NORWEGIAN SEQUENCING CENTRE

PacBio long read sequencing

Ave Tooming-Klunderud, NCS/CEES/UiO,

ave.tooming-klunderud@ibv.uio.no

Oslo universitetssykehus

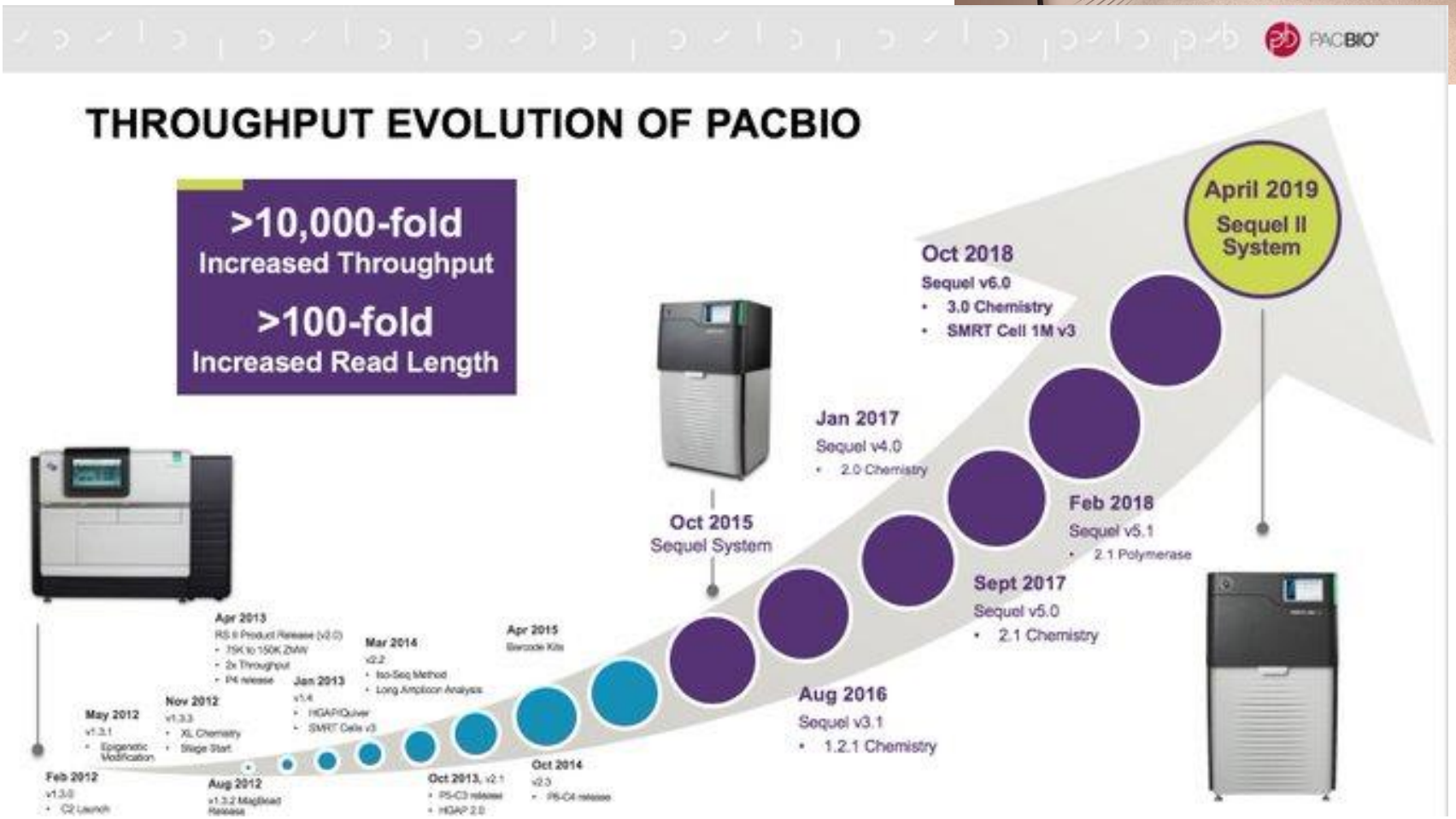
The Research Council of Norway

CEES Centre for Ecological and Evolutionary Synthesis

HELSE SØR-ØST

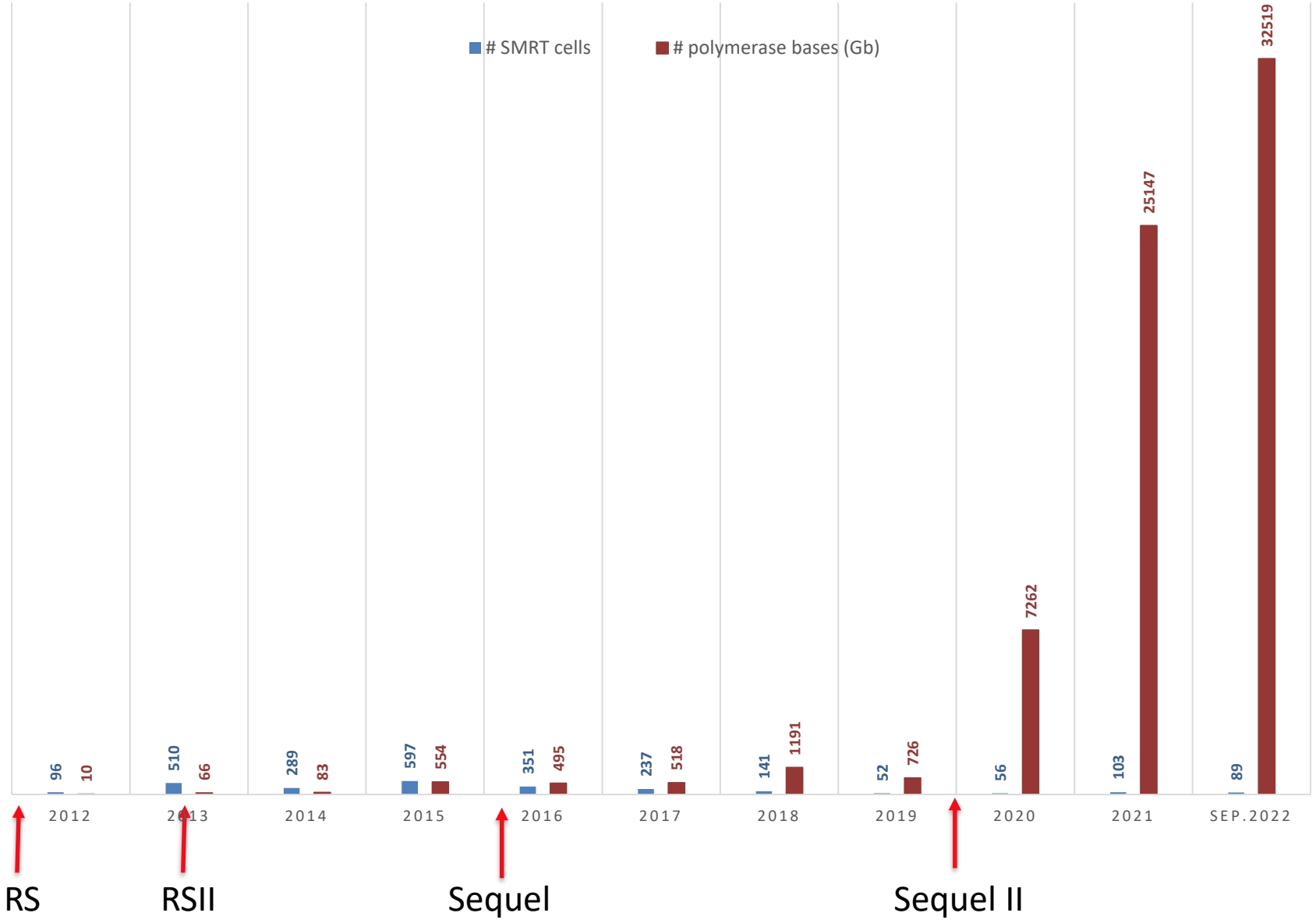
UNIVERSITAS OSLOENSIS MDCCCXI

PacBio sequencing since 2012



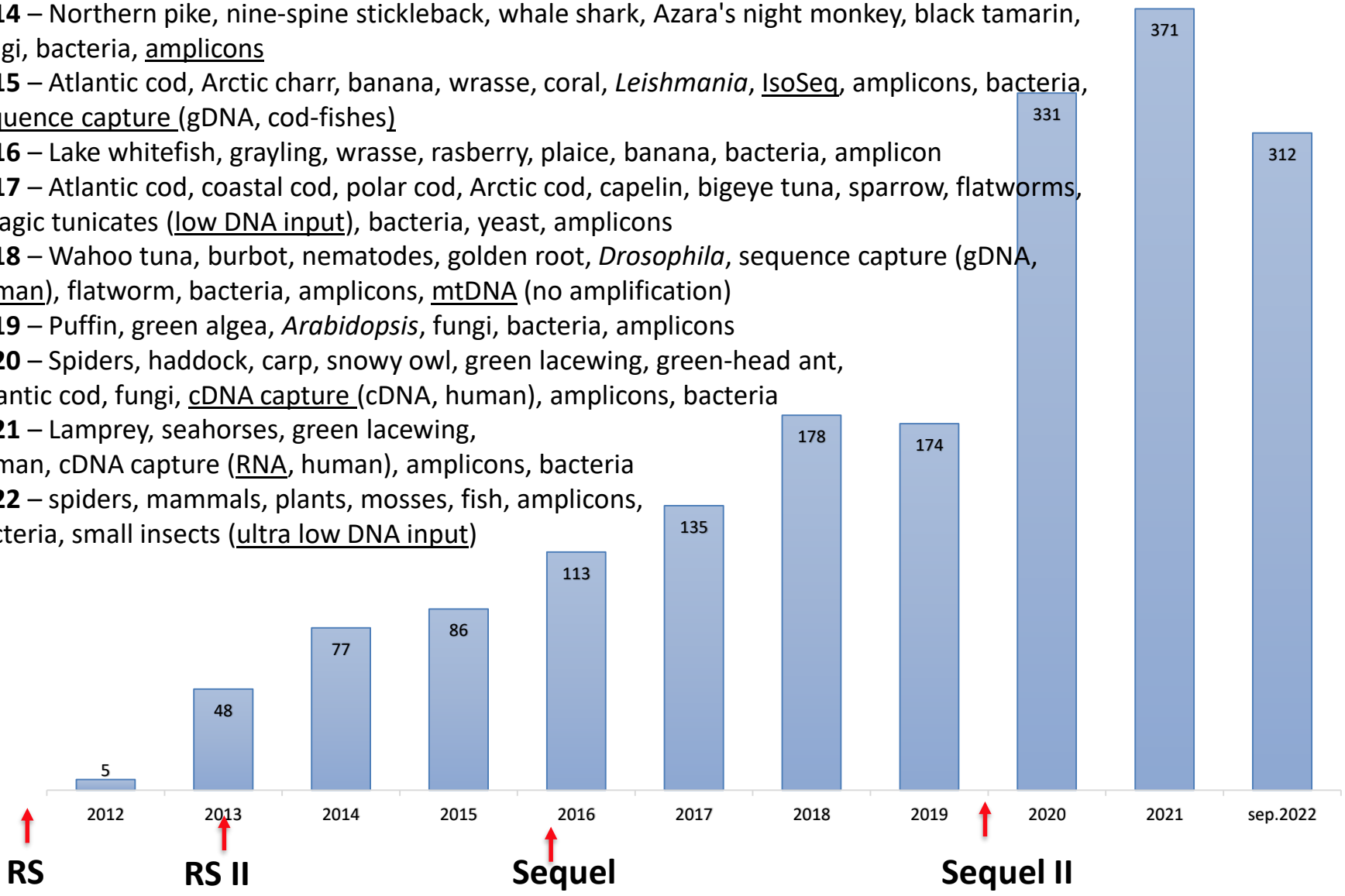
Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA

