NORWEGIAN SEQUENCING CENTRE

SMRT sequencing

Ave Tooming-Klunderud, NCS/CEES/UoO,
ave.tooming-klunderud@ibv.uio.no

Oslo universitetssykehus

The Research Council of Norway

CEES Centre for Ecological and Evolutionary Synthesis

HELSE SØR-ØST

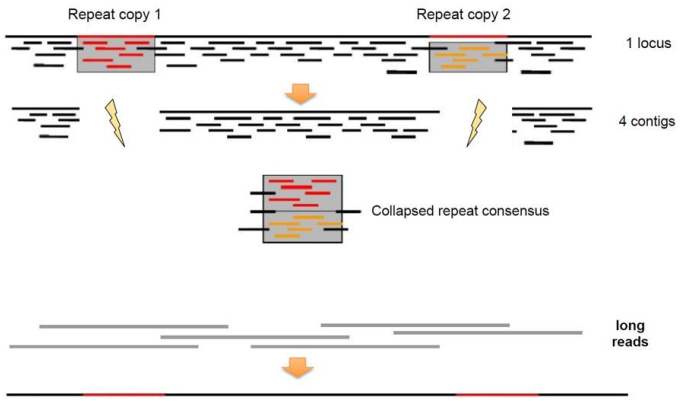
UNIVERSITAS OSLOENSIS MDCCCXI

Short read vs long read sequencing

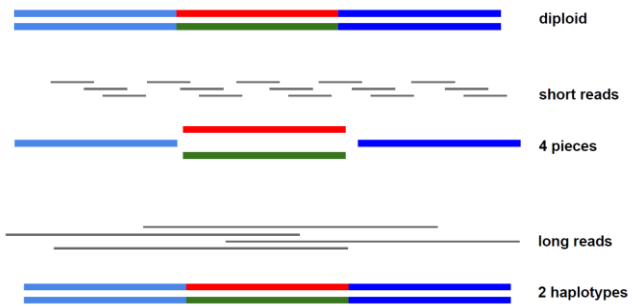
Short read sequencing	Long read sequencing
Amplification during sequencing High read accuracy	No amplification needed Low single read accuracy/high consensus accuracy
Requirements for input DNA:	
Works with almost any DNA sample Low amount Fragmented DNA	High quality DNA needed High amount DNA fragments at least 40 kb long
Price:	
Low	High

Why long reads?

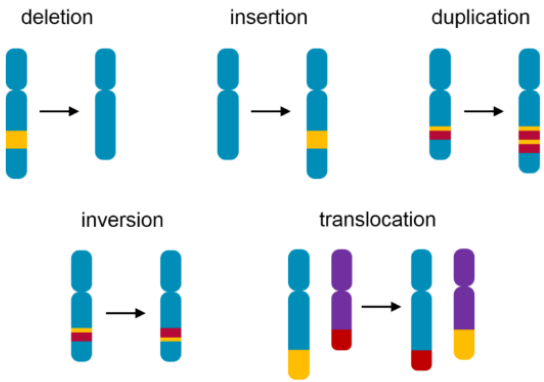
- Long reads span repeats - Complete genomes for small and simple genomes

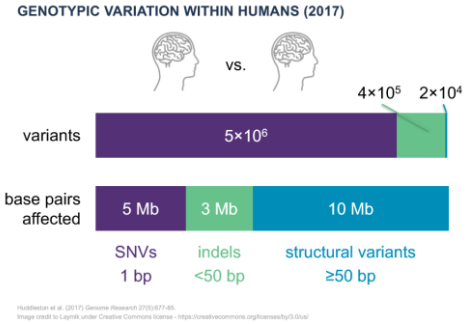


- Phased haplotypes/genomes - can solve heterozygosity



- Structural variation





Long read sequencing

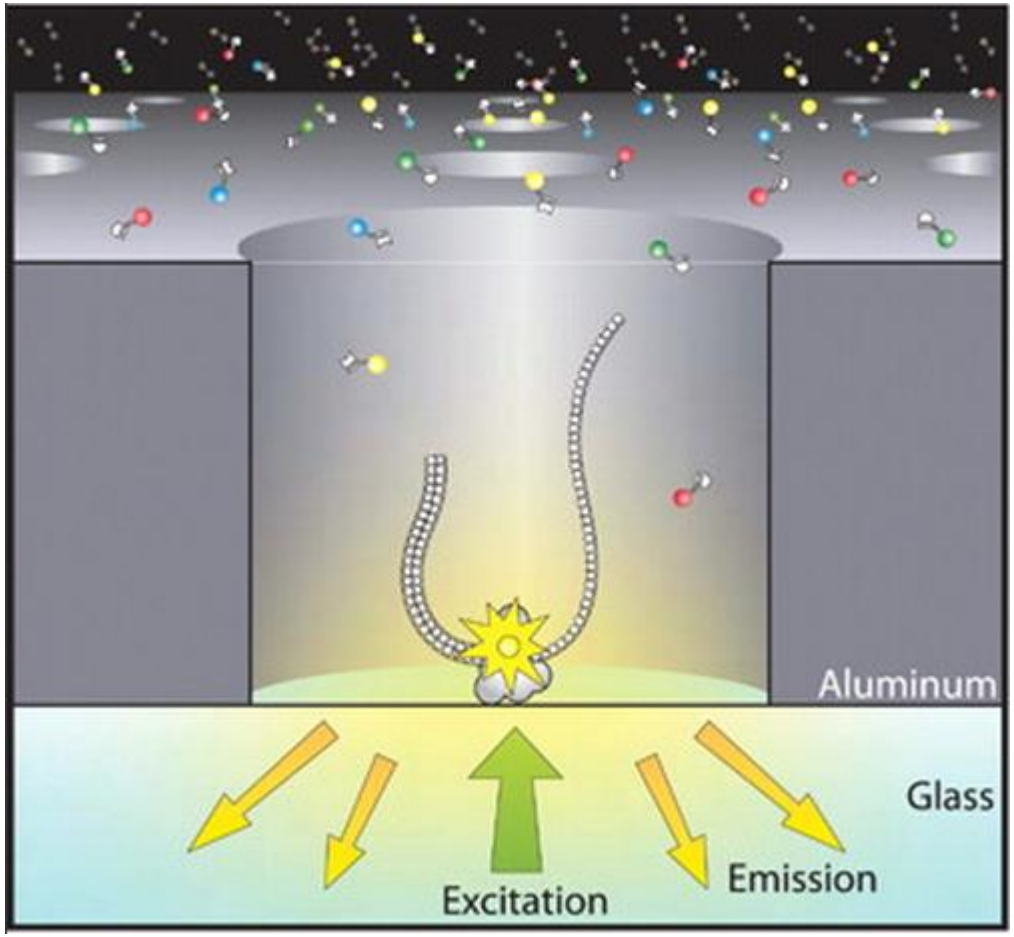
- Single DNA molecule is sequenced (no amplification is involved)
  - Entire genome can be sequenced, also GC/AT rich and repetitive regions
  - Possible to detect DNA modifications

- Two companies – two different technologies:
  - Pacific Bioscience – SMRT sequencing
  - Oxford Nanopore – Nanopore sequencing

The PacBio sequencing technology

PACIFIC BIOSCIENCES™

- Based on the observation of DNA synthesis in real time and involves:
  - The PacBio RS II or Sequel instrument
  - Single DNA molecule bound to the polymerase
  - SMRT cell
  - Phospholinked nucleotides with different coloured fluorophores – light pulse is produced during incorporation of the base

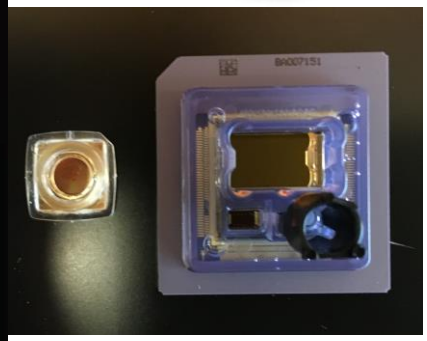


Zero-mode waveguide

PacBio's long-read instruments

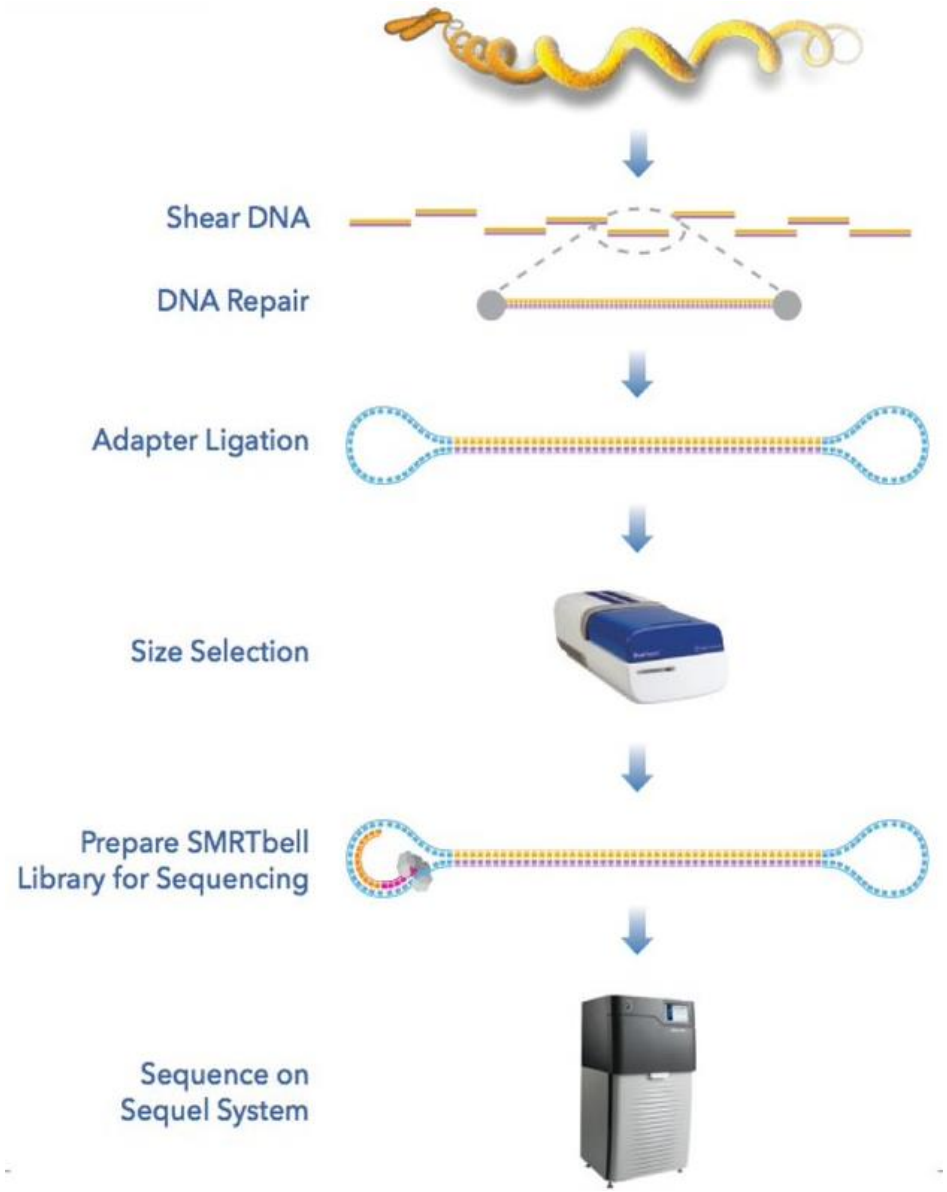
RS II - in service January 2012-March 2018

