MAYO CLINIC

LONG READ SEQUENCING

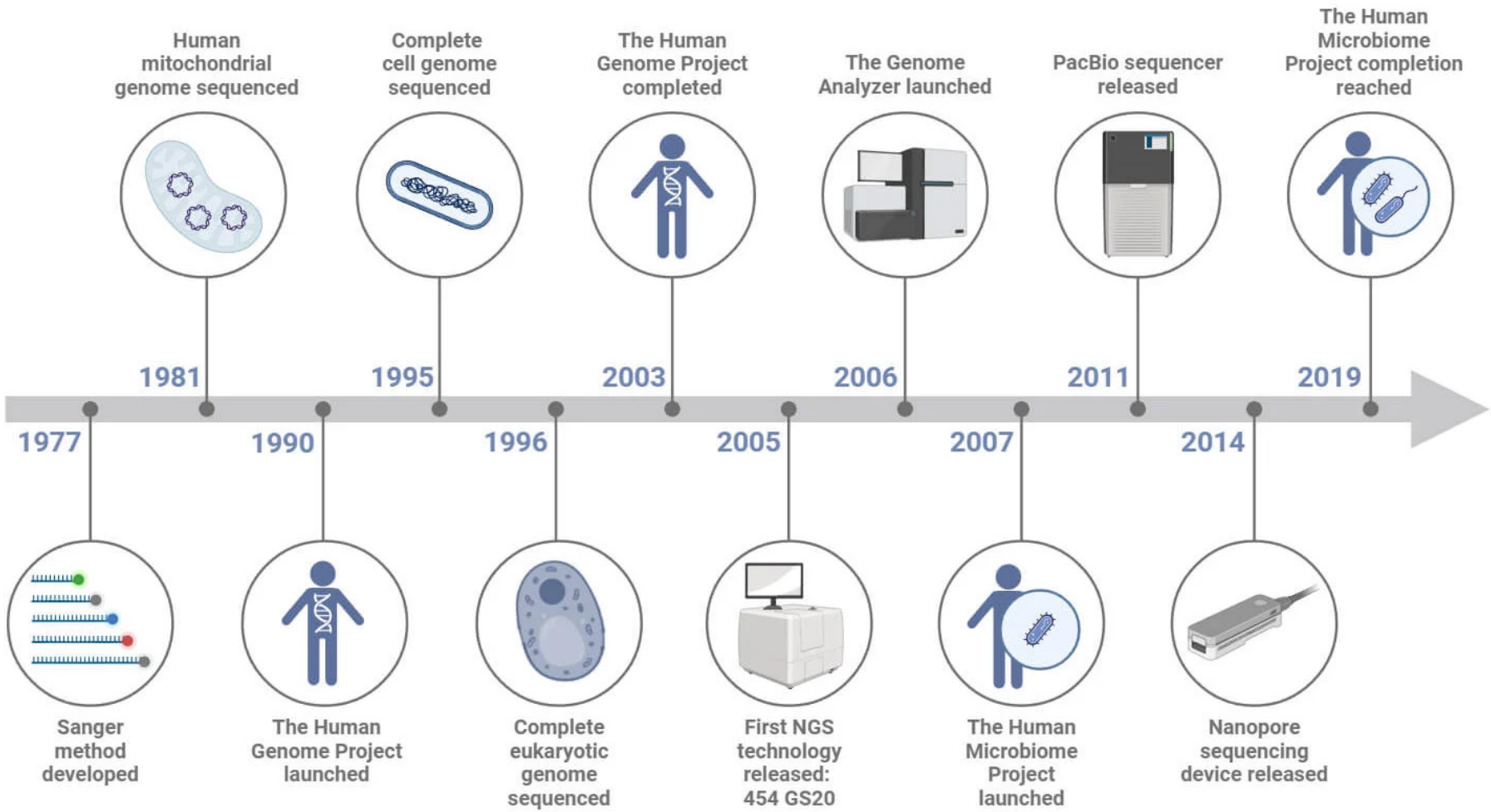
Jesse R. Walsh, Ph.D.
Jag Balan

UIUC – Comp Gen Course
June 25th, 2025

OUTLINE

- Why long reads? (“3rd” generation sequencing)
- Long read platforms
- Bioinformatics tools
- Examples
- Lab/Workshop

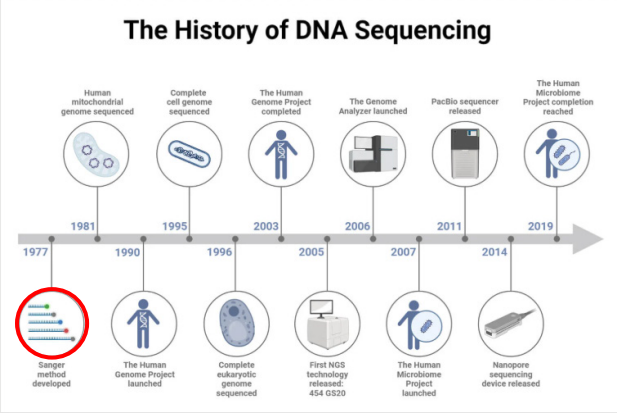
The History of DNA Sequencing



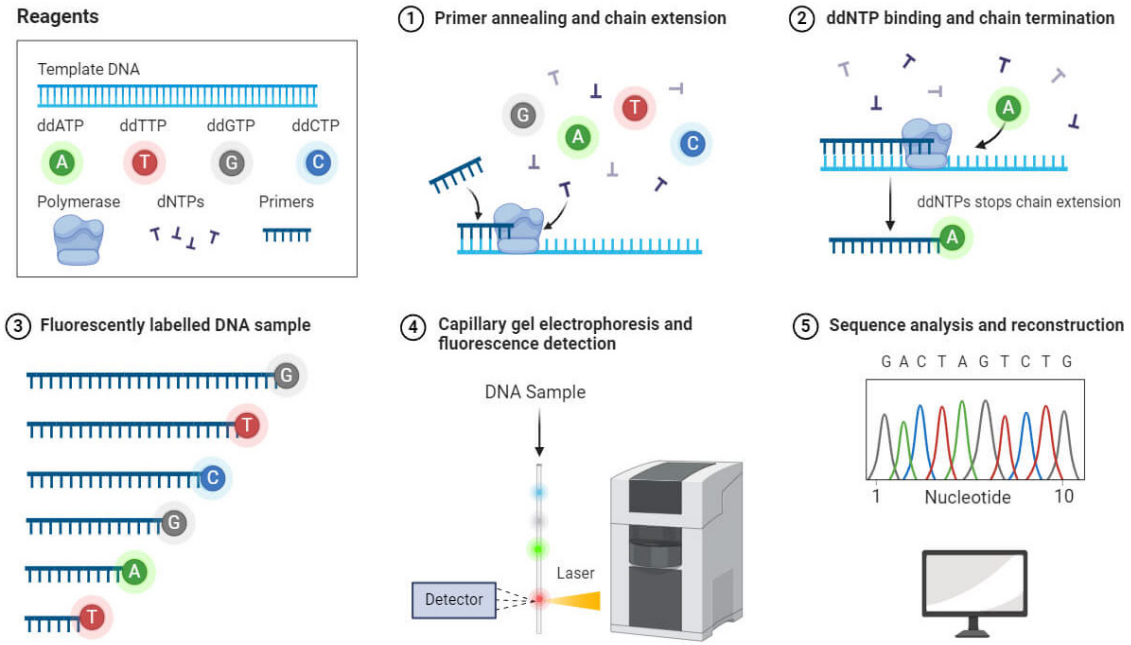
Sanger Sequencing: Principle, Steps, Applications, Diagram

HOW WE GOT TO 3RD GEN SEQUENCING

- 1st Generation sequencing
- Characterized by
  - Chain termination method
  - High accuracy
  - Low throughput
- Example Sequencing Methods
  - Sanger

