

Real Time minION data analysis: Run and Read Until

Matt Loose

#1

The future is here?



"Wondering how you can fit MinION
into your existing workflows is like
sitting in front a space ship and
wondering how you're going to use it
to commute to work; it's like sitting in
front of a time machine, and
wondering if you could get to the
shops before they close."

@BioMickWatson

5-10/20 Gb



50 - 100 Gb



3-6 Tb



#2

Run Monitoring



What is it?

Launched September 2014 (30 b/s)

Slowed down March 2016 (250 b/s)

Struggled with new MinKNOW

Now available at GitHUB for local install....

The screenshot shows a Twitter feed for the account 'minoTour'. The feed contains the following tweets:

- Nick Loman** (@pathogenomenick) · Nov 25
@flashton2003 @AW_NGS @nanopore We can multiplex 12 bacterial genomes on a \$500 MinION cell easily -- output 2-5Gb in our hands.
- Richard Leggett** (@richardmleggett) · Nov 25
@pathogenomenick What percentage of your flowcells are generating that much data?
- Nick Loman** (@pathogenomenick) · Nov 25
@richardmleggett all of them
- mattloose** (@mattloose) · Nov 25
@pathogenomenick @richardmleggett yeah- we're seeing the same now.
- mattloose** (@mattloose) · Nov 19
I think we've just got a 4.7 Gb run on a native sample.

Each tweet includes standard Twitter interaction icons (retweet, like, reply, etc.) and a small image of the user's profile picture.



Real Time Analysis

Processes:

minKNOW

nanonet

metrichor

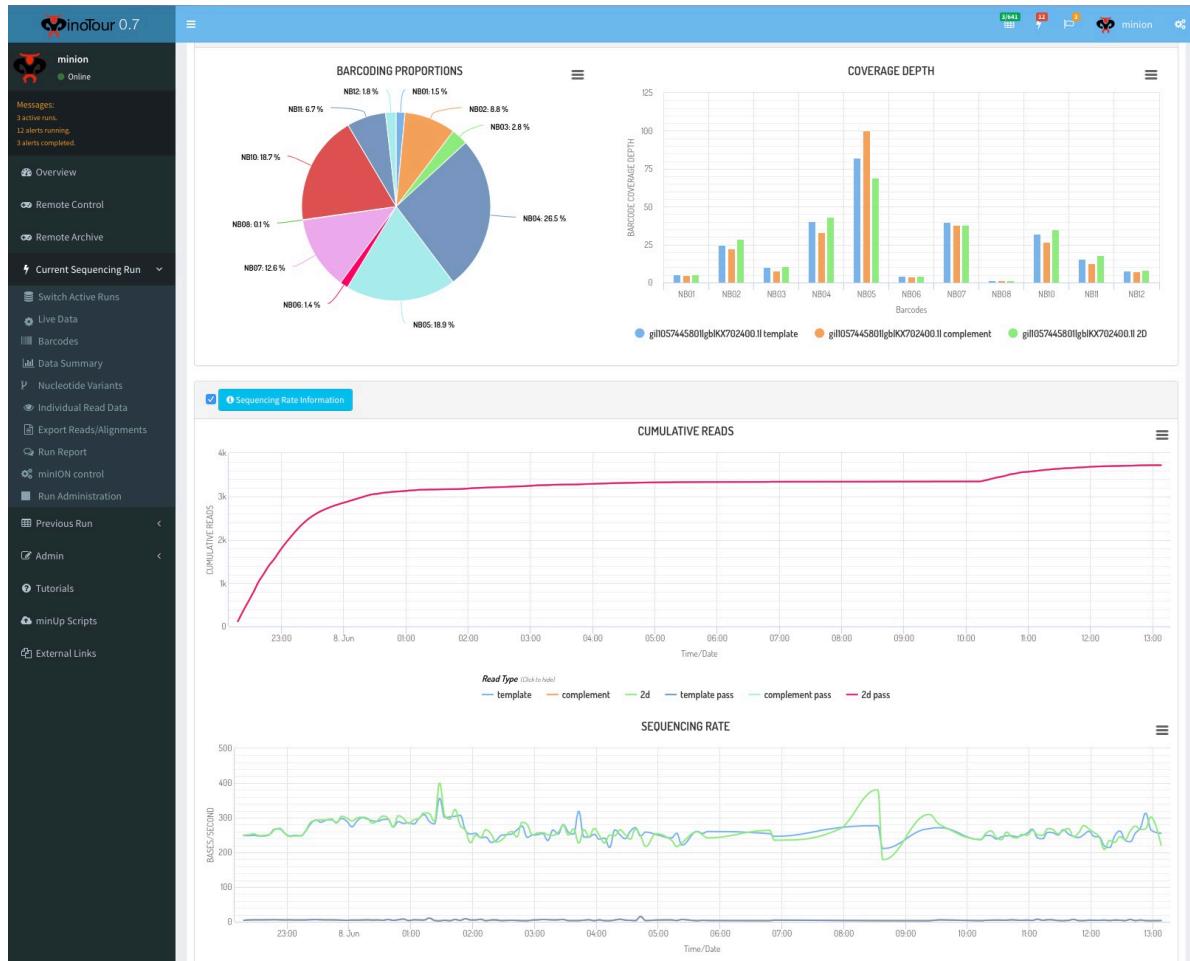
Handles:

Barcodes

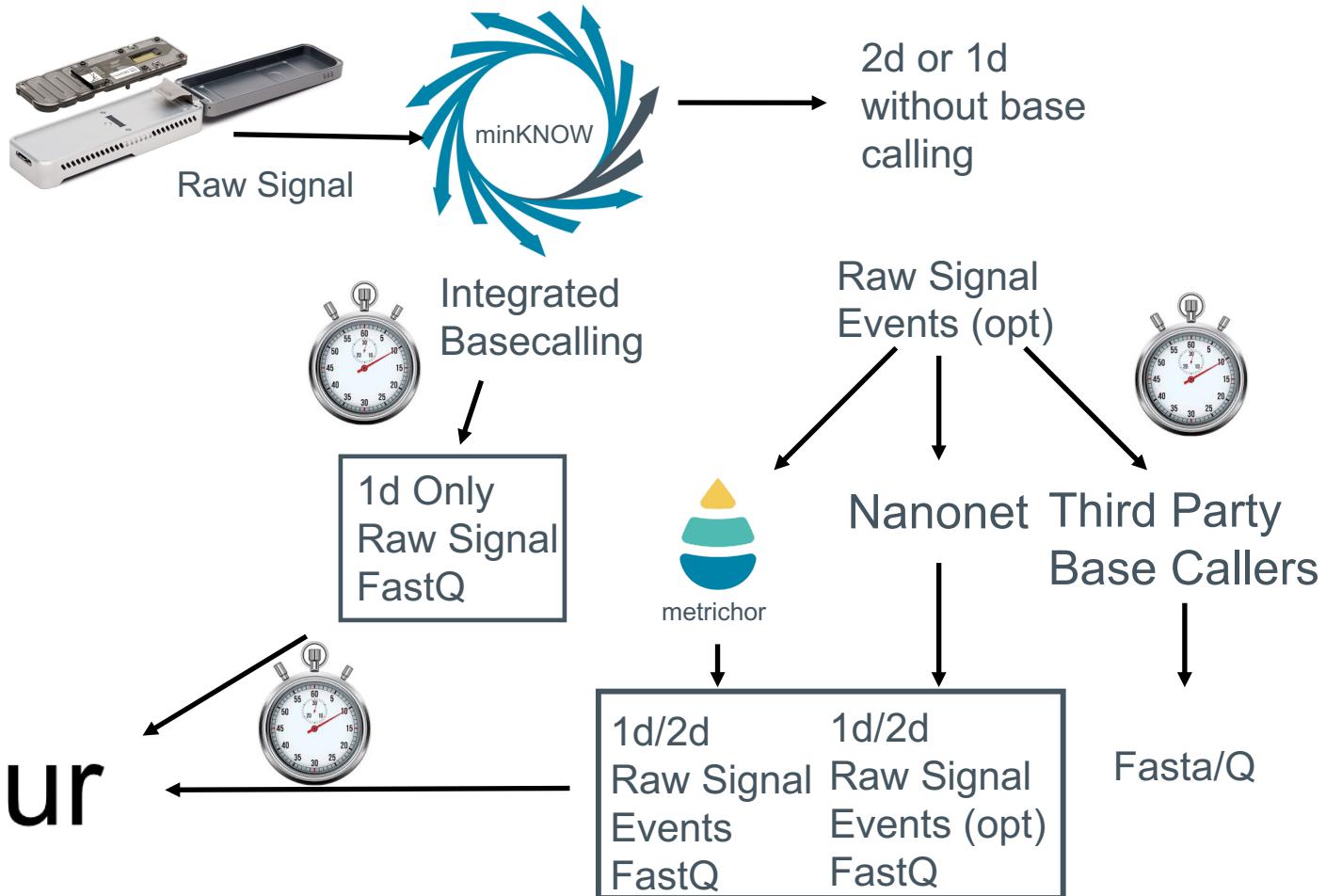
Alignment to Reference

Provides:

Metrics



Monitoring Runs



minoTour keeps up!

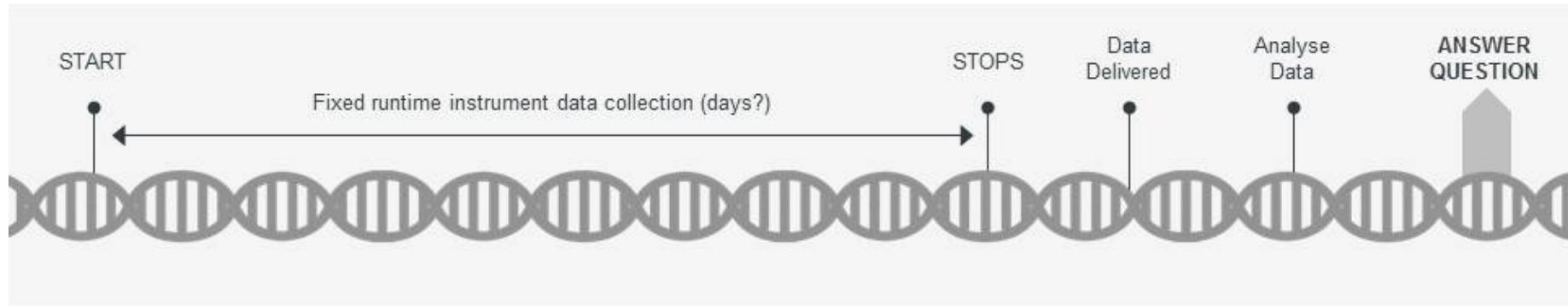
The image displays two side-by-side software interfaces. On the left is the MinKNOW desktop application, which shows a 'processing' status bar at the top. Below it are sections for 'Protocol' (NC_CTC_Run.py), 'MinION ID' (MS00000), 'ASIC ID' (131070), 'Sample ID' (DemoSample), 'Flowcell ID' (fab1234), 'Host' (Oneday.local), 'MinKNOW Version' (1.1.21), and 'Protocols Version' (1.1.21). A large blue progress bar indicates 'Starting Sequencing'. A sidebar on the right lists 'Starting Sequencing', 'Reached target temperature', and 'waiting for temperature to be within acceptable bounds'. In the center, there's a 'Physical layout' section with tables for Temperature, Voltage, and Analysis parameters, and a large heatmap representing the physical layout of the flowcell. On the right is the minoTour web application, showing a 'Welcome to minoTour' message and an 'Important Notice' about bug reporting and usage statistics. The minoTour interface also includes a sidebar with links like Overview, Remote Control, Remote Archive, Previous Run, Admin, Tutorials, minUp Scripts, and External Links.

