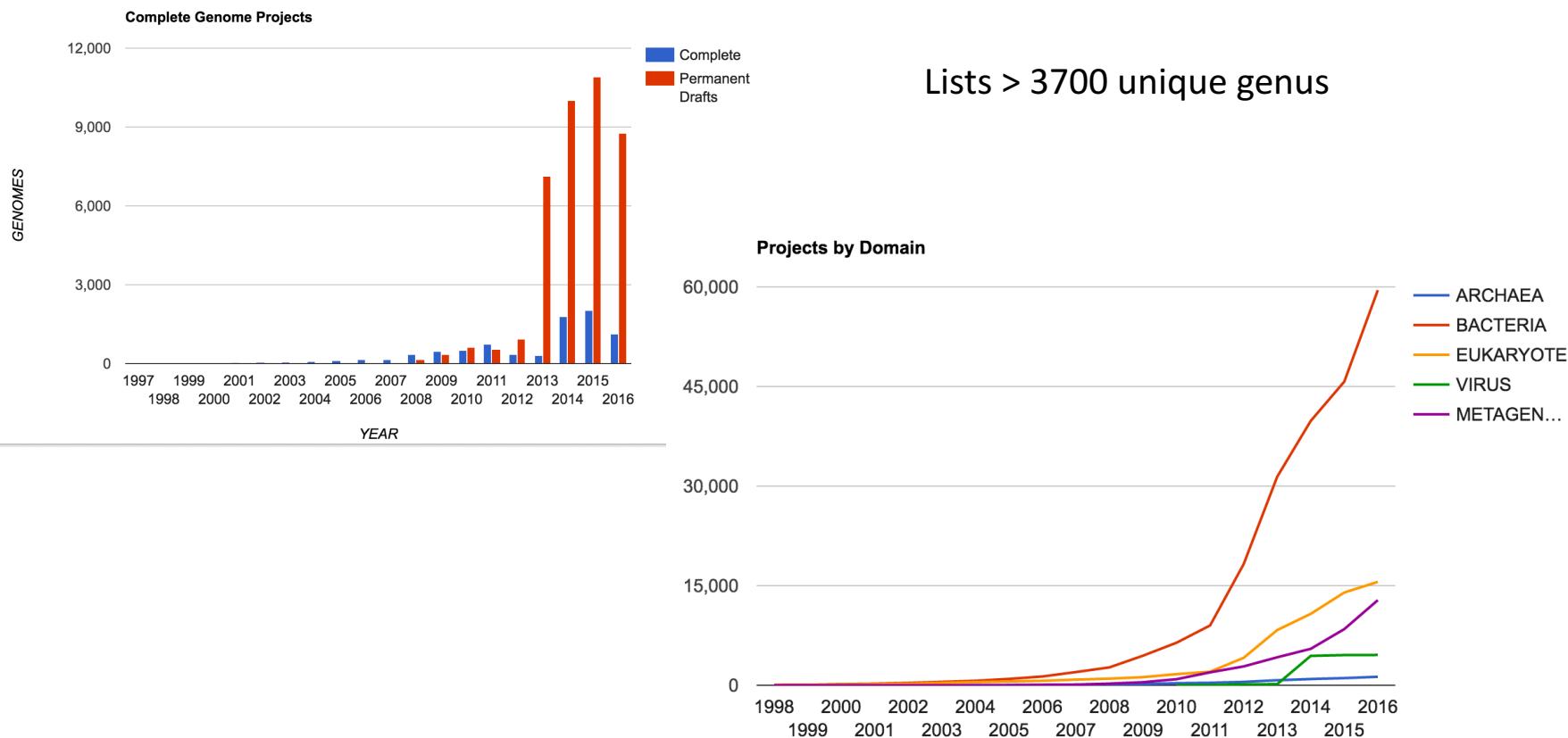


- <http://www.ncbi.nlm.nih.gov/genbank/statistics>

# Short Read Archive (SRA)



# Increase in Genome Sequencing Projects



- JGI – Genomes Online Database (GOLD)
- 67,822 genome sequencing projects

# Brief History

DNA/RNA Sequencing

# Sequencing Platforms

- 1986 - Dye terminator Sanger sequencing, technology dominated until 2005 until “next generation sequencers”, peaking at about 900kb/day



# 'Next' Generation

- 2005 – ‘Next Generation Sequencing’ as Massively parallel sequencing, both throughput and speed advances. The first was the Genome Sequencer (GS) instrument developed by 454 life Sciences (later acquired by Roche), Pyrosequencing 1.5Gb/day

**Discontinued**



# Illumina

- 2006 – The second ‘Next Generation Sequencing’ platform was Solexa (later acquired by Illumina). Now the dominant platform with 75% market share of sequencer and estimated >90% of all bases sequenced are from an Illumina machine, Sequencing by Synthesis > 200Gb/day.



New  
NovaSeq



[Sequencing  
by synthesis](#)

# Complete Genomics

- 2006 – Using DNA nanoball sequencing, has been a leader in Human genome resequencing, having sequenced over 20,000 genomes to date. In 2013 purchased by BGI and is now set to release their first commercial sequencer, the Revolocty. Throughput on par with HiSeq

*NOW DEFUNCT*

Human genome/exomes only.

10,000 Human Genomes per year



# Bench top Sequencers

- Roche 454 Junior
- Life Technologies
- Ion Torrent
- Ion Proton
- Illumina MiSeq



# The ‘Next, Next’ Generation Sequencers (3<sup>rd</sup> Generation)

- 2009 – Single Molecule Read Time sequencing by Pacific Biosystems, most successful third generation sequencing platforms, RSII ~1Gb/zmw, new Pac Bio Seauel ~7Gb/zmw. near 100Kb possible read length.



[SMRT Sequencing](#)