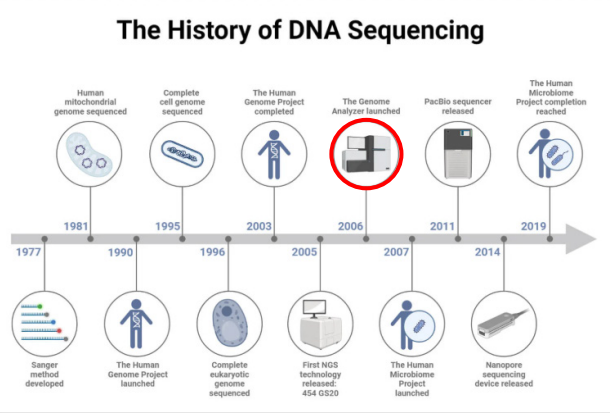


Sanger Sequencing

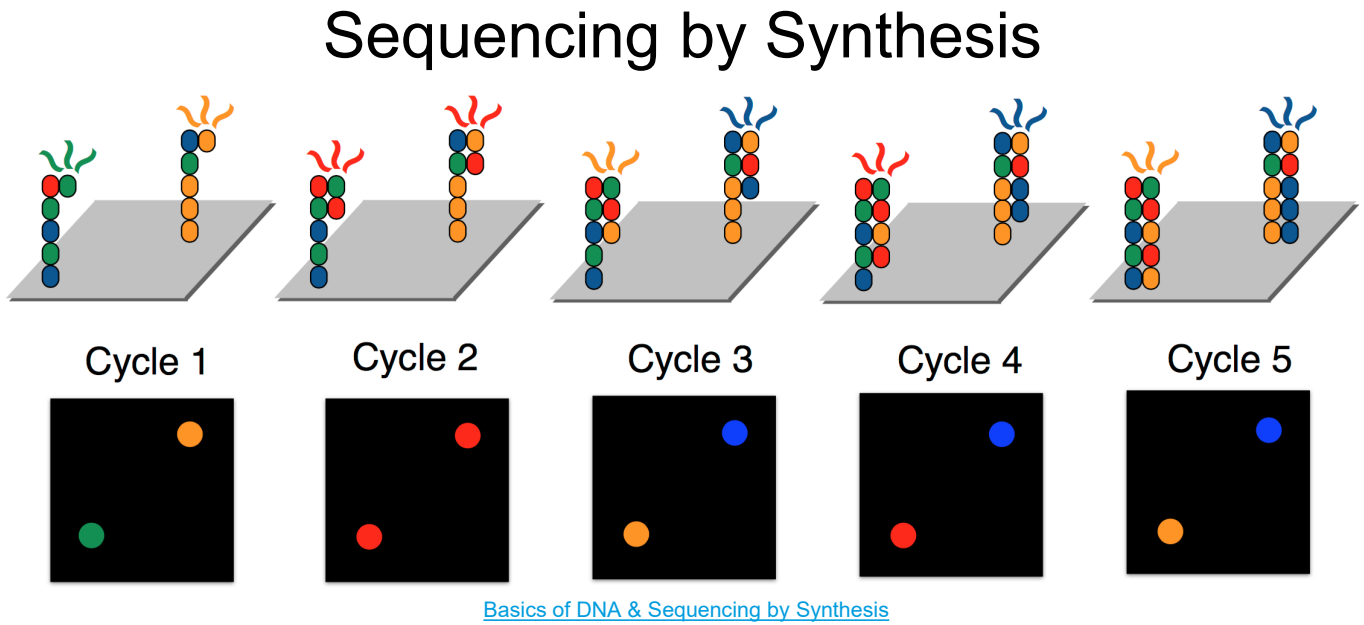


Sanger Sequencing: Principle, Steps, Applications, Diagram

HOW WE GOT TO 3RD GEN SEQUENCING



- 2nd Generation sequencing (NGS)
- Characterized by
  - High throughput
  - Short reads
- Example Sequencing Methods
  - Roche/454
  - Illumina
  - SOLiD

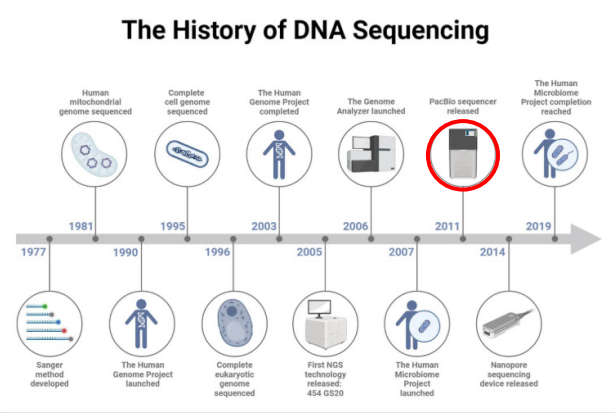


Basics of DNA & Sequencing by Synthesis

WHY LONG READS?

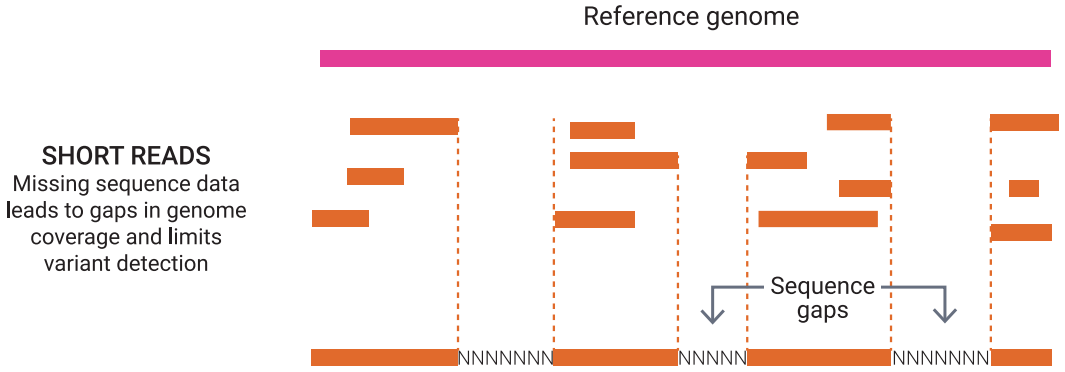
- Eventually, limitations of SRS led to the need to develop LRS
- Short read methods
  - Are unable to span difficult to sequence regions of the genome
  - Haplotyping is a major challenge
  - Additional capabilities, such as methylation, require additional assays

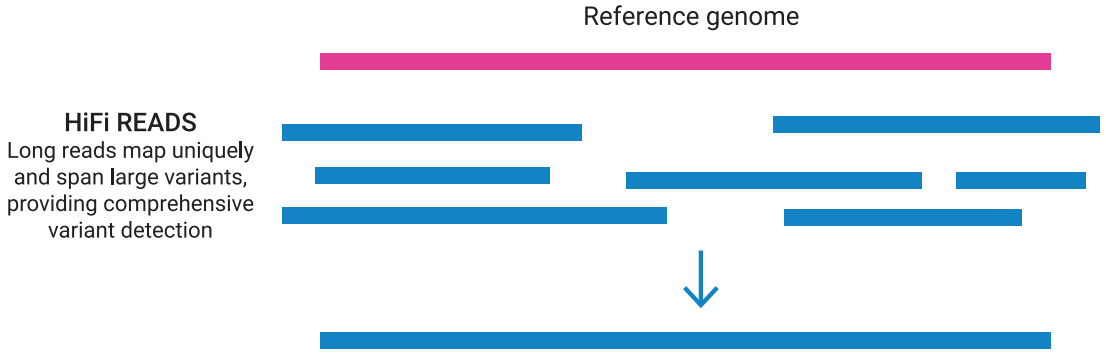
HOW WE GOT TO 3RD GEN SEQUENCING



- 3rd Generation sequencing
- Characterized by
  - Longer reads
  - Single molecule
  - Real time
- Example Sequencing Methods
  - PacBio SMRT
  - ONT Nanopore
- Use Cases for Long Reads

Traditional Panels	Methylation	Dark Genome	Telomeres
BRCA	Fragile-X	SMN1/2	Telomere Length





Sequencing 101: Comparing long-read sequencing technologies - PacBio

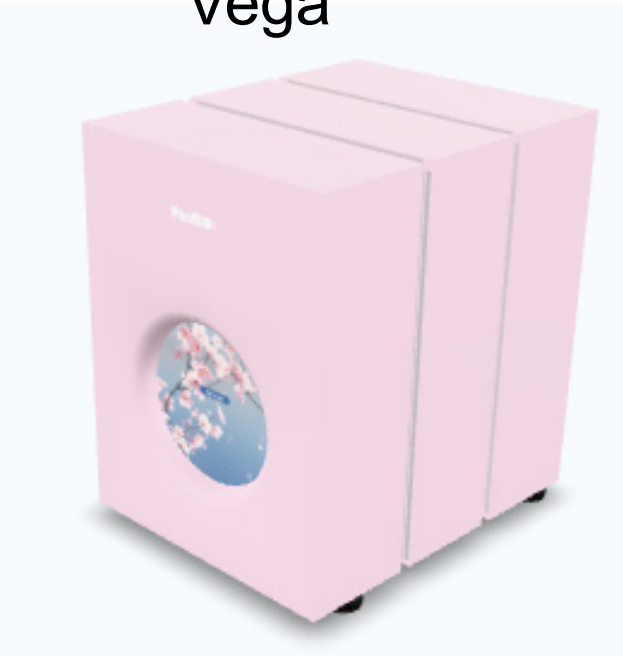
Further Reading The sequence of sequencers: The history of sequencing DNA by James M. Heather, Benjamin Chain