Sequel



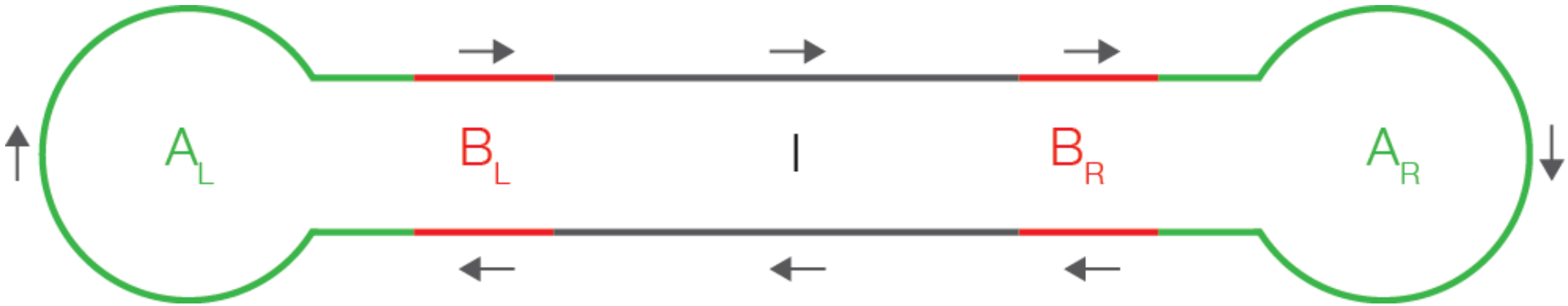
RSII – 150 000 ZMWs
Sequel - 1M SMRT cell = 1 M ZMWs,
8M SMRT cell expected early 2019

Sample prep and sequencing



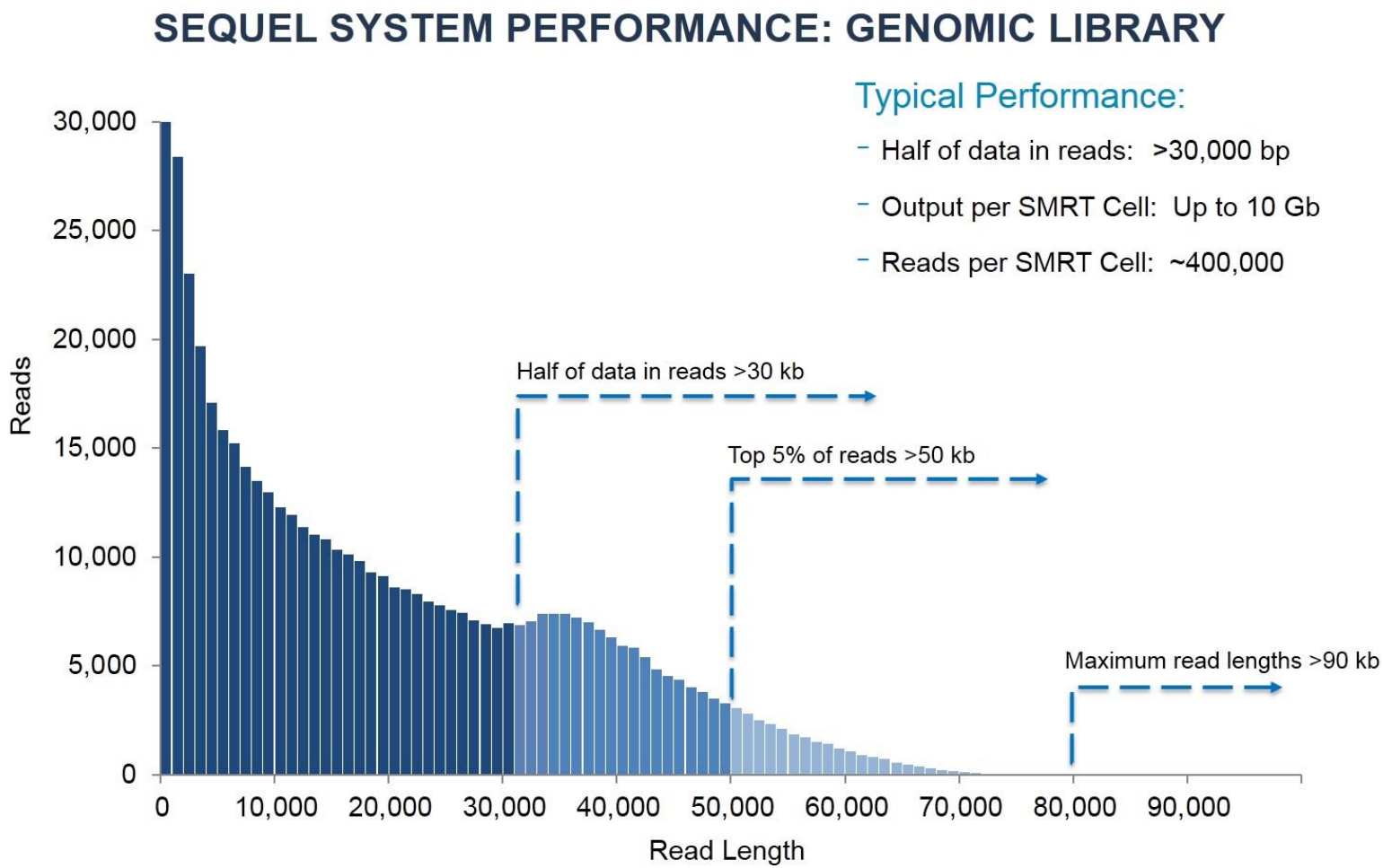
Genomic DNA

SMRTbell library



Sequel performance

SMRT cell 1M v2 LR, polymerase v2.1, run time 20 hours

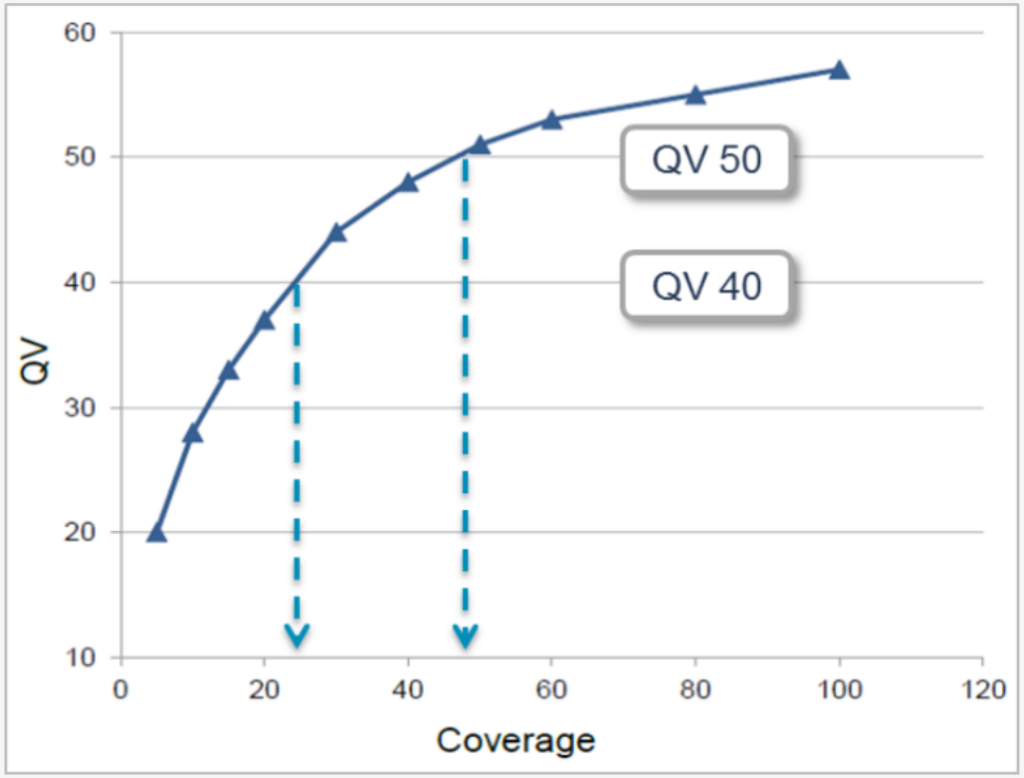


20 Gb with 40 kb average read lengths for amplicon and RNA sequencing projects

Sequence performance: accuracy

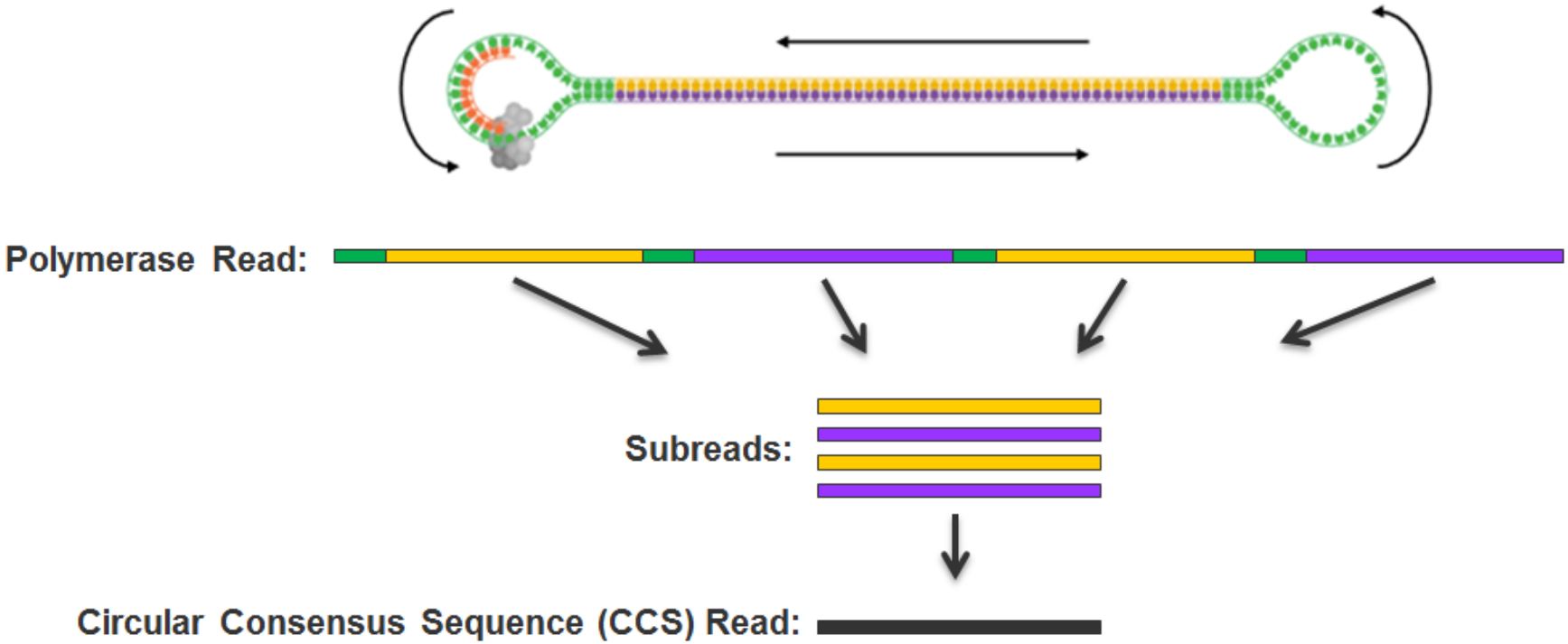
- Single read accuracy 85-87 %
- Lack of systematic sequencing errors
- 99.999% (QV50)

Very low systematic error means PacBio SMRT Sequencing can achieve the highest levels of consensus accuracy



Data generated with 25 kb *E. coli* library on a Sequel System using 2.1 chemistry and SMRT Analysis v 5.1

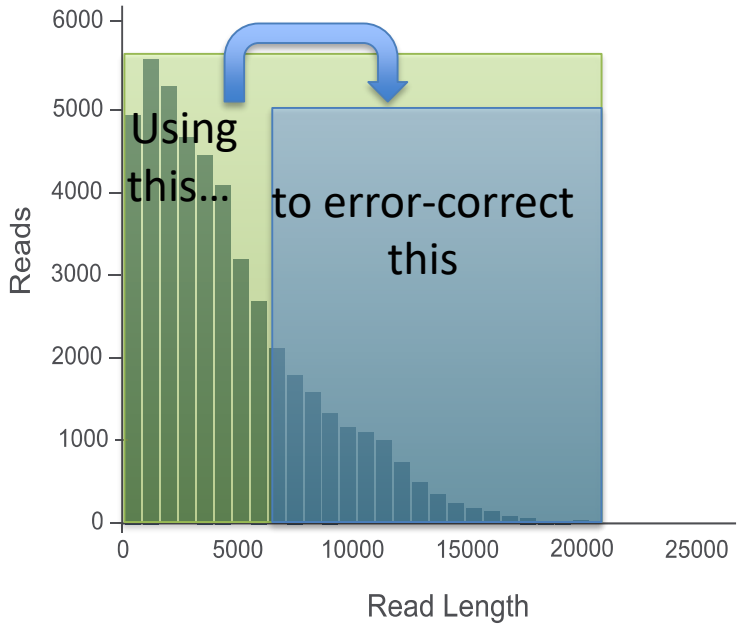
Consensus accuracy – short template



http://files.pacb.com/software/smrtanalysis/2.2.0/doc/smrtportal/help/!SSL!/Webhelp/Portal_PacBio_Glossary.htm