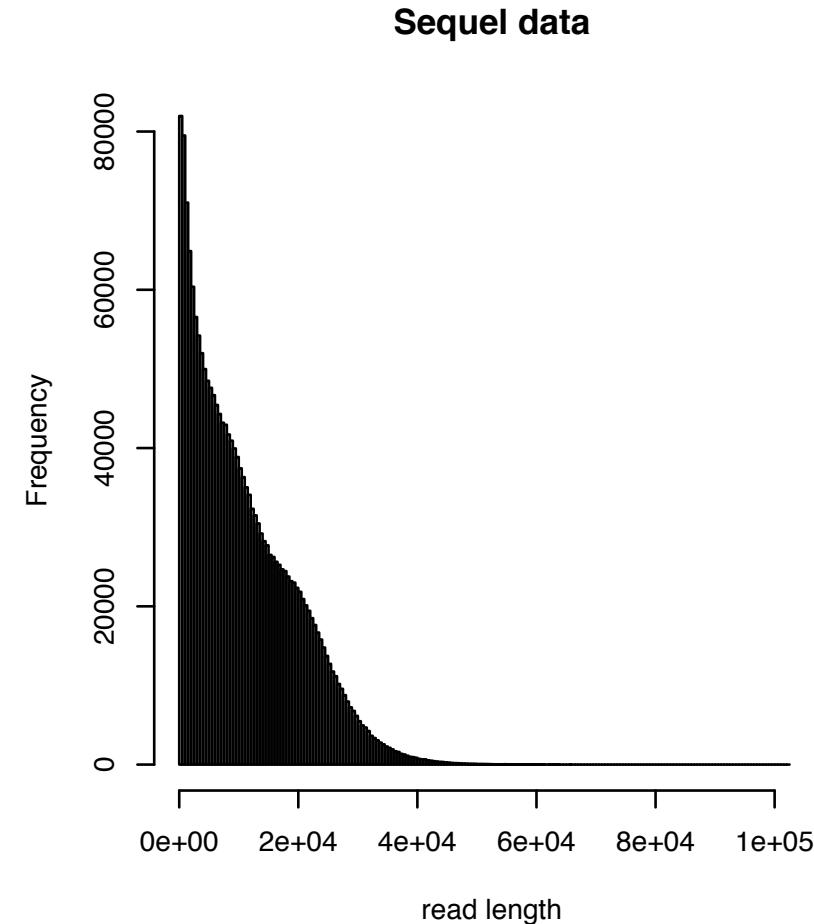
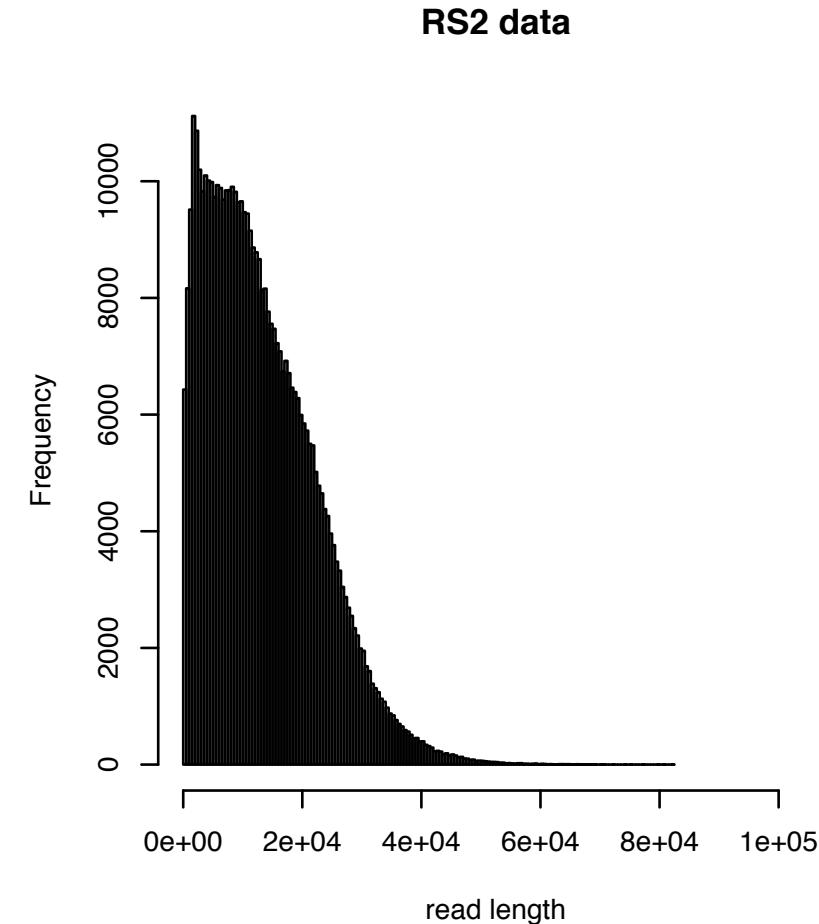


Iso-seq on Pac Bio possible, transcriptome without ‘assembly’

# Pac Bio Advances (RSII vs Sequel)

California Condor data (~1.2Gbp genome) based on 4 SMRT cell in Jan 2017

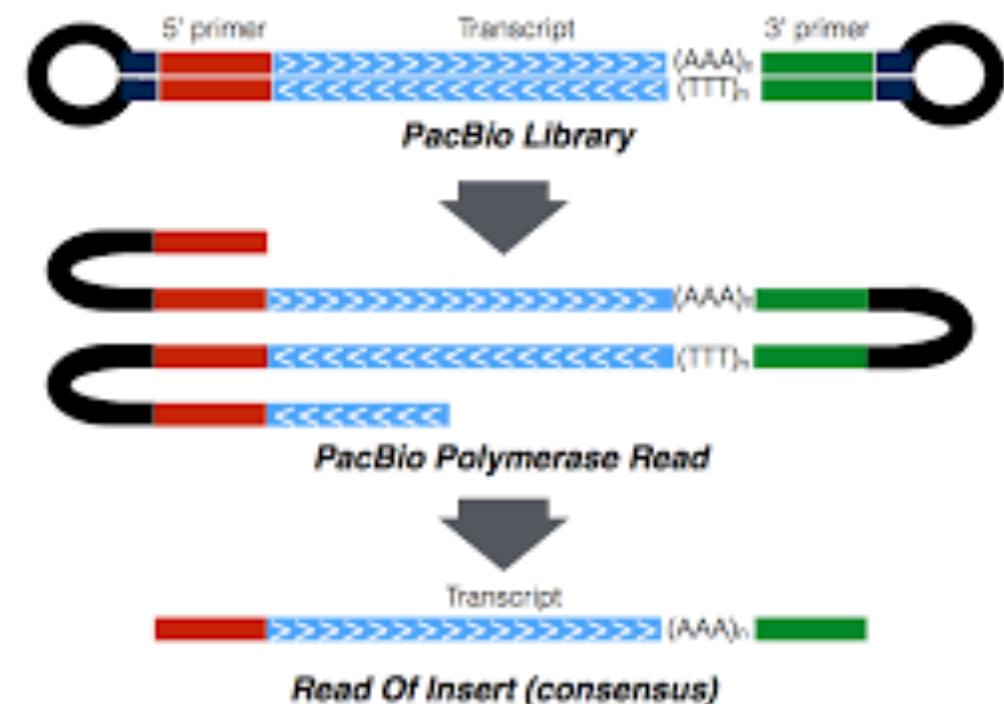
	RS2	Sequel
Read count	448,767	1,947,684
N50	10,426	4,293
Longest Read	82,366	102,310
# reads > 12Kb	217,691	754,157
Coverage > 12Kb	3.64	12.165



# Whole transcripts – Pac bio Iso-seq

**Produce full-length transcripts without assembly**

The isoform sequencing (Iso-Seq) application generates full-length cDNA sequences — from the 5' end of transcripts to the poly-A tail — After Circular consensus sequence (CCS) algorithm produces high quality isoforms.



# Oxford Nanopore

- 2015 – Another 3<sup>rd</sup> generation sequencer, founded in 2005 and currently in beta testing. The sequencer uses nanopore technology developed in the 90's to sequence single molecules. Throughput is about 500Mb per flowcell, capable of near 200kb reads.

**Fun to play with but results  
are highly variable**

[Nanopore Sequencing](#)

FYI: 4<sup>th</sup> generation sequencing is being described as In-situ sequencing