PACBIO LONG READ PLATFORMS

- PacBio
  - Vega
    - 1 SMRT cell
  - Revio
    - Up to 4 SMRT cells

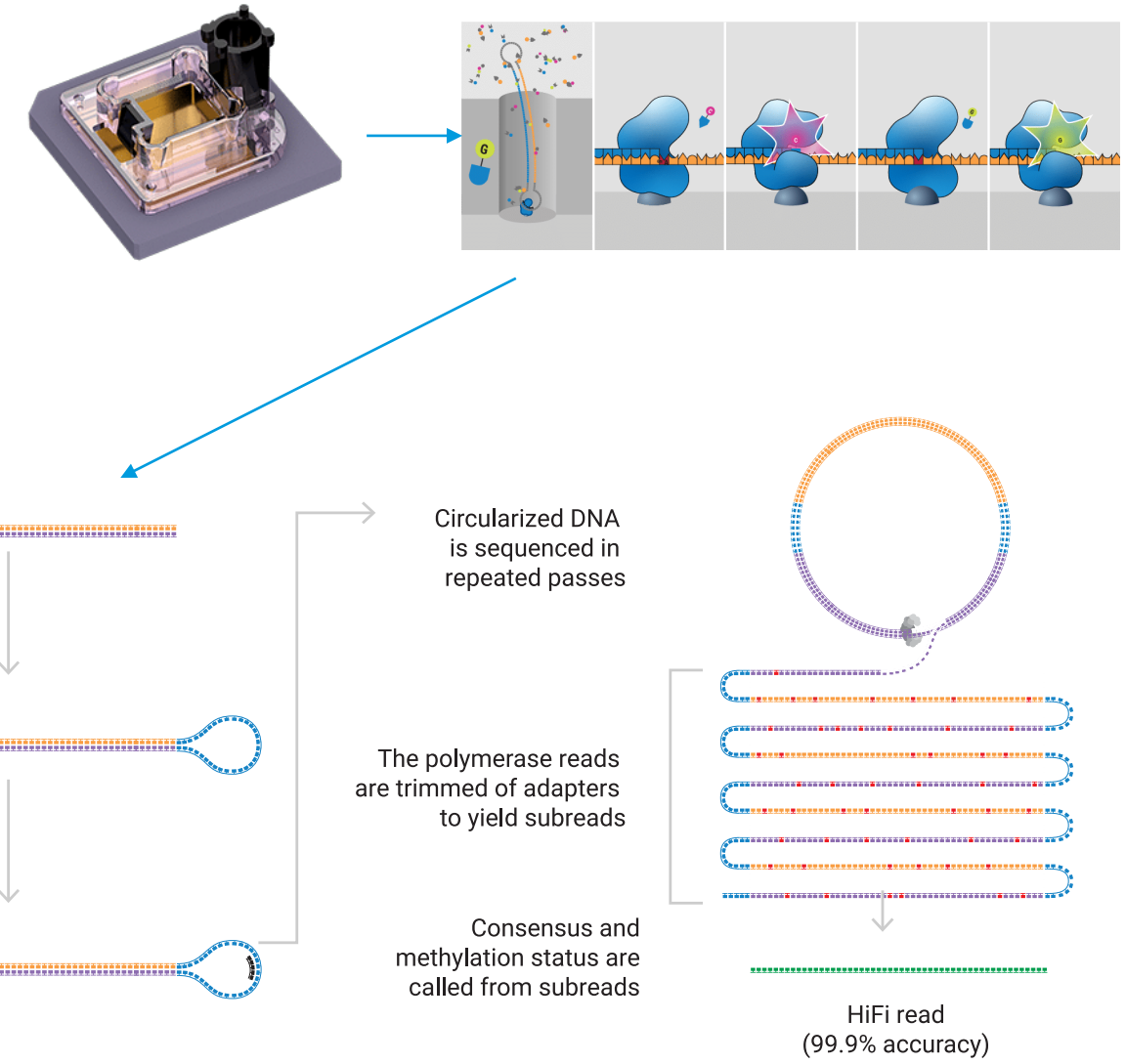
Revio

Vega



HIFI SEQUENCING

- PacBio uses Hi-FI sequencing method
- Each Single Molecule, Real-Time (SMRT) cell contains many zero mode waveguides (ZMWs)



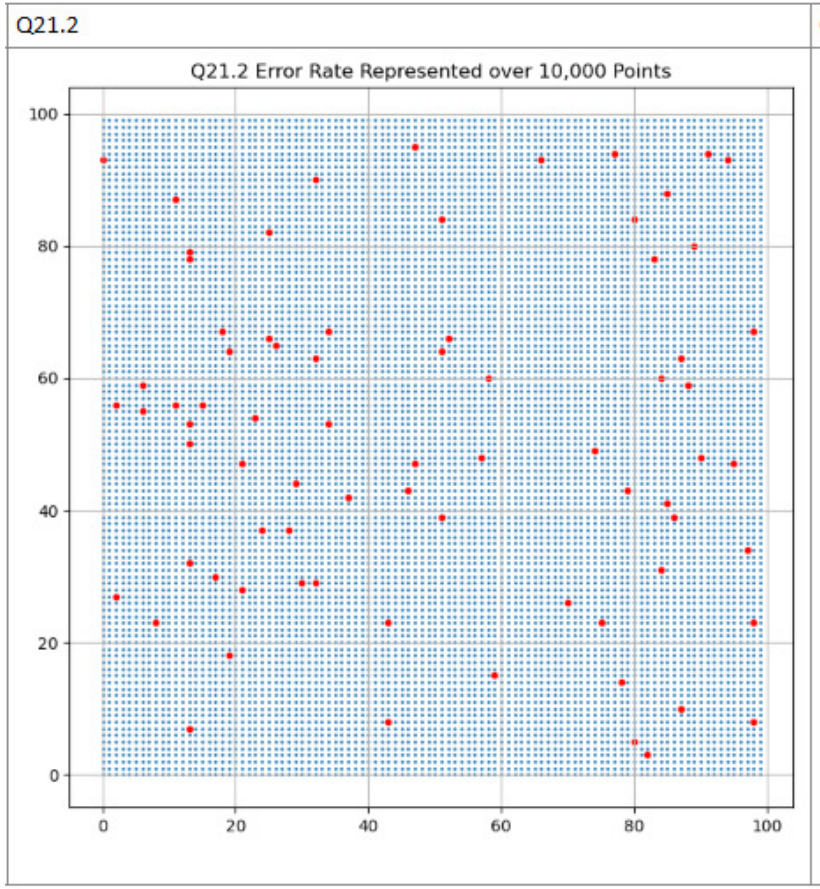
Sequencing 101: Comparing long-read sequencing technologies - PacBio

PACBIO SMRT SEQUENCING

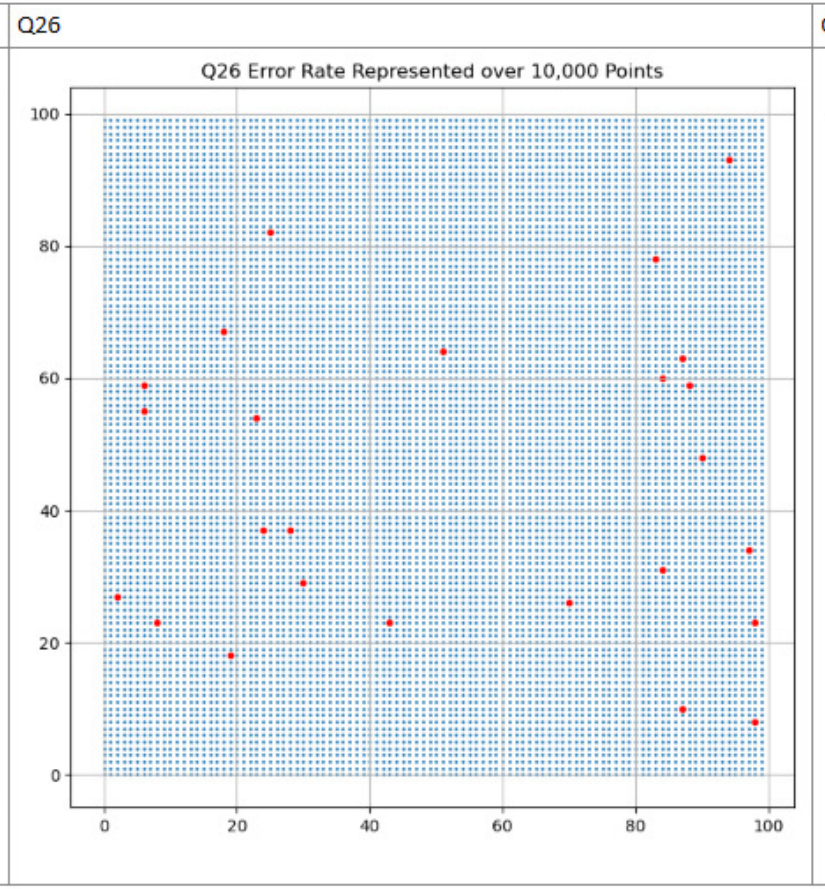
- WGS on human genome
  - 15-20kb library
  - 30-40x per SMRT Cell (1 sample per)
- Error rate in sequencing is advertised as 99.95% (Q33)
- Methylation output available

HOW ACCURATE IS Q33?

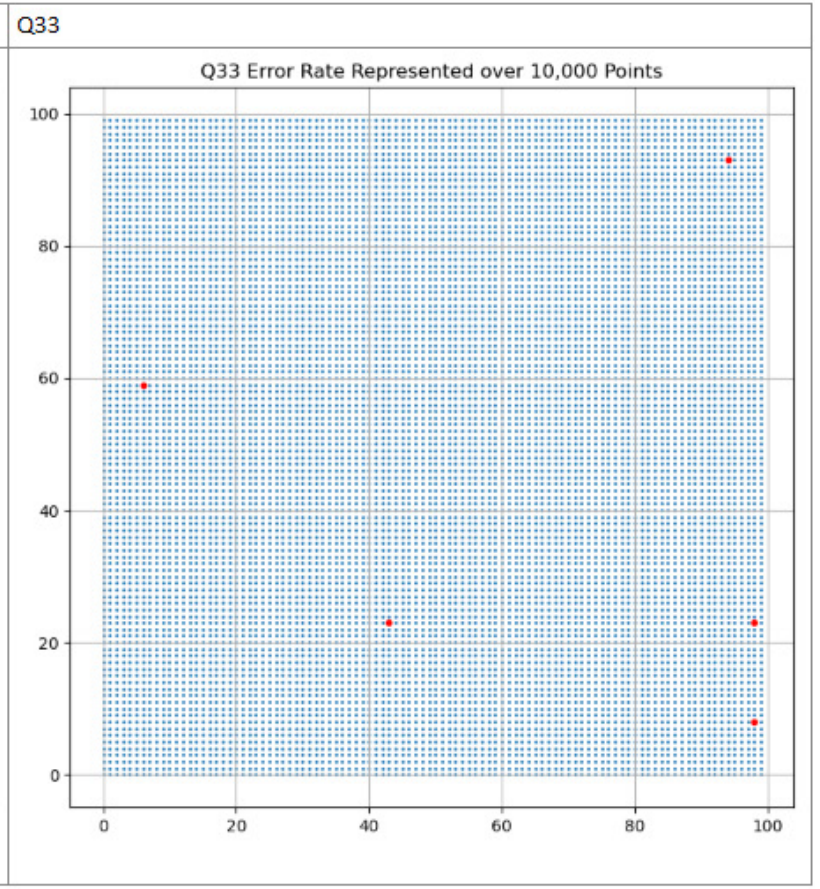
Q21.2 (99.25%) ONT “High-Accuracy”