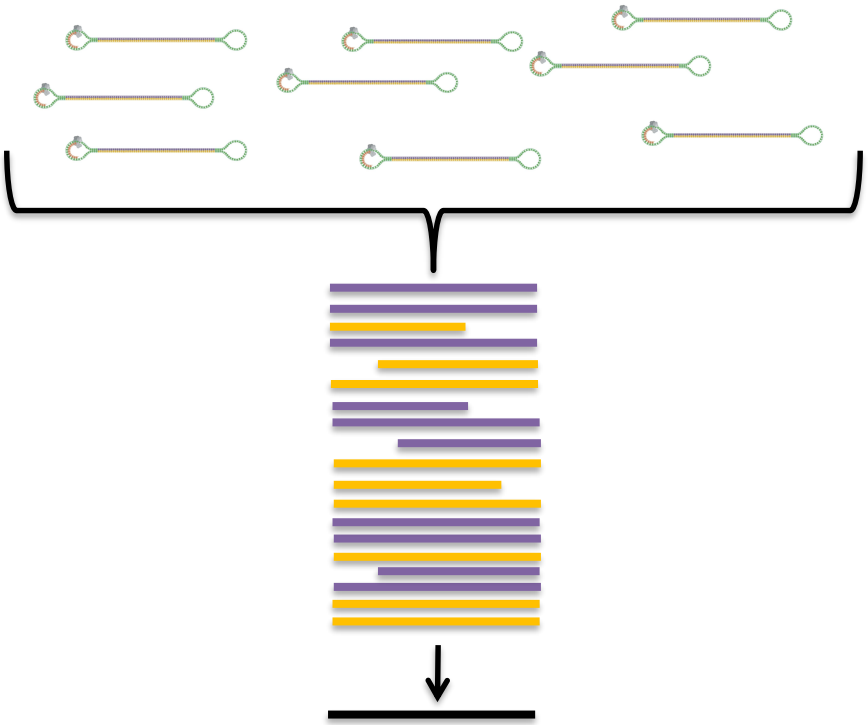
Consensus accuracy – long template

Genomic DNA: *de novo* assembly (60-100 x coverage needed)



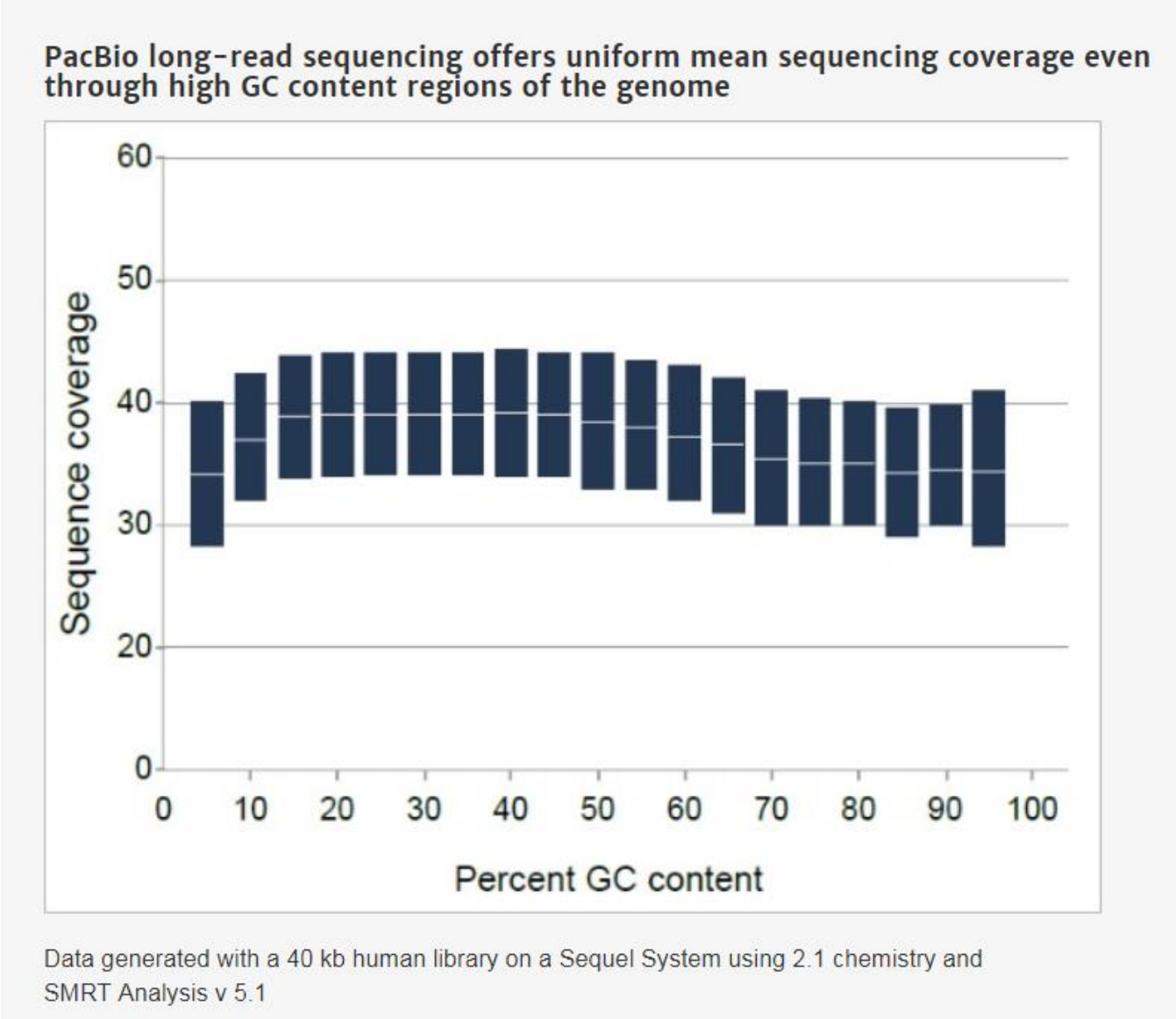
doi:10.6084/m9.figshare.818957

Long amplicons: consensus is built using subreads from different fragments



Sequence performance: uniformity

Lack of GC content or sequence complexity bias



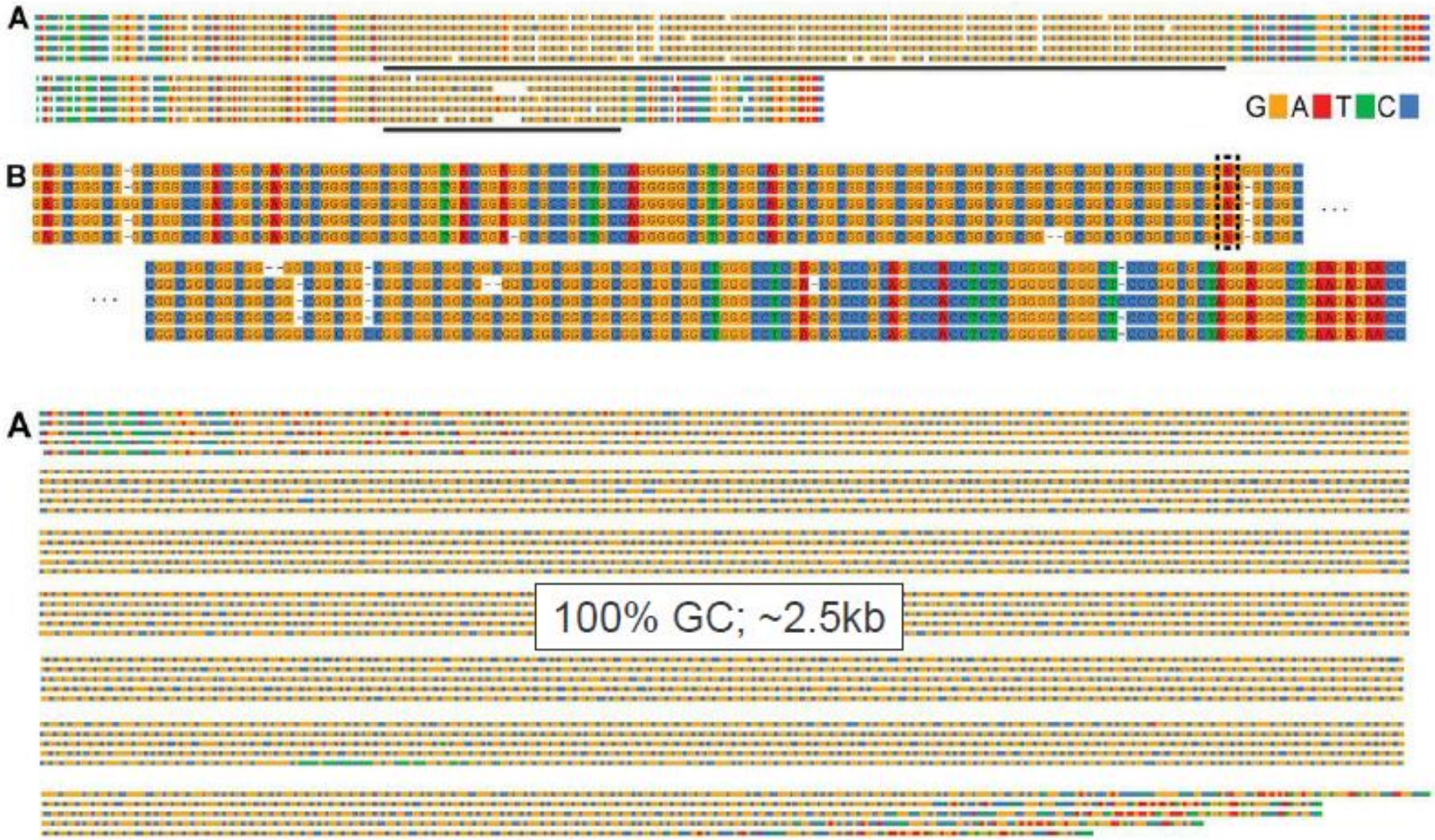
Sequence performance: uniformity

High GC - Trinucleotide Repeat Disorders

- Set of genetic disorders caused by trinucleotide repeat expansion

- A mutation where trinucleotide repeats in certain genes exceed the normal, stable, threshold, which differs per gene
- Disease Examples:
  - Autism
  - Mental retardation (especially males)
  - Huntington's disease
  - Fragile X syndrome

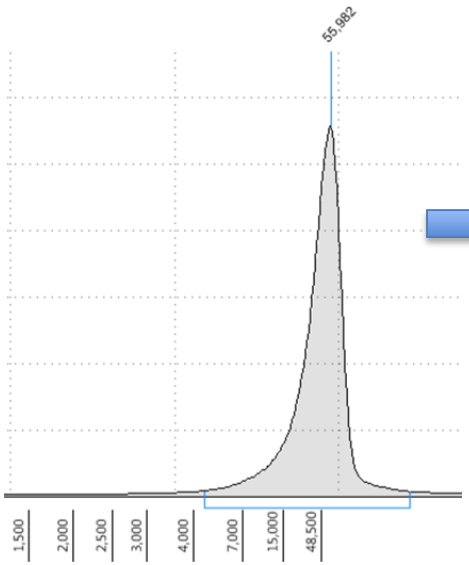
Sequence performance: sequencing the unsequenceable



Loomis et al. (2013) Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome Research 23: 121-128.

