Linked- and Long-Read Sequencing

Sam Kovaka

(Most slides by Michael Schatz)

Feb 11, 2019

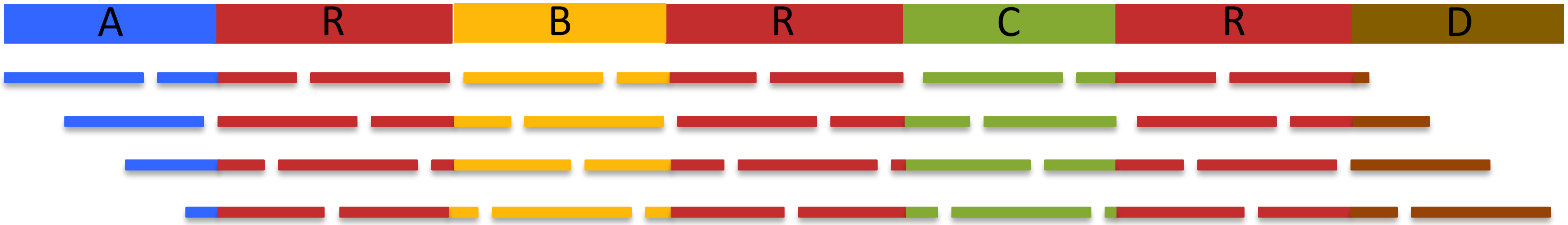
Lecture 5: Applied Comparative Genomics

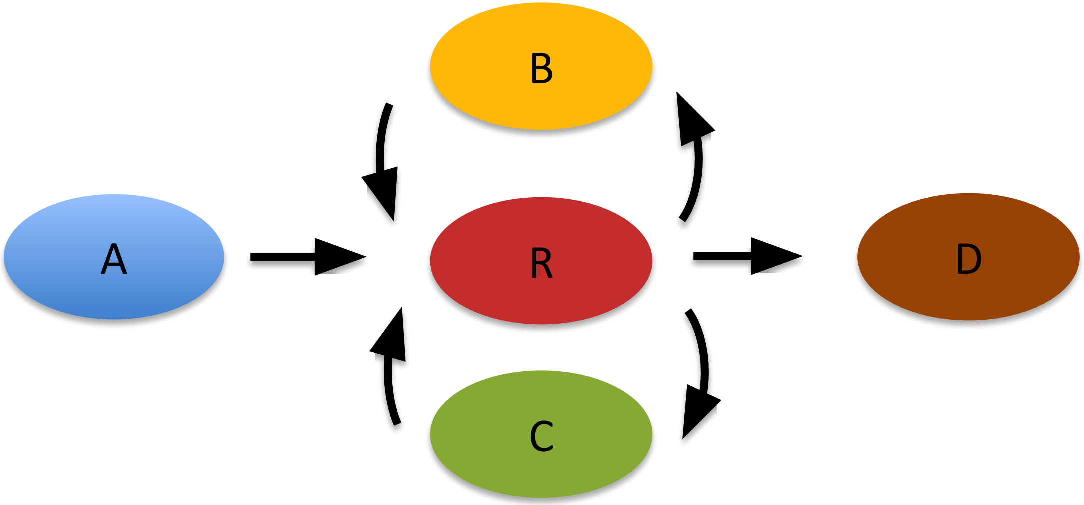
Assignment 1 Feedback

I will try to finish grading today, but here is some general feedback from what I’ve seen:

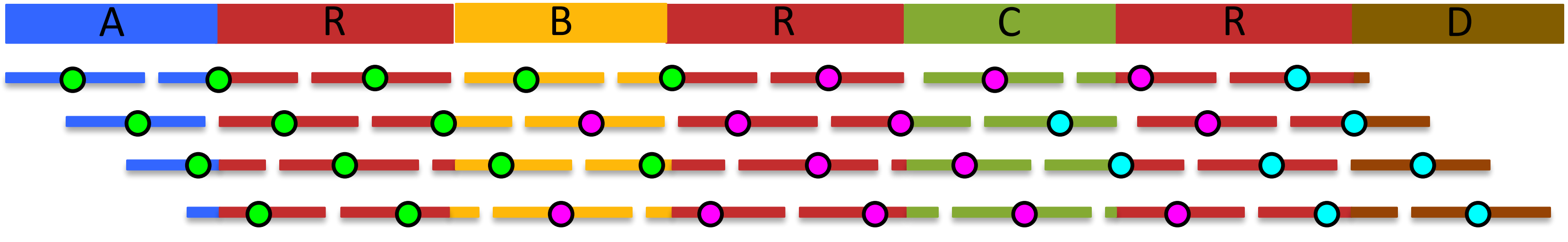
- Provide EXACT commands/code for each question
  - Include it in the corresponding answer, not at the end
- Command-line-based solutions (and brevity in general) is encouraged
  - Many people submitted dozens of lines of code for what could be done with a few piped commands
- Mark the pages for each answer on Gradescope