Q26 (99.75%) ONT “Super High-Accuracy”



Q33 (99.95%) PacBio HiFi



ONT LONG READ PLATFORMS

- Oxford Nanopore Technology (ONT)
  - MinION
    - 1 Flow cell
  - P2 Solo
    - 2 Flow cells
  - GridION
    - 5 Flow cells
  - PromethION 24
    - 24 Flow cells
  - PromethION 48
    - 48 Flow cells

P2 Solo

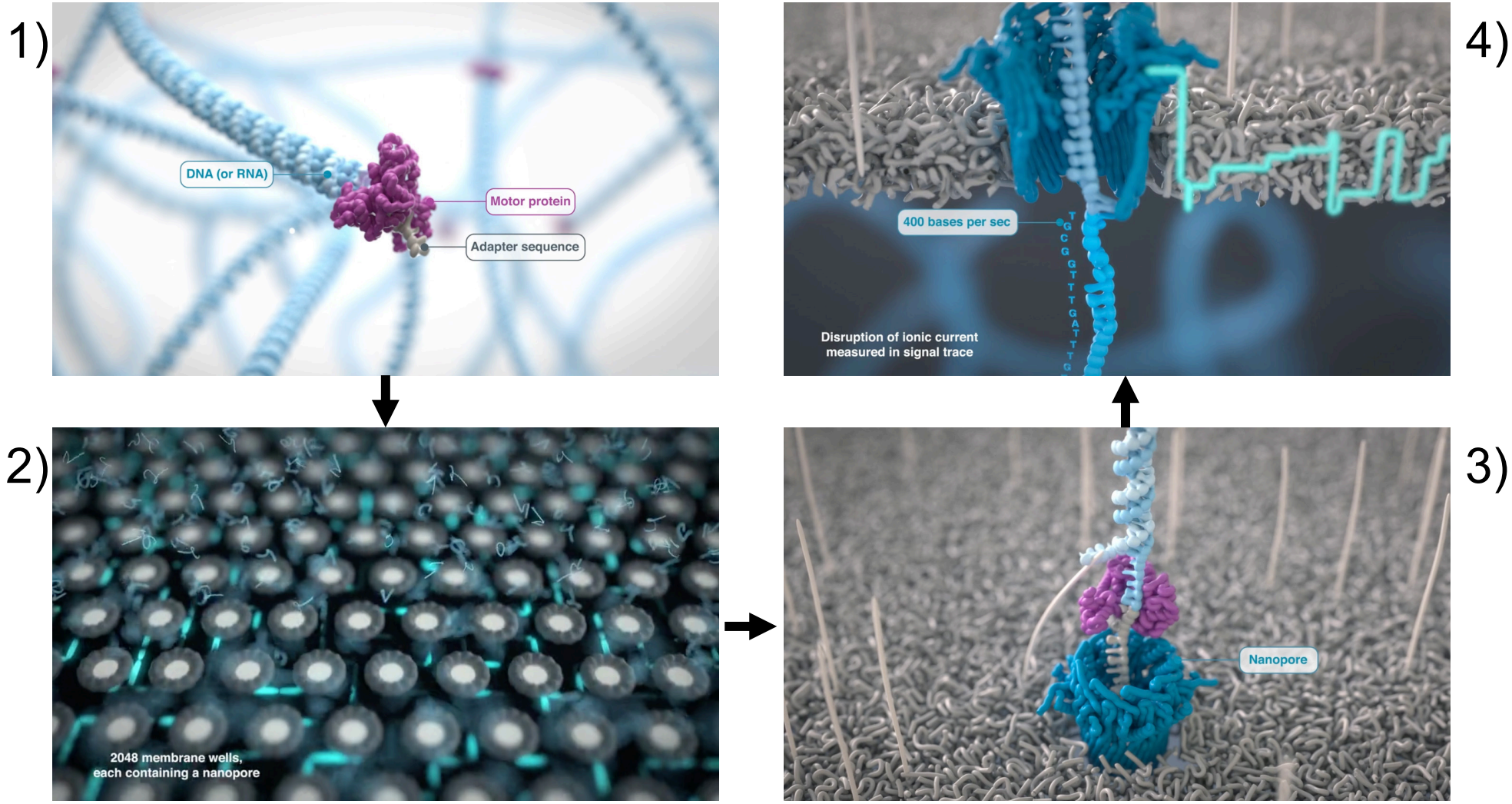
GridION

PromethION 24

PromethION 48

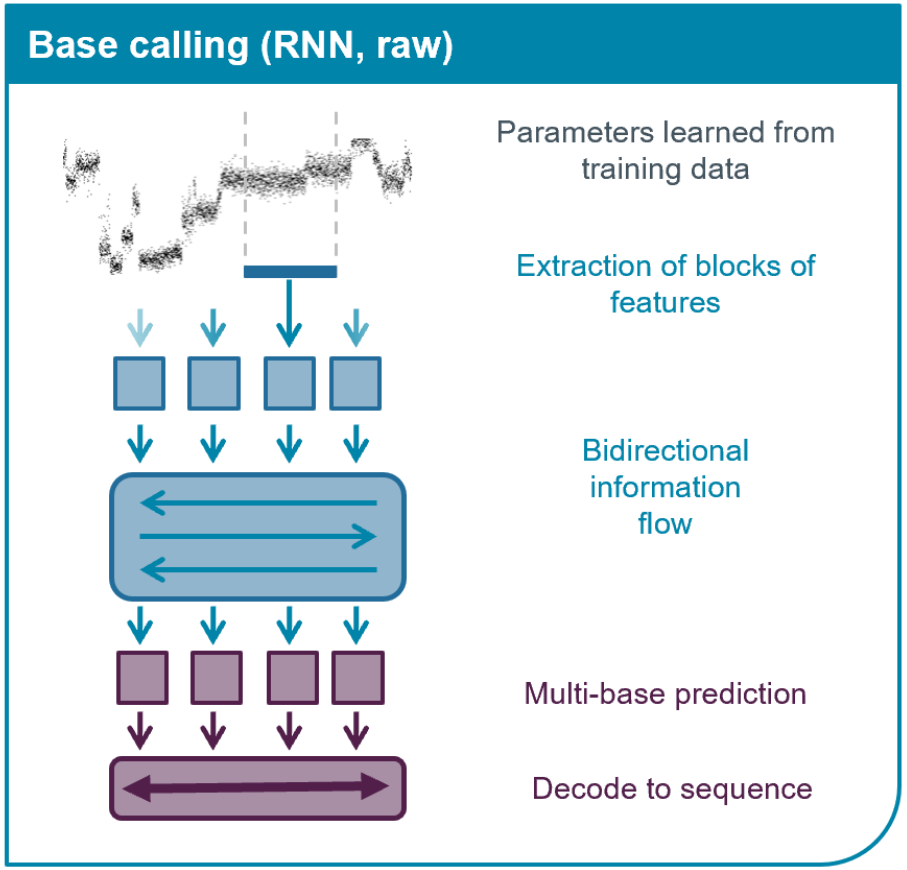
MinION

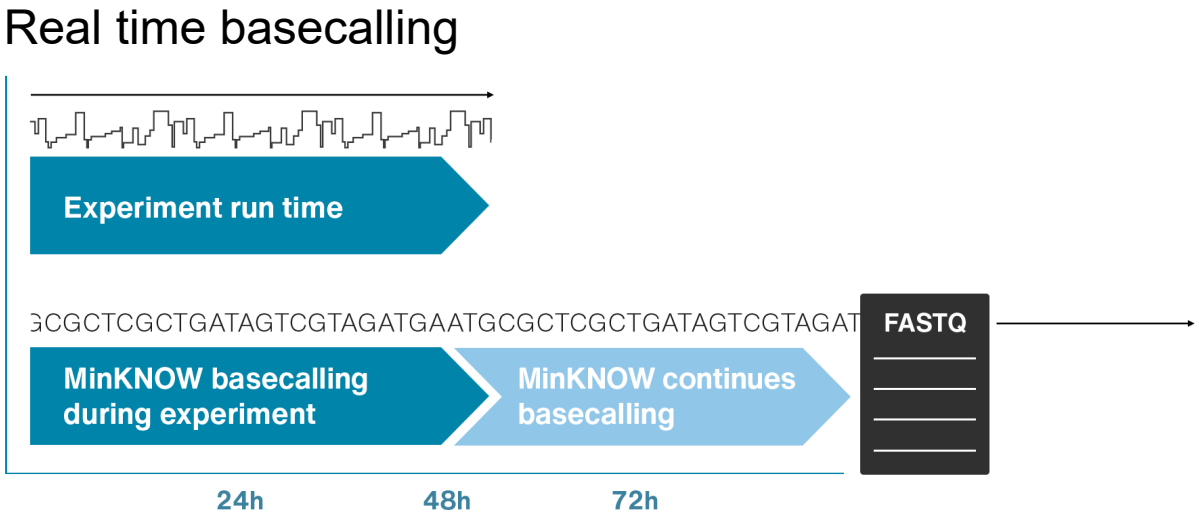
HOW NANOPORE SEQUENCING WORKS



https://nanoporetech.com/

BASE PREDICTION FROM SIGNAL BY PRE-TRAINED NEURAL NETWORK MODELS





ONT NANOPORE SEQUENCING

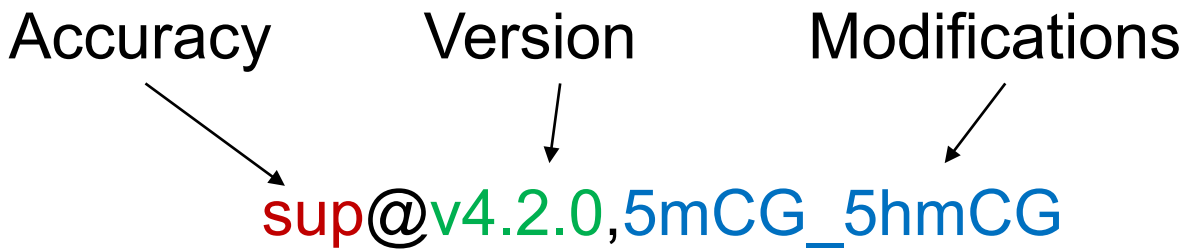
- No set limit to read length
  - Often filter/ignore reads < 200bp
  - Often get reads > 10,000
- Read length is dependent on prep and sample quality
- Some tradeoff between length and coverage
  - i.e. efforts to “clean up” shorter fragments can reduce overall DNA
- Have seen N50’s from ~500bp to 10,000bp
  - Shorter fragments get through nanopore faster, so they can bring down N50 average