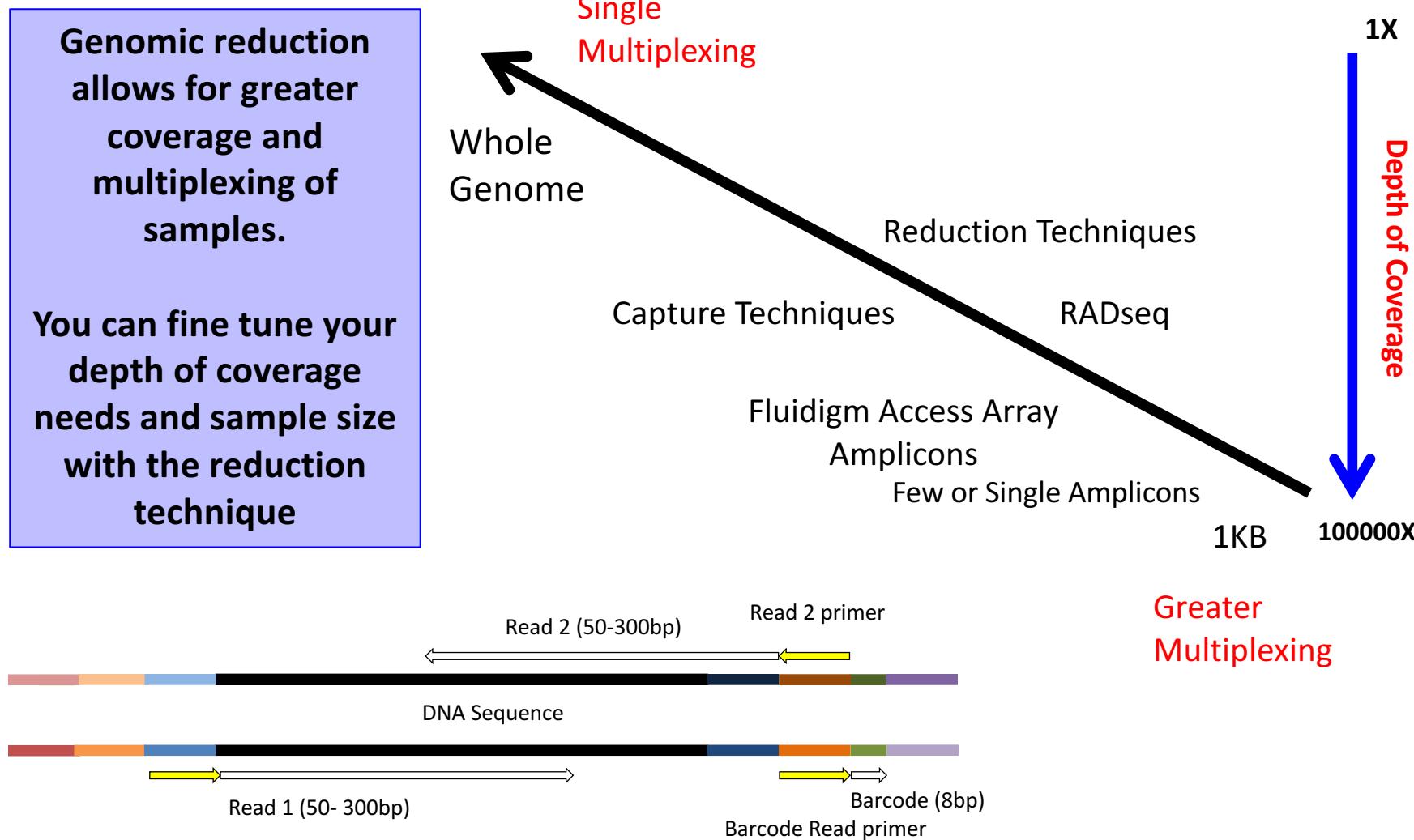


# Complementary Approaches

Illumina	PacBio
Still-imaging of clusters (~1000 clonal molecules)	Movie recordings fluorescence of single molecules
Short reads - 2x300 bp Miseq	Up to 60 kb, N50 23 kb
Repeats are mostly not analyzable	spans retro elements
High output - over 100 Gb per lane	up to 1,3 Gb and 5 Gb per SMRT-cell
High accuracy (< 0.5 %)	Error rate 15 %
Considerable base composition bias	No base composition bias
Very affordable	Costs 5 to 10 times higher (per base)
<i>De novo</i> assemblies results in hundreds of thousands of scaffolds	<i>De novo</i> assemblies results in thousands of scaffolds

**“If you can put adapters on it,  
it can be sequence it!”**

# Flexibility



# Sequencing Libraries

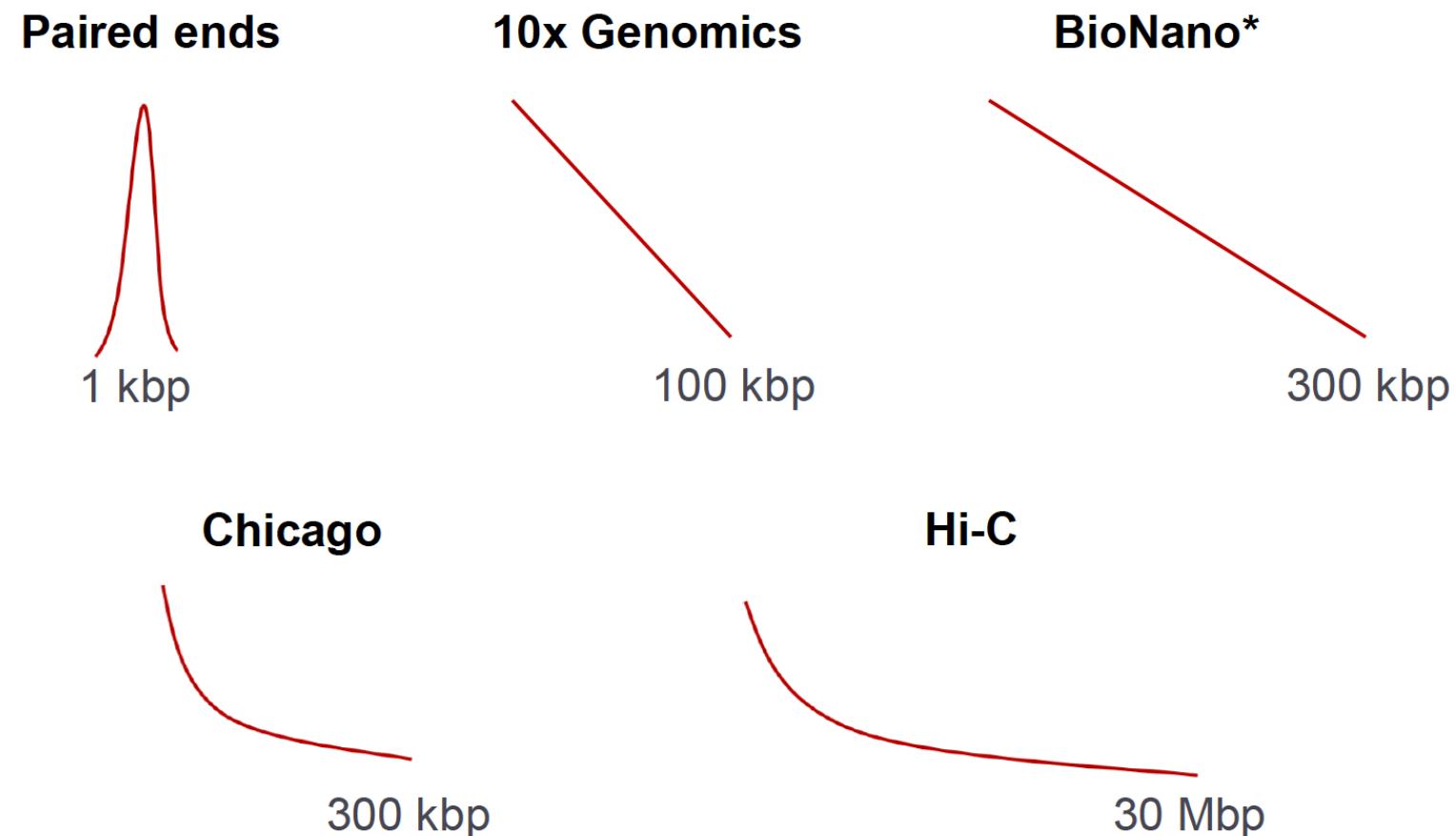
- |             |            |               |
|-------------|------------|---------------|
| • DNA-seq   | DNase-seq  | tagRNA-seq    |
| • RNA-seq   | ATAC-seq   | PAT-seq       |
| • Amplicons | MNase-seq  | Structure-seq |
| • ChIP-seq  | FAIRE-seq  | MPE-seq       |
| • MeDIP-seq | Ribose-seq | STARR-seq     |
| • RAD-seq   | smRNA-seq  | Mod-seq       |
| • ddRAD-seq | mRNA-seq   | BrAD-seq      |
| • Pool-seq  | Tn-seq     | SLAF-seq      |
| • EnD-seq   | QTL-seq    | G&T-seq       |