#3

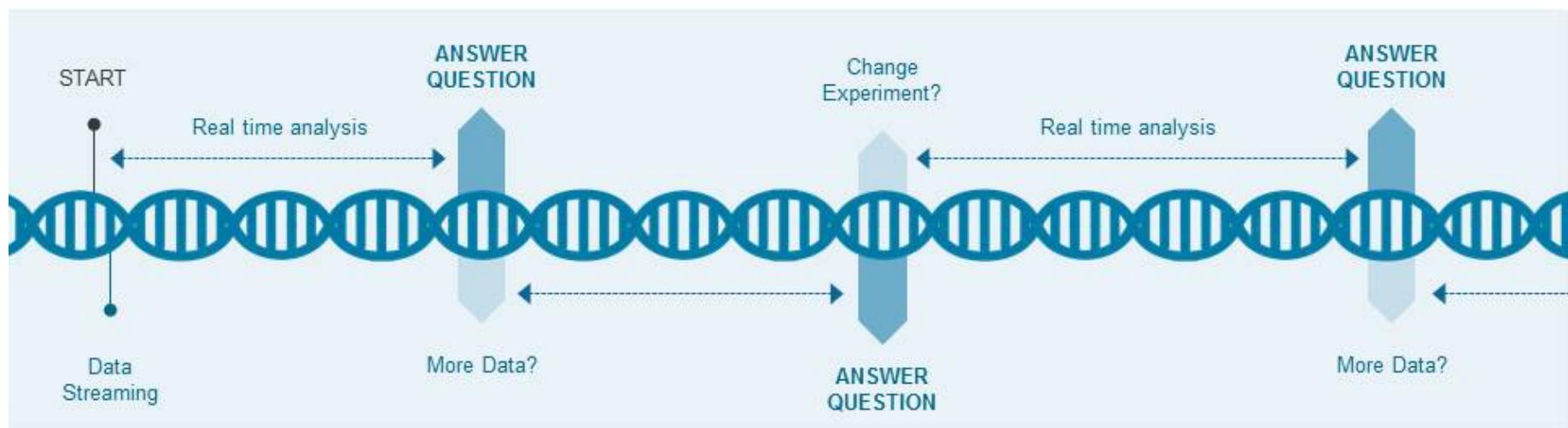
Run Until

Run Until...

Traditional Workflow



Run Until...



MinKNOW

00h 16m 30s processing

Protocol: NC_CTC_Run.py

MinION ID: MS00000

ASIC ID: 131070

Sample ID: DemoSample

Flowcell ID: fab1234

Host: Oneday.local

MinKNOW Version: 1.1.21

Protocols Version: 1.1.21

Starting Sequencing

Reached target temperature

waiting for temperature to be within acceptable range

Physical layout

Temperature	Voltage	Analysis
MinION: 35.00°C ASIC: 35.04°C	0mV	Delay: 1ms
0 unclassified	0 -inf to -5pA	14 zero
39 single pore	298 strand	155 multiple
0 active feedback	0 saturated	6 unavailable
Adapter		

Trace Viewer

Base Calling

Report Available

inolour 0.7

minion Online

Messages: 1 active run, 0 alerts running, 0 alerts completed.

- Overview
- Remote Control
- Remote Archive
- Current Sequencing Run
- Live Data
- Data Summary
- Individual Read Data
- Export Reads/Alignments
- Run Report
- minION control
- Run Administration
- Previous Run
- Admin
- Tutorials
- minUp Scripts

Current Data Summary - run: minion OneDay local 20161126 FNfab1234 MS00000

minion OneDay local 20161126 FNfab1234 MS00000

Processing Activity

READ UPLOAD AND PROCESSING

Reads And Coverage Summary

READ COUNT	YIELD	AVERAGE READ LENGTH	MAXIMUM READ LENGTH
2000	12.5M	8k	40k
1500	10M	6k	30k

'Run Until'

The screenshot shows the miniTour 0.7 software interface. On the left is a sidebar with navigation links: Overview, Remote Control, Remote Archive, Current Sequencing Run (with a dropdown menu for Switch Active Runs, Live Data, Barcodes, Data Summary, Nucleotide Variants, Individual Read Data, Export Reads/Alignments, Run Report, Run Control, Run Administration, Previous Run, Admin, Tutorials, minUp Scripts, and External Links). The main area has a header "Live Control - run: minion COLLES L160692 20160607 FNFAD11740 MN16453 Zika library-4". It includes a "Live Interaction" section with a note about remote control security, a "CUMULATIVE READS" graph showing cumulative reads over time, a "Disk Space Monitoring" section for setting alerts on hard drive space, a "Simple Examples" section for setting alerts on sequencing, a "Coverage Depth" section for defining reference sequences, thresholds, and optional starts/ends, and sections for "Disk Use Notices", "Global Barcode Threshold Set.", and "Individual Barcode Threshold Set.".

Notifications (optional stop): Global Coverage Depth

Targeted Region Depth

Depth Per Barcode (same reference only)

Notifications:

Total number of bases sequenced

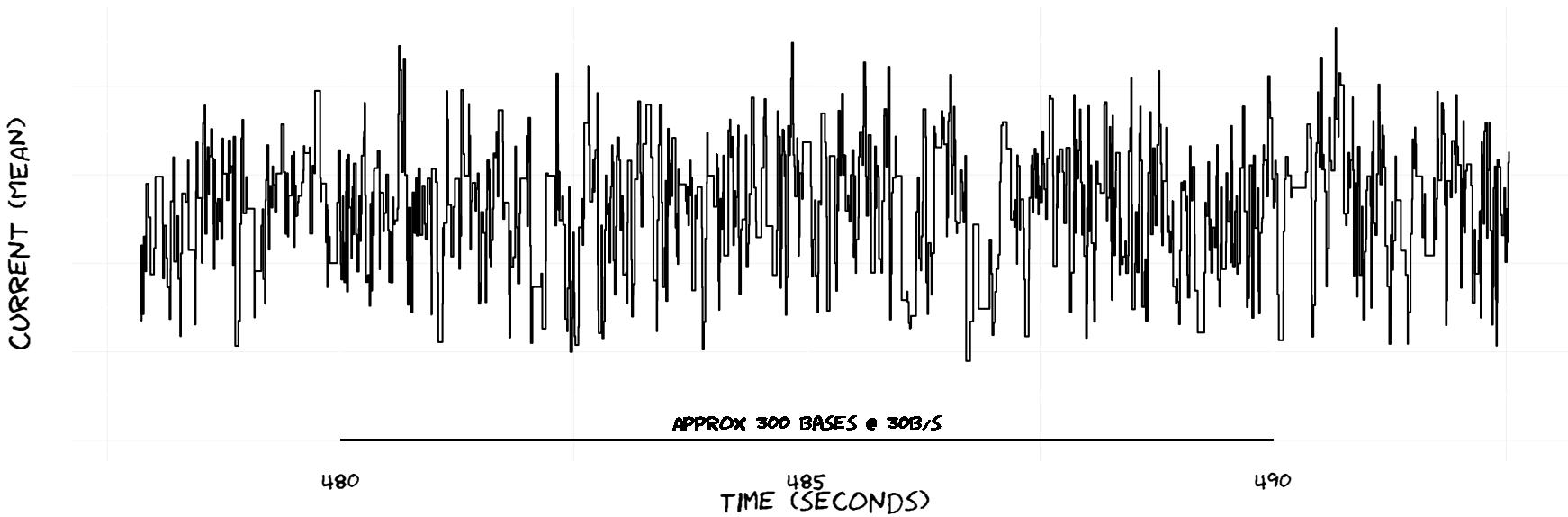
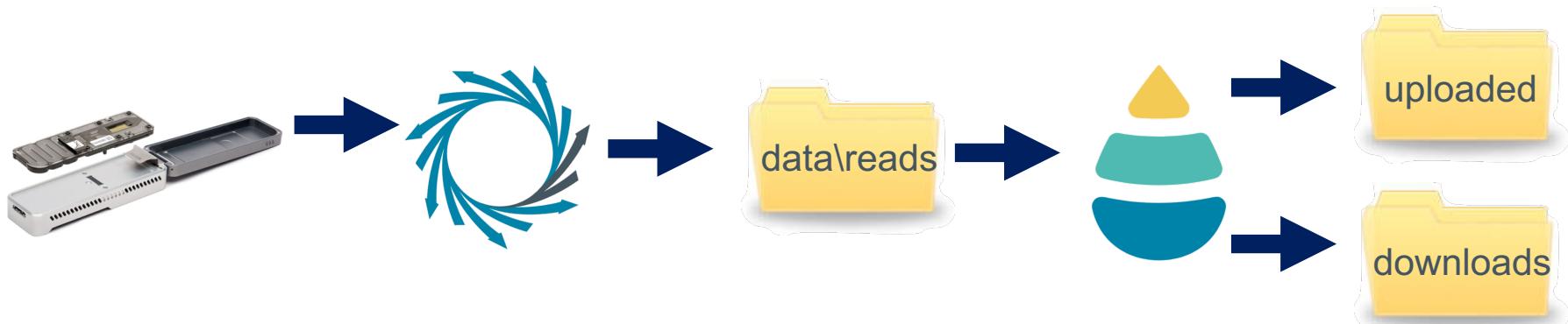
Drive Space Warnings