- There have been some experiments targeting 15kb and 30kb lengths

MANUAL BASE CALLING

- Dorado
  - ONT provided basecaller
  - GPU accelerated
  - Uses pre-trained neural networks
  - Many models (and versions) available
- Model choice matters
  - Accuracy
    - Fast (fast)
    - High (hac)
    - Super-high (sup)
  - Version
  - Modified bases (Methylation)

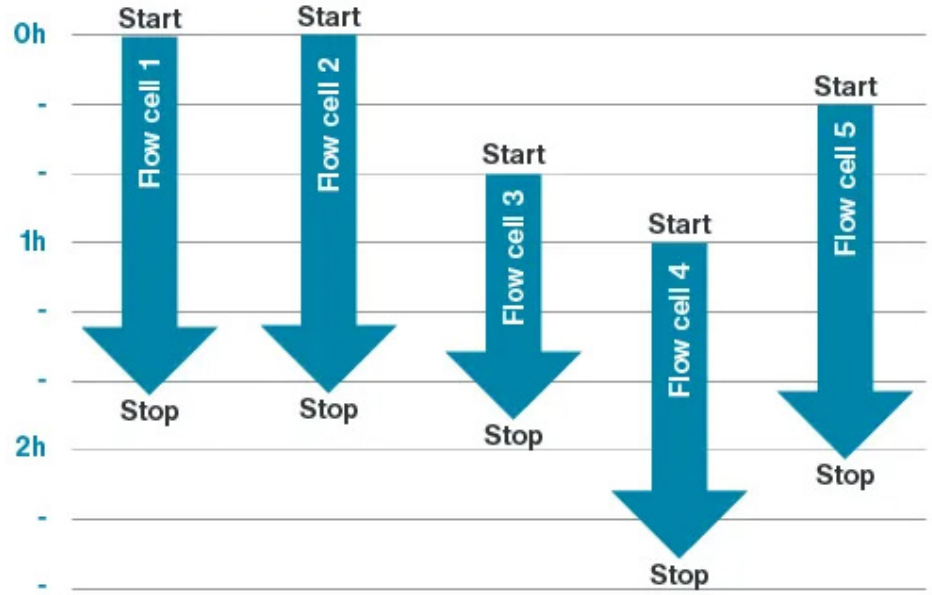
Basecalling Models	Compatible Modifications	Version
dna_r10.4.1_e8.2_400bps_fast@v5.0.0		
dna_r10.4.1_e8.2_400bps_hac@v5.0.0	4mC_5mC 5mCG_5hmCG 5mC_5hmC 6mA	v3 v3 v3 v3
dna_r10.4.1_e8.2_400bps_sup@v5.0.0	4mC_5mC 5mCG_5hmCG 5mC_5hmC 6mA	v3 v3 v3 v3

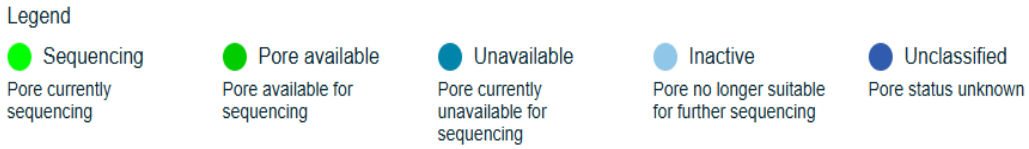
https://github.com/nanoporetech/dorado

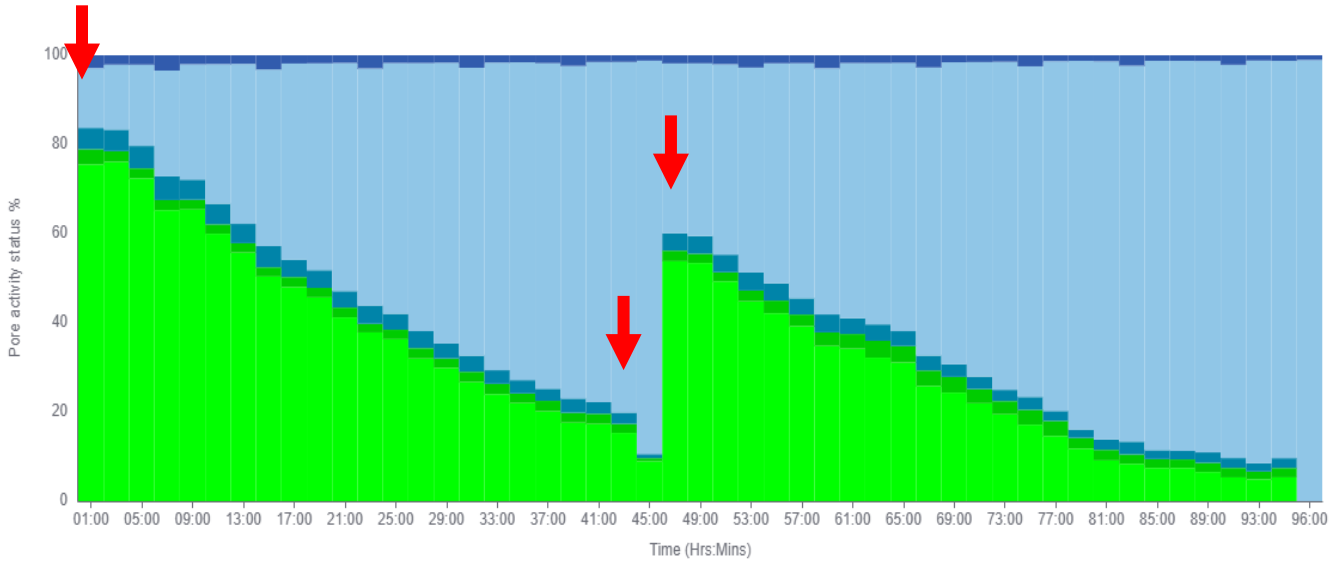


FLEXIBLE LOADING

- Can start and stop individual FC's
- Add an additional sample while machine is already running
- Restart a failed sample without impacting others
- Improve a sample by reloading mid-run
  - e.g. Wash and reload at 48hr