Throughput evolution at NSC



PacBio sequencing at NSC

2012 – Atlantic cod, fungi, bacteria
2013 – Atlantic cod, Atlantic salmon, fungi, bacteria
2014 – Northern pike, nine-spine stickleback, whale shark, Azara's night monkey, black tamarin, fungi, bacteria, amplicons
2015 – Atlantic cod, Arctic charr, banana, wrasse, coral, *Leishmania*, IsoSeq, amplicons, bacteria, sequence capture (gDNA, cod-fishes)
2016 – Lake whitefish, grayling, wrasse, rasberry, plaice, banana, bacteria, amplicon
2017 – Atlantic cod, coastal cod, polar cod, Arctic cod, capelin, bigeye tuna, sparrow, flatworms, pelagic tunicates (low DNA input), bacteria, yeast, amplicons
2018 – Wahoo tuna, burbot, nematodes, golden root, *Drosophila*, sequence capture (gDNA, human), flatworm, bacteria, amplicons, mtDNA (no amplification)
2019 – Puffin, green algea, *Arabidopsis*, fungi, bacteria, amplicons
2020 – Spiders, haddock, carp, snowy owl, green lacewing, green-head ant, Atlantic cod, fungi, cDNA capture (cDNA, human), amplicons, bacteria
2021 – Lamprey, seahorses, green lacewing, human, cDNA capture (RNA, human), amplicons, bacteria
2022 – spiders, mammals, plants, mosses, fish, amplicons, bacteria, small insects (ultra low DNA input)

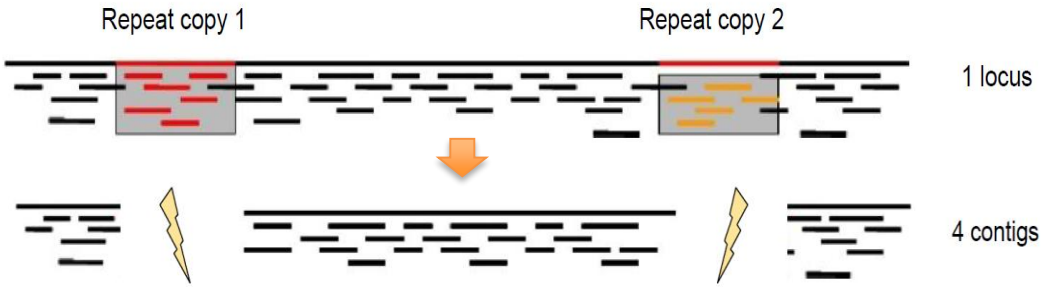


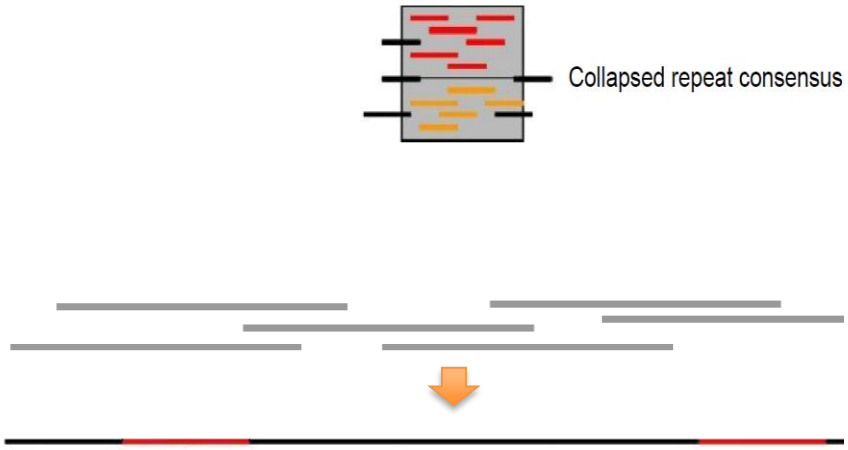
Short read vs long read sequencing

Short read sequencing	Long read sequencing
Amplification during sequencing	No amplification involved
High read accuracy	High consensus read accuracy
DNA requirements	
Works with almost any DNA sample	gDNA: pure HMW DNA needed
Fragmented DNA	DNA fragments at least 40-50 kb long
Low amount of DNA	High amount of DNA
	Low/Ultra-Low DNA input protocols available
Per base price	
Low	Medium/high

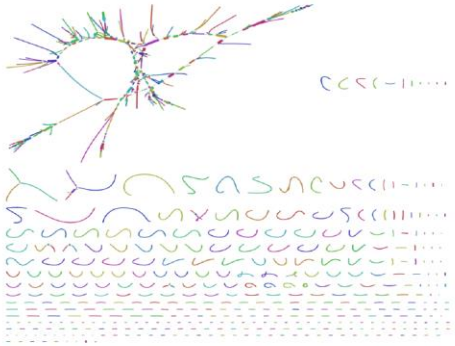
Why long reads?

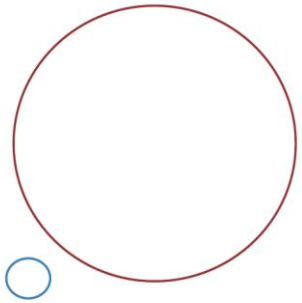
Long reads can span repeats





Complete genomes (small genomes)





Complete genomes (large genomes)

Contig Type	# Polished Contigs	Maximum Contig Length	N50 Contig Length	L50	Sum of Contig Lengths
Primary Contigs	698	29,198,342	12,301,043	20	684,461,580

Draft vs reference quality genome

Type to search GoaT taxon index (e.g. Canidae)

Rangifer tarandus

include descendants: Off | include estimates: On | result columns | query builder | clear all

taxon record 9870

Rangifer tarandus (species) 9870

gc_percent	assembly_span	assembly_level	bioproject	biosample	contig_n50	assembly_date	scaffold_n50	nohit
43.2 median, n=2	2.86G primary, n=2	Scaffold primary, n=2	PRJEB35834 ... list, n=3	SAMEA6417239 ... list, n=3	129k primary, n=2	2022-03-03 primary, n=2	986k primary, n=2	60.2 median, n=2

target	mitochondrion_assembly_span	mitochondrion_gc_percent	haploid_number	chromosome_number	ploidy	genome_size	c_value	long_list
98.2 median, n=2	16.4k median, n=1	36.2 median, n=1	35 median_high, n=1	70 median_high, n=1	2 median_high, n=1	3.33G primary, n=1	3.41 primary, n=1	CANBP ... list, n=2

other_priority	sequencing_status	sequencing_status_zoonomia	sample_collected	sample_acquired	in_progress
CANBP ... list, n=2	published enum, n=3	published enum, n=1	ZOONOMIA list, n=1	ZOONOMIA list, n=1	ZOONOMIA list, n=1

Lineage

Eukaryota Opisthokonta Metazoa Eumetazoa Bilateria Deuterostomia Chordata Craniata Vertebrata Gnathostomata Sarcopterygii Dipnotetrapodomorpha Tetrapoda Amniota Mammalia Theria Eutheria Boreoeutheria Laurasiatheria A Cervidae Odocoileinae Rangifer

Names

caribou - common name | Rangifer spitzbergensis - synonym | Rangifer tarar

ASM2245718v1

Organism name: Rangifer tarandus (reindeer)
Isolate: DF-B-001
BioSample: SAMN07274499
BioProject: PRJNA438286
Submitter: Northwestern Polytechnical University
Date: 2022/03/03
Assembly level: Scaffold
Genome representation: full
GenBank assembly accession: GCA_022457185.1 (latest)
RefSeq assembly accession: n/a
RefSeq assembly and GenBank assembly identical: n/a
WGS Project: JAJJMQ01
Assembly method: SOAPdenovo v. 2
Expected final version: yes
Genome coverage: 200.0x
Sequencing technology: Illumina

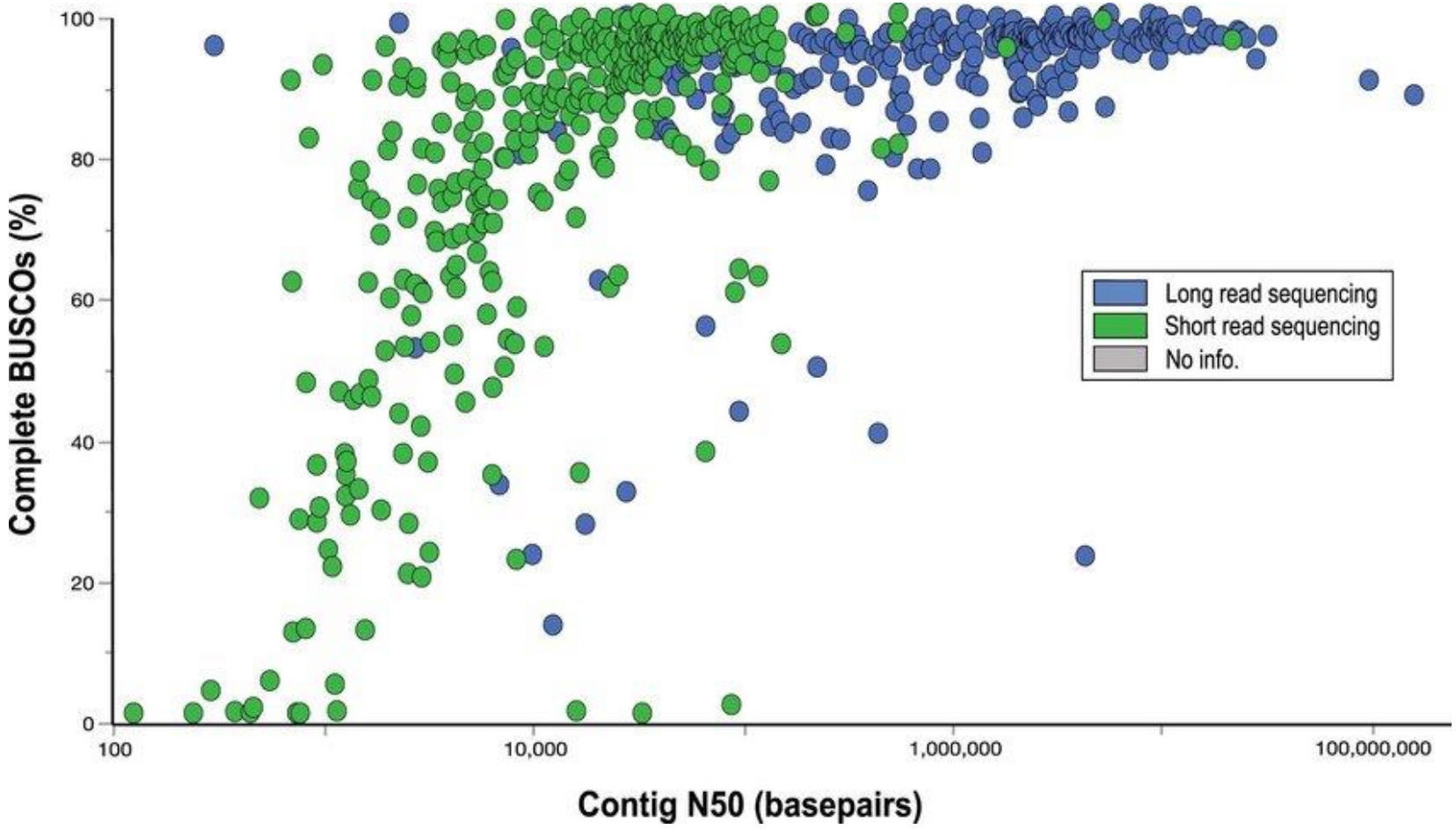
Draft vs reference quality genome II

EBP-Nor sequencing of reindeer from Svalbard:

- 30x PacBio HiFi reads
- 50x Arima Hi-C reads

Assembly	Reinsdyr hap1 (incl X and Y chromosomes)	Reinsdyr hap2
# scaffolds	1395	1291
Total scaffold length:	2.97 Gb	2.82 Gb
Contig N50:	22.48 Mb	25.53 Mb
Scaffold N50:	69.83 Mb	64.92 Mb
Largest scaffold:	157.94 Mb	119.78 Mb
Scaffolds placed in chromosomes (%)	89.22%	82.25%
BUSCOs percentage complete	96.3%	94.1%
BUSCOs complete	8883	8686
BUSCOs single	8548	8381
BUSCOs duplicated	335	305
BUSCOs fragmented	89	81
BUSCOs missing	254	459
BUSCOs total	9226	9226