#4

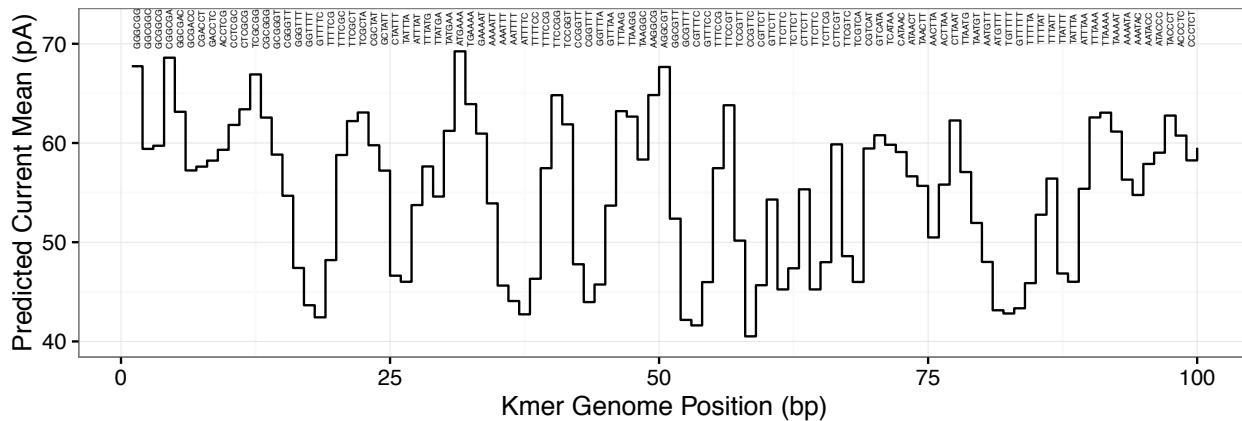
Read Until

Read Until – Selective Sequencing

Sequencing Workflow



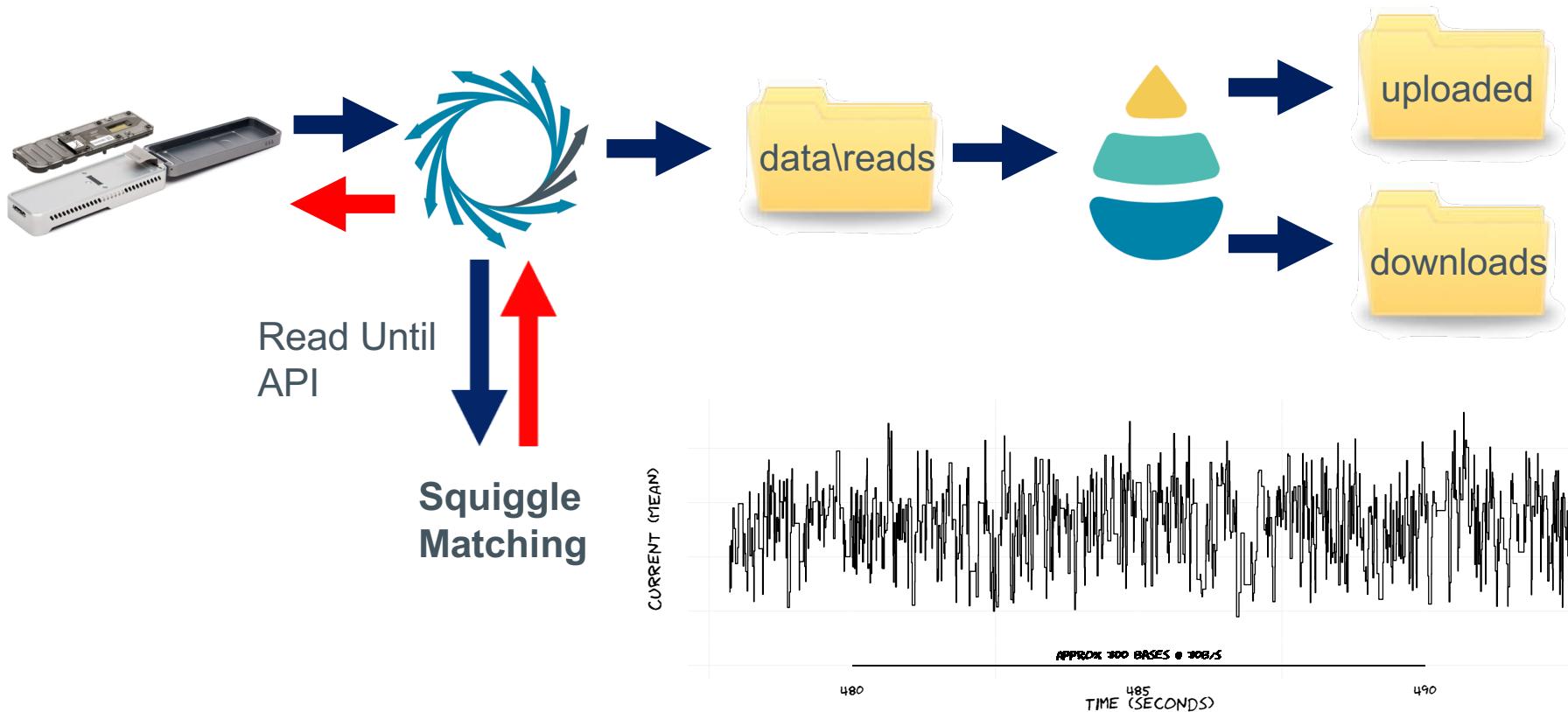
100bp Lambda Genome in Squiggle Space



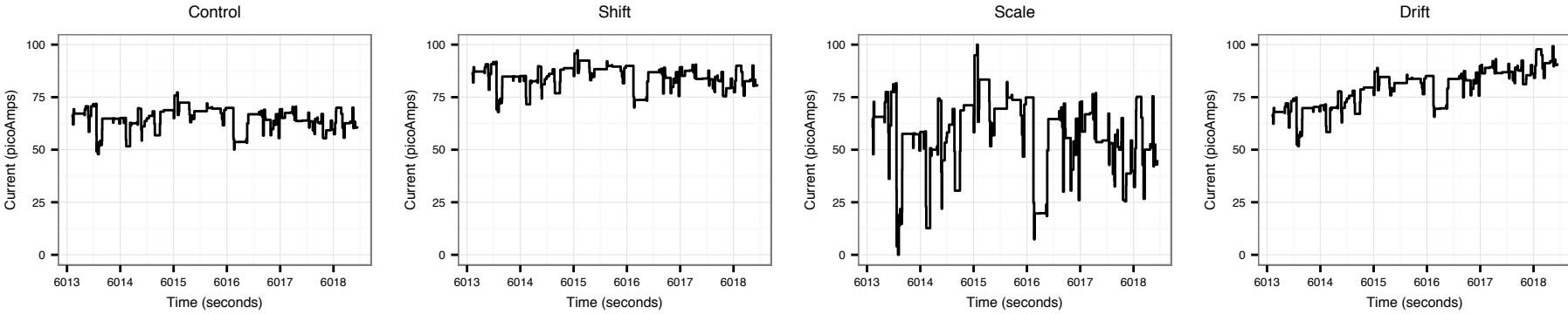
GGCGGG	67.73595
GGCGGC	59.405716
GCGGCG	59.723809
CGGCGA	68.596572
GGCGAC	63.144473
GCGACC	57.233718
CGACCT	57.61982
GACCTC	58.224973
ACCTCG	59.316305
CCTCGC	61.821412



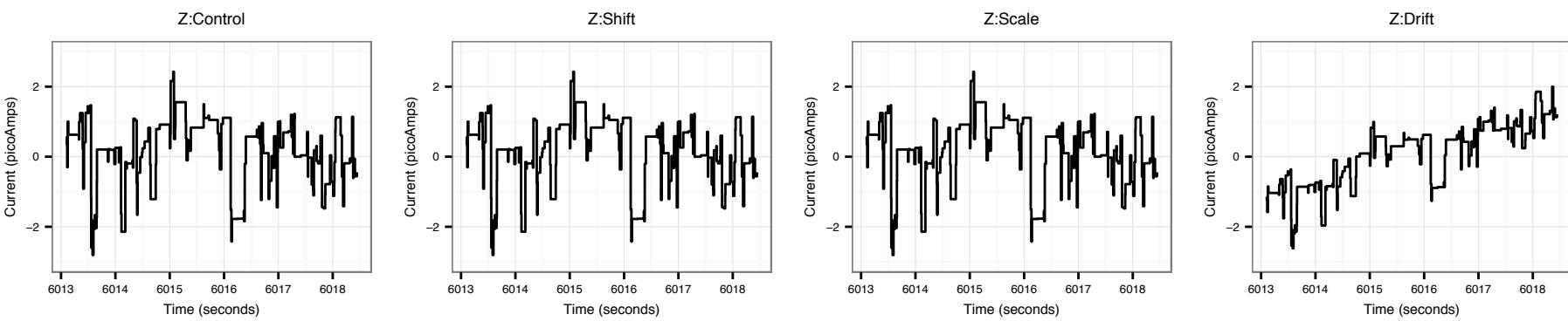
Read Until Workflow



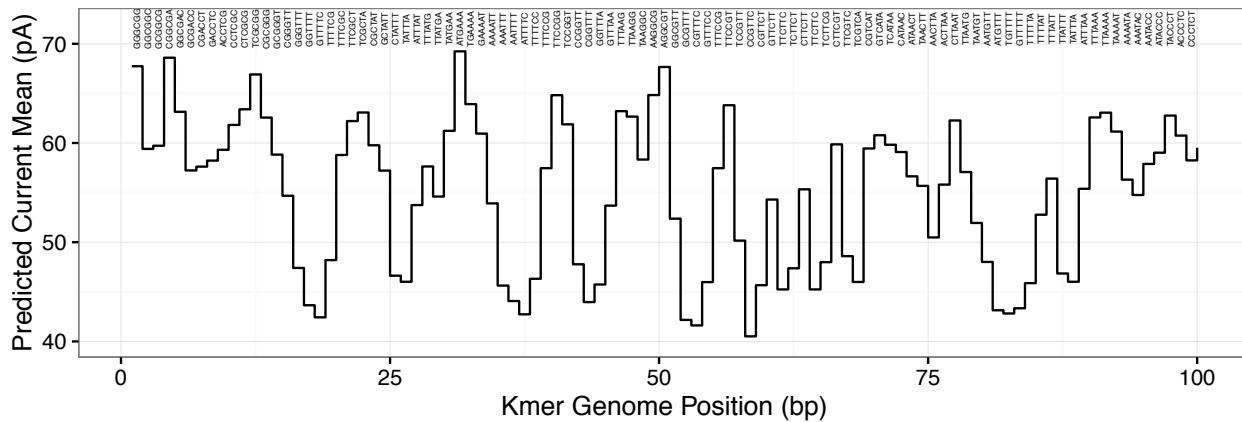
A) 'Raw' Squiggles



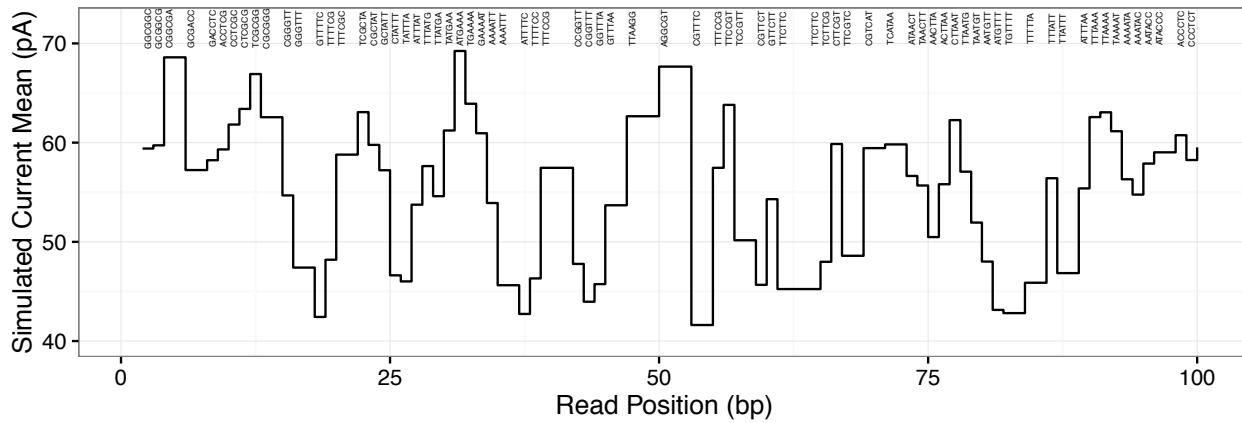
B) Z-Normalised Squiggles



100bp Lambda Genome in Squiggle Space



Simulated Read with Skips, Shift, Scale and Drift



#5

How To Compare Squiggles?