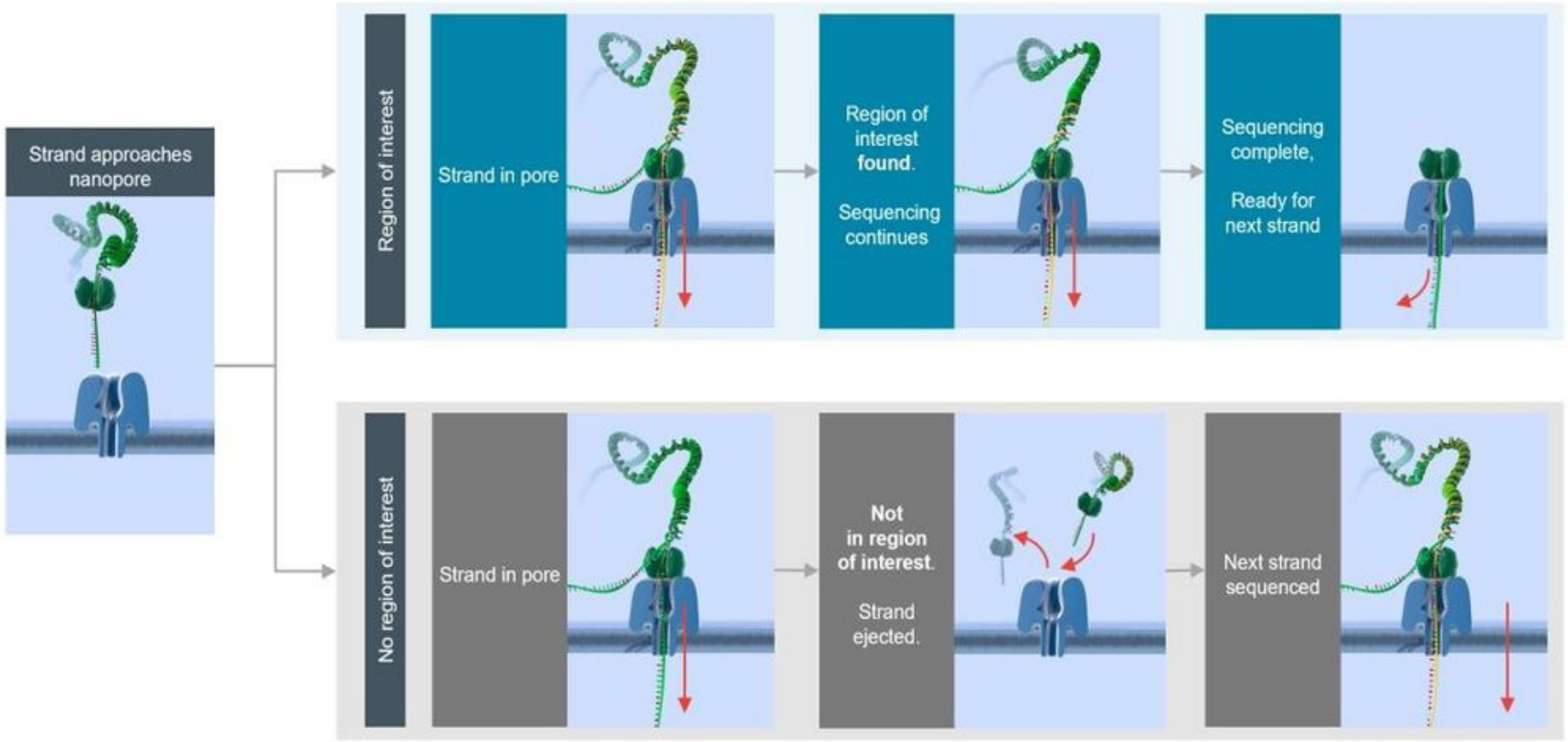






ADAPTIVE SAMPLING

- Whole Genome Sequencing mode (~20-30x depth)
- Adaptive sampling mode can get 3-5x boost



ADAPTIVE SAMPLING CONSIDERATIONS

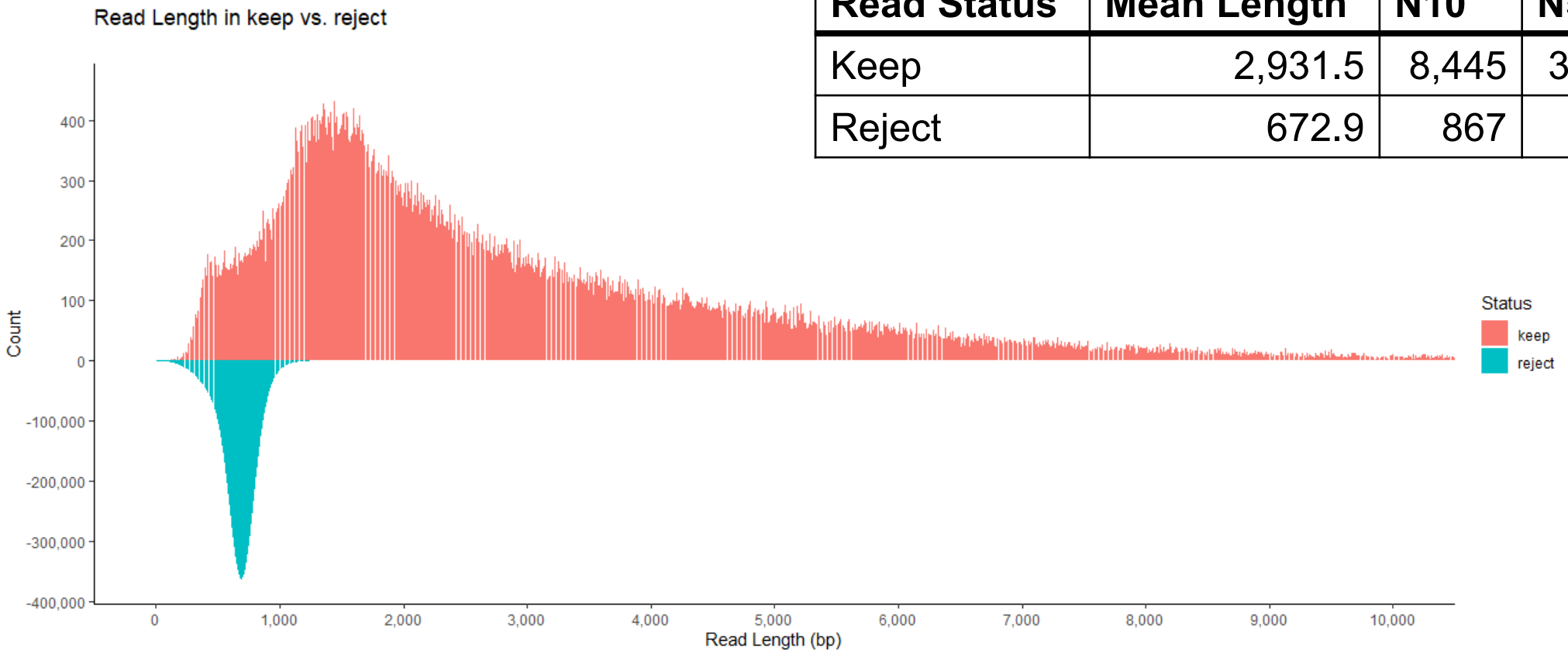
- Need to provide a bed file to define the ROI
- A well-designed bed file is
  - Total size of bed file should cover 1-5% of genome
  - Bigger on-target boost with smaller size
    - Diminishing returns between 1% and 3%
  - Padding size recommended at N90 of fragment library
    - Which is to say, the padding should be larger than most reads
- Size of the ROI impacts the boost in depth
  - A 1% of the genome ROI might see a 5x boost
  - A 10% of the genome ROI might only see a 2x boost
- Off-target reads are rejected after around 200-600bp. These reads can be useful as a low-pass short-read WGS.

ADAPTIVE SAMPLING EXAMPLE OUTPUT

- Target: 3-5x boost in ROI

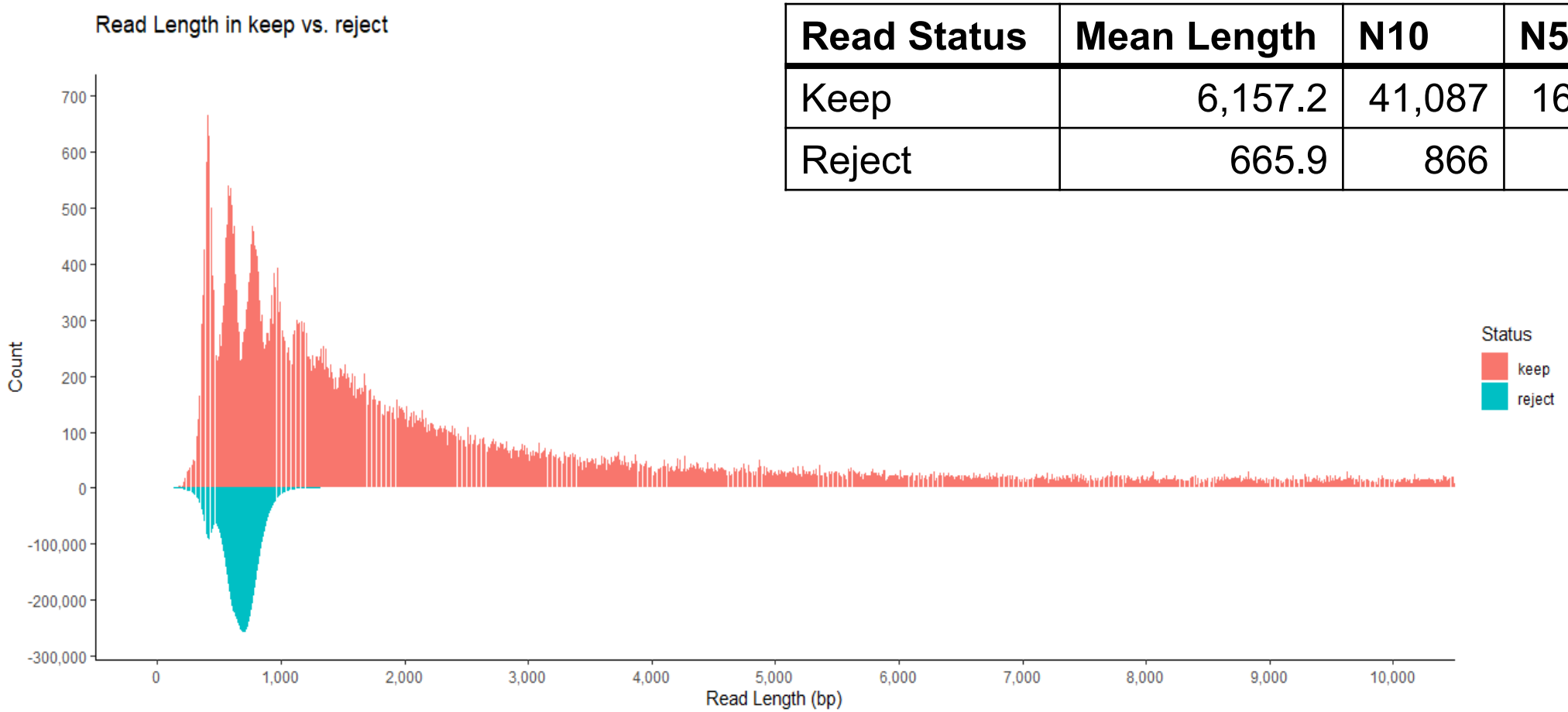
Sample	WGS Depth	On-target Depth	Off-target Depth	%Genome	OnTarget/WGS
Sample 1	32.55	N/A	N/A	100%	1.00x
Sample 2	27.88	41.61	26.12	(Methyl) 10%	1.49x
Sample 1	36.12	60.85	34.88	(Methyl) 5%	1.68x
Sample 1	37.54	70.47	36.62	(Methyl) 3%	1.88x
Sample 1	35.19	73.53	34.89	(Methyl) 1%	2.09x
Sample 1	32.09	65.49	30.27	(CMRG) 1%	2.04x
Sample 1, SRE, Not Sheared	11.84	37.60	11.59	(CMRG) 1%	3.19x
Sample 1, SRE, Sheared	24.04	84.54	23.46	(CMRG) 1%	3.52x
Sample 3, 20kb library	20.41	157.37	19.60	(CMRG) 1%	7.71x