Relationship between assembly contiguity and the percentage of complete BUSCOs

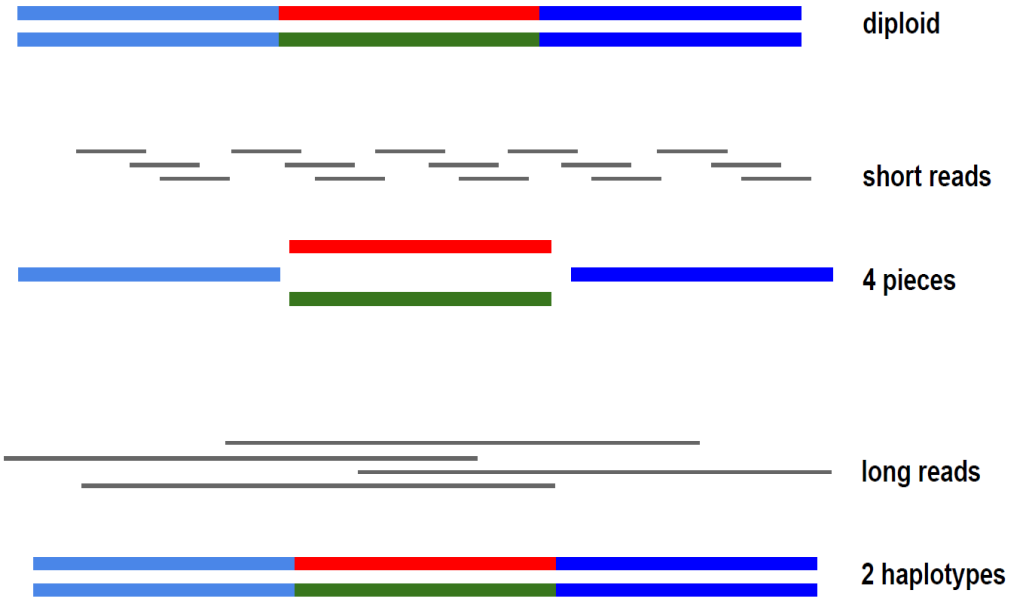


https://www.nature.com/articles/s41477-021-01031-8

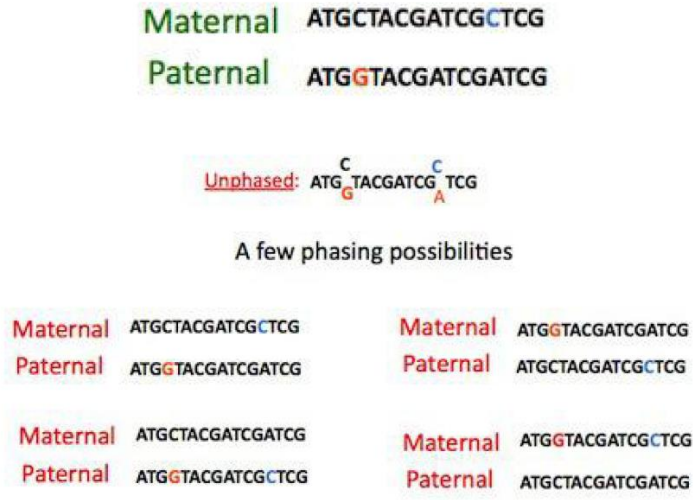
Why long reads?

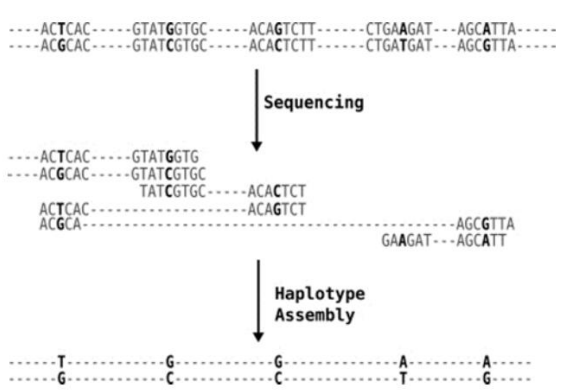
Phased haplotypes

- Genome assembly



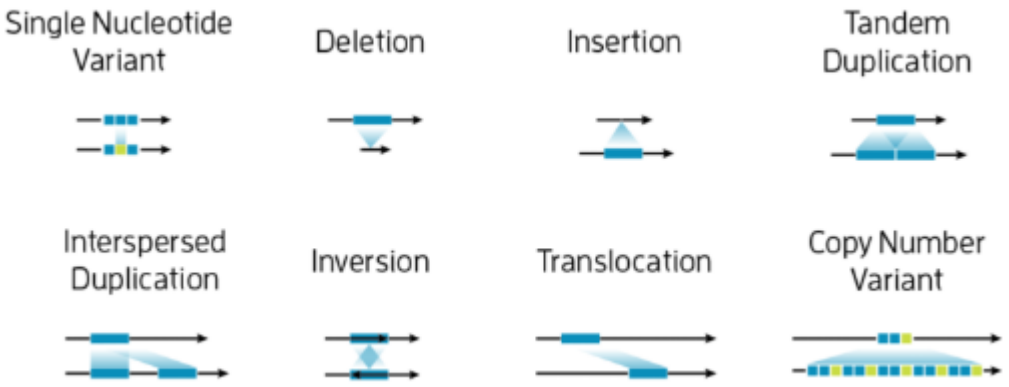
- Shorter regions (for example HLA typing)

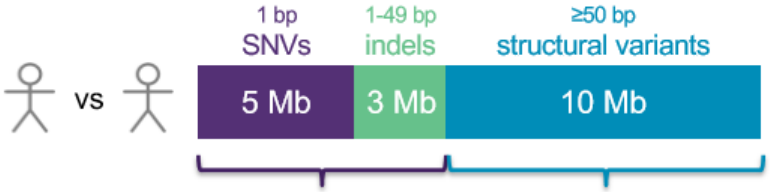




Why long reads?

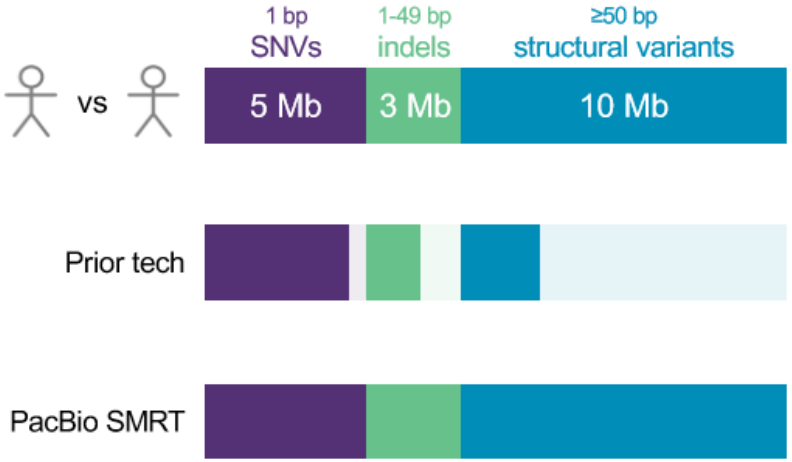
Structural variation – the missing heritability, not just SNVs





"Small variants":
- Single Nucleotide Variants (SNVs)
- Indels <50 bp

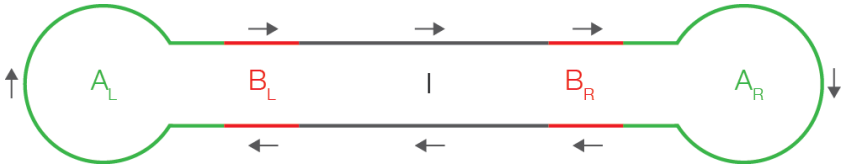
Structural Variants (SVs):
- Indels ≥50 bp
  - Duplications
  - Copy Number Variants (CNVs)
- Translocations
- Inversions



The PacBio sequencing technology

PACIFIC BIOSCIENCES™

SMRTbell library



How SMRT Sequencing Works

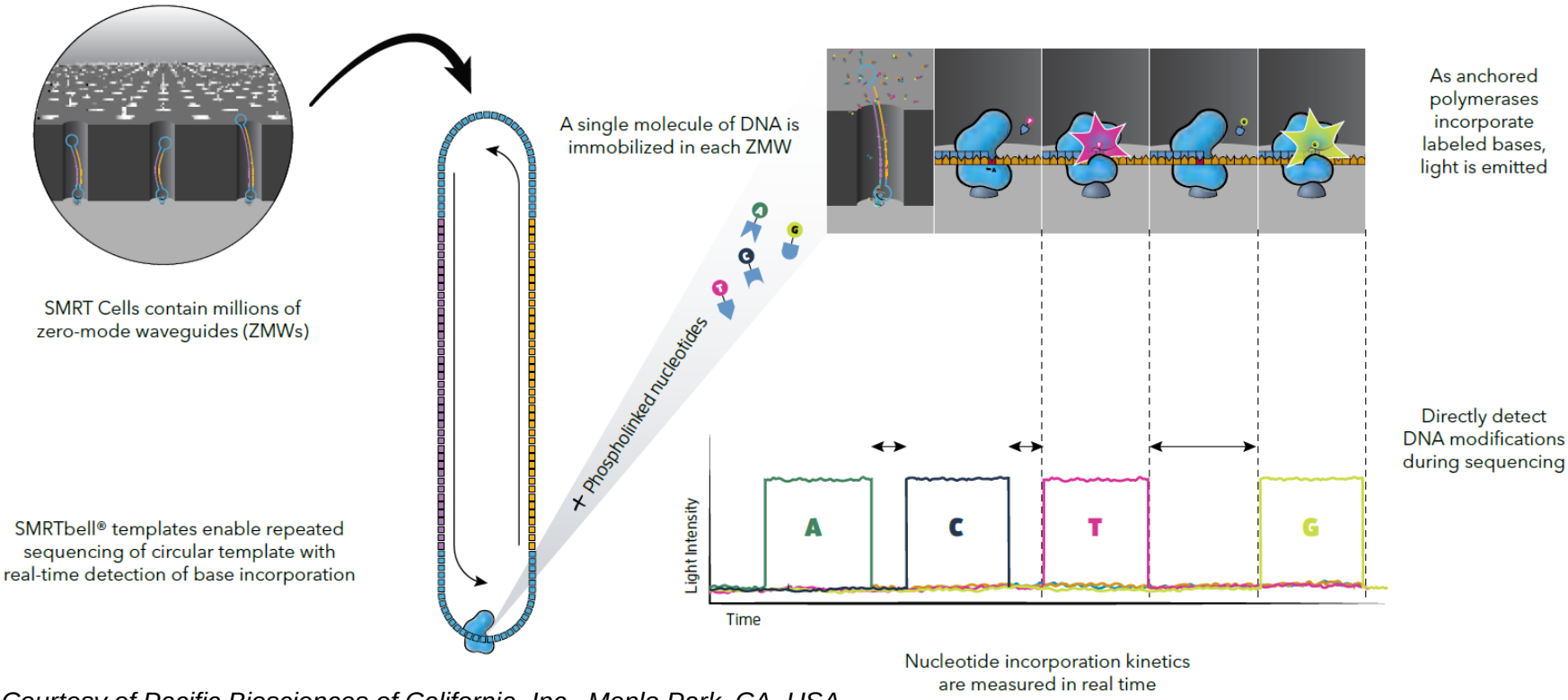


Image: Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA

Measuring DNA methylation

SMRTbell® library

Sequel® IIe system

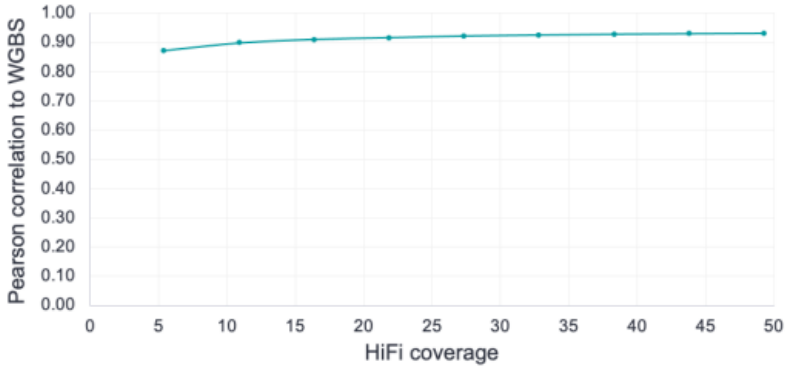
Nucleotide incorporation kinetics are measured in real time

5-base HiFi sequencing with A, C, G, T, +5mC

5mC encoded with standard BAM tags[3]
MM:Z:C+m,4,12,16,4,16,19,44,10
ML:B:C,249,4,247,177,210,228,245,244

The Sequel IIe system directly outputs long, highly accurate HiFi reads with annotation of 5mC methylation at all CpG sites. No special library preparation like bisulfite treatment is required.