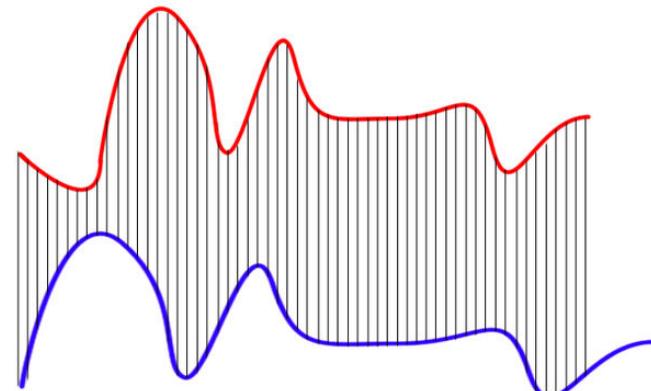
Dynamic Time Warping

Calculates the minimal distance between two paths which vary in time

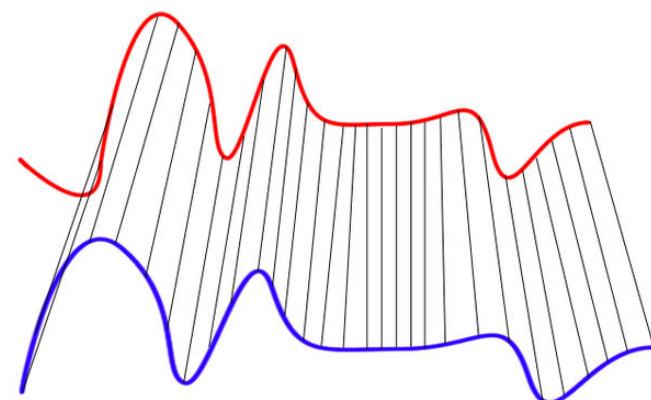
Guaranteed to find the optimal match

Computationally Expensive

Returns: Distance, Cost, Path



Euclidean Matching

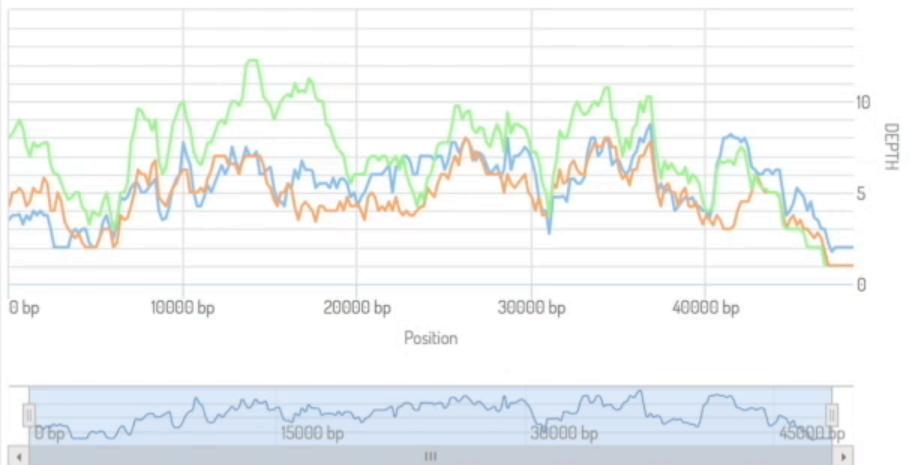


Dynamic Time Warping Matching

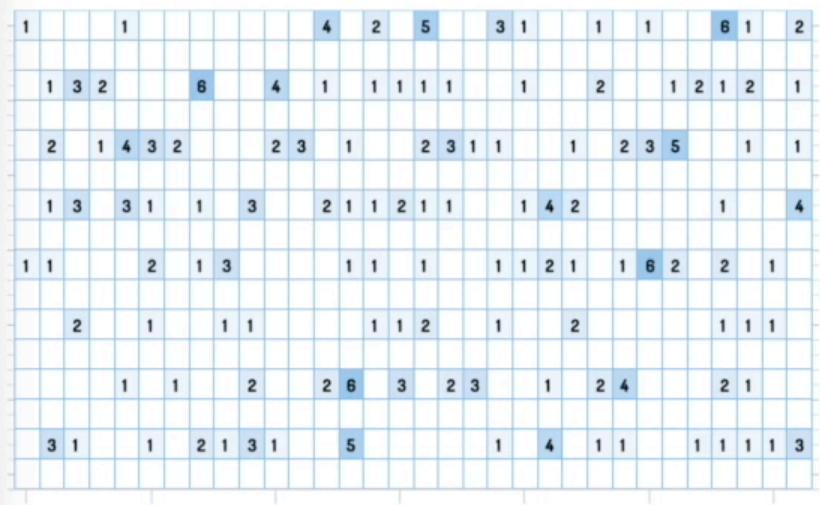
#6

Does Read Until Actually Work?

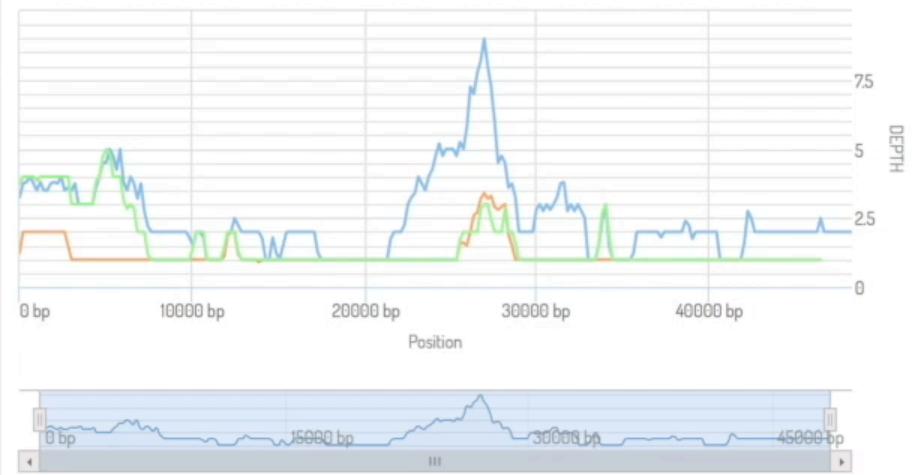
COVERAGE DEPTH FOR GII9626243IREFINC_001416.1I