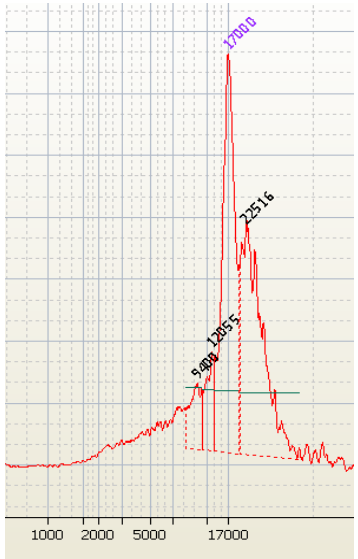
DNA quality: structural integrity

Input: HMW DNA

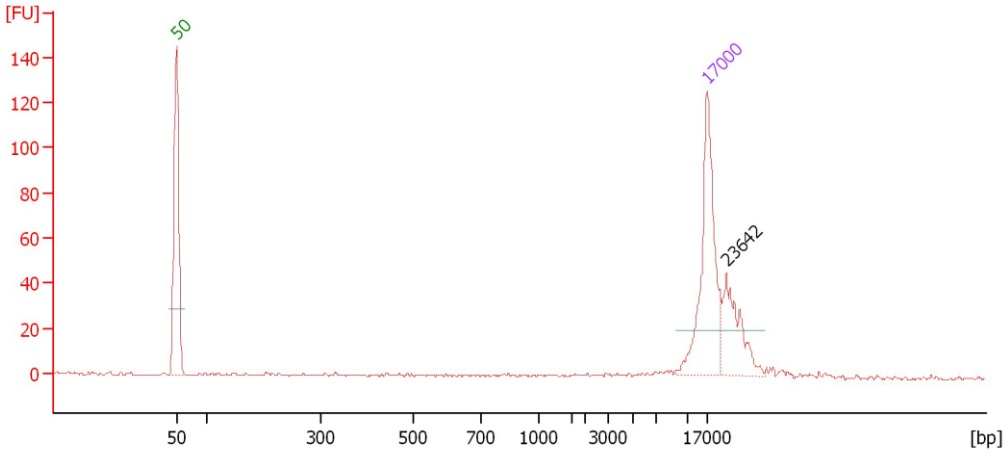


Gentle fragmentation

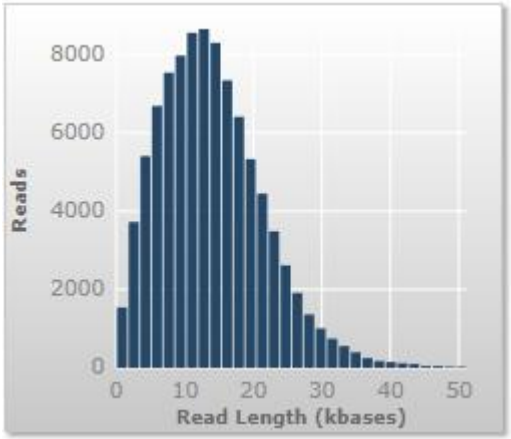
Long library



BluePippin size selection to remove short fragments



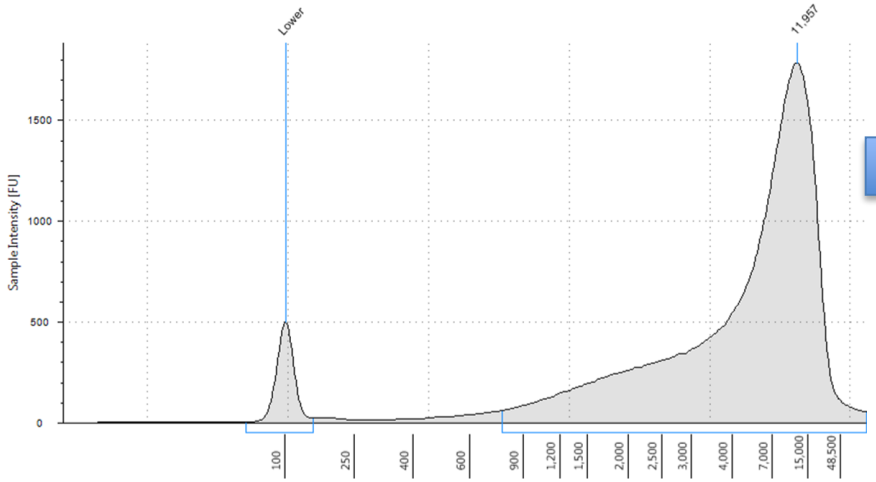
Final library



Average length of insert reads: 15 kb

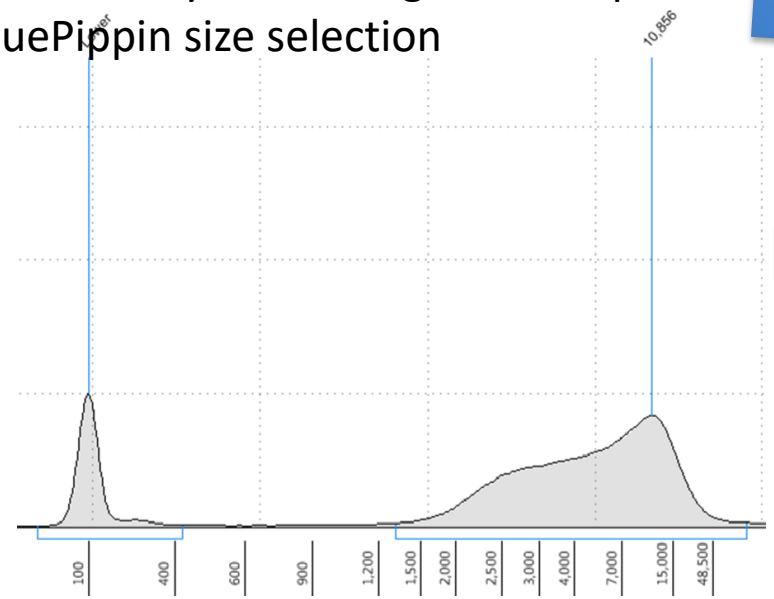
DNA quality: structural integrity

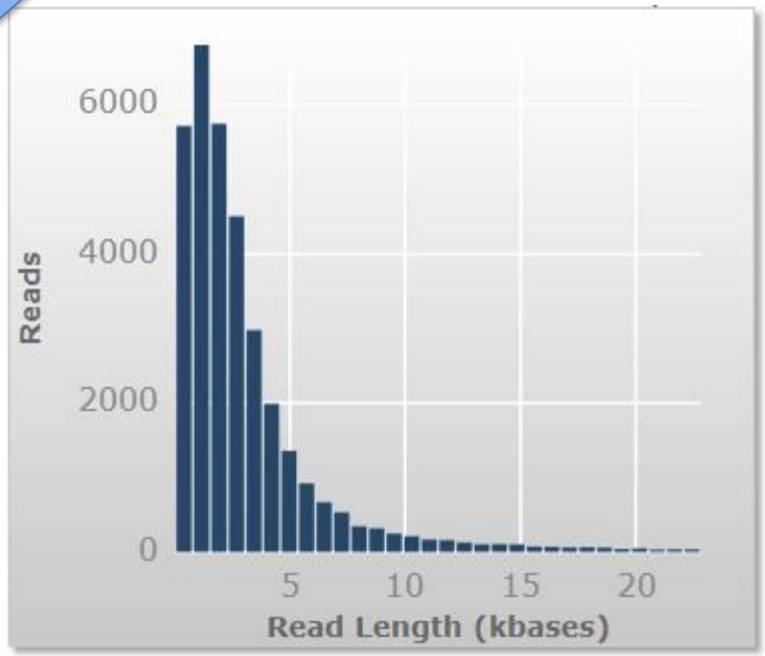
H2 Input: Highly fragmented gDNA



No fragmentation needed before library preparation

Short library, not enough DNA to perform BluePippin size selection

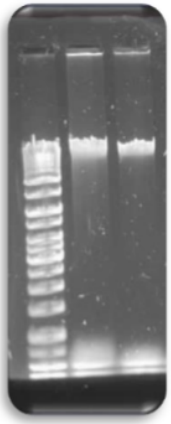


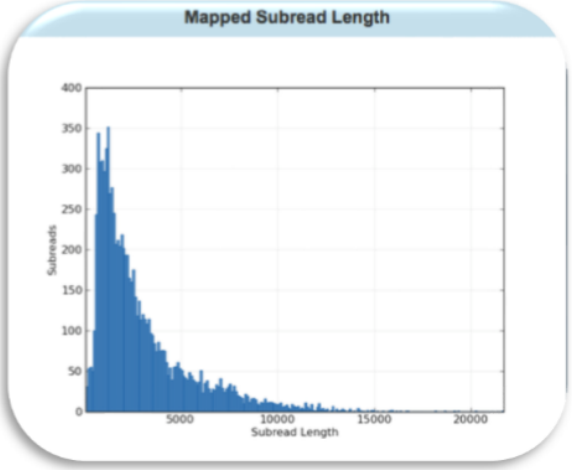


Average length of insert: 2.4 kb

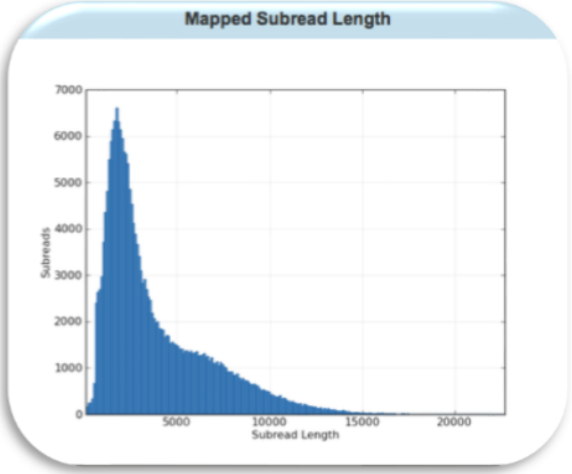
DNA quality: chemical purity

Same yeast, different DNA
FOCUS: chemical purity





Polished Contigs	223	Max Contig Length	36,298
N50 Contig Length	2,932	Sum of Contig Lengths	480,087



Polished Contigs	9	Max Contig Length	1,508,929
N50 Contig Length	1,353,702	Sum of Contig Lengths	7,813,244

For Long Reads one needs to have *long and pure* DNA

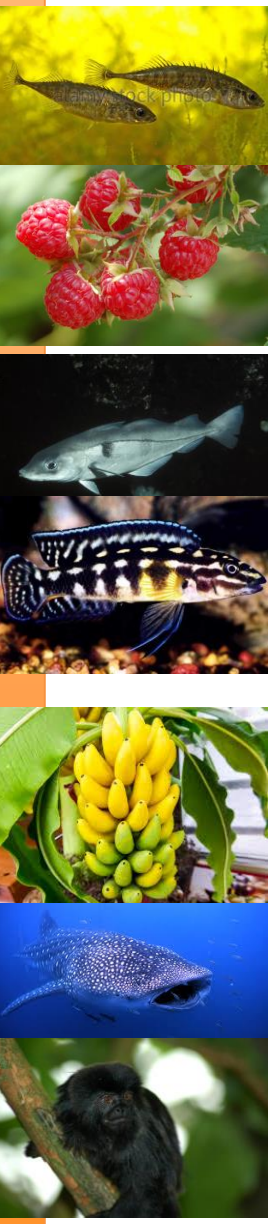
SciLifeLab
Olga Vinnere Pettersson

PacBio applications: WHOLE GENOME SEQUENCING

- *de novo* genome assembly:
  - gold-standard reference genomes
    - Gold genome – a high-quality, highly contiguous representation of the entire genome
  - pan-genomes to characterize the complete genetic diversity within a species
  - population-specific reference genomes to drive precision medicine
  - near-complete microbial genomes and their plasmids in a single experiment

- Structural Variant Calling
  - Mostly used in human research
  - Low-coverage (10-20x) required

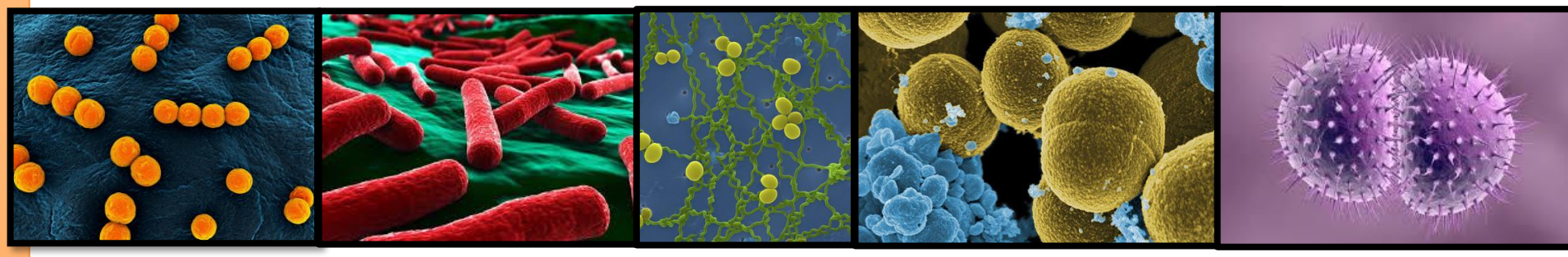
de novo assembly – large genomes



Different options:

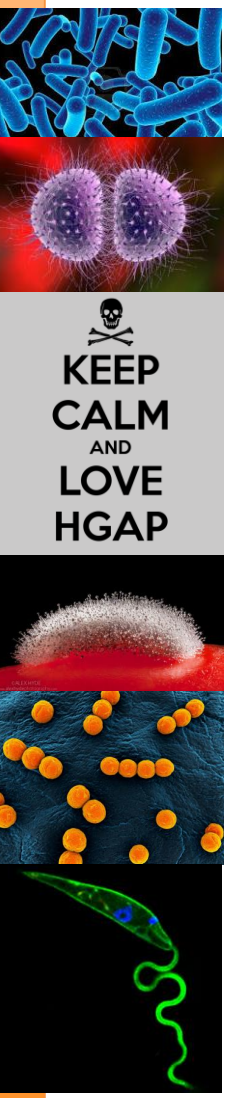
- PacBio-only assembly – at least 60-70 x coverage (short read data still needed for error correction)
- Hybrid *de novo* assembly – at least 30 x PacBio combined with short read data.
- Gap filling – internal gaps in mate-pair assembly are filled using PacBio reads (at least 10x)
- Scaffolding – PacBio reads (at least 10x) are used to join contigs of an existing short read assembly

Microbial whole genome sequencing

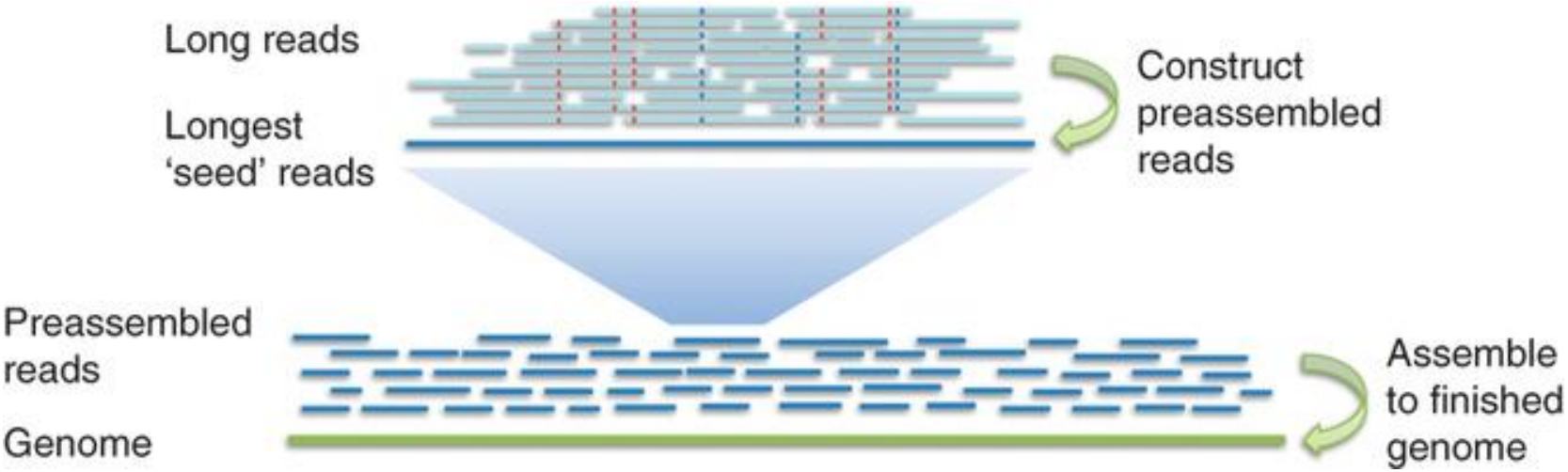


- Generate platinum-standard, closed reference genomes
  - Platinum genome – a contiguous, haplotype resolved representation of the entire genome
  - Ideally - one bacterial genome sequenced on one SMRT cell using long library – very expencive
- Affordably assemble gold-standard genomes by multiplexing up to 16 microbes in one SMRT Cell
  - Shorter library length (10 kb recommended)
  - Max genome size 30 Mb per SMRT cell

de novo assembly – small genomes



- Bacteria and small eukaryotic genomes
  - HGAP assembler

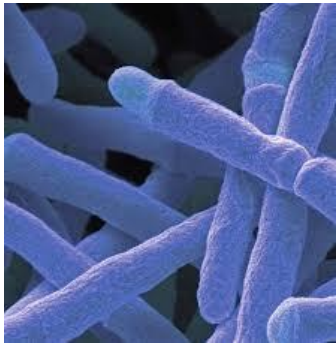


  - Sample quality is crucial.
    - Good quality – an (almost) complete genome
    - poor quality – partial or no genome.

Not all errors are random after all…

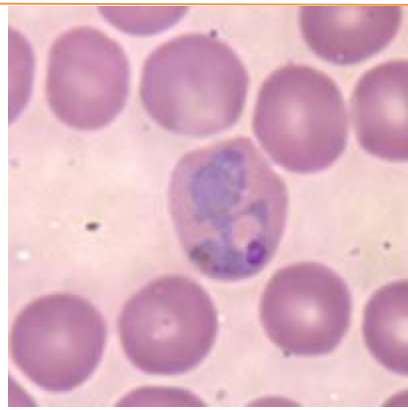
Based on feedback from NSC's customers:

- *Thermus thermophilus*
  - hyperthermophilic bacteria, capable of growing in temperatures up to 85°C
  - 2.2 Mb genome, GC content: 69%
  - Homopolymer errors in PacBio data: deletion of Gs and Cs in homopolymeric regions – 2174 errors corrected by short read sequencing



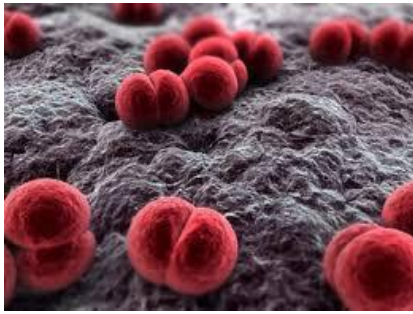
- *Mycobacterium hassiacum*
  - thermophilic mycobacterium that was isolated in human urine in 1997.
  - 5 Mb genome, GC content: 70 %
  - Homopolymer errors in PacBio data: deletion of Gs and Cs in homopolymeric regions – around 500 errors corrected by short read sequencing

Not all errors are random after all…



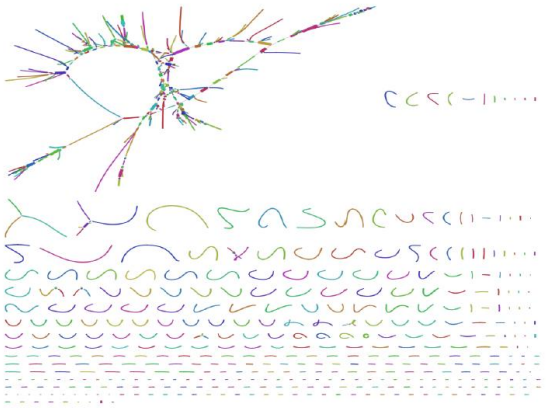
- *Plasmodium*
  - unicellular eukaryotes, obligate parasites of vertebrates and insects
  - 22.9 Mb genome size, GC content: 20%
  - Homopolymer errors in PacBio data: systematic insertions of As and Ts in ALL homopolymeric A/T regions

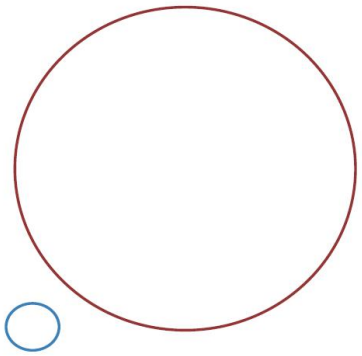
Example on few PacBio sequencing errors:



- *Neisseria meningitidis*
  - Gram-negative bacterium that can cause meningitis
  - 2 – 2.3 Mb genome size, GC content: 51.5%
  - In total, 8 isolates were sequenced using both PacBio and Illumina. Number of bases corrected using Illumina reads:
    - 0 bases (no correction needed) – three isolates
    - 1 base - one isolate
    - 2 bases - two isolates
    - 11 bases – one isolate
    - 14 bases – one isolate

How do I know that my PacBio-only assembly of the bacterial genome needs error correction using Illumina reads?





Is the assembly biologically meaningful?

Annotate the genome and:

- If possible, compare the annotated genome with previously annotated genomes of same/similar species
- Check if number of disrupted ORFs is higher than expected
- Check if genes coding for the proteins essential for life are annotated

Structural Variant Calling in human research

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

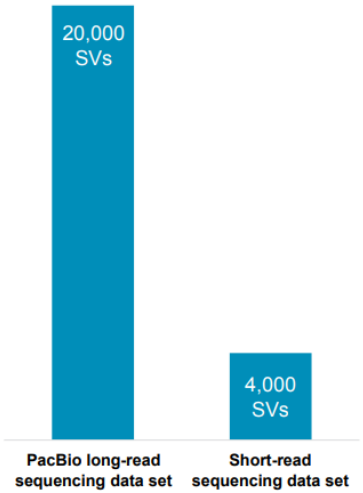
Structural Variant	Disease Examples	PacBio Advantage
INSERTION 	Charcot-Marie Tooth disease, Tay-Sachs disease	- Base pair resolution of breakpoints - Complete inserted sequence
DELETION 	Williams syndrome, Duchenne muscular dystrophy, Smith-Magenis syndrome, Carney Complex	- Base pair resolution of break points - High sensitivity even in repeats
INTERSPERSED DUPLICATION 	APP in Alzheimer's disease, Potocki-Lupski syndrome, Prader-Willi syndrome, Angelman syndrome	- Precise copy number - Base pair resolution of the duplicated sequence - Genomic context of additional copies
TRANSLOCATION 	Down syndrome, XX male syndrome (SRY), schizophrenia (chr 11) , Burkitt's Lymphoma	- Detection of balanced events - Complete sequence information - Unambiguous resolution of genomic context
INVERSION 	Hemophilia A, Hunter Syndrome, Emery-Dreifuss muscular dystrophy	- Detection of balanced events - Continuous sequence information - Base pair resolution of break points
TANDEM DUPLICATION 	FMR1 in Fragile-X, Huntington's disease, Spinocerebellar ataxia	- Complete repeat sequence, including interruptions - Quantitation of repeat expansions

Table 1. Structural variants of all types are known to cause Mendelian disease and contribute to complex disease. *All of these variants can be most robustly detected by PacBio SMRT Sequencing.*

Structural Variant Calling in human research

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

LONG-READ SMRT SEQUENCING PROVIDES HIGHER SENSITIVITY FOR SV DISCOVERY



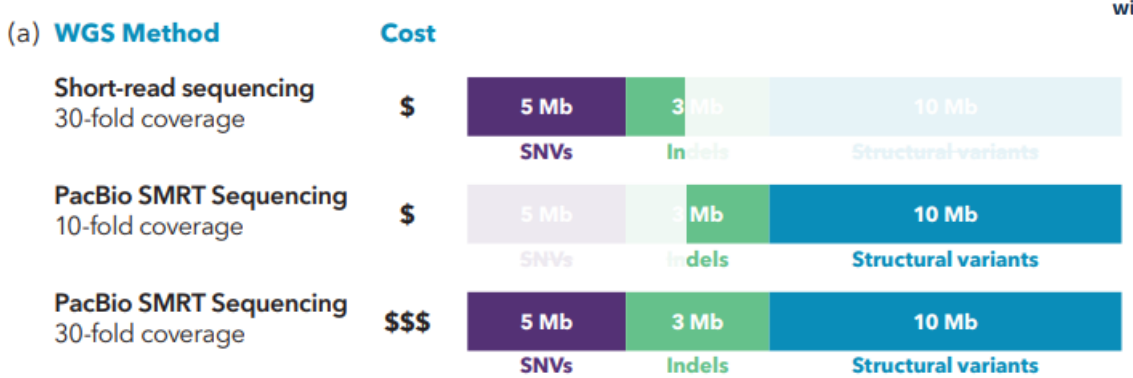
Sensitivity and Reliability of Structural Variant Calling Platforms

PacBio has the highest sensitivity

	Deletions			Insertions		
	Counts	FDR	Sensitivity	Counts	FDR	Sensitivity
PacBio (30-fold)[28]	8,737	3%	95%	12,378	3%	93%
PacBio (10-fold)[28]	6,798	3-10%	83%	11,252	3-10%	83%
ONT (30-fold)[29]	28,791	65%	93%	3,900	65%	11%
10X Genomics (30-fold)	3,166	Not reported	39%	Not reported	N/A	0%
Illumina[30]	1,910	2-4%	24%	1,090	1-4%	9%
BioNano[31,32]	522	3%	6%	769	2%	6%