# comparison of genotyping approaches (Scheben *et al.* 2017)

		Cost per sample <sup>a</sup>	Cost per marker data point <sup>a</sup>	SNP discovery rate	Analysis complexity	Prior genomic knowledge	Preferred population type	Drawbacks	Applications
RADseq	Low	Moderate	Low to moderate	Moderate	No	All	Labour-intensive library preparation; high read depth variation	<i>De novo</i> SNP discovery, genome improvement, genetic mapping	
Elshire GBS	Low	Moderate	Low	Moderate	No	All	High levels of missing data	<i>De novo</i> SNP discovery in simple genomes, genome improvement, genetic mapping	
ddRAD	Low	Moderate	Low to moderate	Moderate	No	All	Sensitive to allele dropout; high-quality sample required	<i>De novo</i> SNP discovery, genome improvement, genetic mapping	
Parental inference WGR	High	Low	High	High	No	Biparental cross	High cost; inference is error-prone	<i>De novo</i> SNP discovery, high-resolution mapping of (complex) plant genomes, genome improvement	
SkimGBS	High	Low	High	High	Yes	Biparental cross	High cost; need for prior genomic information	SNP discovery and high-resolution mapping of (complex) plant genomes, genome improvement	
SNP array	Moderate	High	High	Low	Yes	All	Ascertainment bias; need for prior genomic information	SNP discovery and high-resolution mapping, genetic mapping	
Exome sequencing	Moderate	High	Low	Moderate	Yes	All	Need for prior genomic information	SNP discovery in complex genomes, genetic mapping	
RNA-seq	Moderate	High	Low	Moderate	No	All	Biases in transcript abundances	SNP discovery in complex genomes, genetic mapping, expression analysis	

# Scaffolding, Structure, Haplotypes

# Scaffolding Options

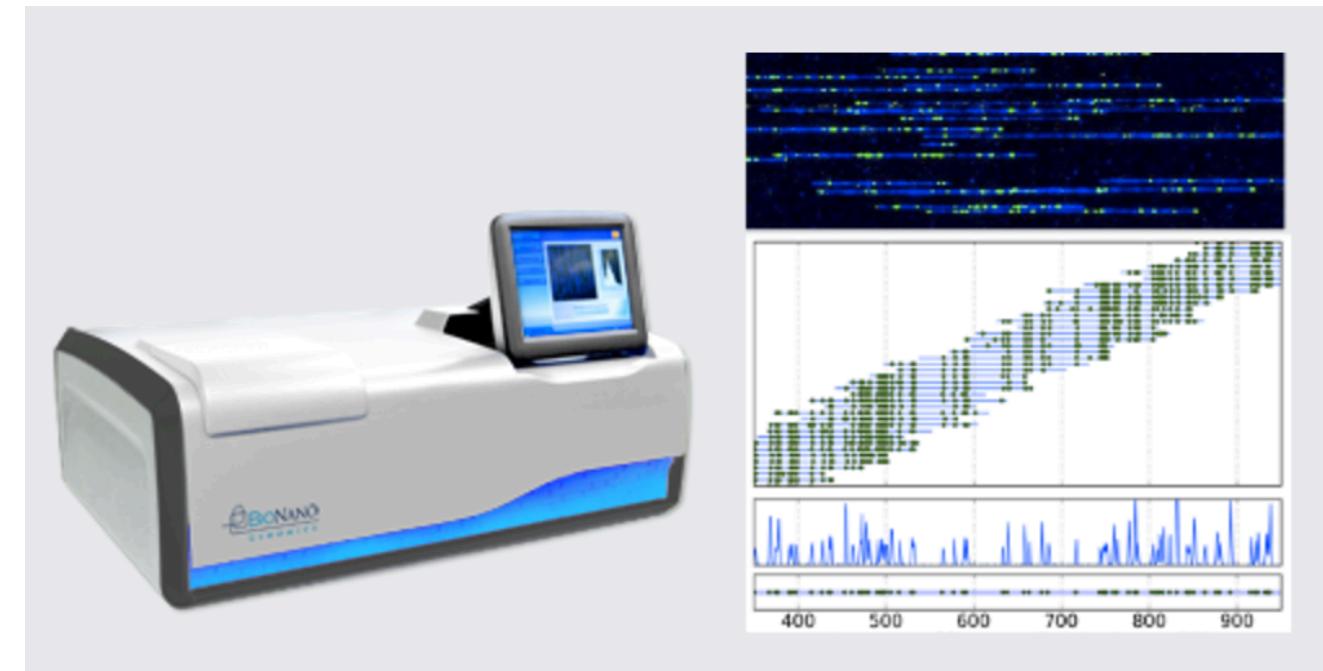


Borrowed from Sergy Koren talk from PacBio Informatics Developer Meeting in Jan 2017

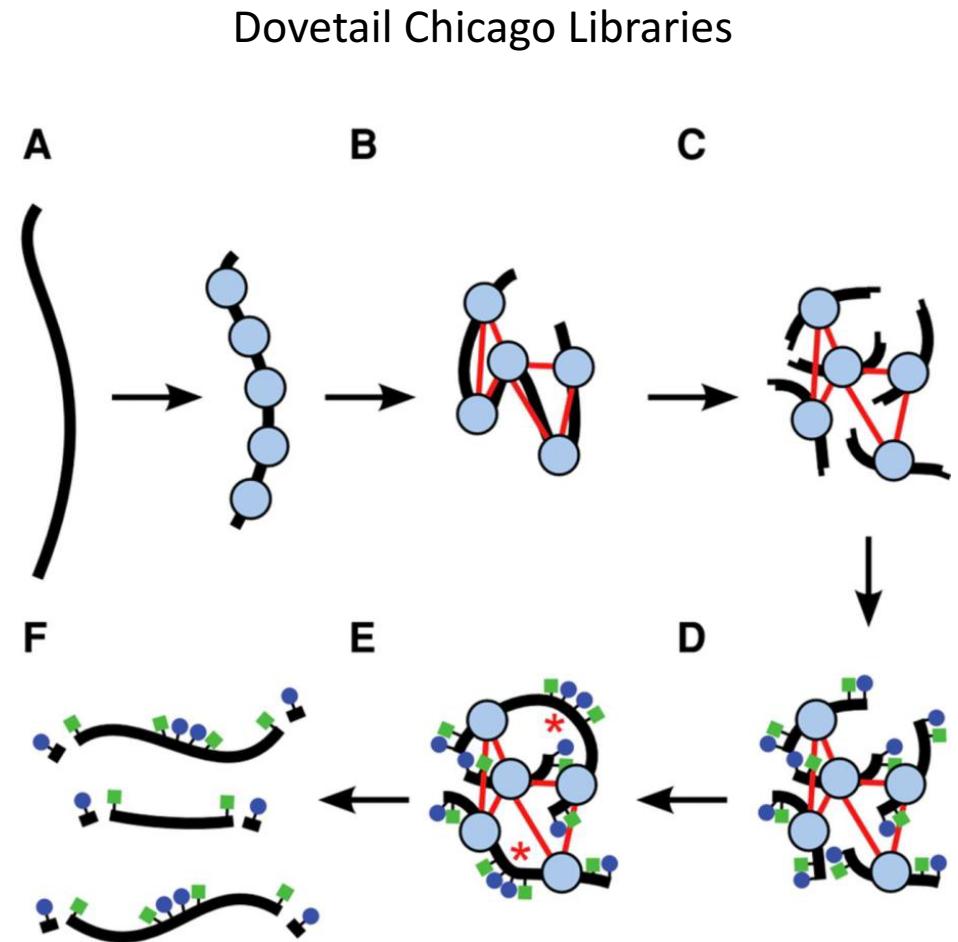
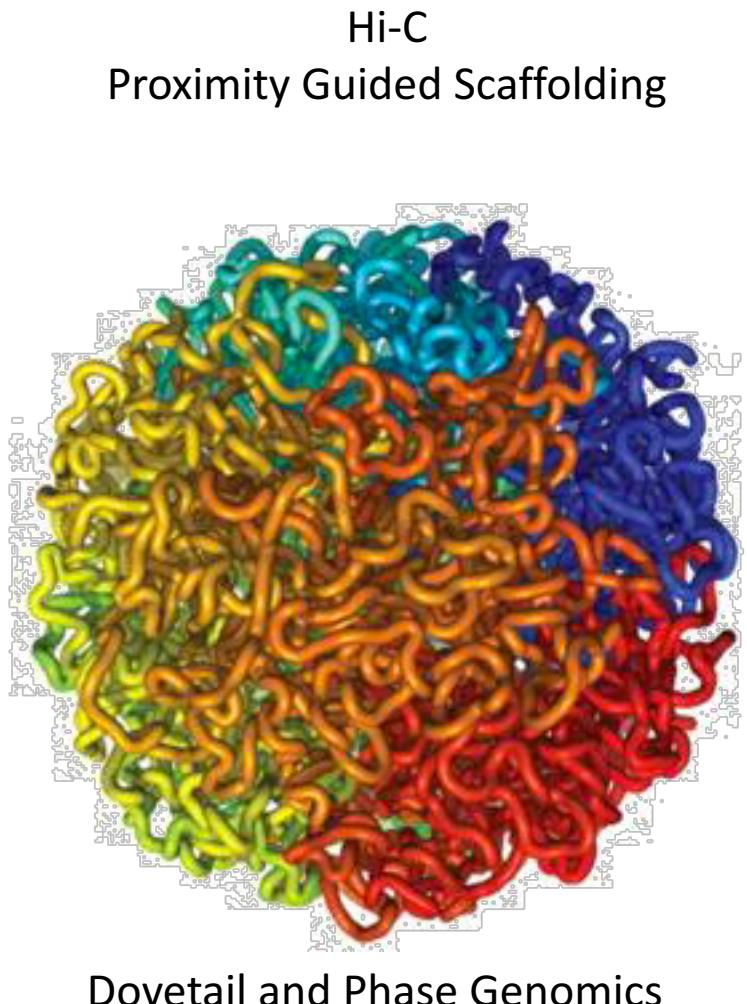
# Bionano Irys/Saphyr