Coverage



Correlation of methylation calling in HiFi reads to whole-genome bisulfite sequencing (WGBS) of the human sample HG002.[4,5,6]

Applicability

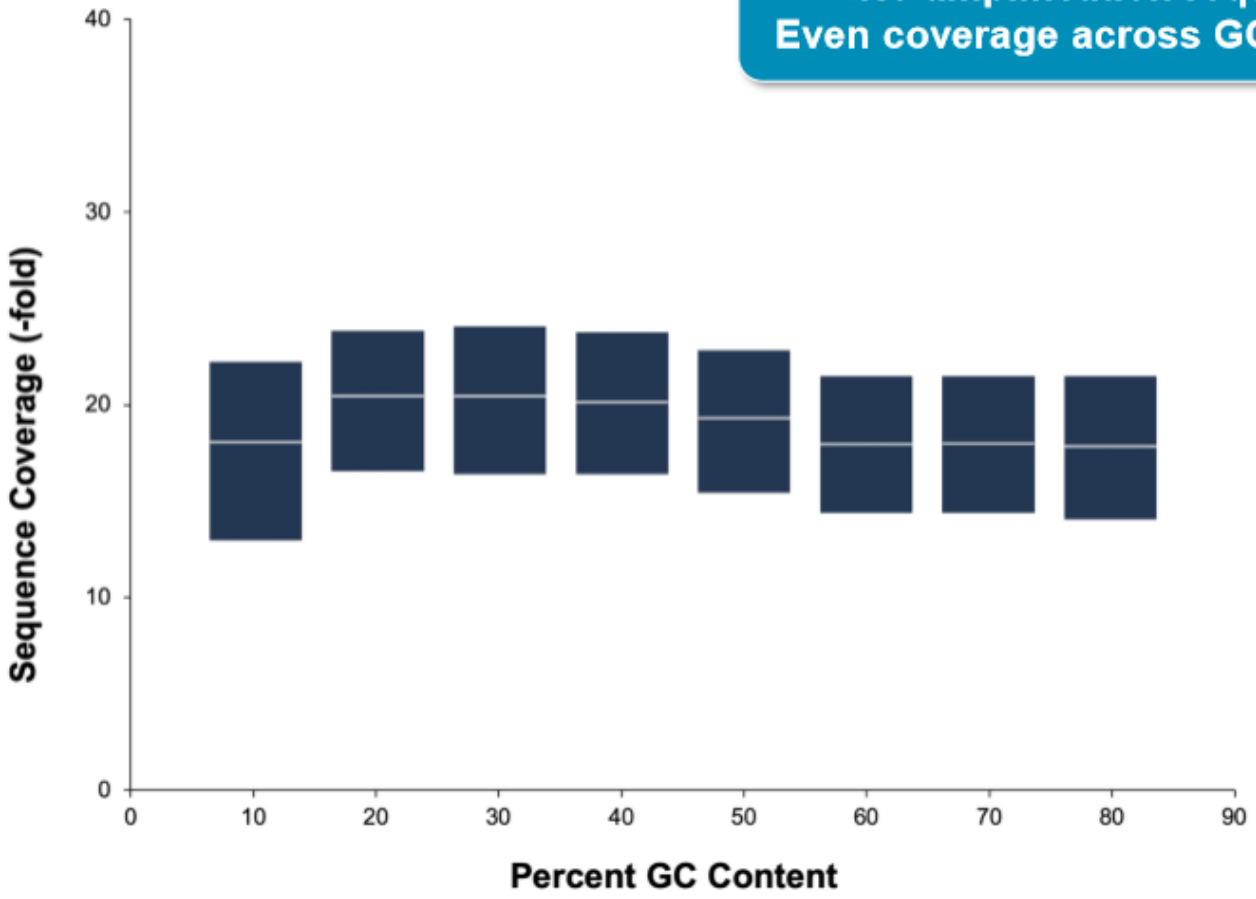
Methylation	Species	5-base HiFi sequencing
5mC at CpG sites	Human and other vertebrates	✓
5mC at various motifs	Other eukaryotes, including plants	✓ Useful though partial view
4mC and 6mA	Microbes	Enabled through SMRT® Link microbial genome analysis

PacBio

Sequence performance: uniformity

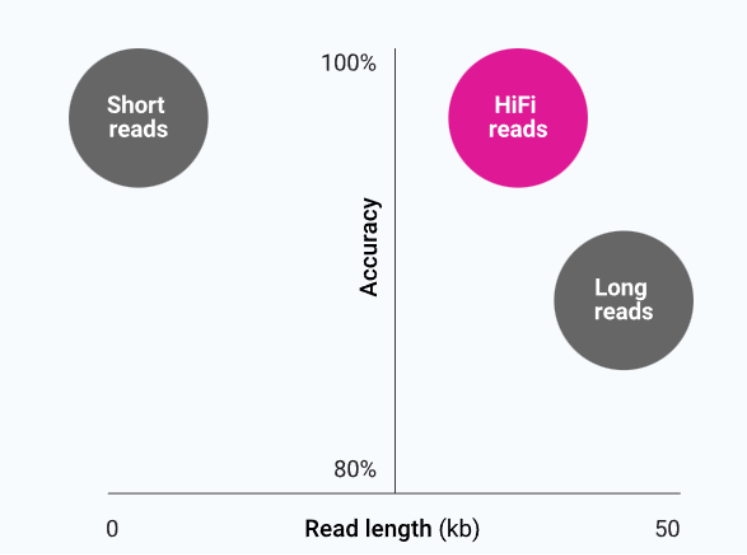
UNIFORM COVERAGE

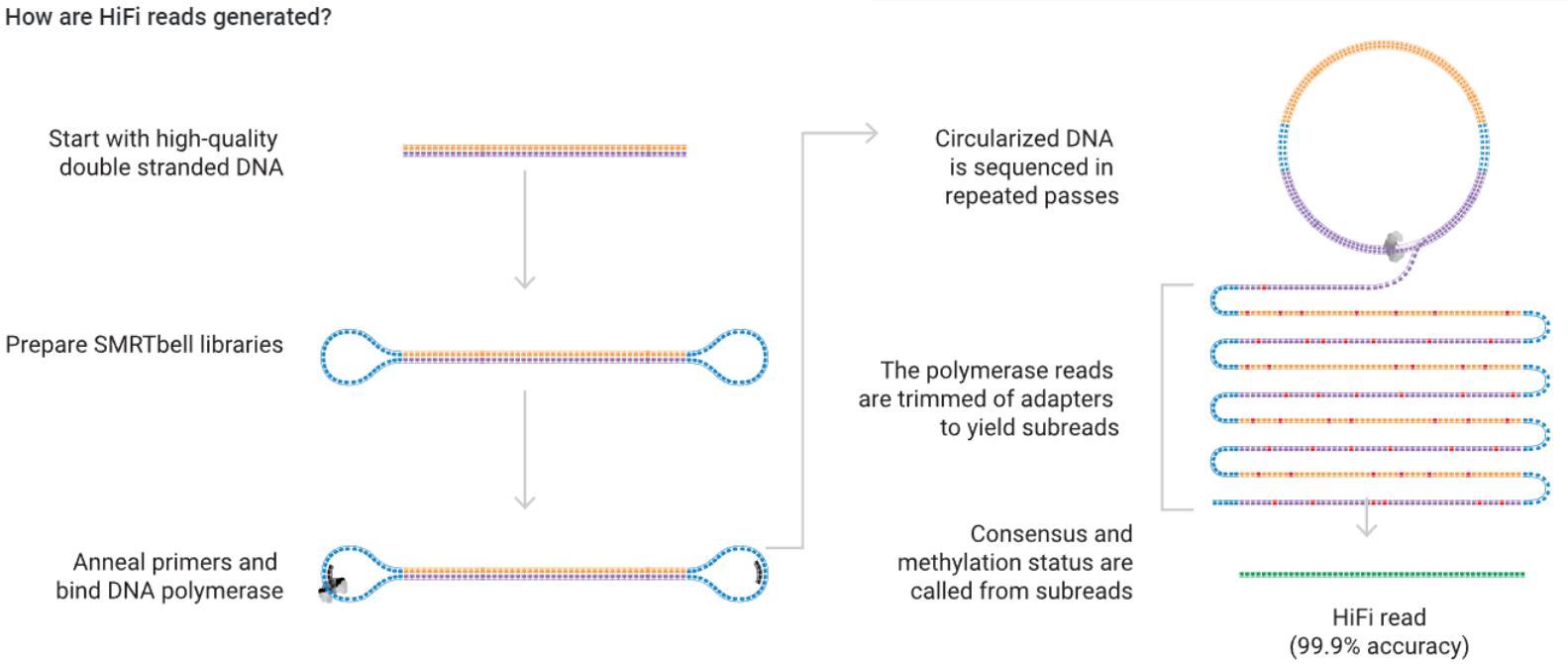
No amplification required
Even coverage across GC content



Mean coverage per GC window across a human sample. Data generated with a 15 kb human HiFi library on a Sequel II System using 2.0 Chemistry and Sequel II System Software v8.0

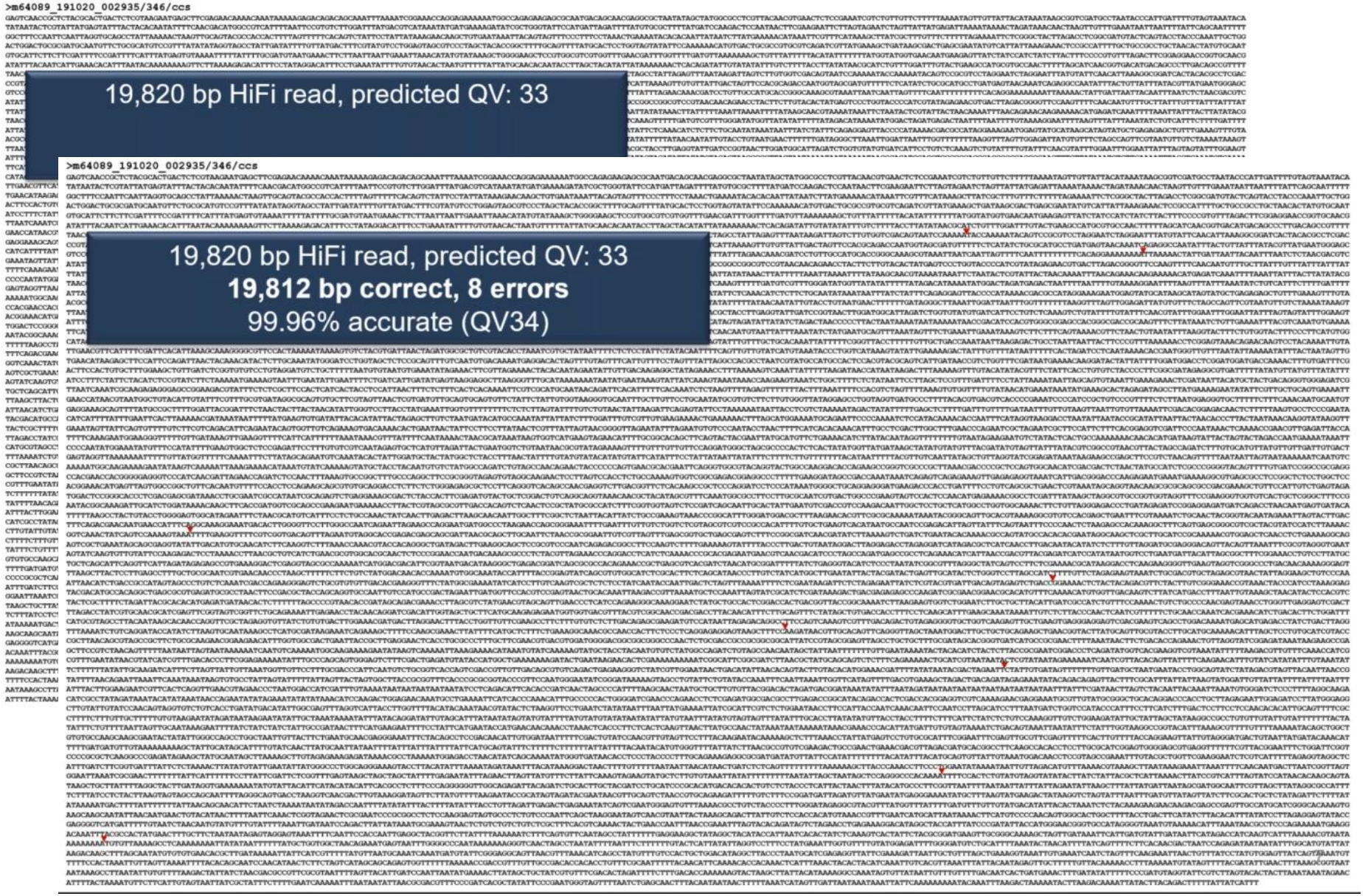
HiFi sequencing:



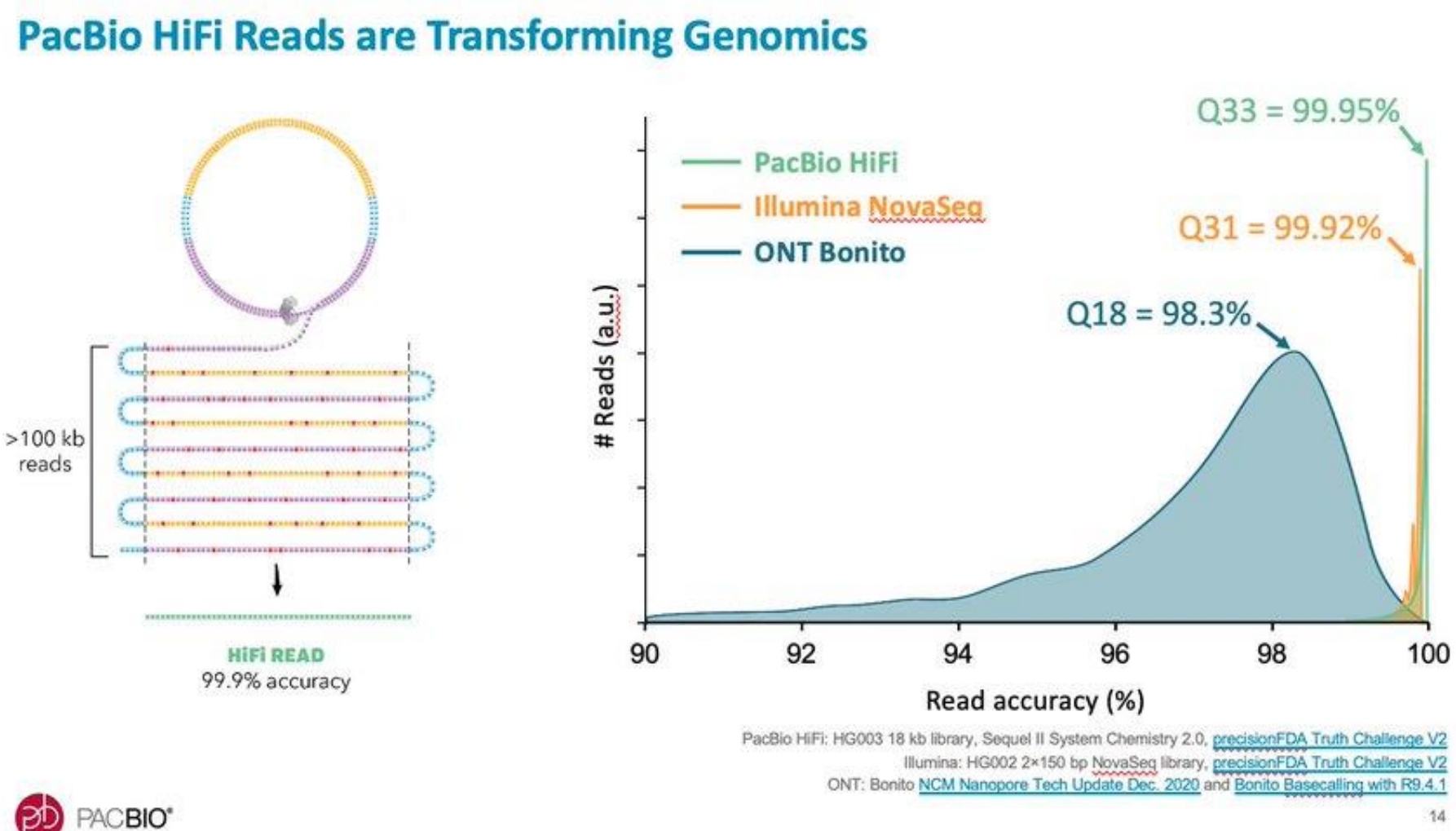


Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA

HiFi read – long and accurate!



Read accuracy comparison between different sequencing platforms:



Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA

Sequel II/IIe applications

SEQUEL II SYSTEM KEY APPLICATIONS

Whole Genome Sequencing for *De novo* Assembly

- Single Molecule, Real-Time (SMRT) Sequencing on the Sequel II System enables easy and affordable generation of high-quality de novo assemblies. With megabase size contig N50s, accuracies >99.99%, and phased haplotypes, you can do more biology – capturing undetected SNVs, fully intact genes, and regulatory elements embedded in complex regions.

With one 8M SMRT cell you can:

- Produce reference quality assemblies for genomes around 1 Gb – HiFi reads (10- to 15-fold HiFi read coverage per haplotype)
- Produce reference quality assemblies for genomes up to 500 Mb – Ultra-Low DNA input (5-20 ng DNA)
- Sequence up to 96 microbes

SAMPLE & PROJECT CONSIDERATIONS	STANDARD HIFI SEQUENCING	LOW DNA INPUT SEQUENCING (2-PLEX)	LOW DNA INPUT SEQUENCING (SINGLE SAMPLE)	ULTRA-LOW DNA INPUT SEQUENCING
Minimum DNA Input	≥5 µg for a 3-Gb genome	300 ng for each genome	400 ng	5 ng
Amplification Based?	No	No	No	Yes
Genome Size Limit	N/A	600 Mb for each genome	1 Gb	500 Mb
Supported Applications	*De novo* Assembly Human Variant Detection	*De novo* Assembly	*De novo* Assembly	*De novo* Assembly Human Variant Detection

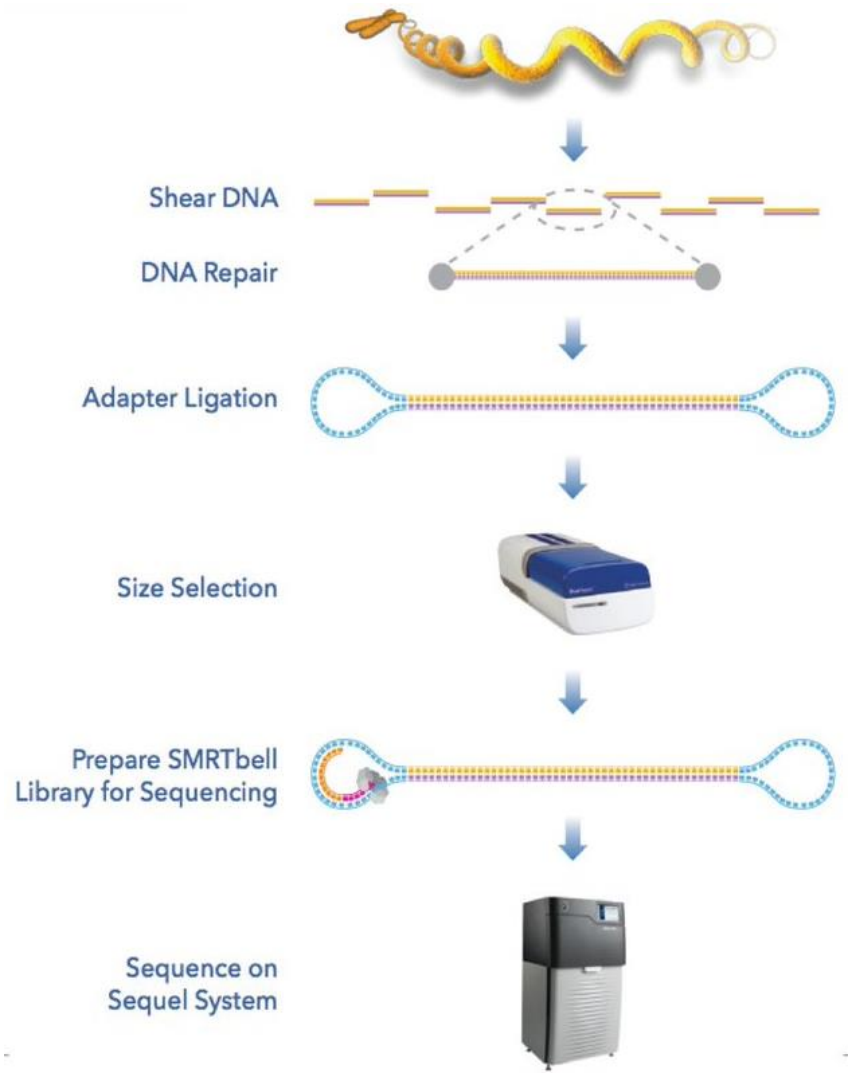
Ultra-Low DNA Input: SUPPORTED APPLICATIONS

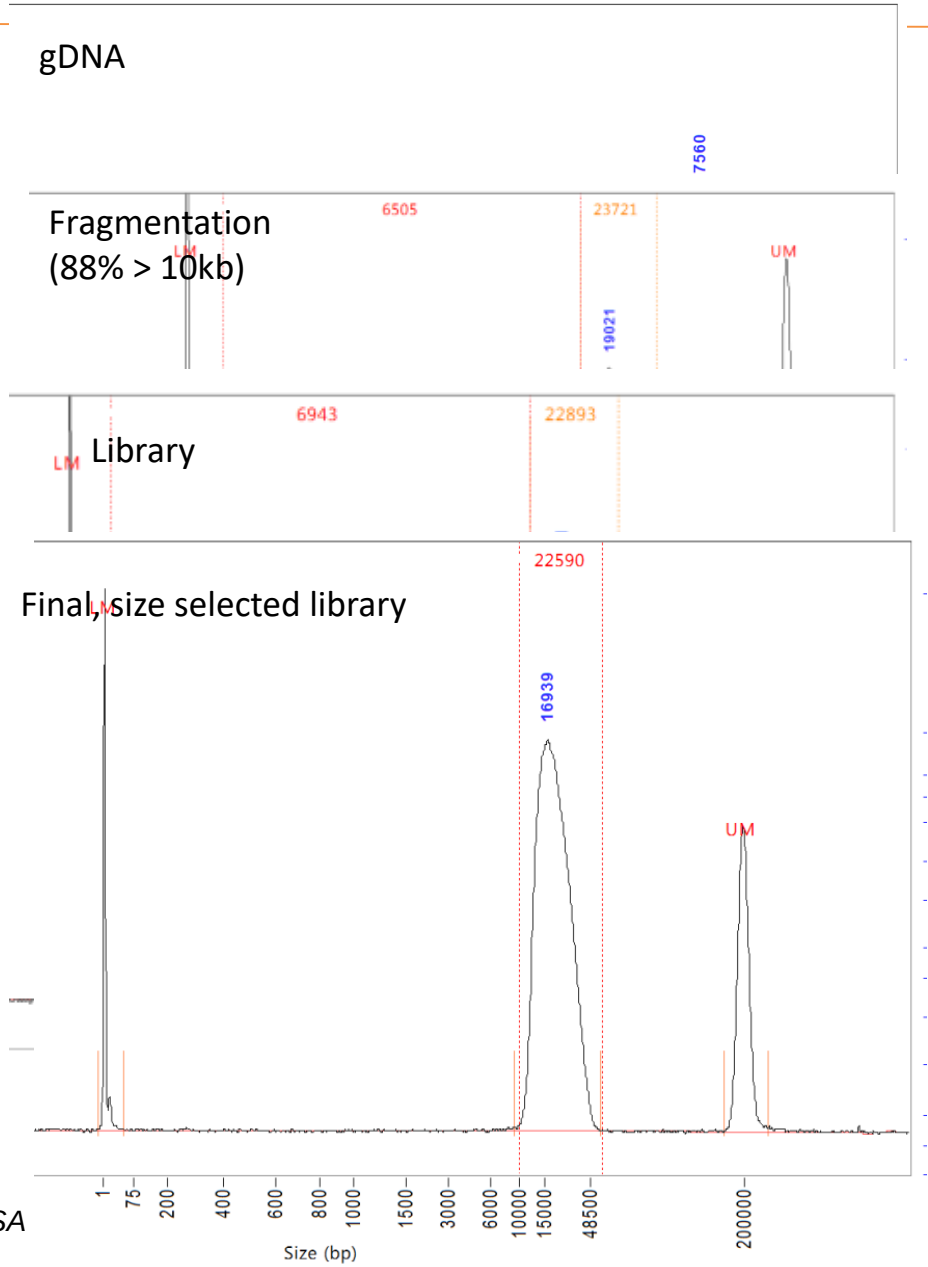
- ASSEMBLY: *De novo* assembly of insect/arthropod genomes (Up to 500 Mb)
- VARIANT DETECTION: Variant detection (SNPs, Indels, SVs) in human genomes (3 Gb)