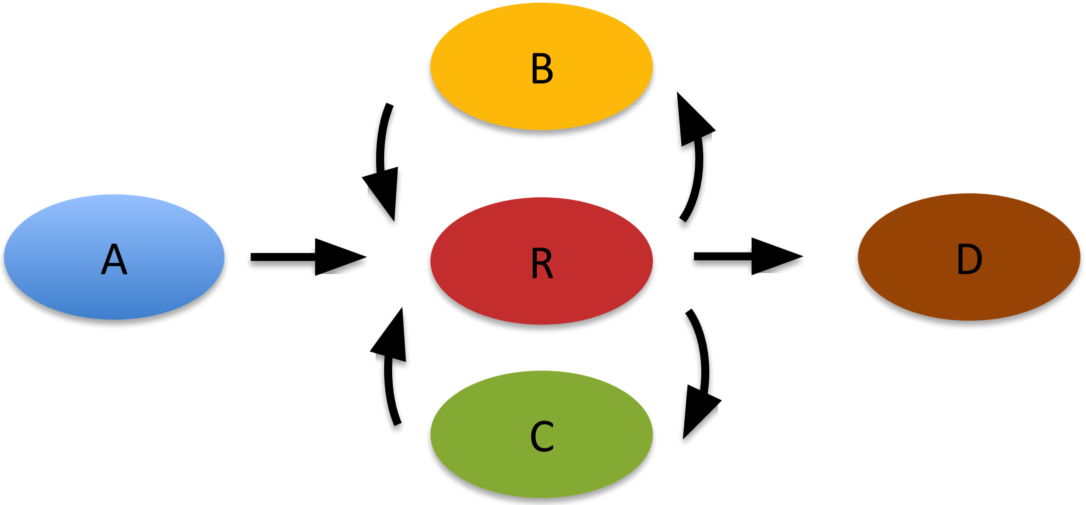
Assembly Complexity



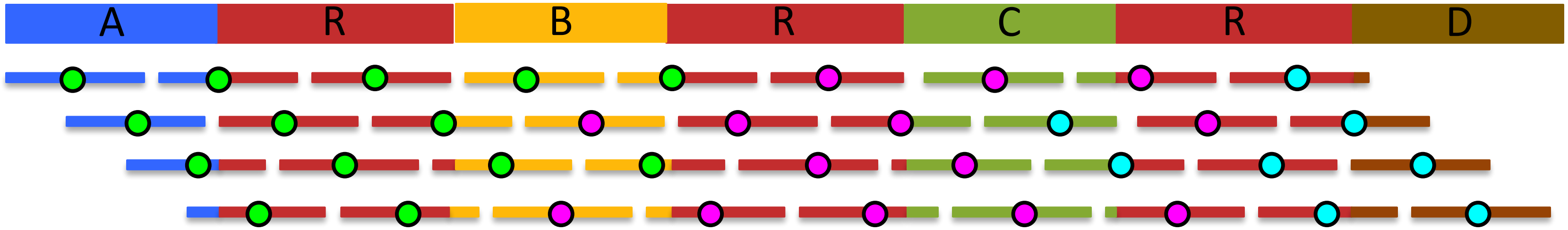


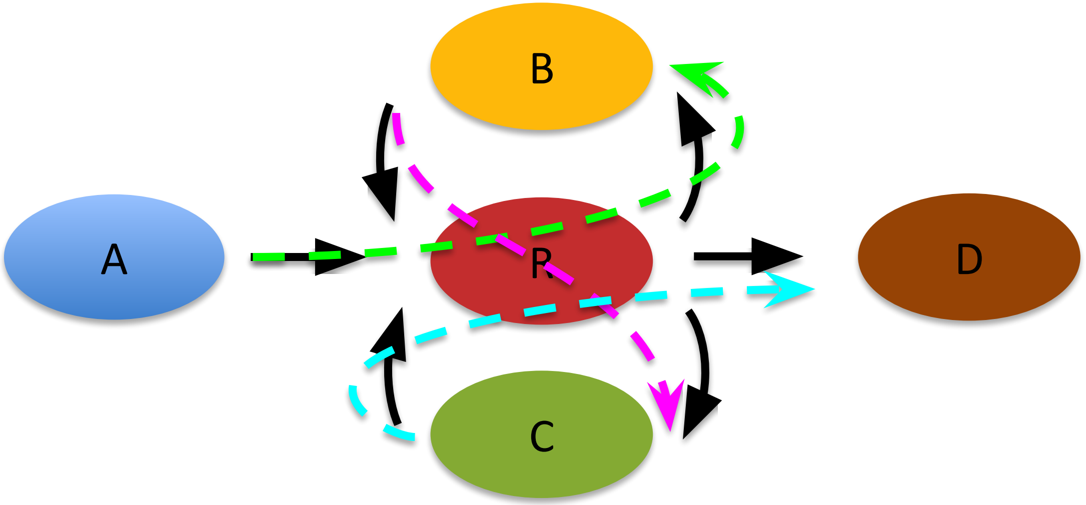