- The Irys/Saphyr System puts the power of optical genome mapping. No more waiting for months to get a physical genome map. Bionano Next-Generation Mapping (NGM) provides long-range information to reveal true genome structure. Assists genomes assemblies to near chromosomal arms.

**Not sequencing based**



# Dovetail and Hi-C (Cross Linking) on Illumina



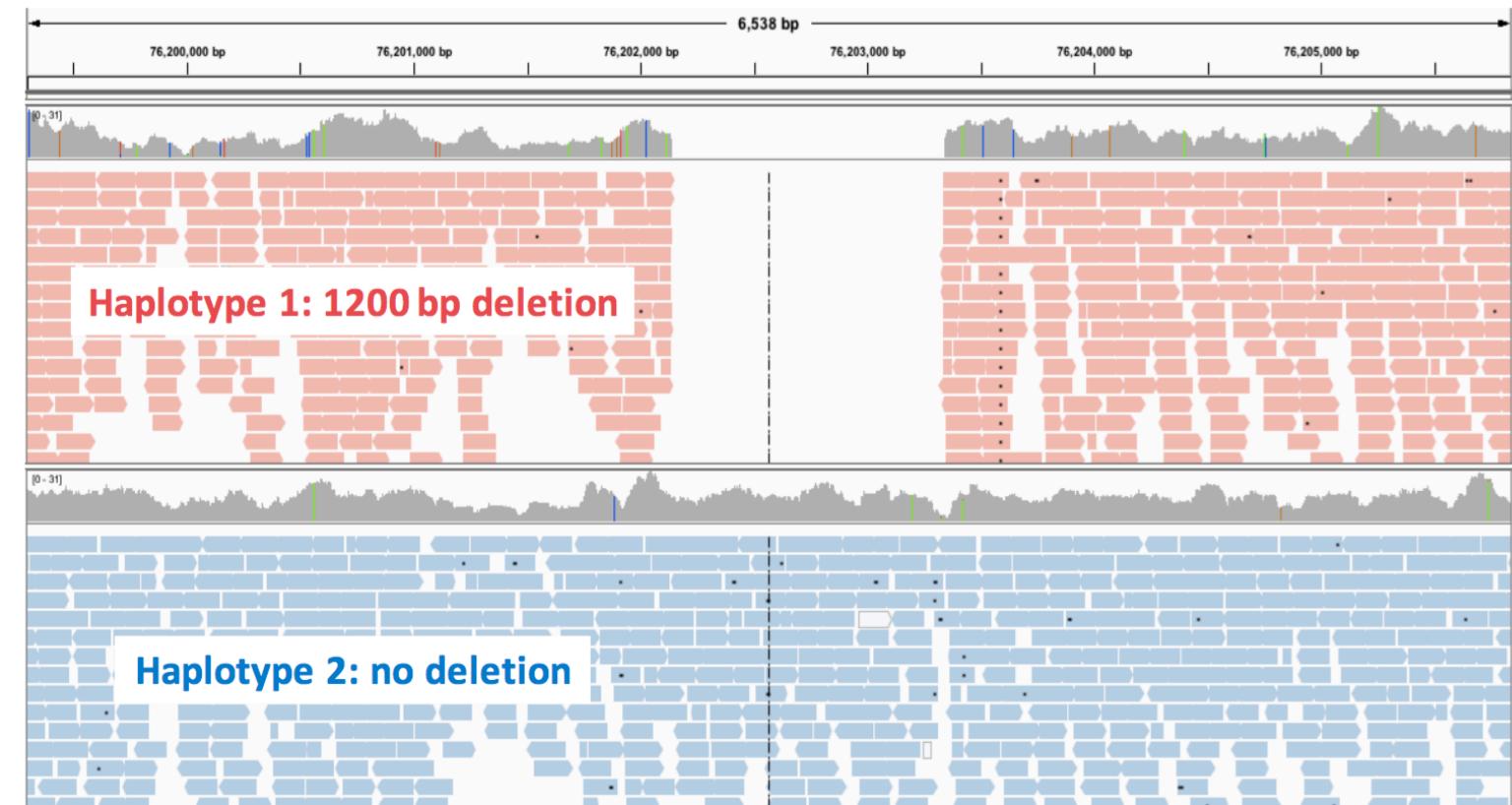
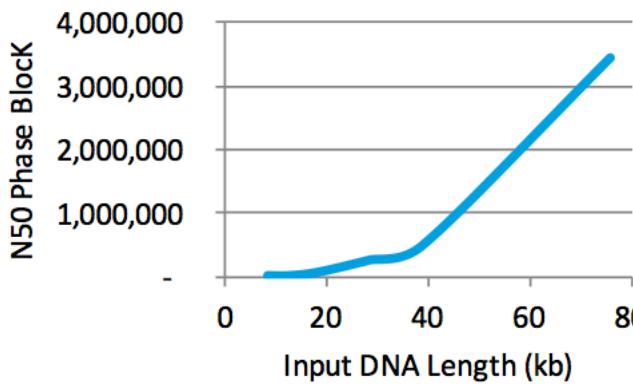
# 10x genomics on Illumina

- 10x Genomics, Linked reads technology
- Illumina machines, Sequencing by Synthesis ~ 120Gb/lane, 2x150bp reads.