Ultra-Low DNA Input: UNSUPPORTED APPLICATIONS

- ASSEMBLY: *De novo* assembly for microbes, plants, vertebrates, or other non-DNA limited sample types
- COMPLEX POPULATIONS: Metagenomics sequencing

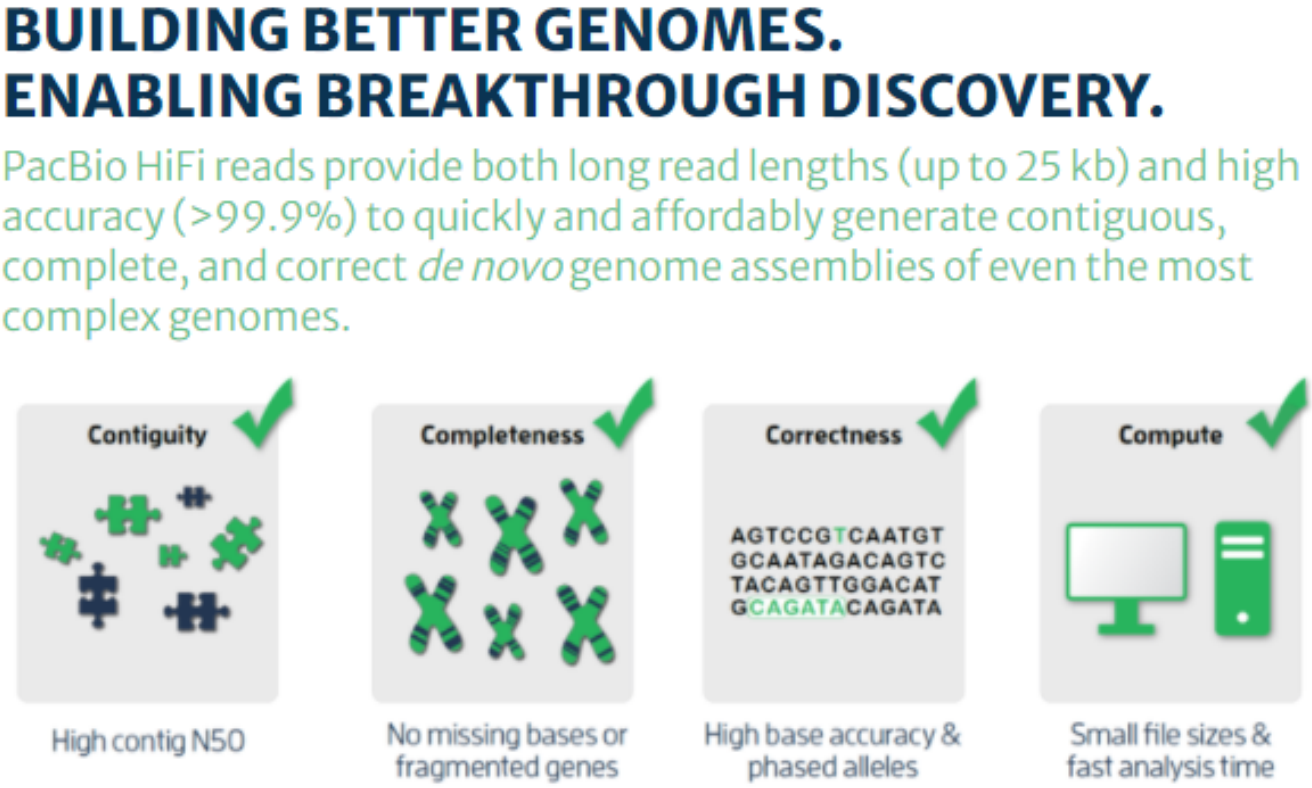
Library prep for reference quality assemblies





Courtesy of Pacific Biosciences of California, Inc., Menlo Park, CA, USA

Whole genome sequencing



What about hybrid approaches?

- ✗ **HiFi reads + short reads:** no benefit for contig building or polishing
- ✗ **HiFi reads + long reads:** may have marginal benefit to contiguity, but no readily available tools
- ✓ **HiFi + scaffolding:** technologies like optical maps and HiC help assign your high-quality HiFi genome assemblies into chromosomes

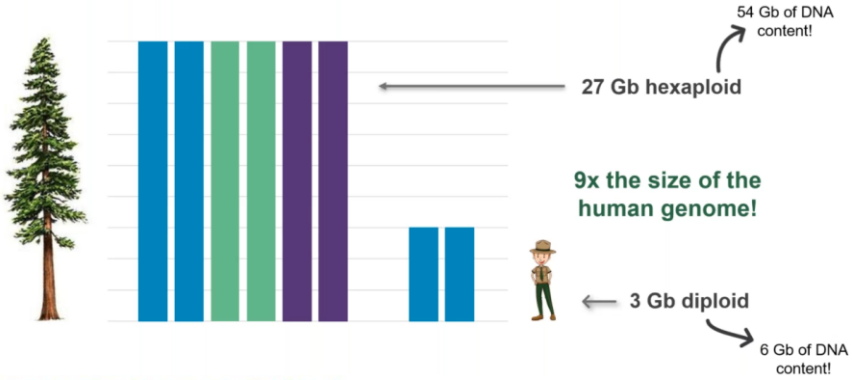
HiFi assembly of large genomes - redwood

THE CALIFORNIA (COASTAL) REDWOOD GENOME

Sequoia sempervirens

- One of the world's fastest-growing conifers
- Live for thousands of years
- Only 5% of the original old-growth coast redwood forest remains
- 27 Gb hexaploid genome
- Genome assemblies by ONT in 2019 and PacBio in 2020

THE REDWOOD GENOME IS LARGE AND COMPLEX



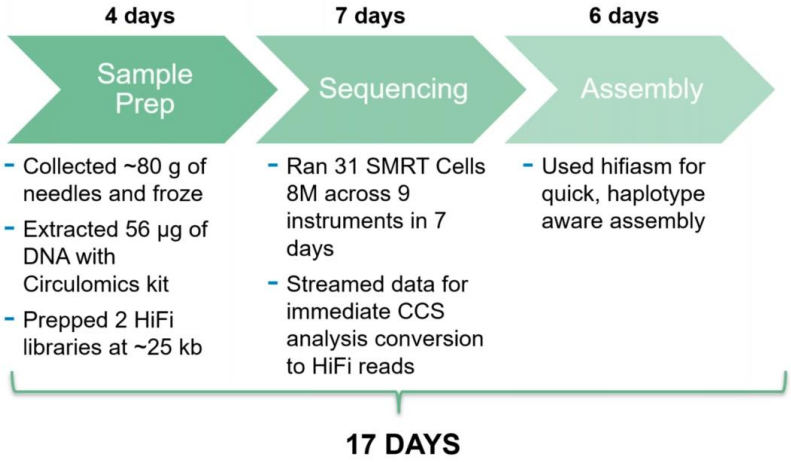
RESULTING GENOME ASSEMBLY

- Standard running parameters – no iteration
- Run on 64 cores with 512 Gb of RAM – no specialized or particularly large compute cluster

California Redwood Genome Assembly Results		
Methodology	**PacBio HiFi reads**	ONT + short reads[1]
Genome Coverage	**22-fold**	23-fold + 122-fold
Assembly Size (Gb)	**47.7**	26.5
Contig N50 (Mb)	**1.92**	0.11
BUSCO Complete	**59%**	56%
Mapped transcripts with frameshift errors[2]	**0.12%**	1.97%

BUSCO does not work well in conifers due to very long introns

THE PROJECT WORKFLOW

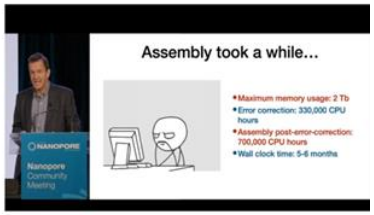


PacBio HiFi reads[1]

- 64 cores with 512 Gb of RAM
- ~46,000 CPU hours for HiFi generation ("error correction")
- 6 days wall time, ~7,200 CPU hours for assembly

6 days vs ***5-6 months*** of wall time for just genome assembly

ONT + short reads[2]



1. Sequencing and assembling mega-genomes of mega-trees: the giant sequoia and coast redwood genomes
2. Using transcript set of *Abies alba* from Neale, D. et al. consisting of 22,561 transcript sequences

Courtesy of Pacific Biosciences of California, Inc.

HiFi sequencing data available

NEW HIFI DATASETS – "TRY BEFORE YOU BUY"

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

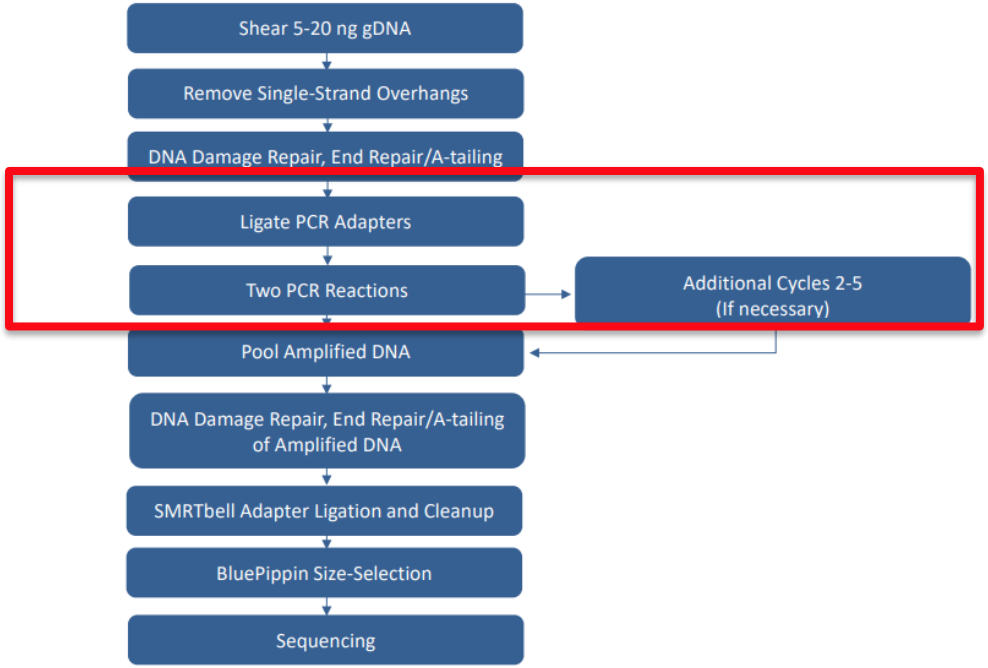
Highly accurate long-read HiFi sequencing data for five complex genomes

Ting Hon, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, David Kudrna, Michael A. Hardigan, Cynthia C. Steiner, Steven J. Knapp, Doreen Ware, Beth Shapiro, Paul Peluso, David R Rank

doi: https://doi.org/10.1101/2020.05.04.077180

Procedure & Checklist - Preparing HiFi SMRTbell® Libraries from Ultra-Low DNA Input

Required gDNA Input Amount	Required Quality of Input gDNA	gDNA Shearing Method	Target Sheared Fragment Size Distribution Mode	Amplification Target Size Distribution Mode	Total Mass of Pooled PCR Product Required for Library Construction	Required SMRTbell Library Input for BluePippin Size-Selection
5-20 ng	Majority of gDNA >20 kb	Megaruptor or g-TUBE	10 kb sheared DNA is optimal	8-10 kb	≥500 ng	≥400ng