Assembly Complexity



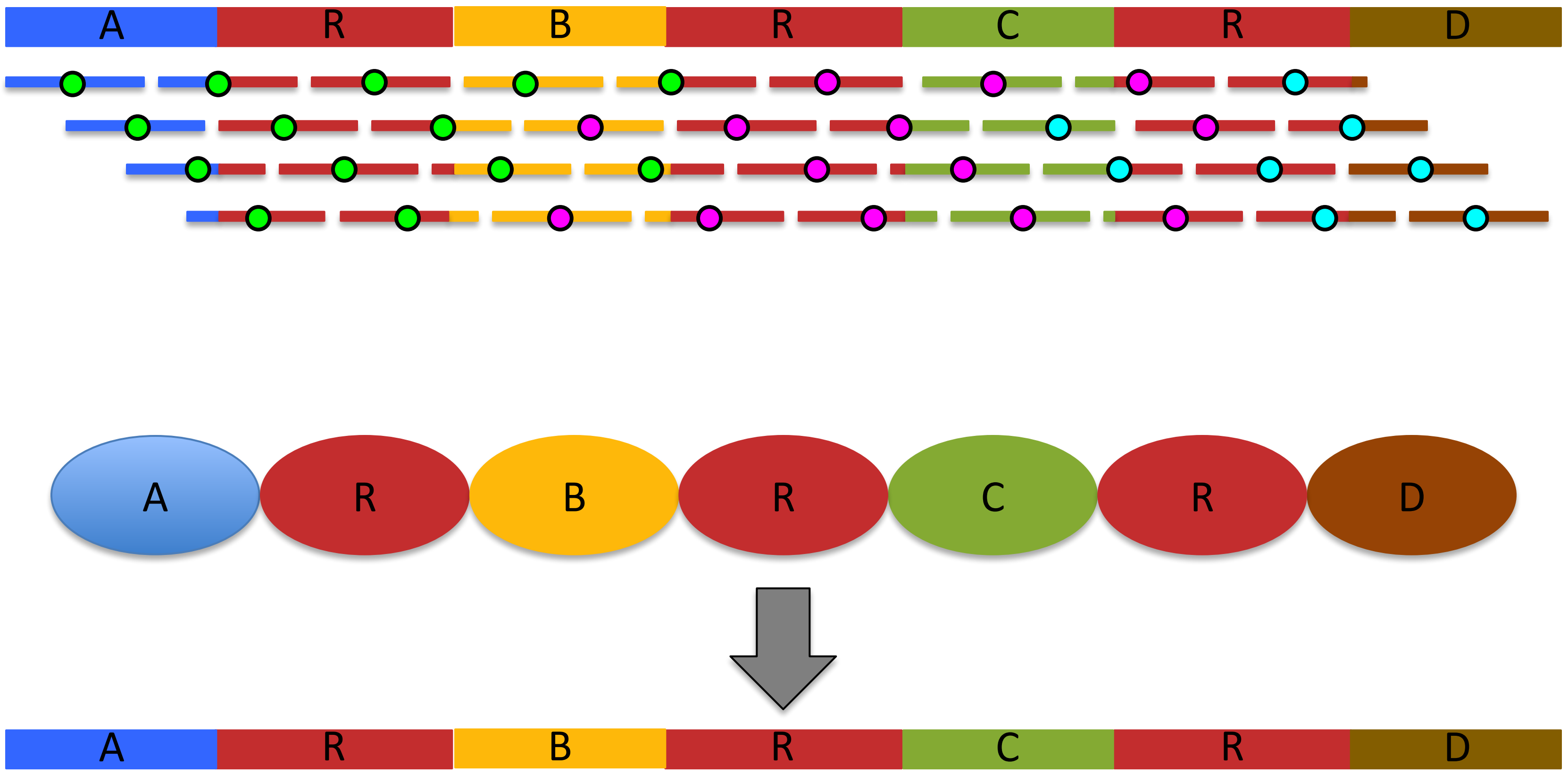


Assembly Complexity





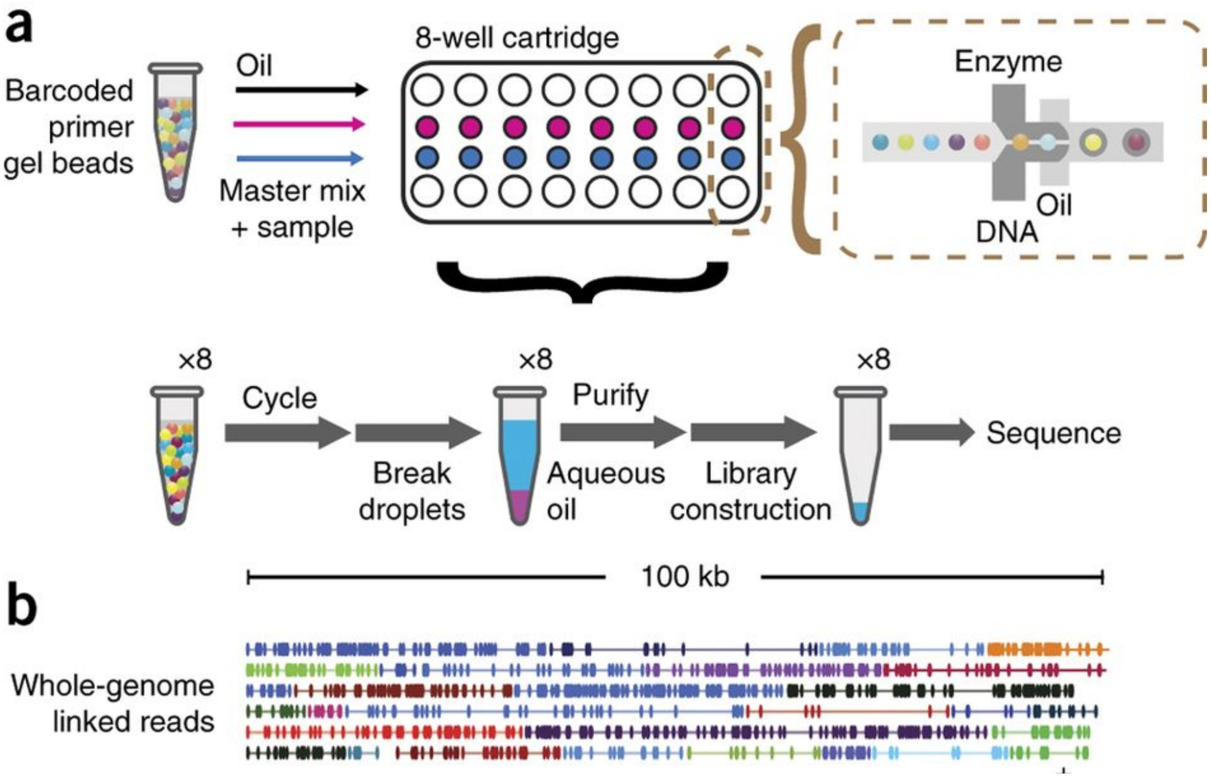
Assembly Complexity



The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

10X Genomics Linked Reads

10X Genomics Linked Reads



(Zheng *et al*, 2016)