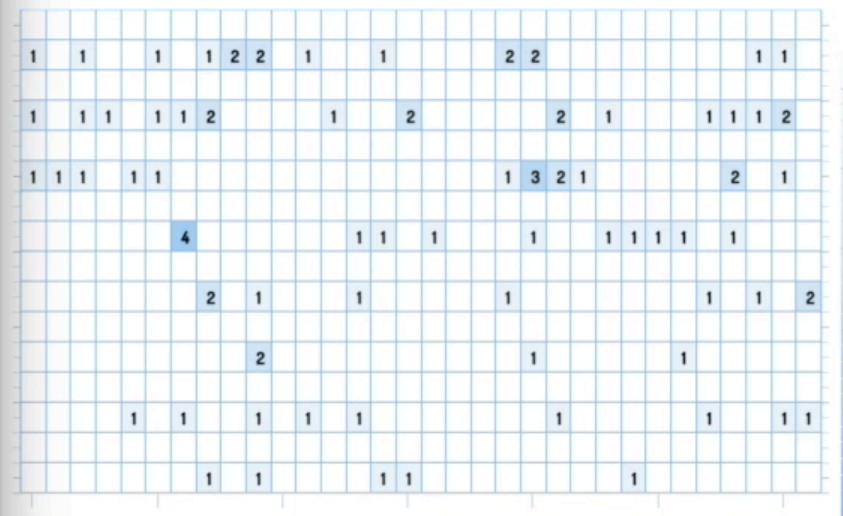
READS PER PORE



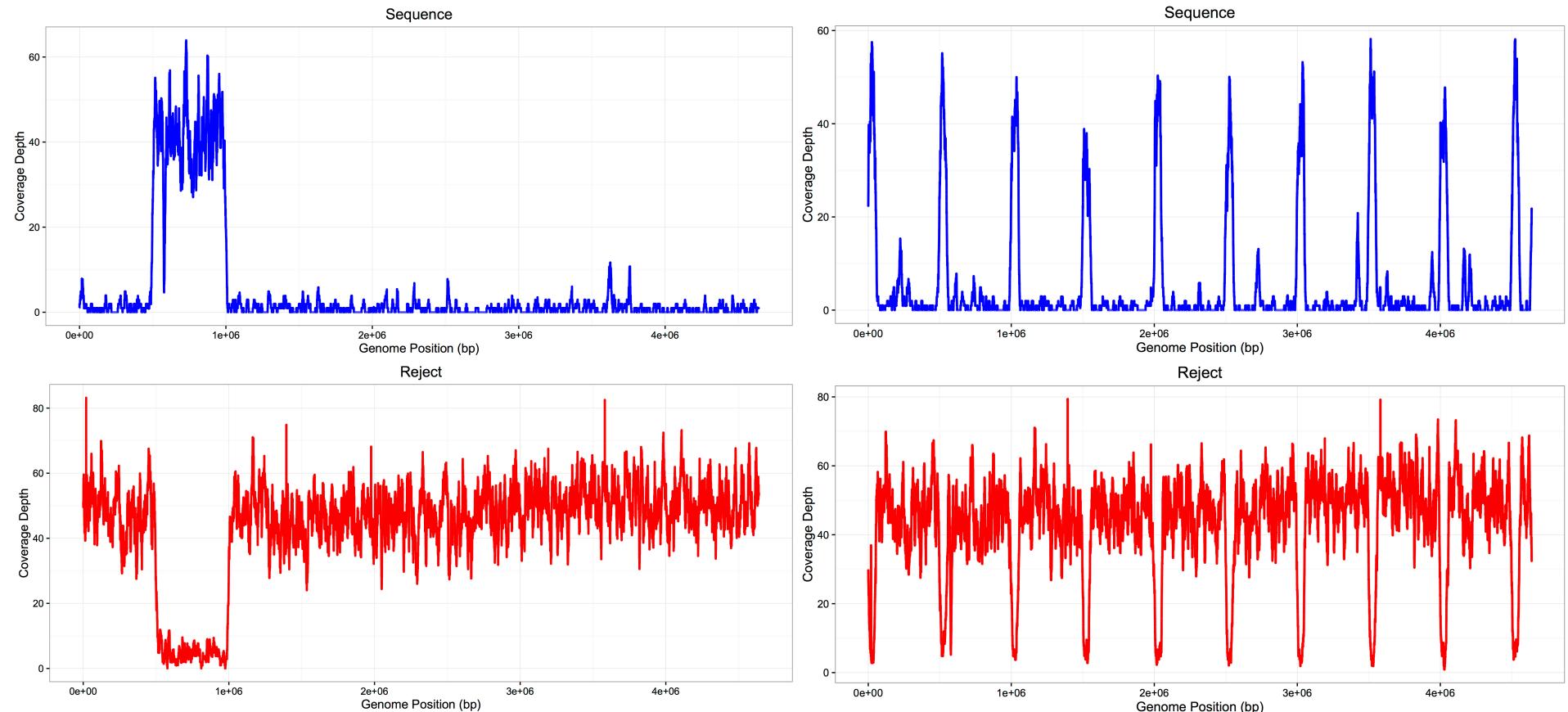
COVERAGE DEPTH FOR GII9626243IREFINC_001416.II



READS PER PORE



Selective Sequencing from a Background



#7

Potential Applications



Bill Gates

@BillGates



Following

From Ebola to Zika, this “lab in a suitcase” provides crucial data for outbreaks: [b-gat.es/1XIKQkZ](http://bit.ly/1XIKQkZ) via @verge



RETWEETS
1,377

LIKES
2,691



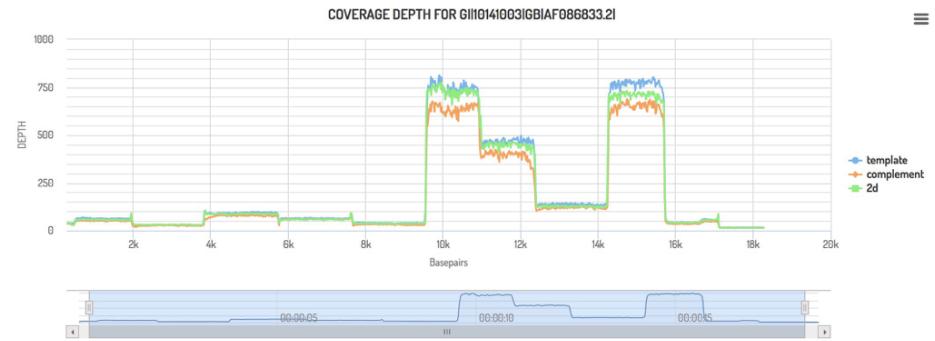
2:19 PM - 12 Feb 2016



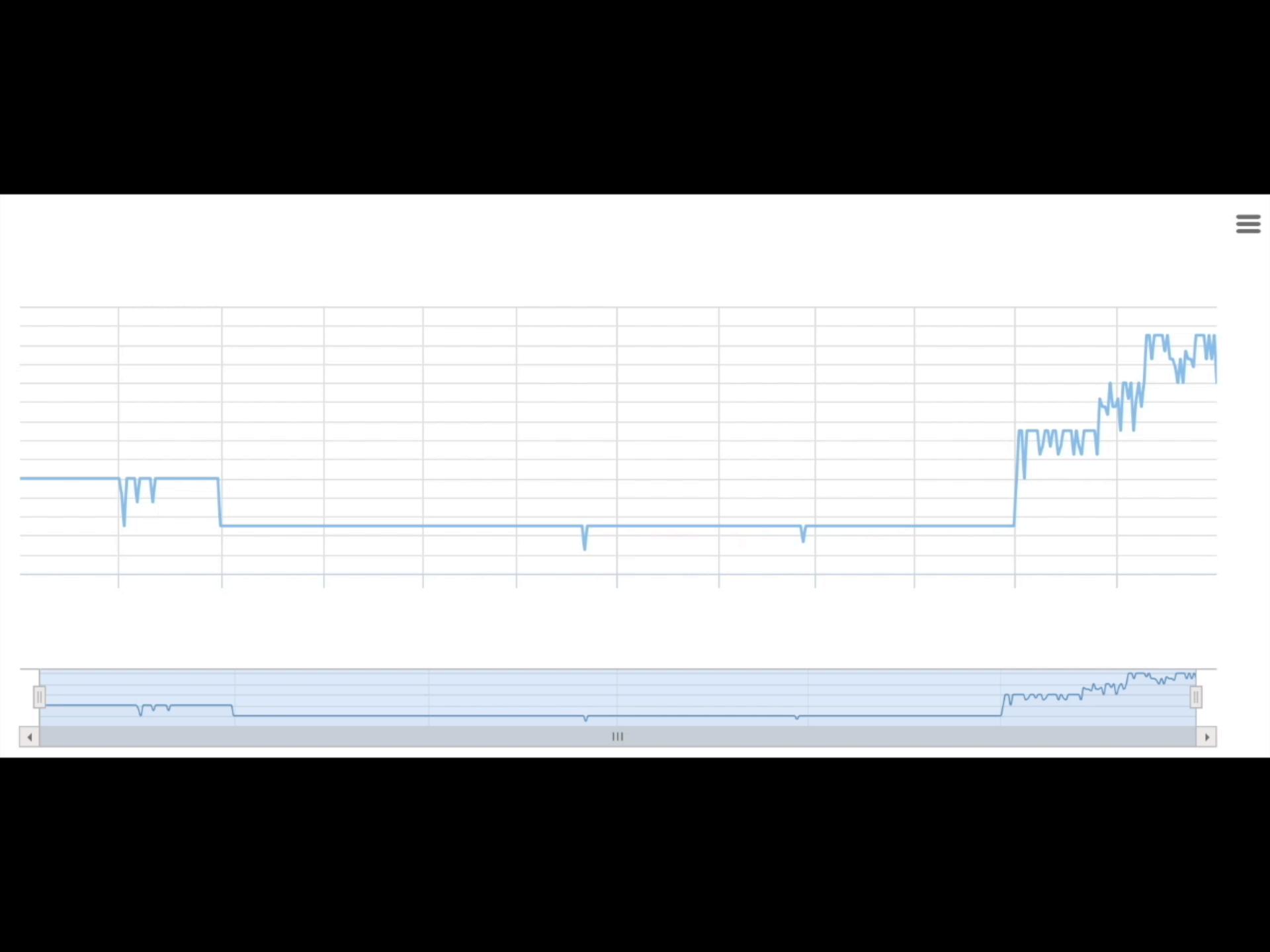
...