Sequel II/IIe applications

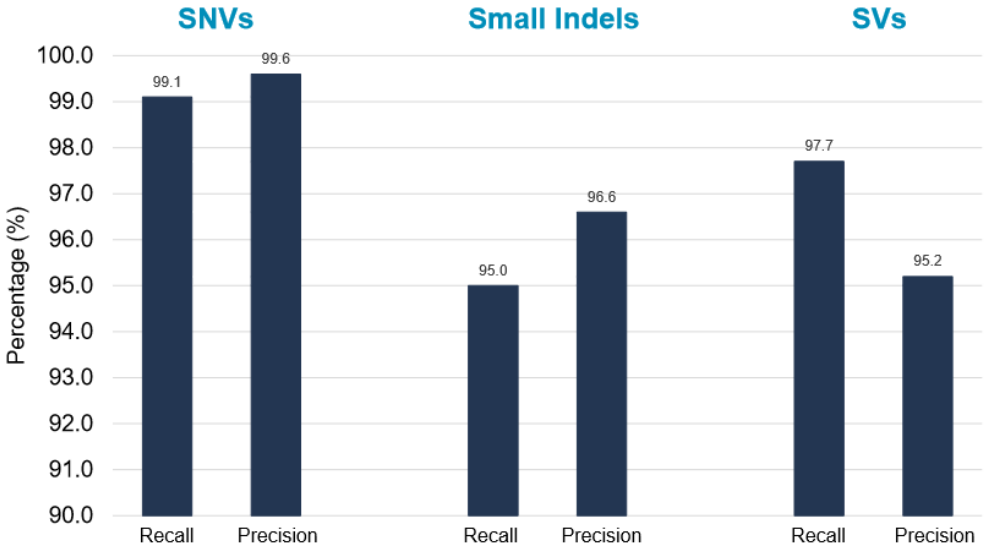
SEQUEL II SYSTEM KEY APPLICATIONS

Variant Detection Using Whole Genome Sequencing with HiFi Reads

- With highly accurate long reads (HiFi reads) from the Sequel II System, powered by SMRT Sequencing technology, you can comprehensively detect variants in a human genome. HiFi reads provide high precision and recall for single nucleotide variants (SNVs), indels, structural variants (SVs), and copy number variants (CNVs), including in difficult-to-map repetitive regions.

HiFi READS

EXAMPLE: VARIANT CALLING WITH HIFI READS



Variant calls from ~15-fold HiFi read coverage of a human genome (HG002) were measured against the Genome in a Bottle small variant benchmark (v3.3.2) for SNVs and indels using Deep Variant and SMRT Link 8.0 for SVs. Libraries were generated using a 15 kb insert and sequenced using Chemistry 2.0.

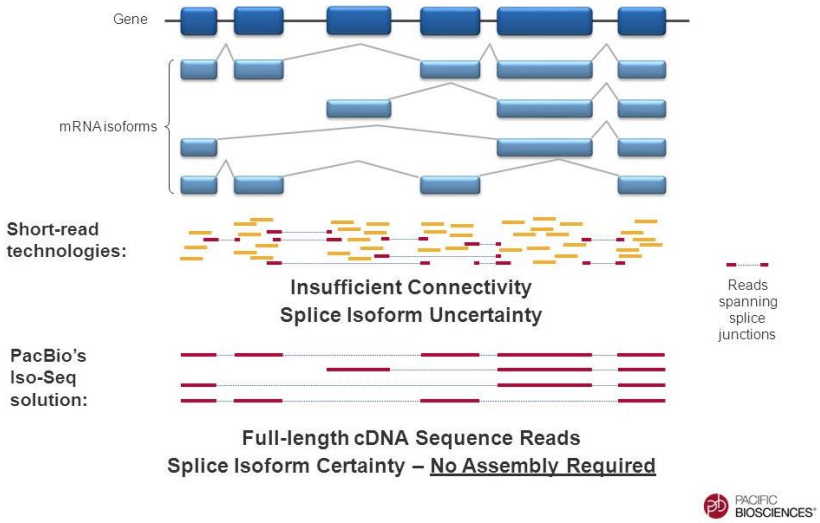
- With two 8M SMRT Cells you can call SNVs, InDels, and SVs in a 3 Gb genome
- ≥15-fold HiFi read coverage of a human genome

Sequel II/IIe applications

SEQUEL II SYSTEM KEY APPLICATIONS

Long-Read RNA Sequencing (Iso-Seq Analysis)

- With SMRT Sequencing and the Sequel II System, you can easily and affordably sequence complete transcript isoforms in genes of interest or across the entire transcriptome. The Iso-Seq method allows users to generate full-length cDNA sequences up to 10 kb in length – with no assembly required – to confidently characterize full-length transcript isoforms.



Output: 2-4 M full length reads

- One human transcriptome per 8M SMRT Cell
- Multiplex up to 12 samples
- Single-Cell Iso-Seq:
  - 1000 unique reads/ single cell for 3000 cells
  - 10 000 unique reads/ single cell for 300 cells

cDNA capture - project for identification of novel isoforms/ splice sites of 35 genes (input: total RNA)

Procedure & Checklist – Preparing Single-Cell Iso-Seq™ Libraries Using SMRTbell® Express Template Prep Kit 2.0

Intact (un-sheared) RT-PCR products initially generated using a third-party single-cell preparation system used as input.

- Although PacBio does not have a specific single-cell partner or system recommendation, in principle, practically **any single-cell platform should be compatible with single-cell Iso-Seq library preparation so long as that platform generates cDNA**.
  - For the Iso-Seq method to achieve full-length cDNAs, it is recommended to use a template-switching oligo (TSO). This is a common technique and is currently used in single-cell platforms and PacBio's current bulk Iso-Seq methods.

- For optimal analytical results, **PacBio recommends combining matching short-read and Iso-Seq datasets** (generated for the same exact single-cell library sample).
  - We recommend that the post-reamplification cDNA yield allow for parallel processing of both short-read sequencing and SMRT Sequencing.
  - The **Sequel System requires >80 ng of cDNA**, while the **Sequel II System requires >160 ng cDNA**. These are target DNA amounts for the PCR re-amplification steps for the Iso-Seq library construction workflow (see Page 4 of the procedure).

Sequel II/IIe applications

SEQUEL II SYSTEM KEY APPLICATIONS

Metagenomic Sequencing of Complex Populations

- The ability to identify and understand the functions of the complex microbial populations living in, on, and around us requires comprehensive characterization of each community member. Highly accurate long reads – HiFi reads – with single-molecule resolution make SMRT Sequencing and the Sequel II System ideal for full-length 16S rRNA sequencing, long-read metagenomic profiling, and shotgun metagenomic assembly

With one 8M SMRT cell you can:

- Characterize up to 96 samples using full-length 16S rRNA with strain level resolution (8000 reads per sample)
- Generate near-complete assemblies of high-complexity sample(s) (e.g. gut microbiome) - up to 4 communities per 8M SMRT cell

Metagenomics: 16S rRNA

The ISME Journal (2016) 10, 2020–2032

www.nature.com/ismej

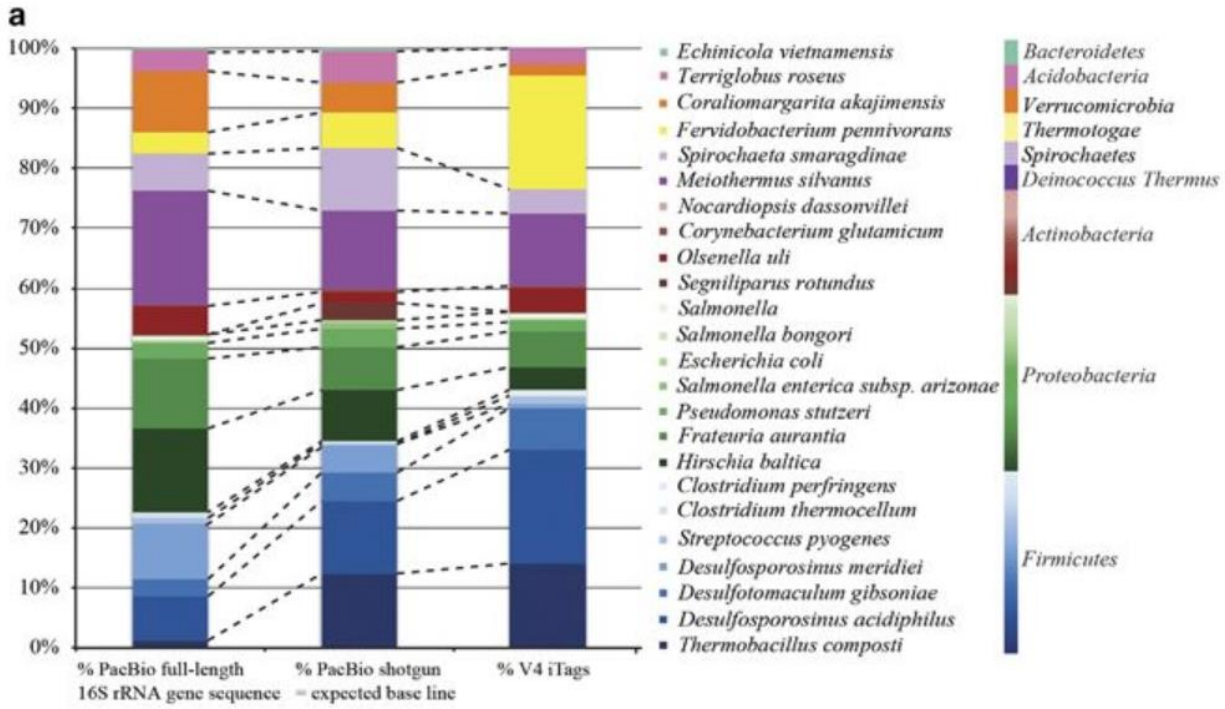
OPEN

ORIGINAL ARTICLE

High-resolution phylogenetic microbial community profiling

Esther Singer[1], Brian Bushnell[1], Devin Coleman-Derr[1,2], Brett Bowman[3], Robert M Bowers[1], Asaf Levy[1], Esther A Gies[4], Jan-Fang Cheng[1], Alex Copeland[1], Hans-Peter Klenk[5], Steven J Hallam[4], Philip Hugenholtz[6], Susannah G Tringe[1] and Tanja Woyke[1]
[1]*US Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA;* [2]*USDA-ARS, Albany, CA, USA;* [3]*Pacific Biosciences, Menlo Park, CA, USA;* [4]*University of British Columbia, Vancouver, BC, Canada;* [5]*Newcastle University, School of Biology, Newcastle upon Tyne, UK and* [6]*Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia*

- Study of **mock** community made up of 23 bacterial and 3 archaeal species and microbial community in Sakinaw Lake.
- **Conclusion**: Comparison with V4 iTag, using PacBio sequencing enables more accurate phylogenetic resolution of microbial communities and predictions on their metabolic potential.