1. High molecular weight DNA is diluted and isolated within oil emulsion droplets

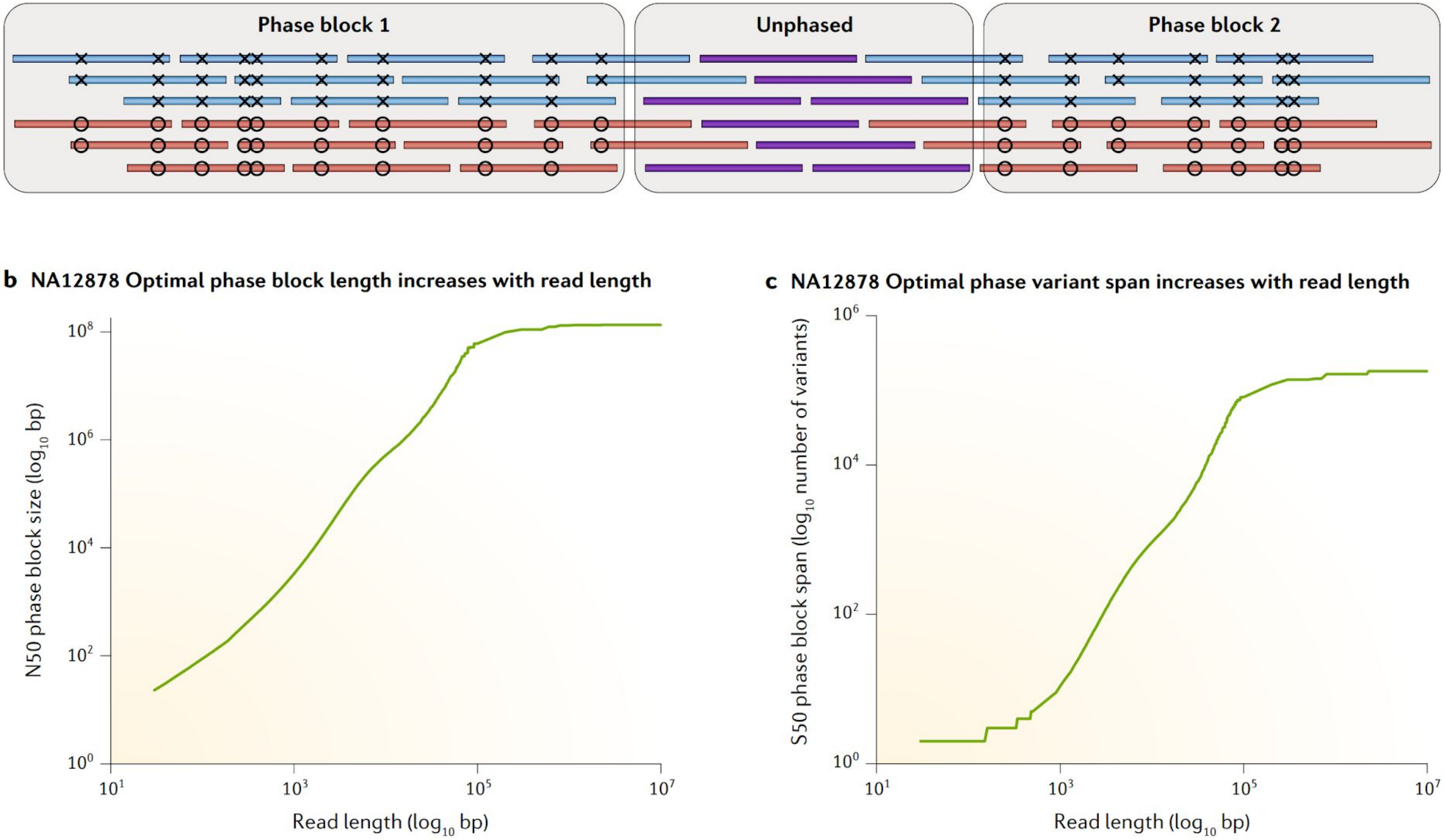
2. Within each droplet, short fragments are randomly amplified and tagged with barcode sequences for standard Illumina sequencing