Oxford Nanopore has the highest false discovery rate

Other technologies struggle with poor sensitivity



Structural Variant Calling - PacBio

Sequence to desired coverage based on study needs:

- 5 to 10-fold: population genetics studies –

  sensitivity limited per individual, but high for variants shared in the population using joint calling

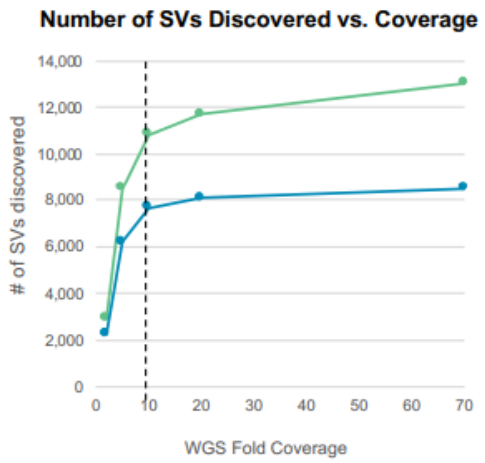
- 10-fold: rare undiagnosed disease studies –

  sensitivity high per individual allowing discovery of pathogenic SVs

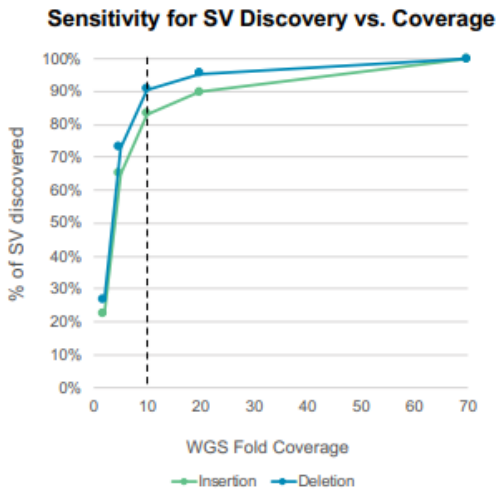
- 10 to >20-fold: genetic disease studies –

  identify a variant or gene that causes disease in a cohort of individuals with a shared phenotype; higher coverage required for de novo SV detection in trios

SV DISCOVERY POWER AT VARIOUS COVERAGE LEVELS



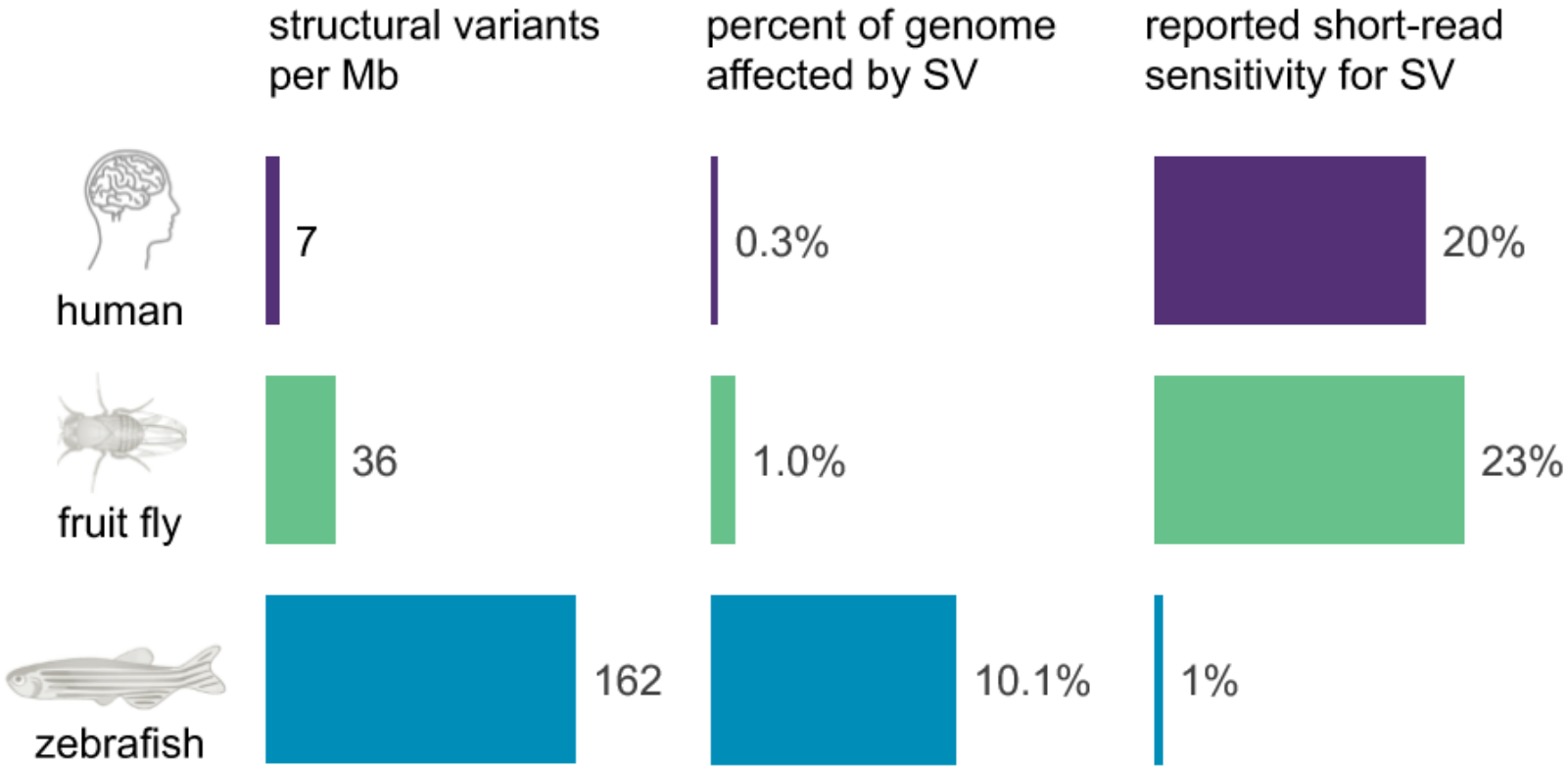
A diploid human (HG00733) was sequenced to 70-fold coverage on the Sequel System. The reads were randomly sampled to various coverage levels, and the SV calls at each coverage were evaluated against the calls at full 70-fold coverage.



Sensitivity increases sharply with coverage until about 10-fold, where it begins to level off. At 10-fold coverage, 10,854 insertions and 7,692 deletions are called (83% and 90.5% sensitivity, respectively)[4].

Structural Variant Calling

THE HUMAN GENOME IS NOT PARTICULARLY SPECIAL

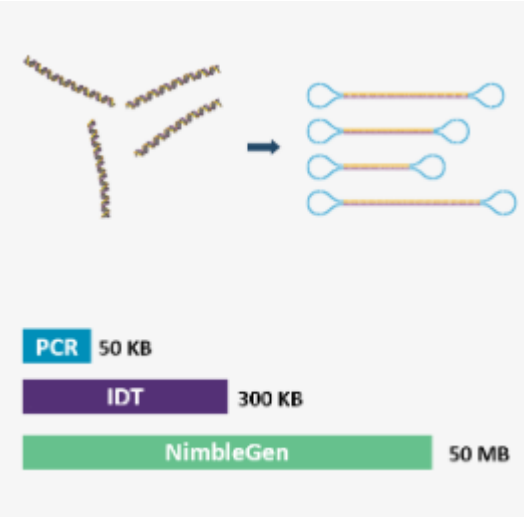


Huddleston et al. (2017) *Genome Research* 27(5):677-85.
Zichner et al. (2013) *Genome Research* 23(3):568-79.
Patowary et al. (2013) *Zebrafish* 10(1):15-20.

Zebrafish image courtesy of Lizzy Griffiths

PacBio applications: TARGETED SEQUENCING

- PCR-mediated targeted sequencing
  - Amplicon sequencing – multiplex up to 384 samples/PCR products on one SMRT cell
    - Insert sizes from 250 bp to 40 kb
    - HLA – sequencing – span majority of HLA class I and II genes
    - Sequencing of viral genomes
    - Identification of individual members of complex metagenomic populations – 16S rRNA, ITS
    - Multi-locus sequence typing (MLST)
  - Targeted enrichment using probe based capture technologies
    - Capture up to 5 kb genomic DNA fragments
    - Multiplex up to 12 samples in a single capture reaction



When targeting >50 kb genomic regions – use probe-based capture using DNA oligo hybridization.
Protocols available for:
- IDT xGen Lockdown probes
- Nimbelgen SeqCap EZ

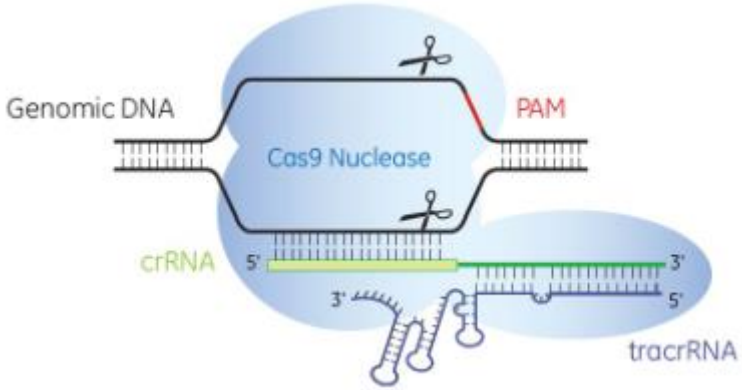
PacBio applications: TARGETED SEQUENCING

- No-amplification targeted sequencing using CRISPR/Cas9 system
  - Why?
    - Challenging regions for PCR amplification (repeat expansions, low complexity regions)
    - No PCR bias
    - Preserves epigenetic modification signals

How does CRISPR/Cas9 system work?

Cas9 nuclease, programmed by the tracrRNA:crRNA complex, cuts both strands of genomic DNA 5' of the PAM motif

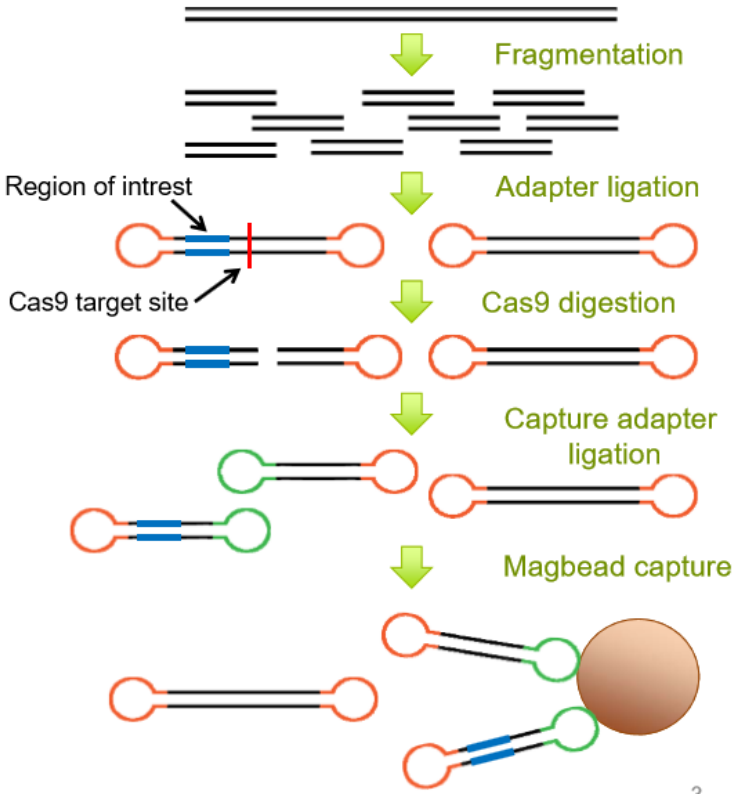
- crRNA – target spesific CRISPR RNA
- tracrRNA – trans-activating crRNA
- PAM – Protospacer-adjacent motif – needed for Cas9 nuclease to bind and cleave the target DNA sequence



https://dharmacon.horizondiscovery.com/applications/gene-editing/

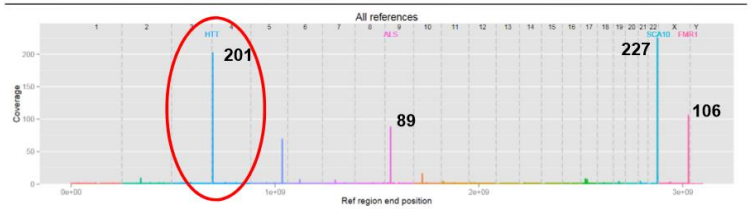
No-Amp targeted sequencing

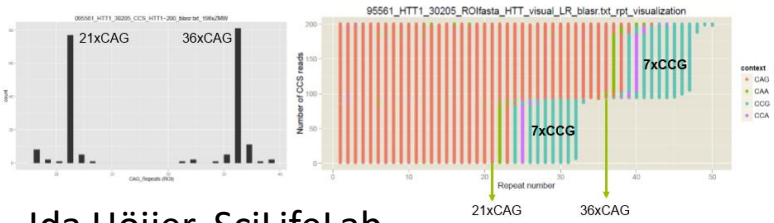
- RE fragmentation of gDNA
- Adapter ligation to form standard SMRTbell
- Digestion of SMRTbell using cas9
- Ligation of capture adapter to enable enrichment of SMRTbells containing region of intrerest
- Capture using magbeads
- Sequencing of enriched SMRTbells



SciLifeLab

Results – Huntington's Patient 1





Ida Höijer, SciLifeLab

PacBio applications: RNA SEQUENCING

Iso-Seq method generates high-quality, full-length transcripts – no assembly required.