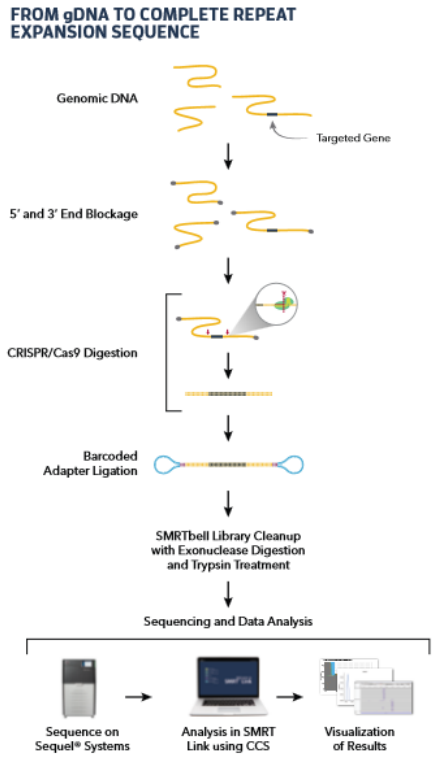


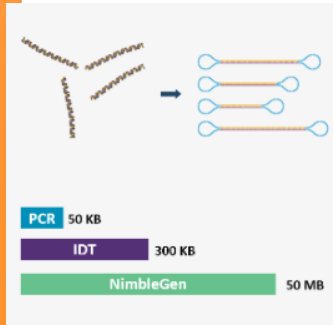
Sequel II/IIe applications – targeted sequencing

Application	Targeted Sequencing	
	Amplicon Sequencing	No-Amp Targeted Sequencing
With 1 SMRT Cell 8M you can:	Sequence 384 barcoded amplicons	Sequence 5 targeted regions in a multiplex of 10 samples
Minimum Recommended Coverage	30-fold ≥Q20 CCS read coverage for variant detection 6,000-fold ≥Q20 CCS read coverage for minor variant detection (1% sensitivity)	≥100-fold ≥Q20 CCS read coverage per target locus
Library Insert Size	500 bp - 15 kb	4-6 kb or larger
Minimum Input Amount	250-500 ng for 250-1000 bp 500-1000ng for 1-3 kb bp 1000-2000 ng for 3-10 kb 3000 ng for 15kb	5 to 10 µg (represented by either a single sample or the total of multiple samples that will be multiplexed)
Multiplexing/SMRT Cell	Up to 1,000+ samples/ SMRT Cell 8M or SMRT Cell 1M	Up to 10 samples/SMRT Cell
Sequencing Mode	CCS	CCS

No-amplification targeted sequencing using CRISPR/Cas9 system:

- Challenging regions for PCR amplification (repeat expansions, low complexity regions)
- No PCR bias
- Preserves epigenetic modification signals



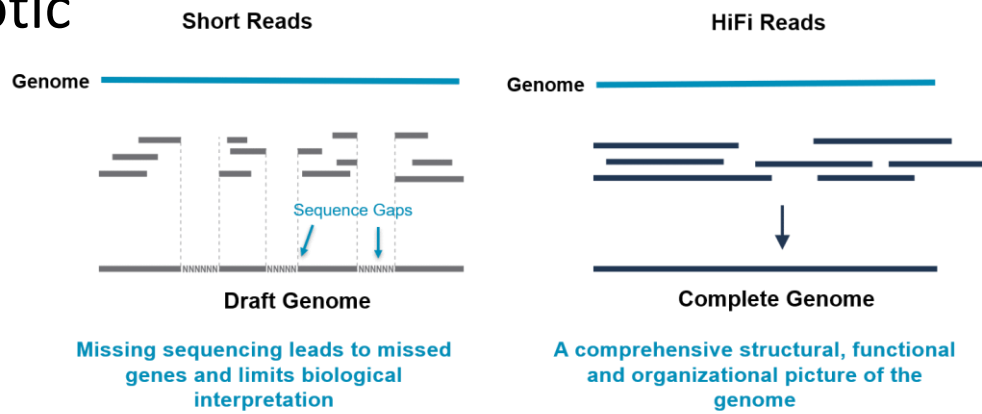


When targeting >50 kb genomic regions – use probe-based capture using DNA oligo hybridization.
Protocols available for:

- IDT xGen Lockdown probes
- Nimbelgen SeqCap EZ

PacBio applications at NSC

- *de novo* sequencing of prokaryotic and eukaryotic organisms –
  - Multiplexing up to 96 bacterial samples
  - Mostly HiFi library prep and sequencing for large genomes



- Sequencing of full-length transcriptomes – IsoSeq
  - Multiplexing up to 12 samples for genome annotation

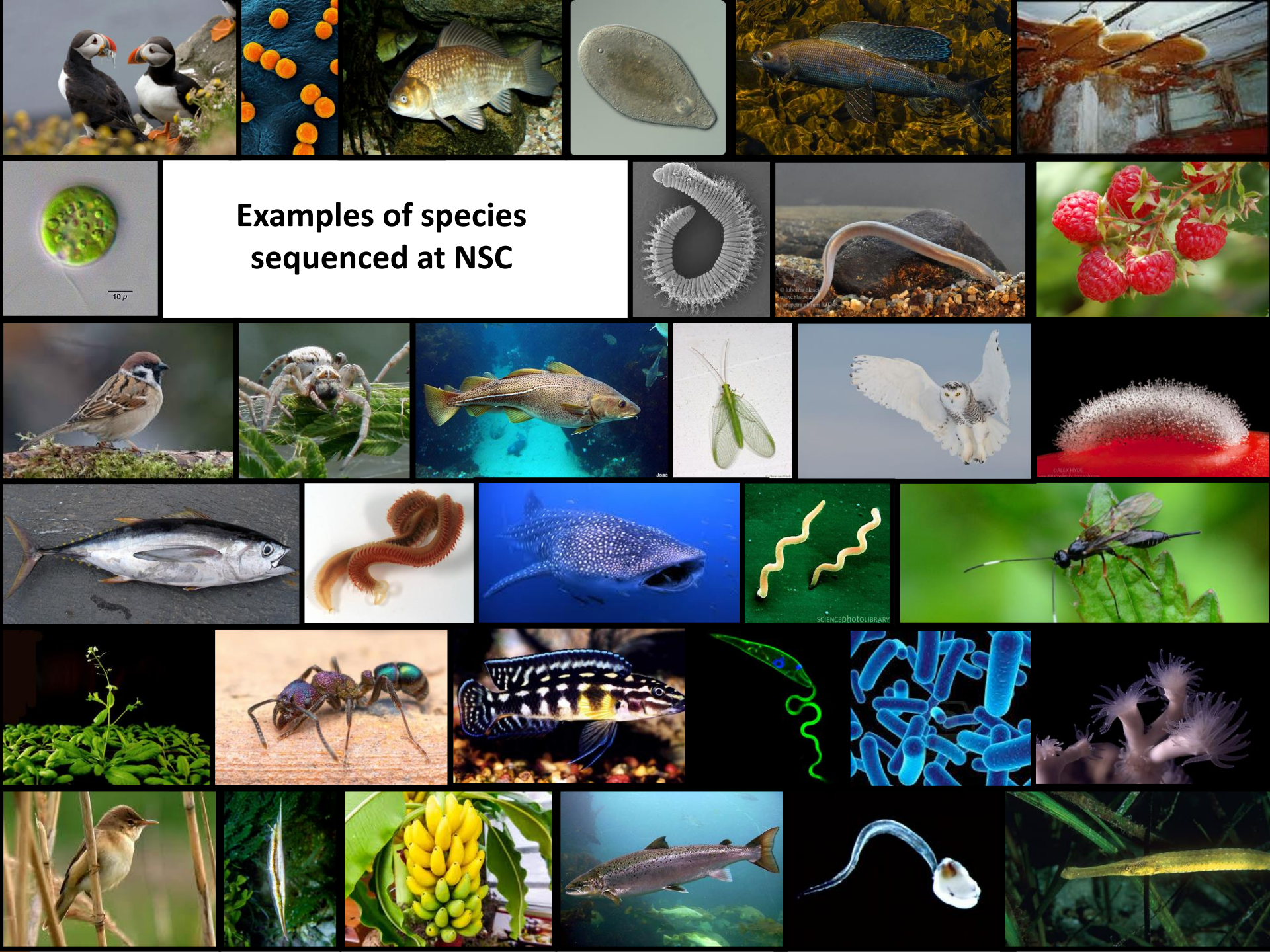
"The way we do RNA-seq now is… you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again**… If you think about it, it's kind of a **crazy way to do things**"

Michael Synder
Professor and Chair of Genetics
Stanford University

Tal Nawy, End to end RNA Sequencing, *Nature Methods*, v10, n10, Dec . 2013, p1144–1145

- Targeted sequencing – amplicons and sequence capture

post@sequencing.uio.no
https://www.sequencing.uio.no/


Examples of species sequenced at NSC

Largest ongoing project: EBP-Nor

The Norwegian EBP initiative (EBP-Nor)

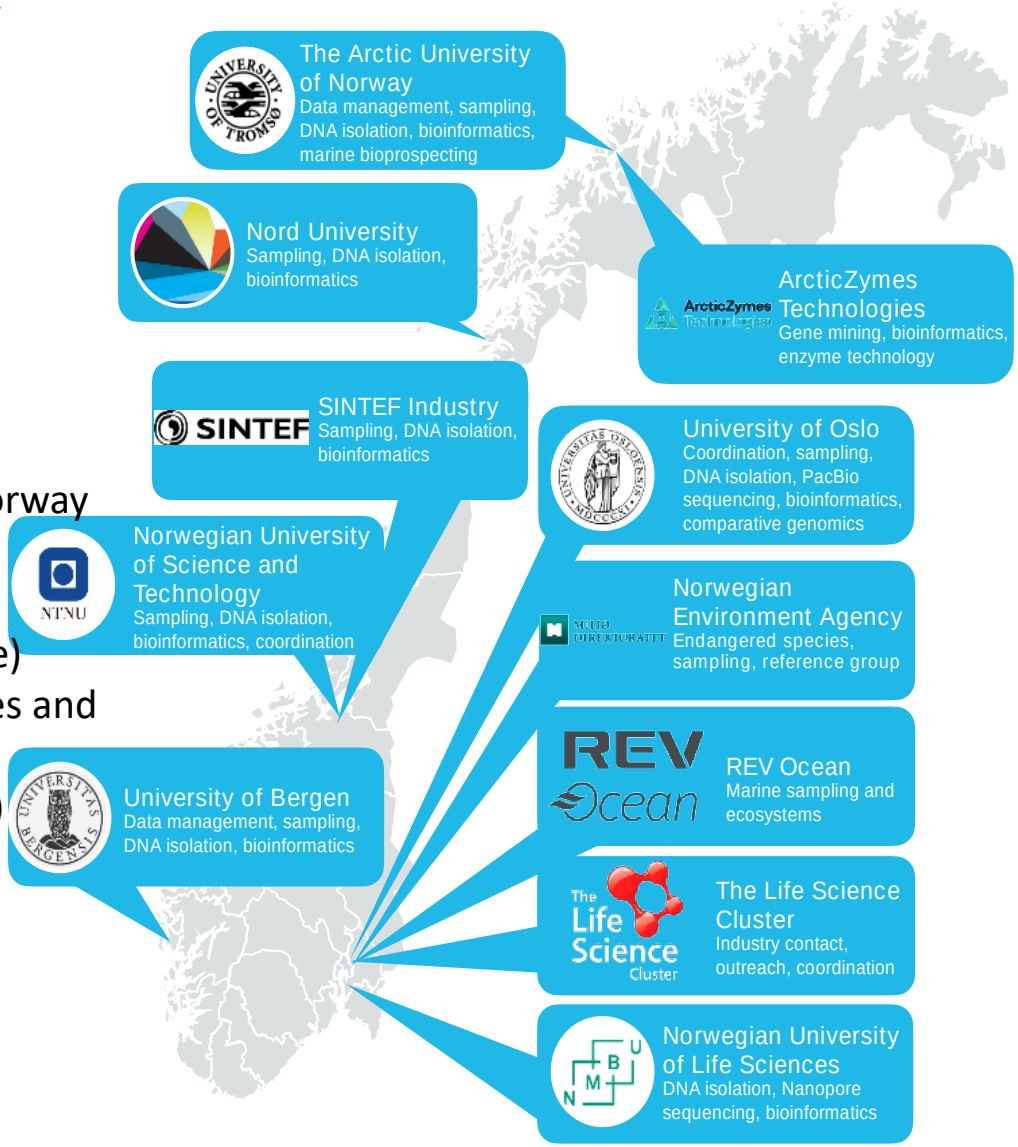
Planned in 3 phases

Phase 1 2021-2024 (30 million NOK)

- Funded by the Research Council of Norway
- Planned 150 species
- Norwegian and arctic species
- Marine species (sampling competence)
- Coordination with the Nordic countries and ERGA (and EBP, VGP, DToL etc..)
- Several genomes underway (HiFI, HiC)

Preparation for 2 phase has begun

The Research Council of Norway



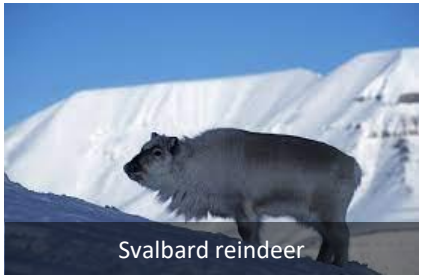
The first EBP-Nor genomes are finished


Brook lamprey


River lampreye


25 fungal species


4 moss species


Svalbard reindeer


Snowy owl

In progress: insects (bumble bee, lacewing), Atlantic puffin, cod and salmon (improved), bird cherry (*Prunus padus*), cloudberry ++

Conference in 2023:
Norwegian Biodiversity & Genomics

Feb. 8, 2023 12:00 PM–Feb. 10, 2023 1:00 PM,
Gamle festsal

https://www.ebpnor.org/

cees-drift@sequencing.uio.no
NSC team at CEES: Ave Tooming-Klunderud
Morten Skage
Spyros Kollias