Applications for Squiggle Matching

Amplicon Sequencing

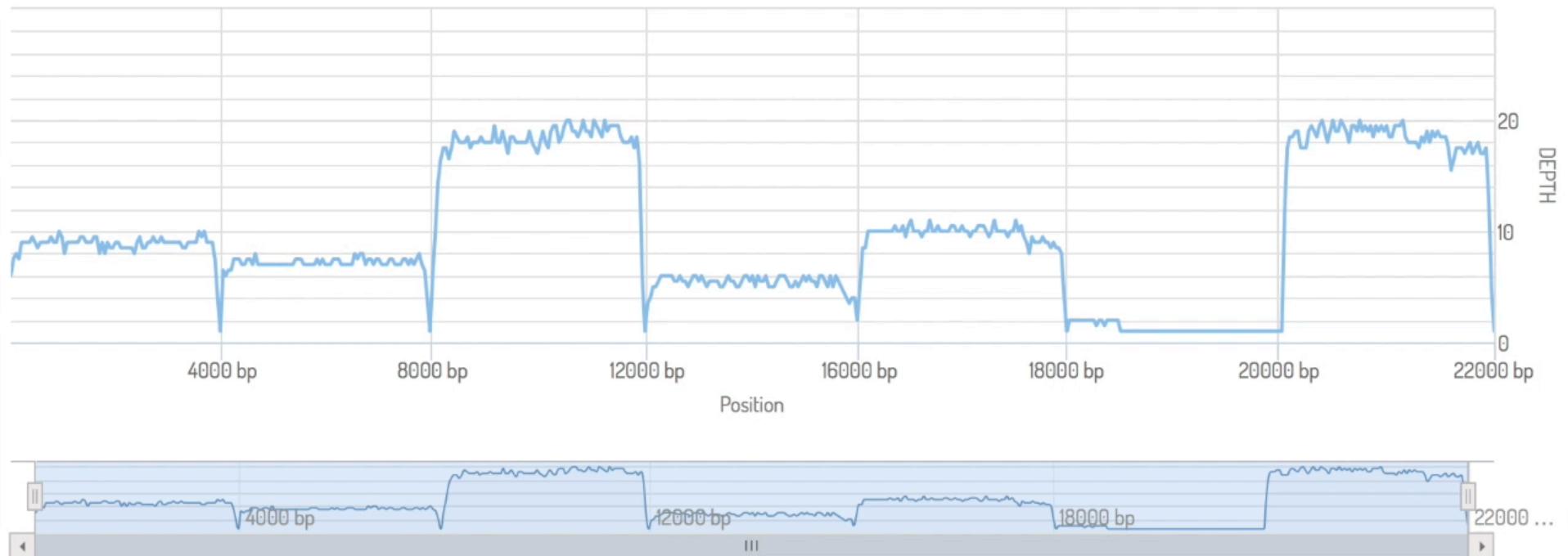


11 amplicons spanning 1 viral genome
Aligned in basespace by BWA in minoTour

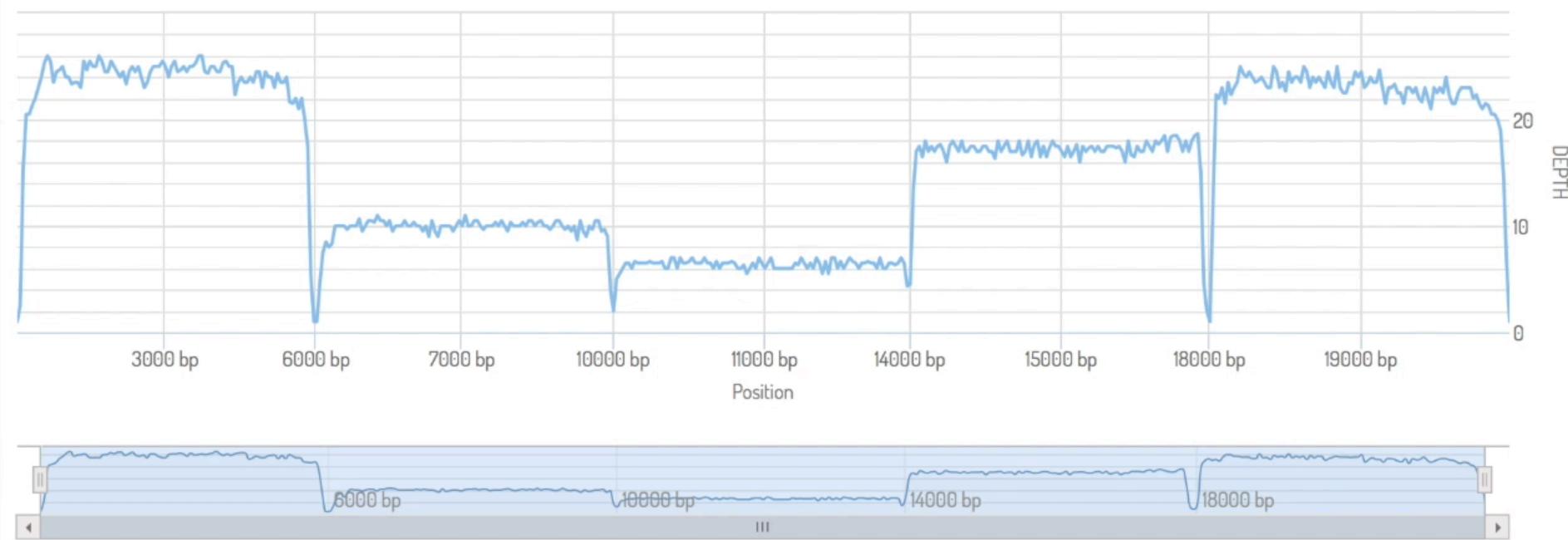


COVERAGE DEPTH FOR GII9626243|REFINC_001416.1I

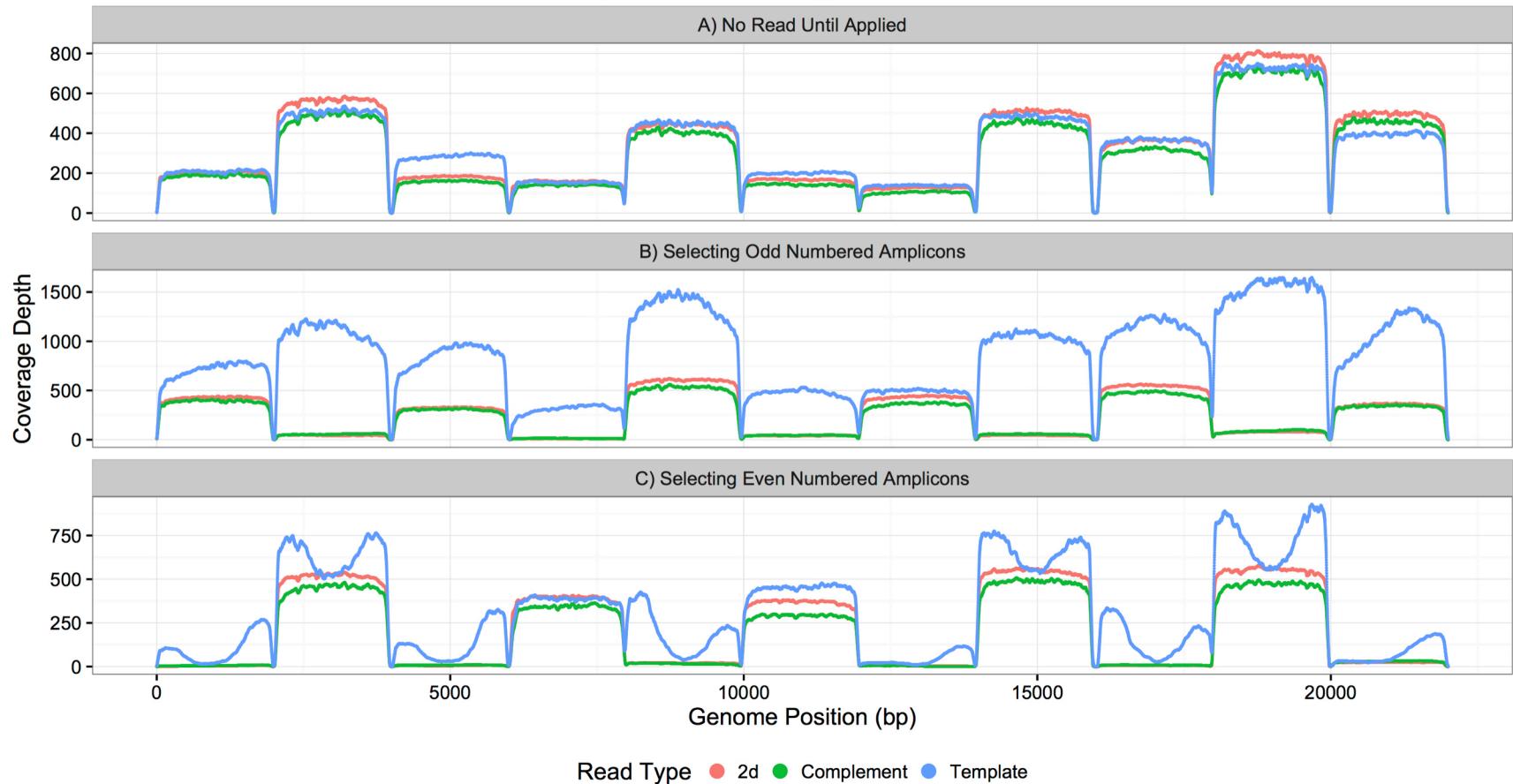
≡



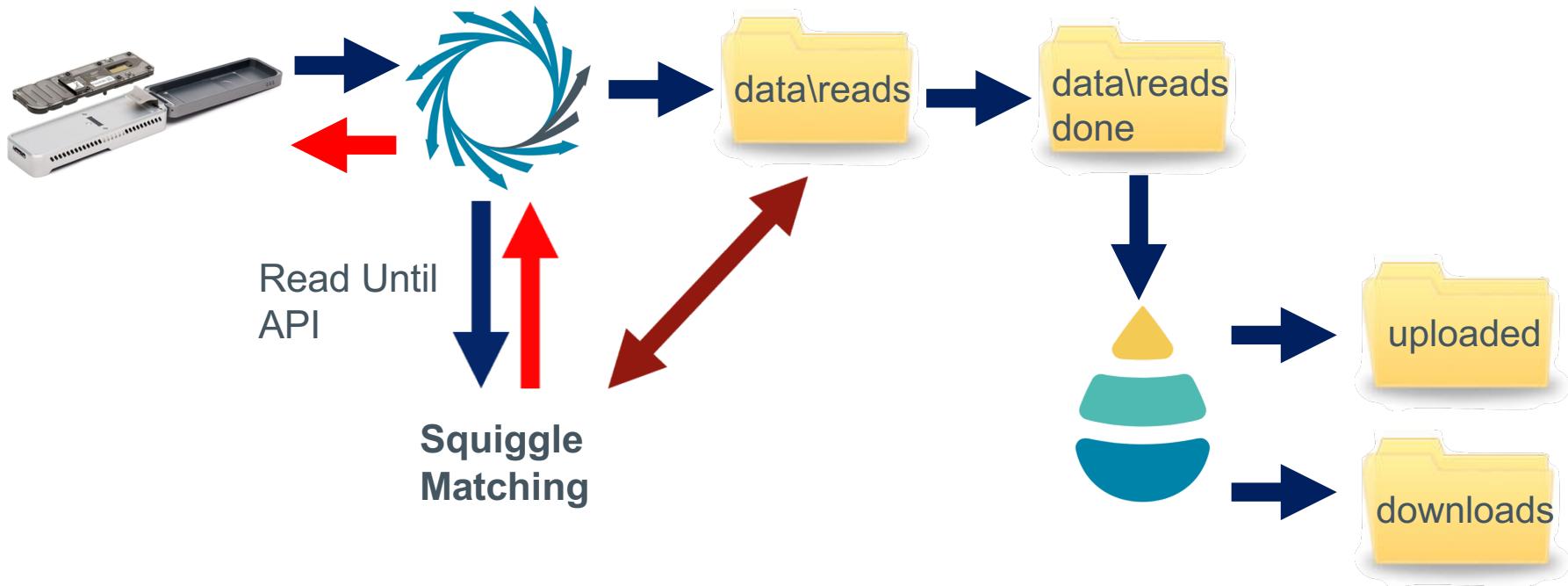
COVERAGE DEPTH FOR GII9626243|REFINC_001416.1I