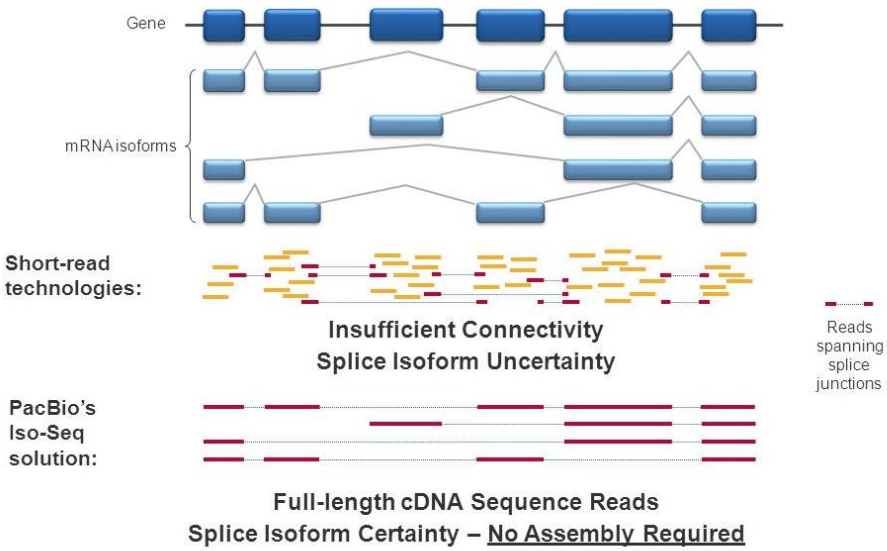
"The way we do RNA-seq now is… you take the transcriptome, you **blow it up into pieces** and then you try to figure out **how they all go back together again**… If you think about it, it's kind of a **crazy way to do things**"

Michael Synder
Professor and Chair of Genetics
Stanford University

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

Consider Iso-Seq if you need to transcriptome data for:

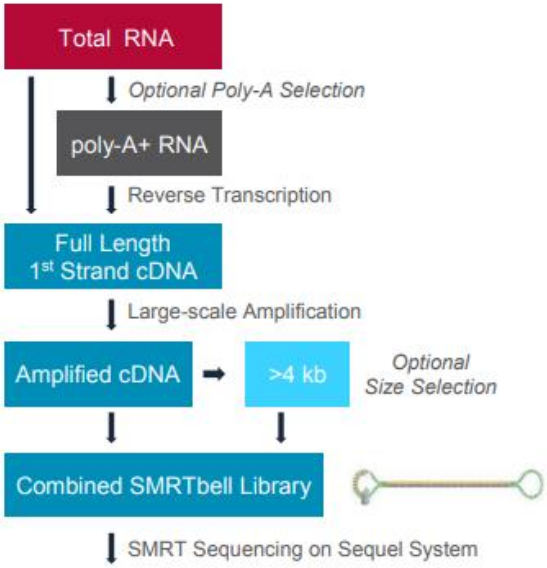
- Whole genome annotation
- Isoform discovery
- Fusion gene detection
- Creating *de novo* reference transcripts for RNA-seq quantification



PACIFIC BIOSCIENCES

PacBio applications: RNA SEQUENCING

FROM RNA TO ACCURATE GENE MODELS



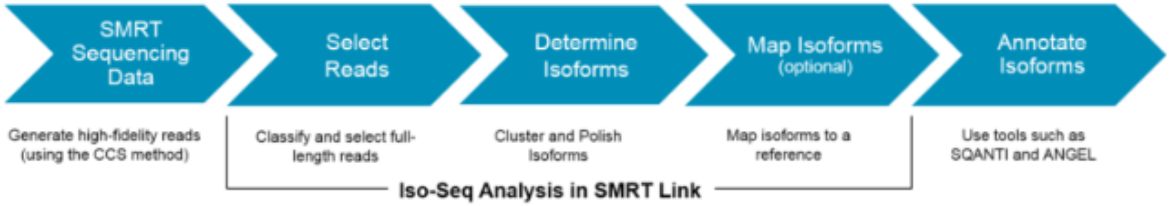
- Compatible with standard target enrichment methods
- Multiplex with sample barcoding
- Scalable throughput:
  - 400k-500k reads/up to 40 Gb per 1M SMRT Cell
  - Targeted genes: < 1 SMRT cell needed – multiplexing recommended
  - Whole transcriptome: 2-4 SMRT cells

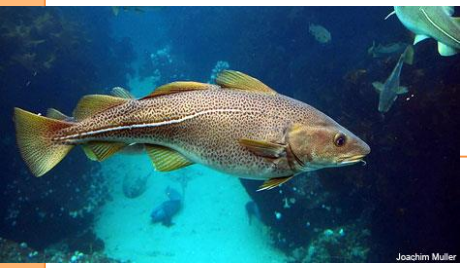
Analyze with SMRT Analysis Software Suite

SMRT Analysis:

Iso-Seq analysis uses high-fidelity long reads to cluster and generate full-length, high-quality transcript isoforms. Isoforms can be then mapped to a reference. Iso-Seq analysis is an easy-to-use application accessible via the SMRT Link graphical user interface.

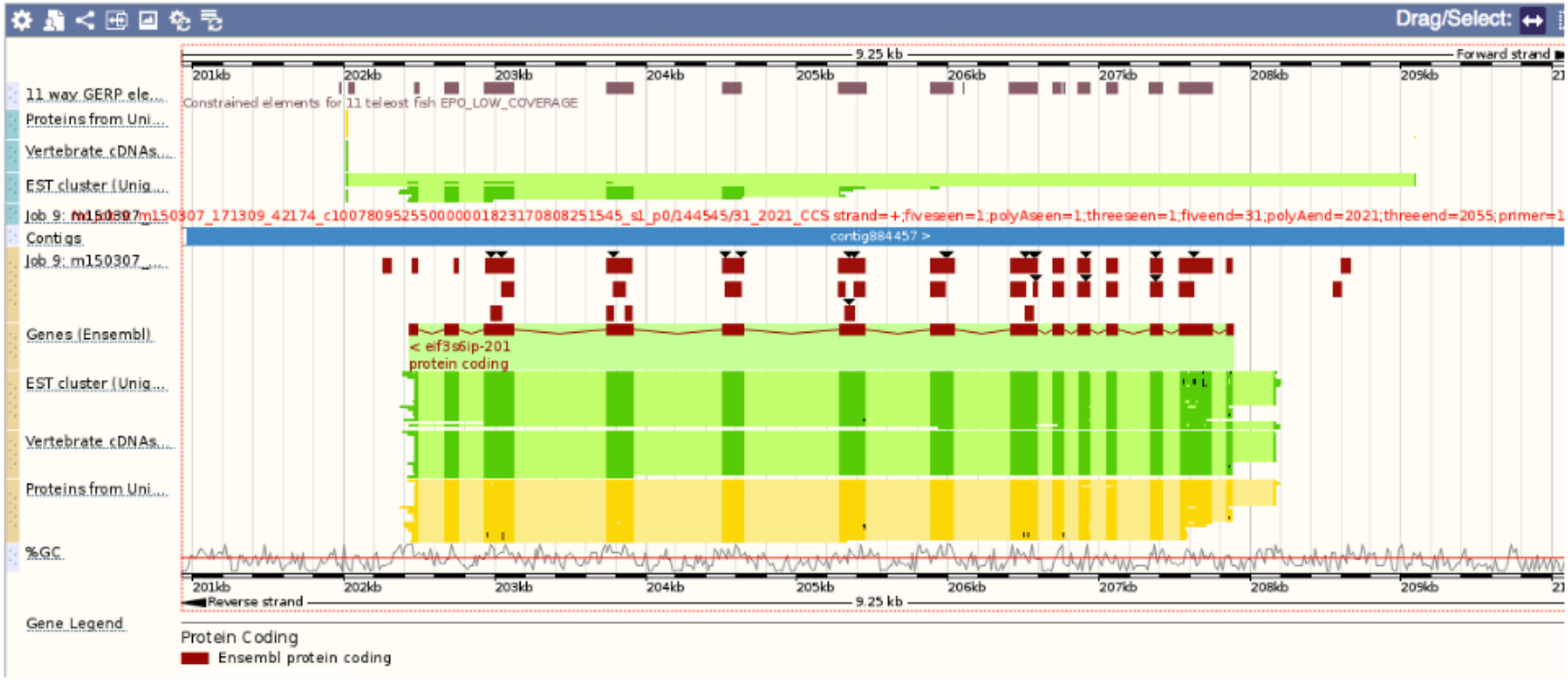
How it works: Full-length, high-fidelity reads (Q20, single-molecule resolution) are identified and clustered at the isoform level, then polished to create high-quality consensus.





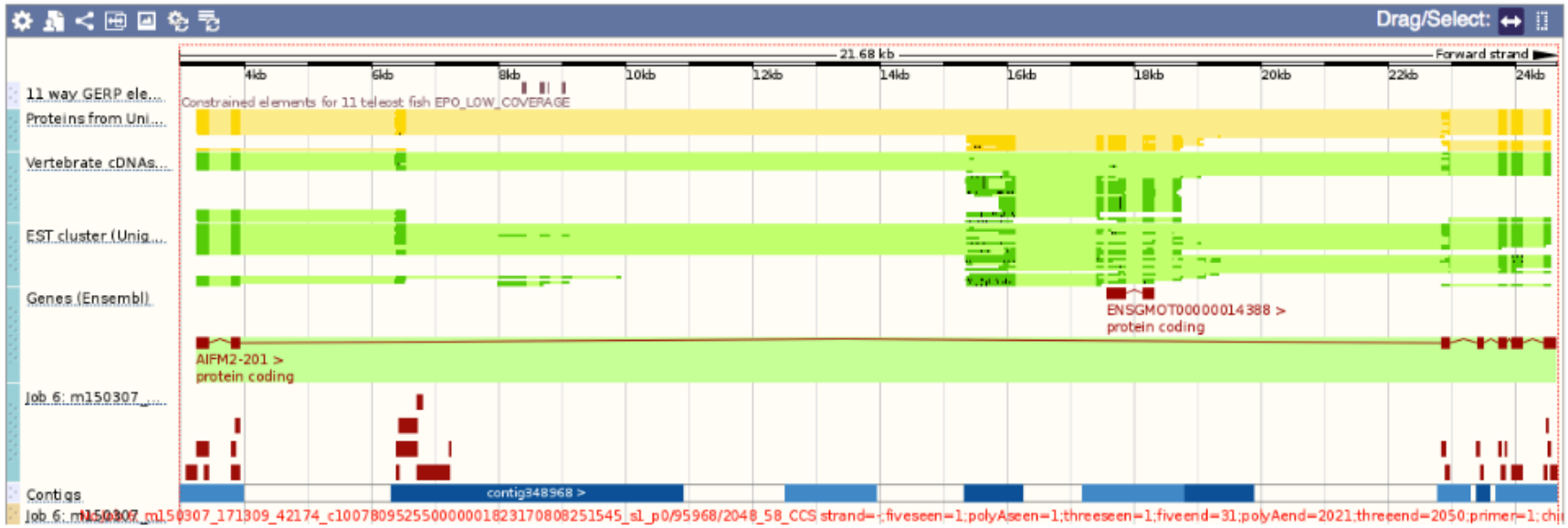
Iso-Seq example: Mapping isoforms to Ensamble annotation of *Gadus* genome