10x has its own assembler, Supernova

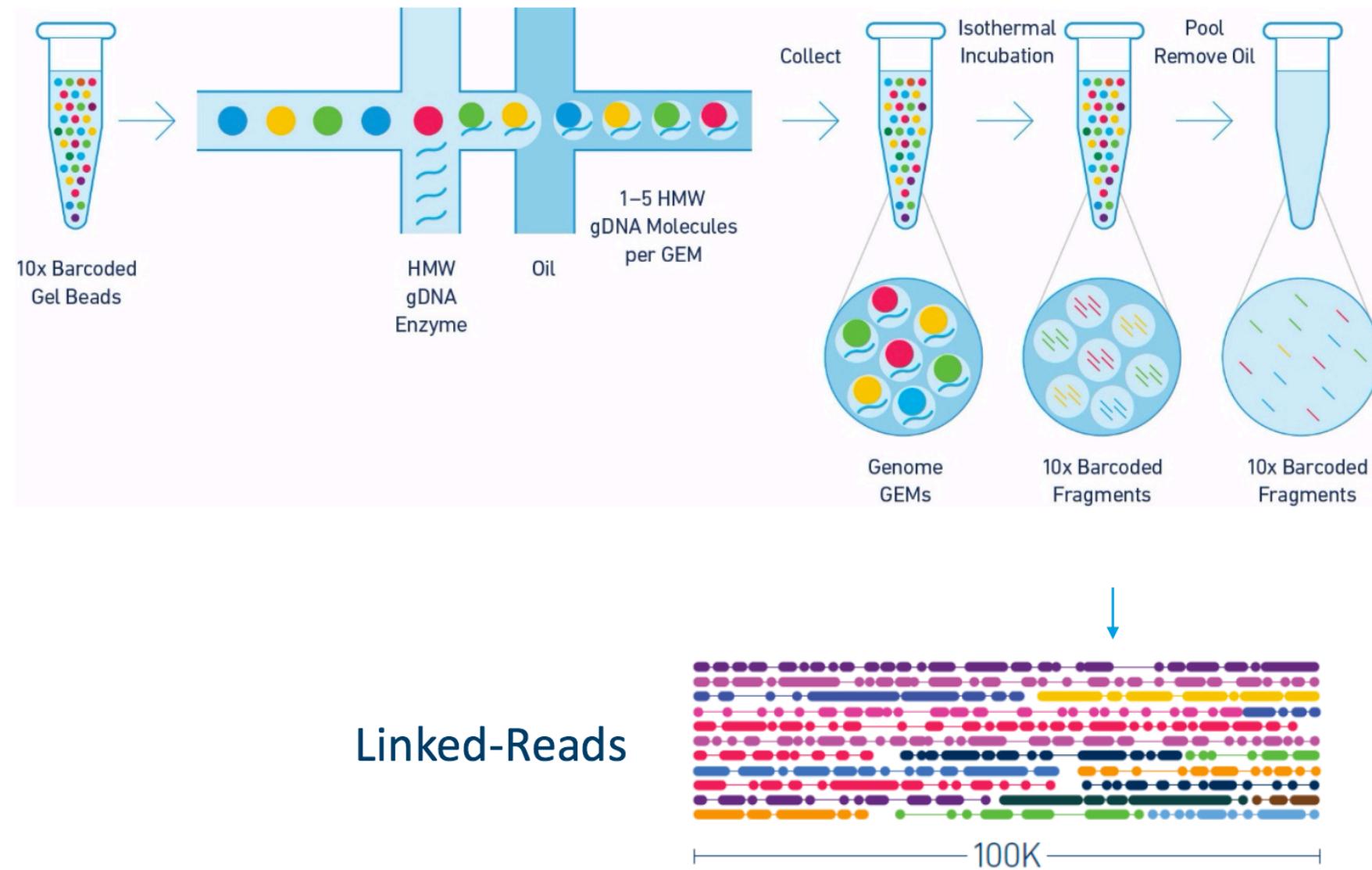
# 10x Genomics phasing + high quality Illumina data



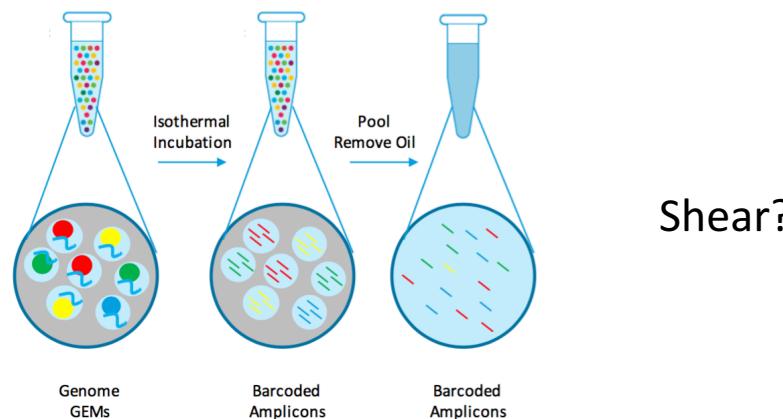
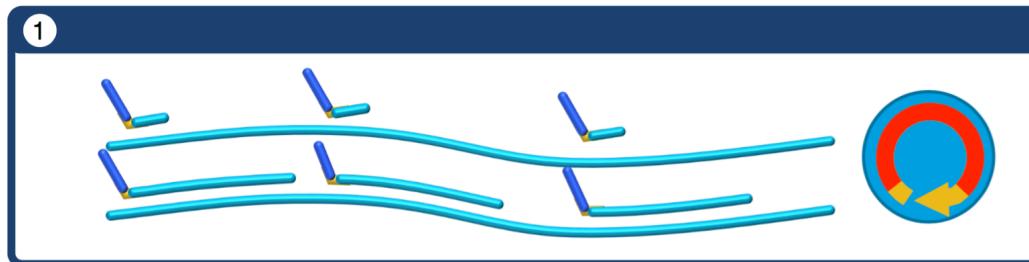
# 10x Genomics, Supernova Assembly Stats

Genome	Size (Gb)	DNA size(Kb)	N50 contig(Kb)	N50 scaffold(Mb)	N50 phase block (Mb)
NA12878	3.2	95.5	85.0	12.8	2.8
NA24385	3.2	111.3	90.0	10.4	3.9
HGP	3.2	138.8	104.9	19.4	4.6
Yoruban	3.2	126.9	100.5	16.1	11.4
Komodo dragon	1.8	85.4	95.3	10.2	0.4
Spotted owl	1.5	72.2	118.3	10.1	0.2
Hummingbird	1.0	86.2	87.6	12.5	10.1
Monk seal	2.6	92.3	93.8	14.8	0.6
Chili pepper	3.5	53.3	84.7	4.0	2.1
<b>CowPea</b>	<b>0.38</b>	<b>46.5</b>	<b>28.3</b>	<b>0.83</b>	<b>0.35</b>
<b>Walnut</b>	<b>0.89</b>	<b>55.0</b>	<b>48.0</b>	<b>0.60</b>	<b>0.25</b>
<b>California Condor</b>	<b>1.19</b>	<b>67.0</b>	<b>147.5</b>	<b>17.9</b>	<b>1.0</b>