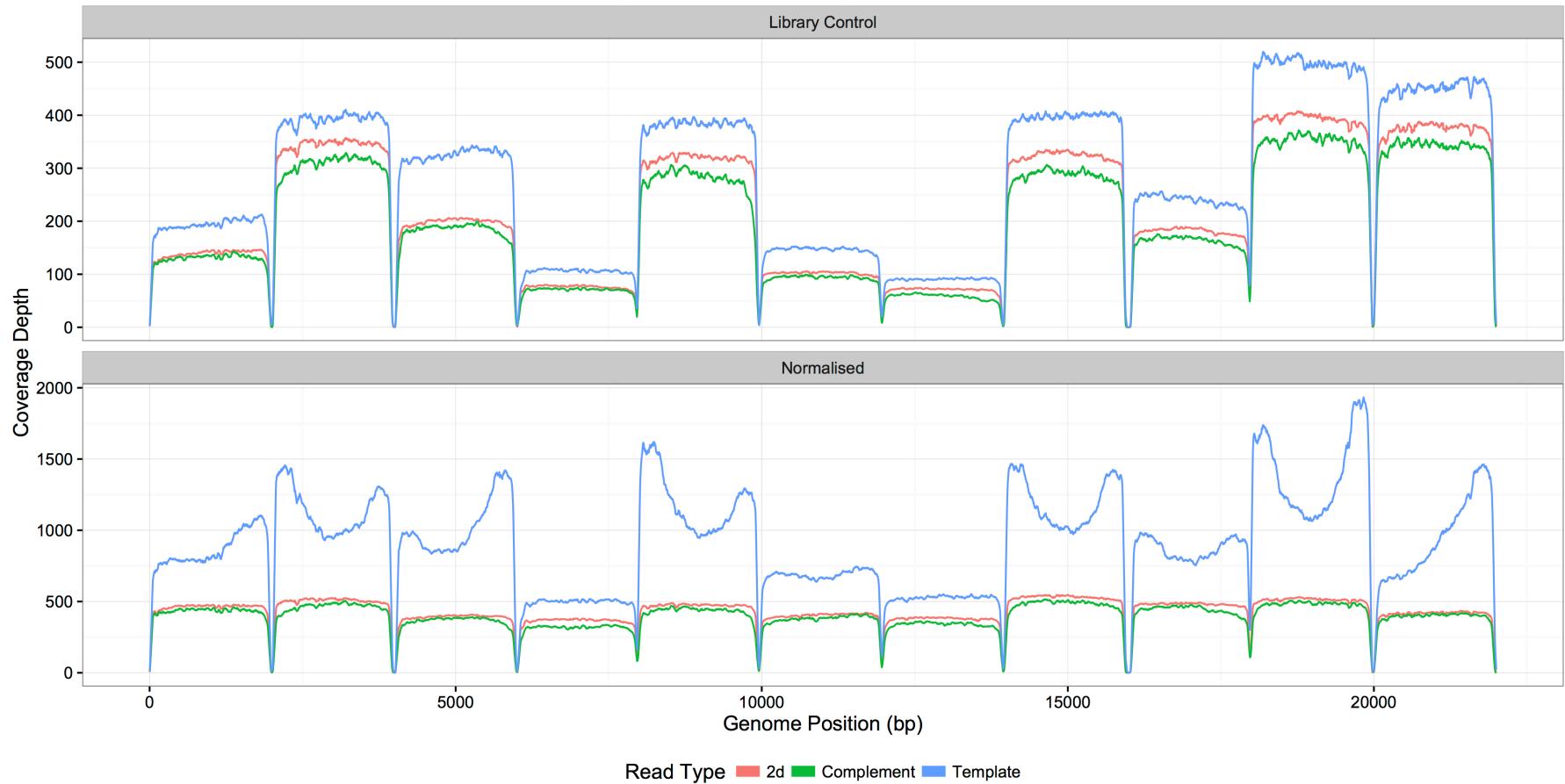
Amplicon Sequencing: Lambda ~70 b/s



Read Until Workflow – Counting Amplicons



A) Amplicon Balancing: Lambda ~70 b/s



#8

Challenges

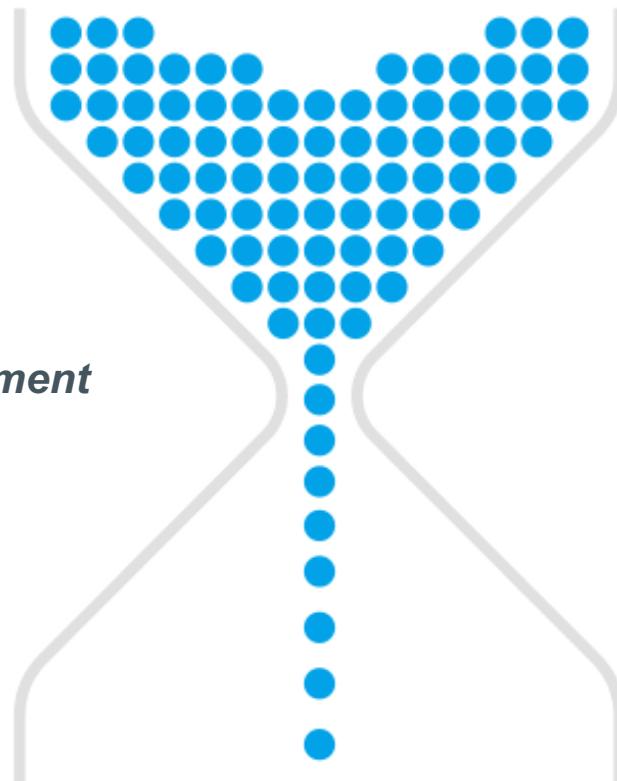
Challenges:

Yield

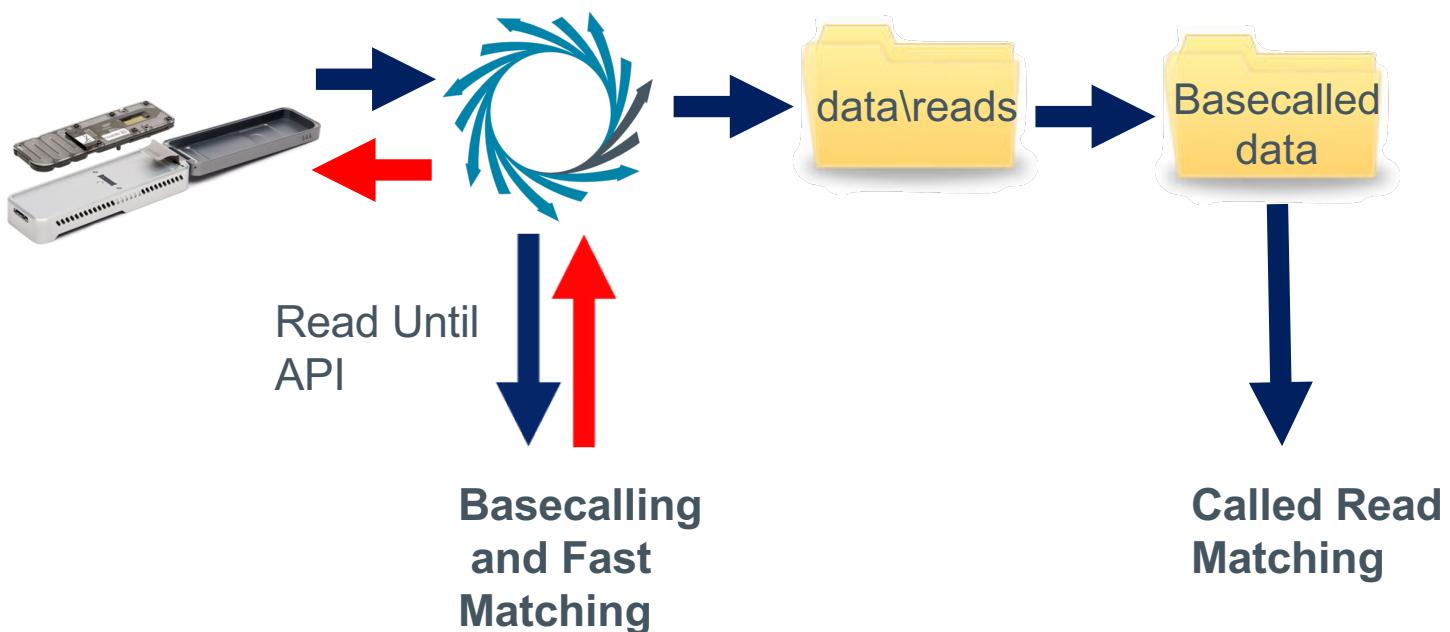
Rapid Asynchronous Read Alignment

Live Basecalling

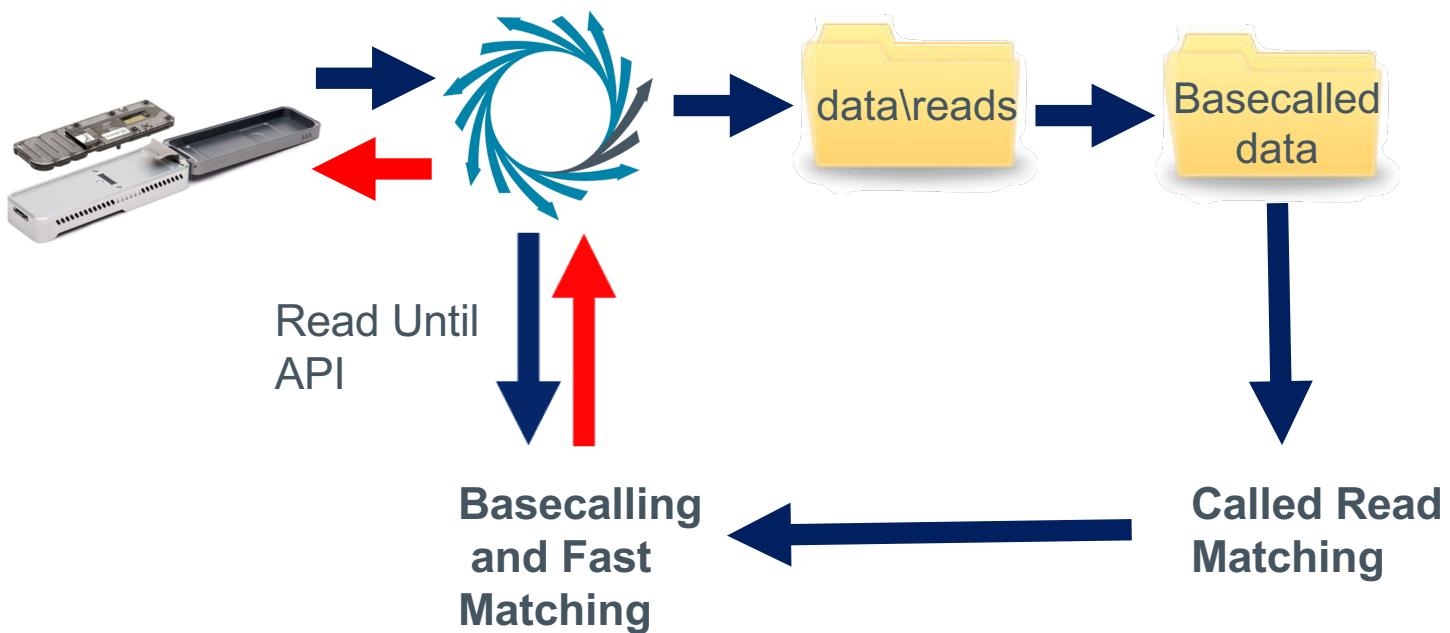
minoTour currently suited to small
genomes (<150 mB)