“KEEP/REJECT” READS



Read Status	Mean Length	N10	N50	N90
Keep	2,931.5	8,445	3,934	1,504
Reject	672.9	867	707	542

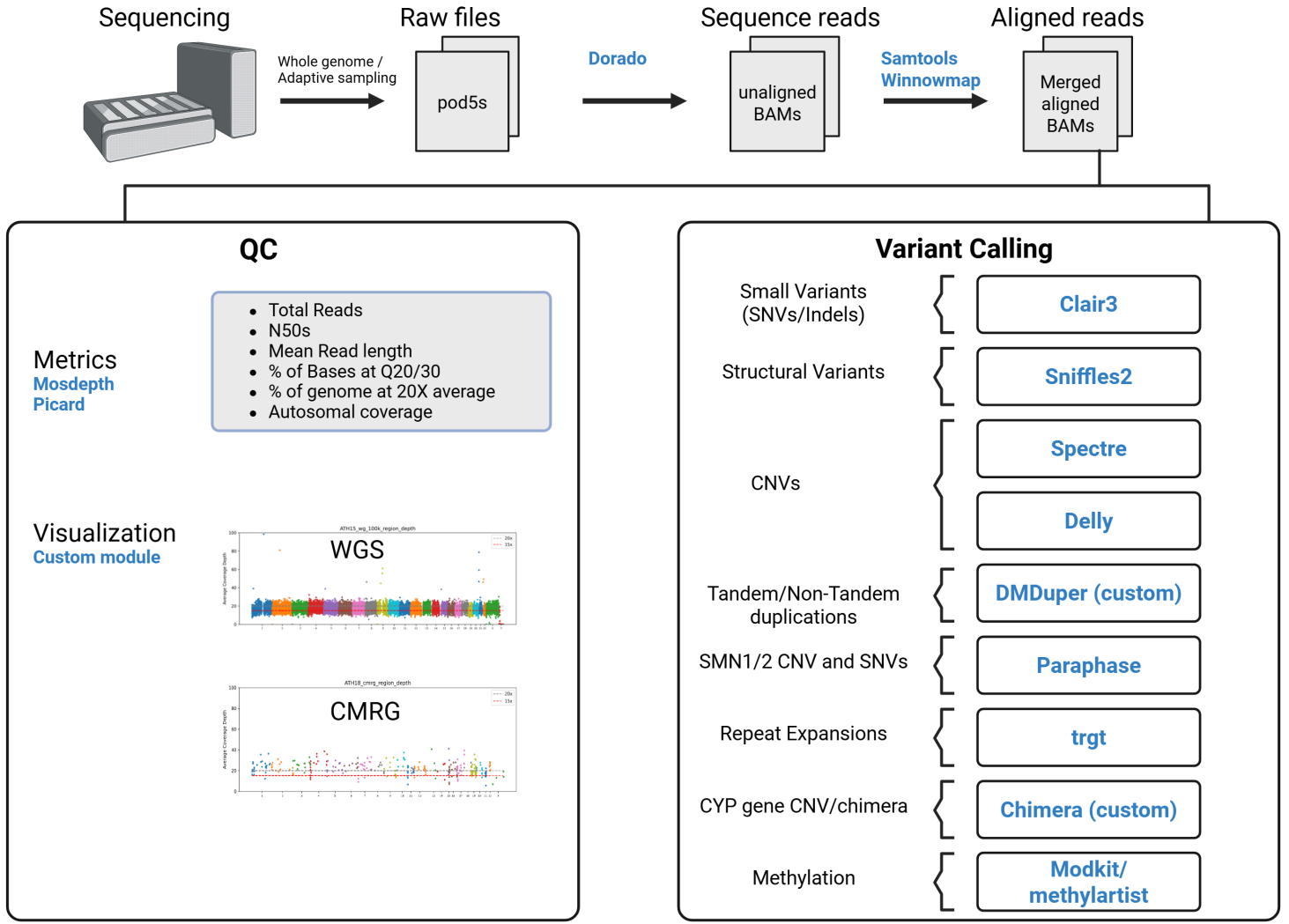
“KEEP/REJECT” READS



Read Status	Mean Length	N10	N50	N90
Keep	6,157.2	41,087	16,922	2,446
Reject	665.9	866	703	519

PIPELINE FOR ONT DATA

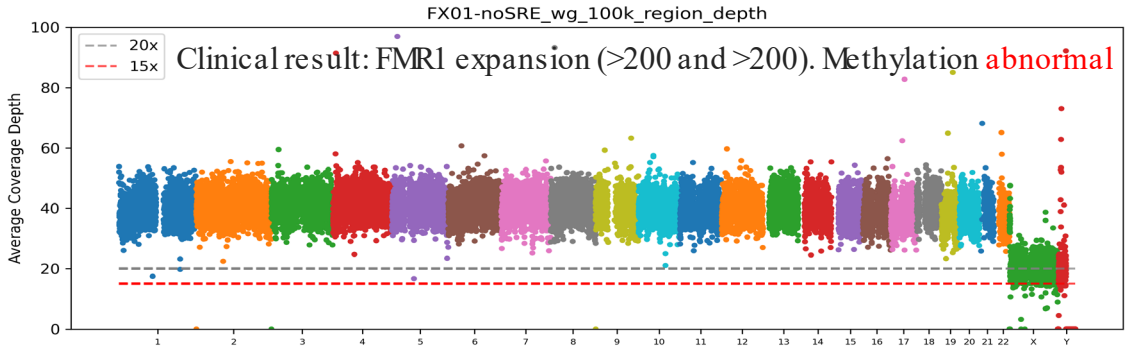
- Internally developed using mostly published tools
- Designed for DNA variant calling
- Includes methylation
- Several long-read specific tools doing familiar bioinformatics tasks
  - e.g. Alignment

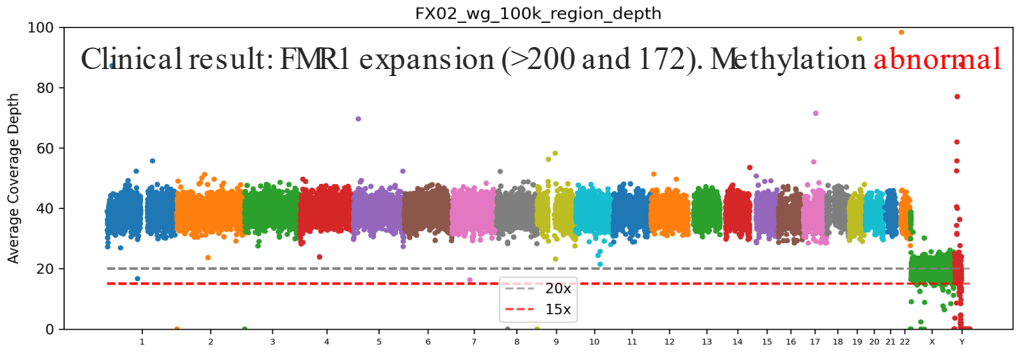


MODKIT

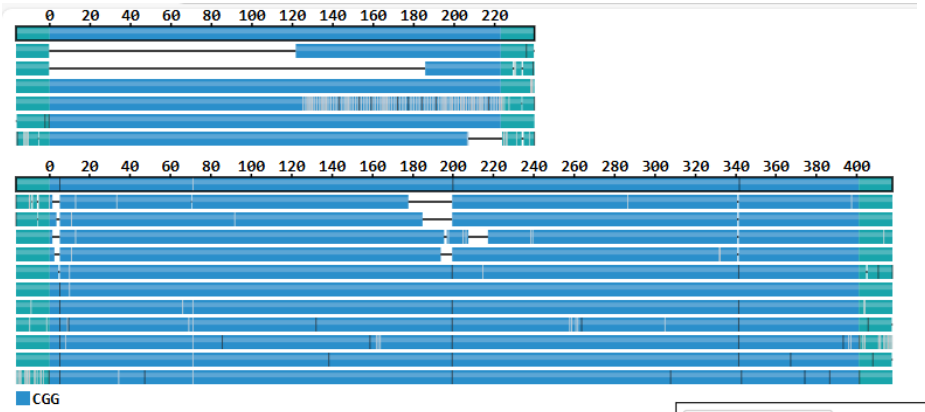
- ONT developed tool for converting modBAM to bedMethyl format
- If base calling was performed with an appropriate modification model, then the resulting bam has MM: and ML: tags representing methylation
- The bedMethyl format is the base 3-column bed file with additional columns describing methylation
  - Count of reads at this position
  - Count of methylated reads at this position
- Caveats
  - If there are no reads at a CpG position, then there is no output at this position from modKit