- IsoSeq reads are nicely fitting the gene model



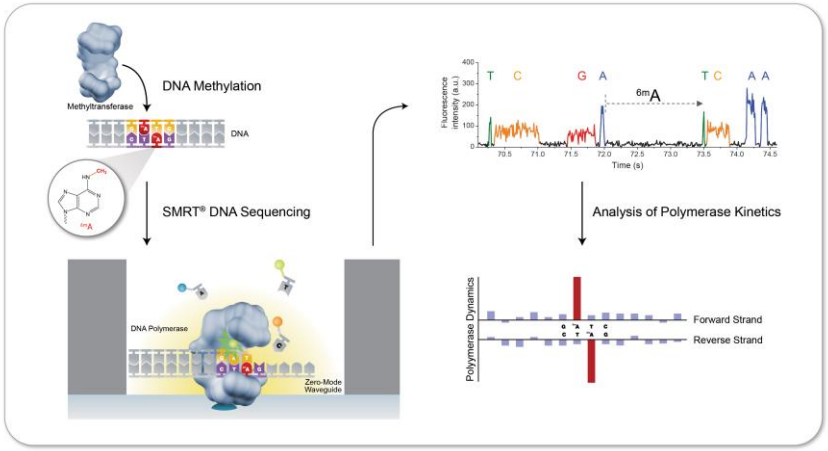
IsoSeq example: Mapping isoforms to Ensamble annotation of *Gadus* genome

- IsoSeq reads add a missing exon:

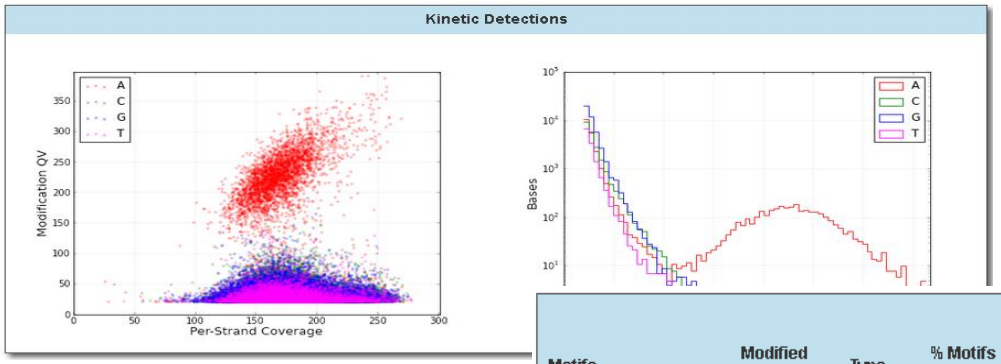


PacBio applications: EPIGENETICS

How it works?



- Microbial epigenetics –
  - detect genome wide m6A and m4C R-M system motifs
  - Determine m6A and m4C methylation status at all genomic positions



Motif Summary

Motifs	Modified Position	Type	% Motifs Detected	# Of Motifs Detected	# Of Motifs In Genome	Mean Modification QV	Mean Motif Coverage	Partner Motif
RAYCNNNNNNTTRG	2	m6A	99.81%	1047	1049	236.64	167.23	CYAANNNNNNGRTY
CYAANNNNNNGRTY	4	m6A	99.24%	1041	1049	224.77	168.52	RAYCNNNNNNTTRG
RTCANNNNNNNTRRG	4	m6A	99.3%	565	569	237.81	173.89	CYYANNNNNNNTGAY
CYYANNNNNNNTGAY	4	m6A	99.12%	564	569	239.43	173.93	RTCANNNNNNNTRRG

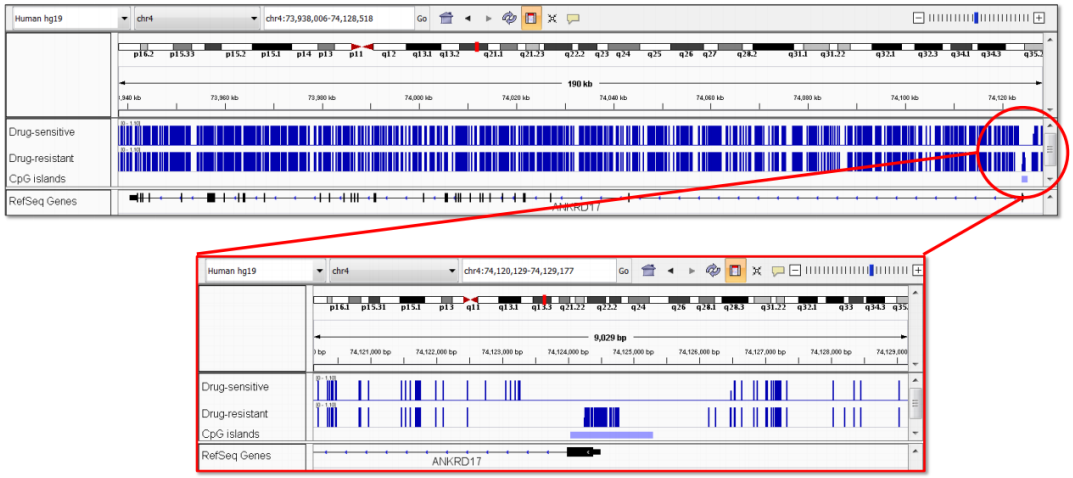
PacBio applications: EPIGENETICS

- Eukaryotic Epigenetics
  - Identify hyper- and hypo-methylated CpG islands to explore gene expression and regulation
    - 20 kb library
    - > 20 x coverage needed
    - "AgIn: measuring the landscape of CpG methylation of individual repetitive elements" Y. Suzuki et al, 2016
    - Software designed and tested with RSII, not Sequel

Cancer genomes of both drug-sensitive and drug-resistant PC9 cells show differential methylation status:

https://www.pacb.com/wp-content/uploads/Whole-Genome-Sequencing-and-Epigenome-Characterization-of-Cancer-Cells-using-the-PacBio-Platform.pdf

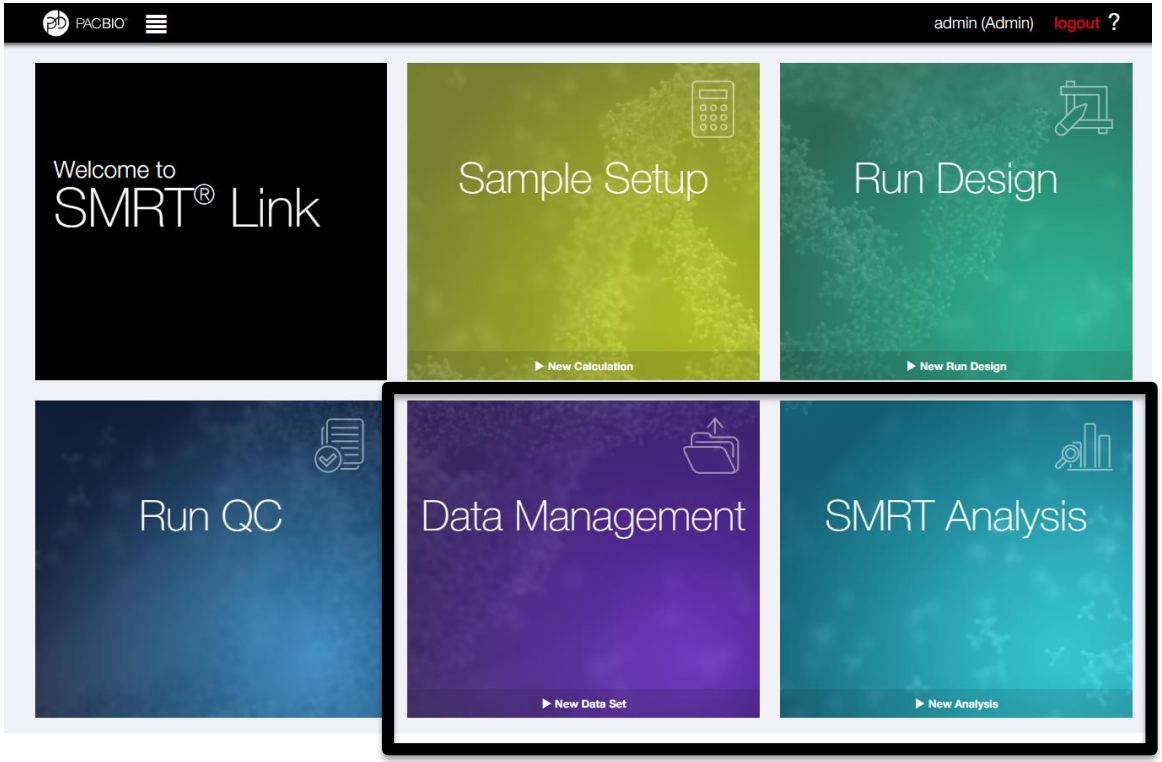
Chr. 4: ANKRD17 (already implicated in breast cancer)



- Targeted applications – SMRT Bisulfite sequencing
  - SMRTbell libraries are generated from long amplicons (1-2 kb) following bisulfite treatment of gDNA.

SMRT Analysis software

- Open source software
  - https://www.pacb.com/support/software-downloads/
- data can be analyzed, visualized, and managed through an intuitive GUI or command-line interface.



SMRT Analysis – different applications

PACBIO SMRT Analysis admin (Admin) logout ?

Create New Analysis - Settings

CANCEL START

Analysis Application

- Assembly (HGAP 4)
- Base Modification Detection
- Base Modification and Motif Analysis
- CCS Mapping
- Circular Consensus Sequences (CCS)
- Convert BAM to FASTX
- Convert RS to BAM
- Demultiplex Barcodes
- Iso-Seq 1
- Iso-Seq 1 Classify Only
- Iso-Seq 1 with Mapping
- Iso-Seq 3
- Iso-Seq 3 with Mapping
- Long Amplicon Analysis (LAA)
- Minor Variants Analysis [Beta]
- Resequencing
- Site Acceptance Test (SAT)
- Structural Variant Calling

Analysis Name

Data Sets

Advanced Search Search...

	Data Set Name	Demultiplexed Subsets	Well Sample Name	Bio Sample
	PB_0649_4pM-Cell3		PB_0649_4pM	PB_0649_4
	lambda/0007_tiny		Inst42267-040315-SAT-10...	unknown
	PB_0644_4pM-Cell2		PB_0644_4pM	PB_0644_4
	PB_0635_5pM-Cell1		PB_0635_5pM	PB_0635_5
	PB_0620_10pM-Cell3		PB_0620_10pM	PB_0620_1
	PB_0633_5pM-Cell1		PB_0633_5pM	PB_0633_5
	PB_0624pool_4pM-Cell2		PB_0624pool_4pM	PB_0624po
	PB_0603pool_4pM-Cell1		PB_0603pool_4pM	PB_0603po
	PB_0138_4pM-Cell3		PB_0138_4pM	PB_0138_4

EDIT OUTPUT FILE NAME PREFIX Example: analysis-<Bio Sample Name>-<Job Id>

DevNet analysis tools

Advanced bioinformatics methods and novel applications for PacBio data have been developed through continuous collaboration between PacBio and the bioinformatics community.

More information available at:

https://www.pacb.com/products-and-services/analytical-software/devnet/devnet-analysis-tools/

Bioinf at NSC/CEES

- Included to sequencing price: demultiplexing

- Analyses performed for small fee:
  - HGAP assembly of small genomes – 1000 kr/sample
  - Base modification and motifs analysis – 500 kr/sample
  - Iso-Seq – 1000 kr per sample

- For large genomes – bioinf service including assembly (PacBio+Illumina) and annotation can be ordered for an extra fee (please contact post@sequencing.uio.no for more information)

- Data delivered:
  - raw data
  - .fastq (or fasta+fastq)
    - (demultiplexed) subreads
    - consensus reads (CCS/LAA)
    - polished assembly (HGAP)
    - isoforms (Iso-Seq)
  - Base modification:
    - Motifs Summary.csv
    - Motifs and Modifications.gff