# in a nut shell

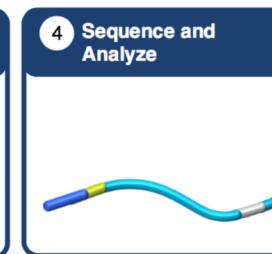
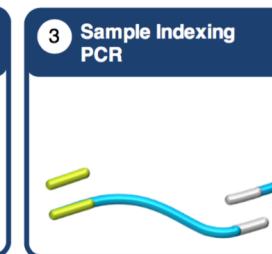
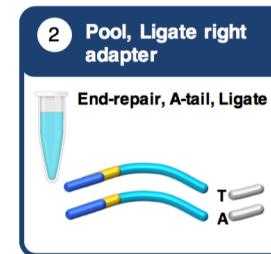


# Laboratory Workflow

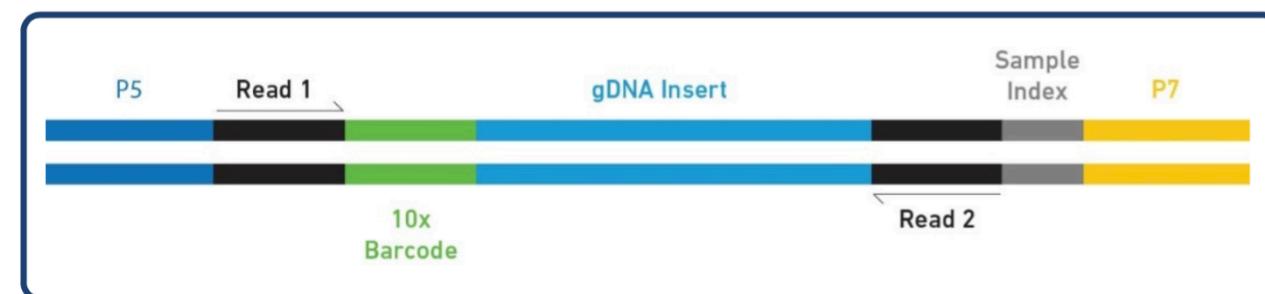
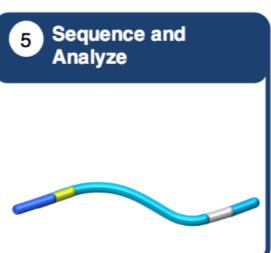
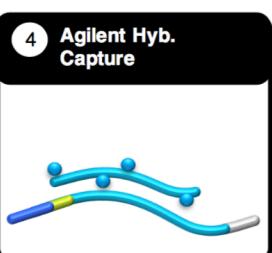
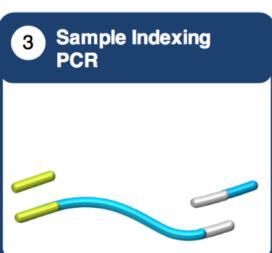
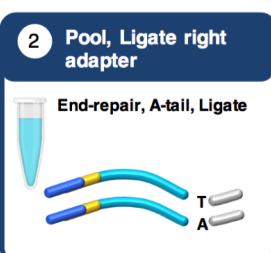


Shear?

## Whole Genome Sequence



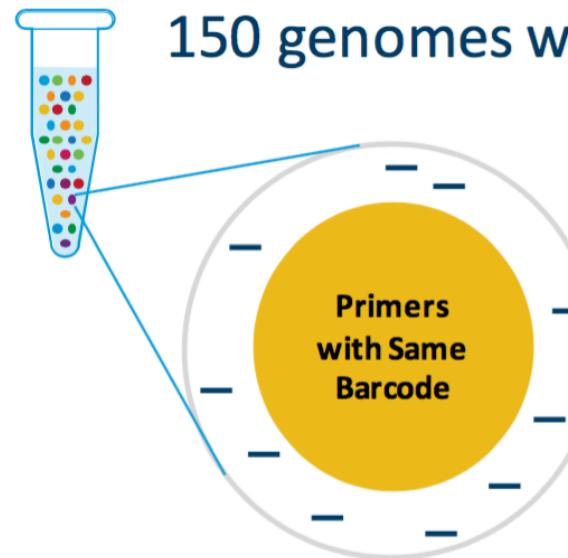
## Whole Exome Sequence



# The Math

1 ng Input DNA  
= 300 genomes  
copies of the genome

Calculations imply that  
about 50% of all possible  
fragments end up in a bead



150 genomes went into 1M partitions

Each GEM contains:

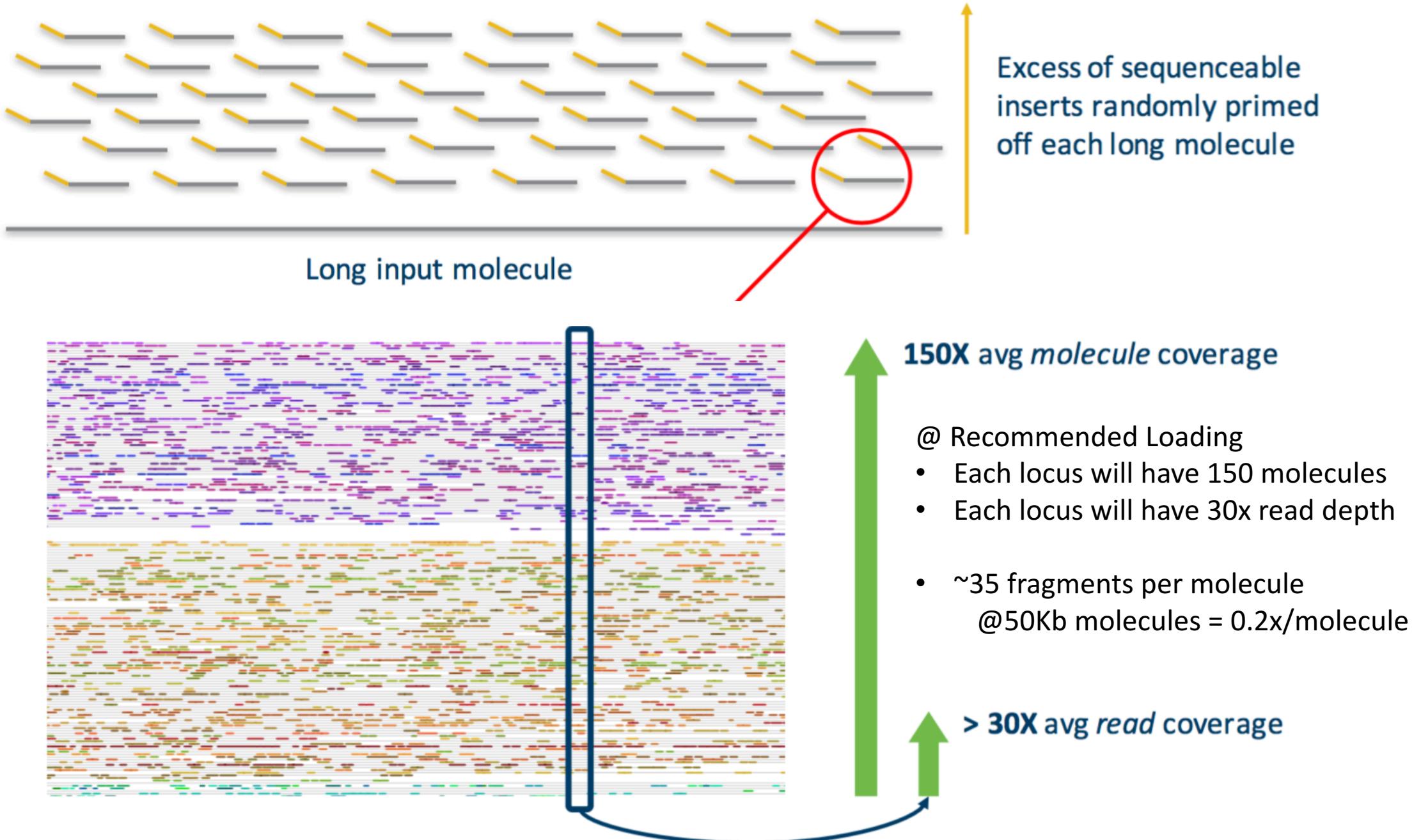
- One barcode (many copies)
- 1/6000 of the genome (500 Kb)
- At 50Kb length, 10 molecules