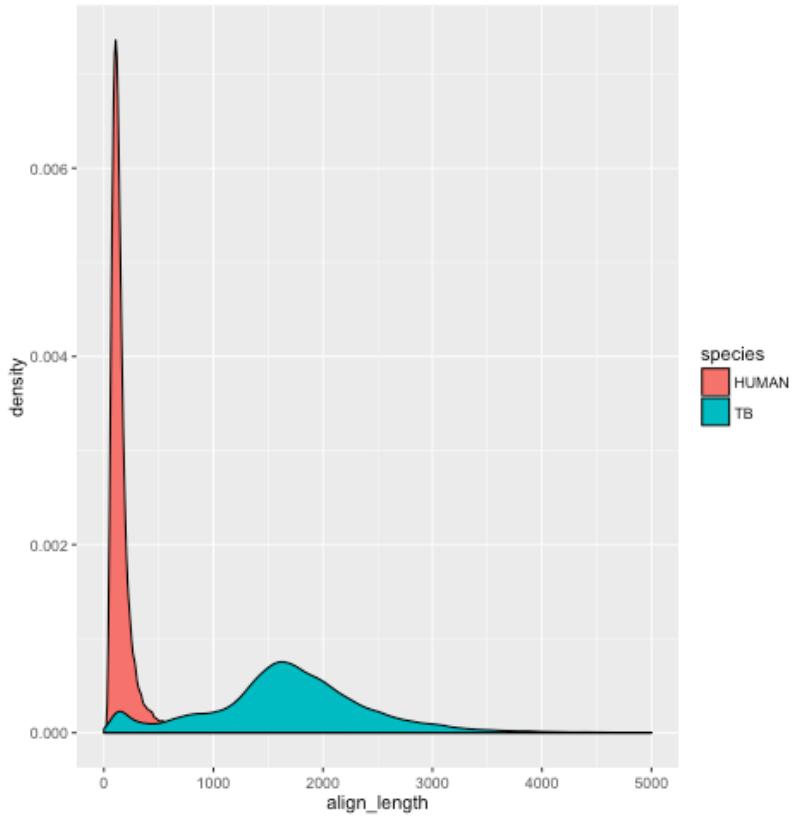
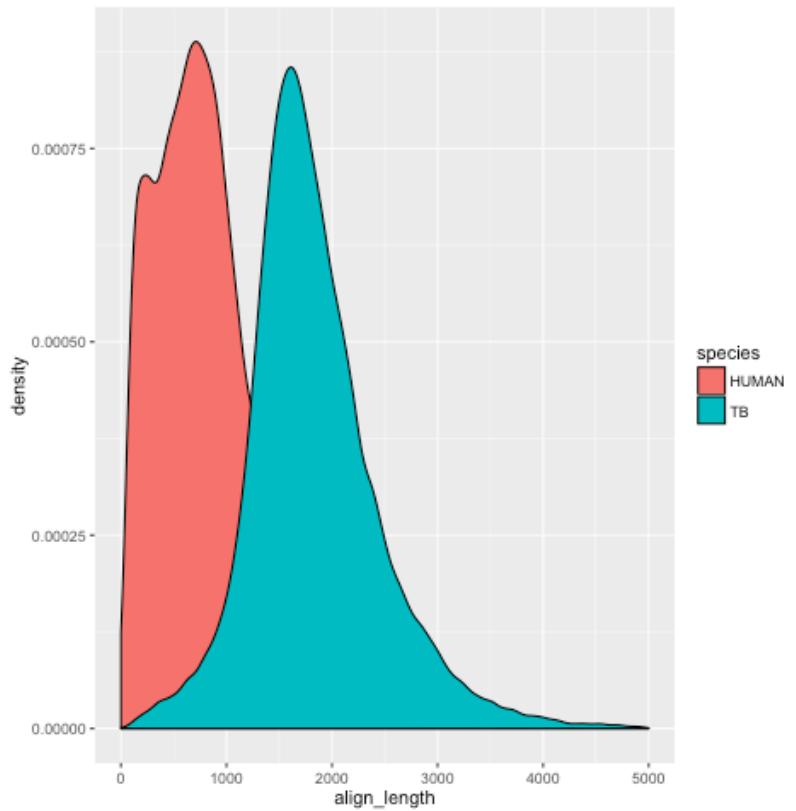
Read Until Workflows – Short and Long Loop



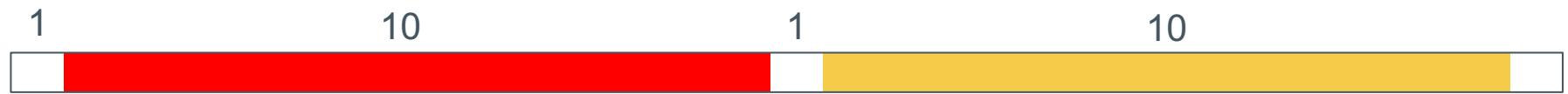
Read Until Workflows – Short and Long Loop



Alternative Workflows – rejecting host.



How does Read Until help?



Lower Bound ~ 400 bases
Allow 0.5 Seconds for match
Approximately 625 bases



Mean 10 kb library

Theoretical Max
Enrichment – 16 Fold



Pinned Tweet

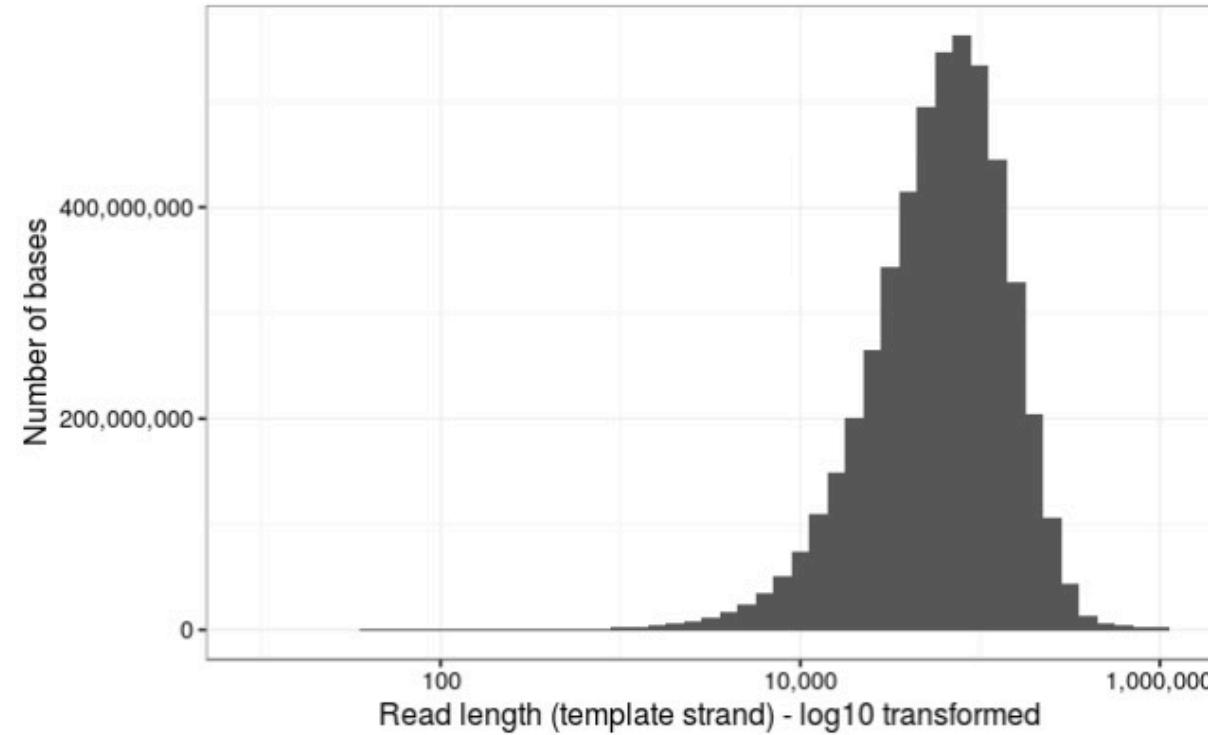


Nick Loman @pathogenomenick · Mar 10

New blogpost: "Thar she blows! Ultra long read method for nanopore sequencing"
lab.loman.net/2017/03/09/ult... full protocol included by @Scalene !

22 184 179

16x > 50x enrichment



Alignment stats

Wow! The longest 10 reads in this dataset are:

1113805 916705 790987 778219 771232 671130 646480 629747 614903 603565

!!!



Whale > Mass

Killer whale: 3,600 – 5,400 kg



Blue whale: 140,000 kg



Short-finned pilot whale: 1,000 – 3,0...



Humpback whale: 30,000 kg



Beluga whale: 1,400 kg



Sperm whale: 15,000 kg



North Pacific right whale: 50,000 – 8...



Narwhal: 940 kg



North Atlantic right whale: 40,000 – ...





Human Genome Data Release

Link

<http://github.com/nanopore-wgs-consortium/NA12878>

13M reads from 29 flowcells (9M up with ~4M reads remaining to post)

>20x sequence coverage of NA12878 Genome

Contributors

Mark Akeson ⁽¹⁾, Andrew D. Beggs ⁽²⁾, Thomas Nieto ⁽²⁾, Miten Jain ⁽¹⁾, Nicholas J. Loman ⁽³⁾, Matt Loose ⁽⁴⁾, Sunir Malla ⁽⁴⁾, Justin O'Grady ⁽⁵⁾, Hugh E. Olsen ⁽¹⁾, Josh Quick ⁽³⁾, Hollian Richardson ⁽⁵⁾, Jared T. Simpson ^(6,7), Terrance P. Snutch ^(8,9), Louise Tee ⁽²⁾, John R. Tyson ^(8,9)

1 University of California, Santa Cruz, Santa Cruz, CA, USA

2 University of Birmingham, Birmingham, B15 2TT

3 Institute of Microbiology and Infection, School of Biosciences, University of Birmingham, Birmingham, UK

4 DeepSeq, School of Life Sciences, University of Nottingham, Nottingham, UK

5 Norwich Medical School, University of East Anglia, Norwich, UK

6 Ontario Institute for Cancer Research, Toronto, Canada

7 Department of Computer Science, University of Toronto, Toronto, Canada

8 Michael Smith Laboratories, University of British Columbia, Vancouver, Canada

9 Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada



DASHBOARD DISCUSSION

Nicholas Loman

WARP- Nick Loman

Progress update • 4 hours ago

Human genome data release (R9.4 450 b/s 1D ligation chemistry)

Hi

We're pleased to announce that a consortium of five academic centres (Birmingham, Nottingham, Norwich, UBC and UCSC) have managed to generate >20X sequence coverage of the NA12878 human genome reference sequence.

We've posted ~9M reads from 29 flowcells up here, with around ~4M reads remaining to be posted:

<https://github.com/nanopore-wgs-consortium/NA12878>

We will be posting more reads, analysis, alignments and signal-level data over the coming days.

It turns out managing this number of files is quite difficult, so there is a new toolkit to help. Called poredb; this tool tracks the files on the filesystem and keeps a SQLite database of basecalls and metadata for easy extraction. It's very much alpha at the moment but if you are interested in looking at it (or contributing) please see:

<https://github.com/nickloman/poredb>

Regards

Nick

Acknowledgements

minoTour.nottingham.ac.uk
github.com/minoTour/minoTour

University of Nottingham: Martin Blythe, Sunir Malla, Mike Stout, Teri Evans

University of Birmingham: Nick Loman, Josh Quick

Nanopore WGS Consortium: github.com/nanopore-wgs-consortium/NA12878

EBI: Ewan Birney, Guy Cochrane

Oxford: Zam Iqbal

@mattloose matt.loose@nottingham.ac.uk