FRAGILE-X SYNDROME

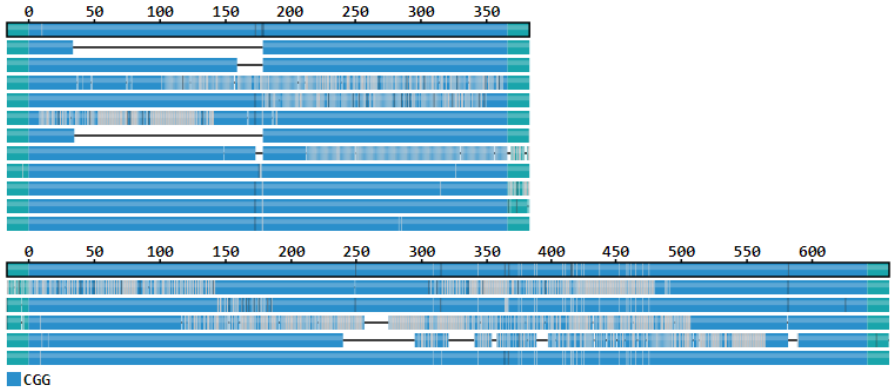




Tool: trgt

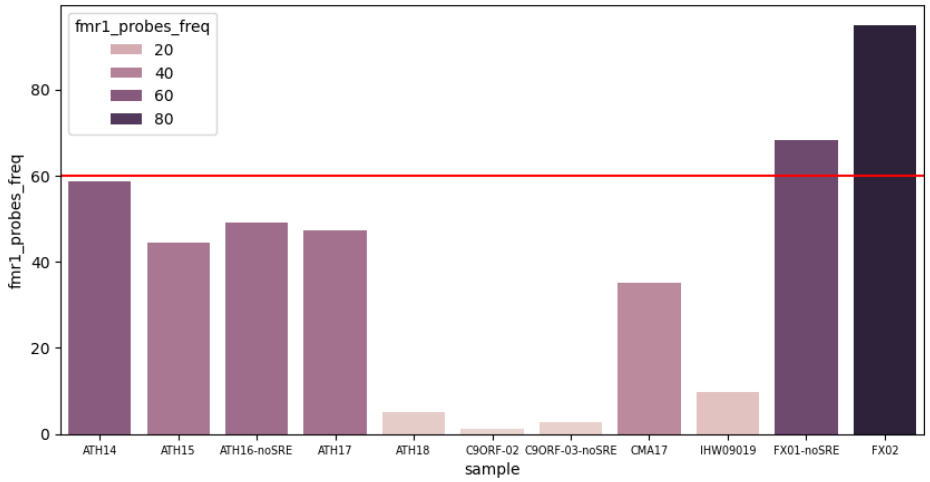


Tool: trgt



Methylation result: Hypermethylation in FMR1 promoter region

Used average of methylation frequencies from modkit output in FMR1 promoter region



IMPROVED DARK GENOME

Research | Open access | Published: 20 May 2019

Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight

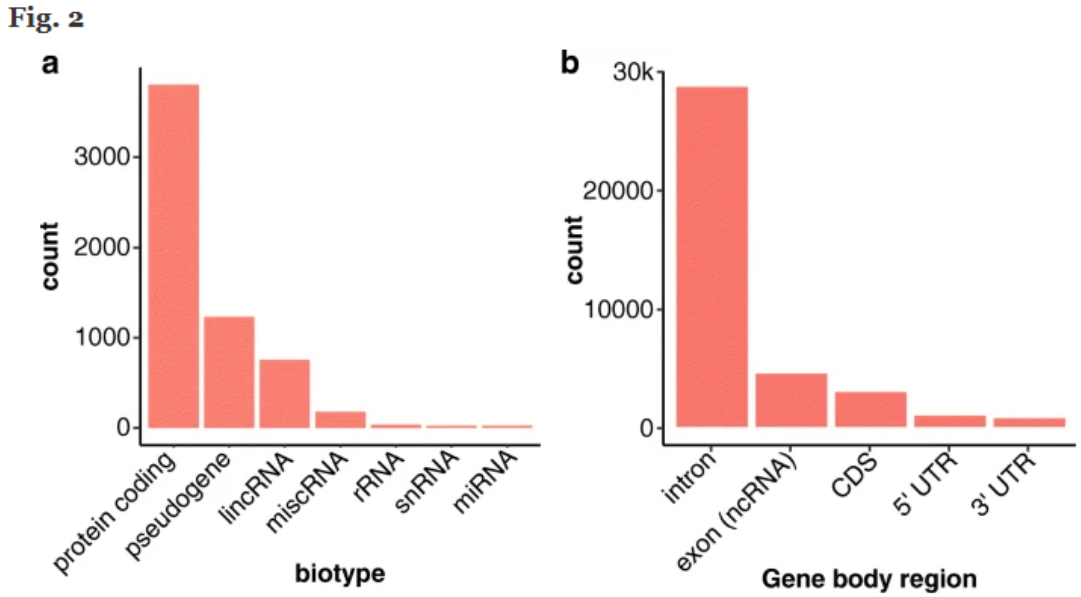
Mark T. W. Ebbert, Tanner D. Jensen, Karen Jansen-West, Jonathon P. Sens, Joseph S. Reddy, Perry G. Ridge, John S. K. Kauwe, Veronique Belzil, Luc Pregent, Minerva M. Carrasquillo, Dirk Keene, Eric Larson, Paul Crane, Yan W. Asmann, Nilufer Ertekin-Taner, Steven G. Younkin, Owen A. Ross, Rosa Rademakers, Leonard Petrucelli & John D. Fryer

Genome Biology **20**, Article number: 97 (2019) | Cite this article

- Genomic regions that cannot be adequately assembled or aligned using short reads.
- Dark by “depth” – Lack of coverage for regions. (<5x).
- Dark by “ambiguous alignments” of sequence reads. (>90% of reads have low MQ, <9)

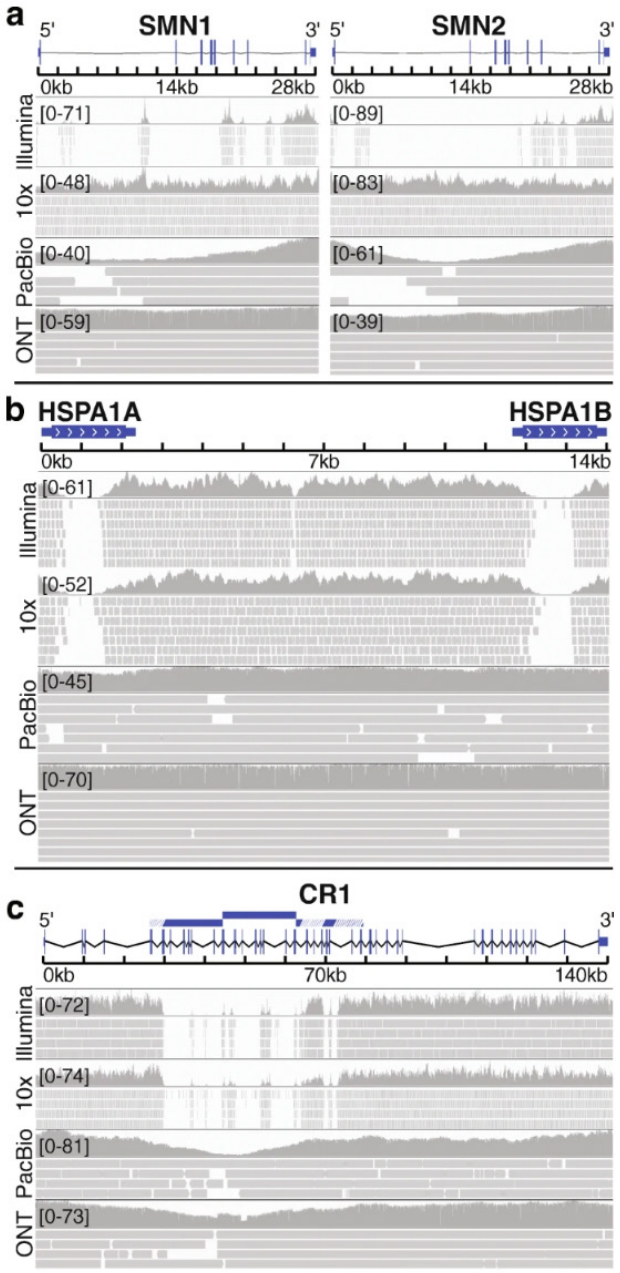
DARK GENOME AND WHAT IT CONSTITUTES?

- https://github.com/mebbert/DarkRegionFinder
- 36,794 dark regions characterized
- Implicated in pathways important to human health, development, and reproduction.
- 76 dark genes with known mutations associated with 326 human diseases



Ebbert, M.T.W., Jensen, T.D., Jansen-West, K. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol 20, 97 (2019). https://doi.org/10.1186/s13059-019-1707-2

LONG READ SEQUENCING RESOLVES DARK REGIONS



- Authors showed good alignments in
  - Pseudogenes
  - Low complexity regions
  - Repeats
- Resolves 77% of dark regions

Ebbert, M.T.W., Jensen, T.D., Jansen-West, K. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol 20, 97 (2019). https://doi.org/10.1186/s13059-019-1707-2

FINAL THOUGHTS

- Find some data from ONT/PacBio NIST benchmarking that shows LR can basically do anything that SR can do and some more
  - Like CMRG regions
- And this is something our internal investigations have so far supported

ACKNOWLEDGMENTS

QUESTIONS & ANSWERS