Chance that 2 molecules covering a locus are in same GEM:

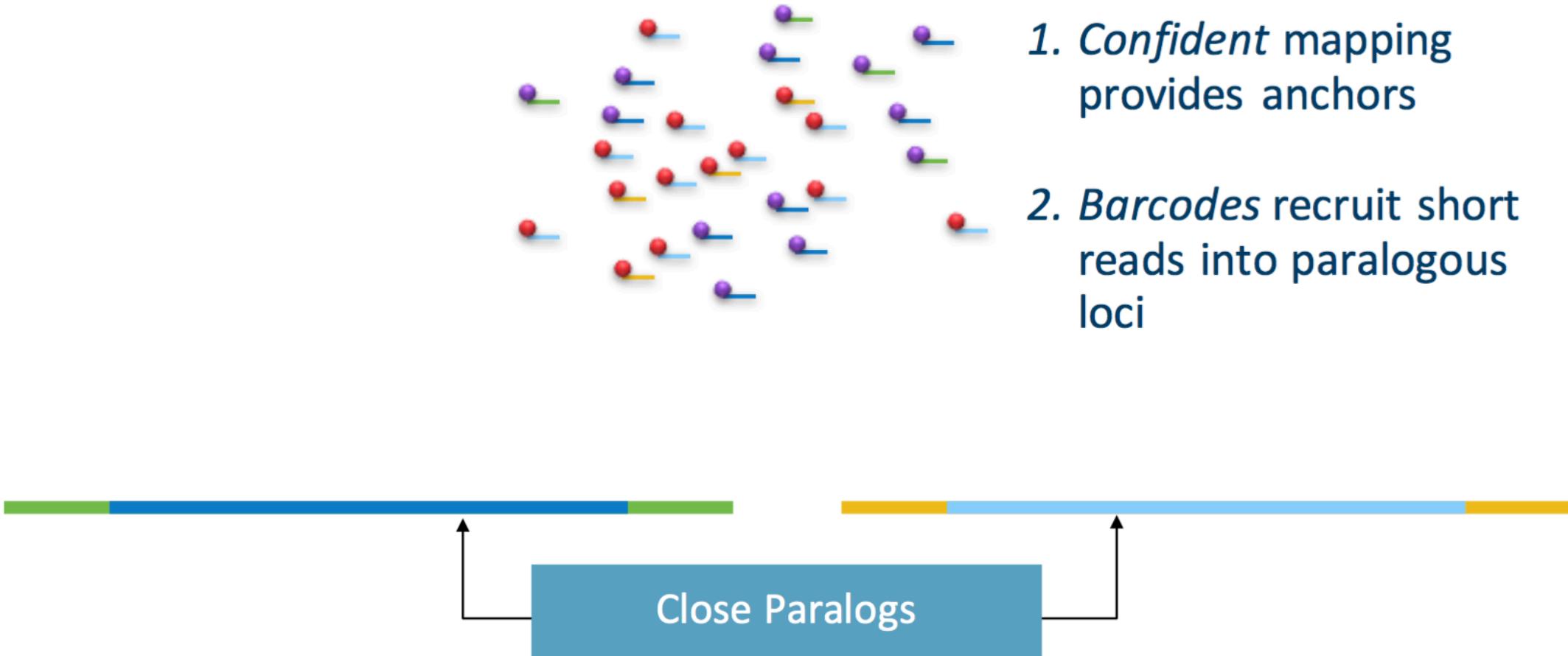
**1 in 6000**

Percent unique barcodes at any genomic locus:

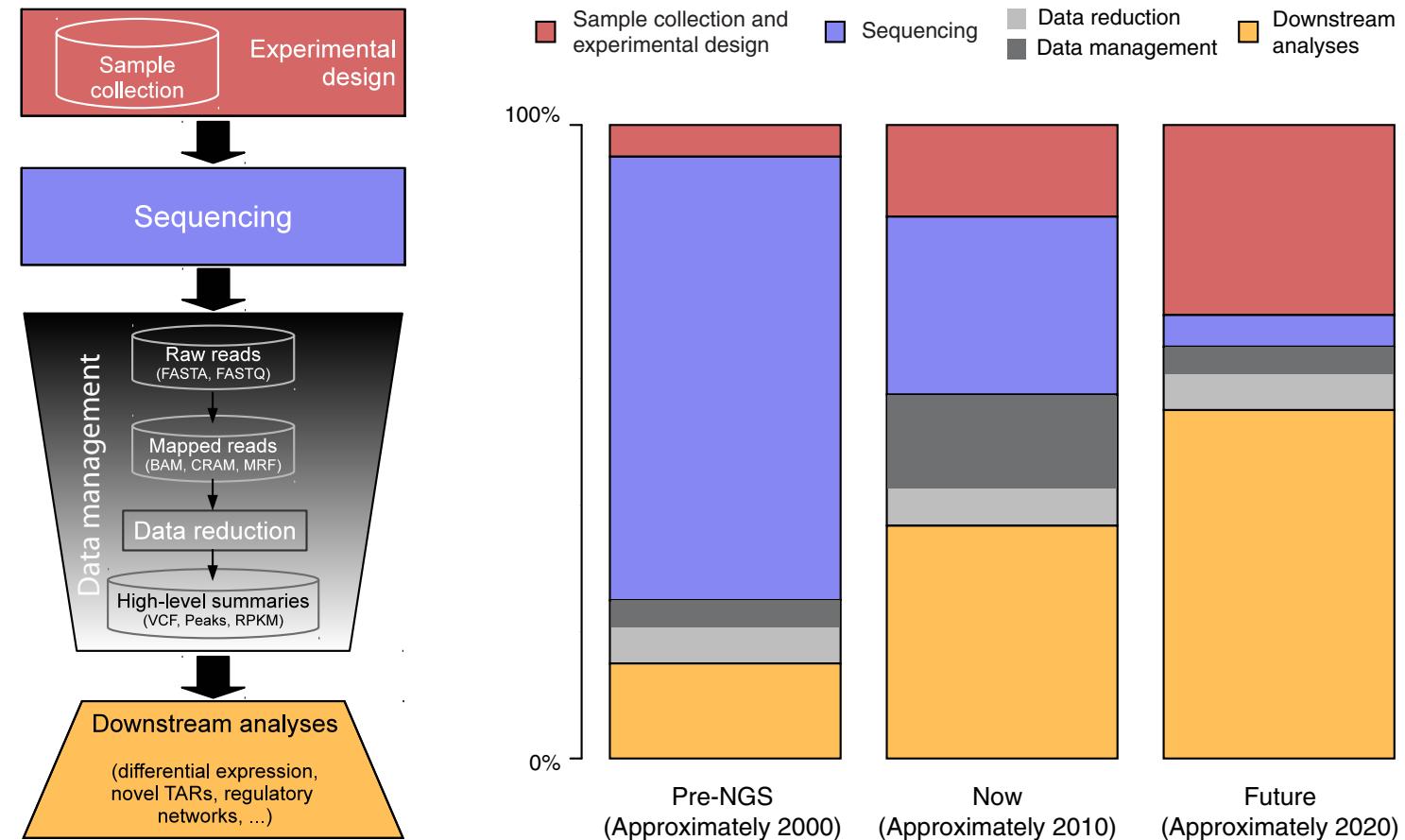
**99.98%**



# Increased Map-ability



# The real cost of sequencing



omicsmaps.com