3. Reads sharing the same barcode can then be localized to the same original template molecule

4. The resulting “linked reads” can be used for phasing variants or identifying SVs

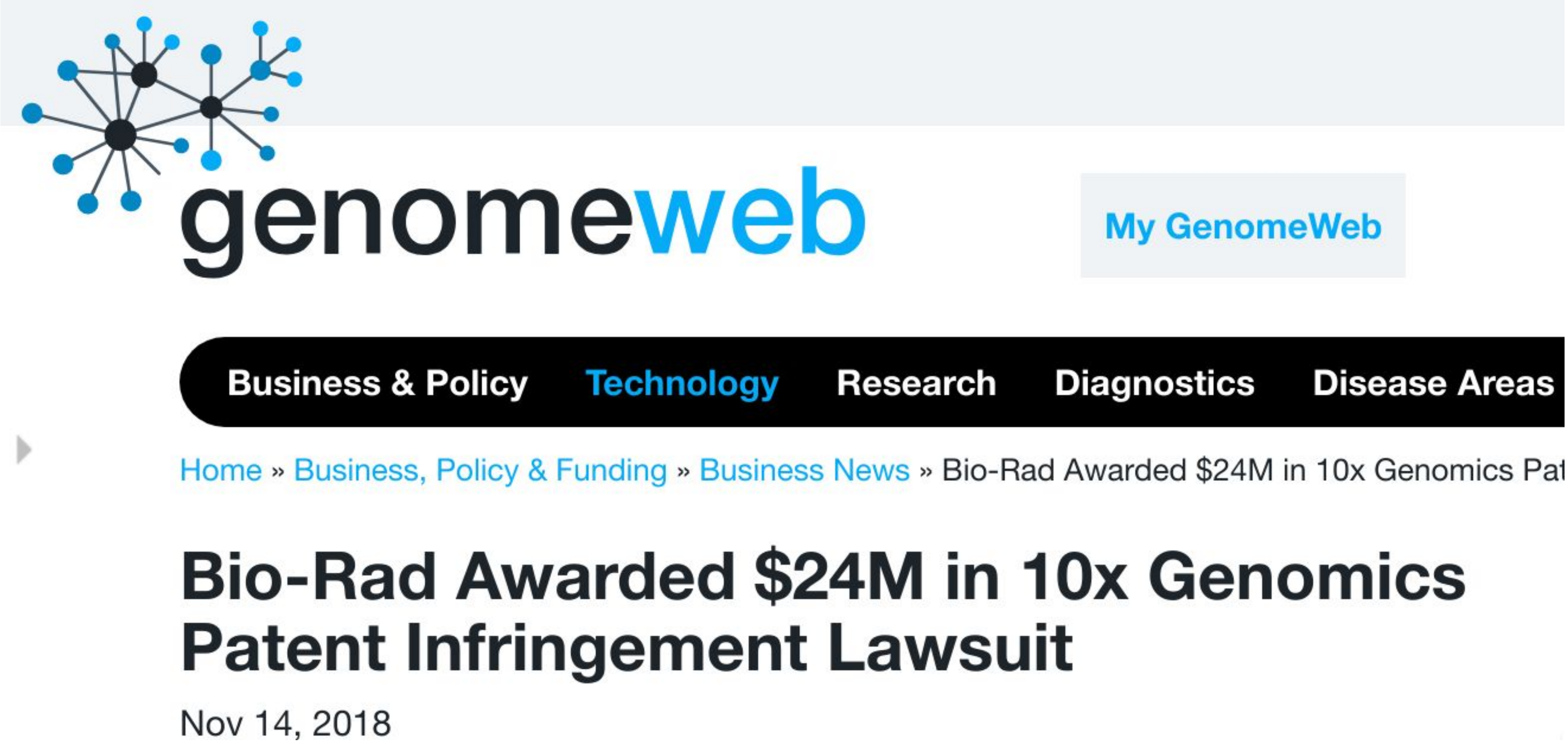
https://www.youtube.com/watch?v=nk2kXM59LRM

Haplotype Phasing



Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

Uncertain Future for 10X



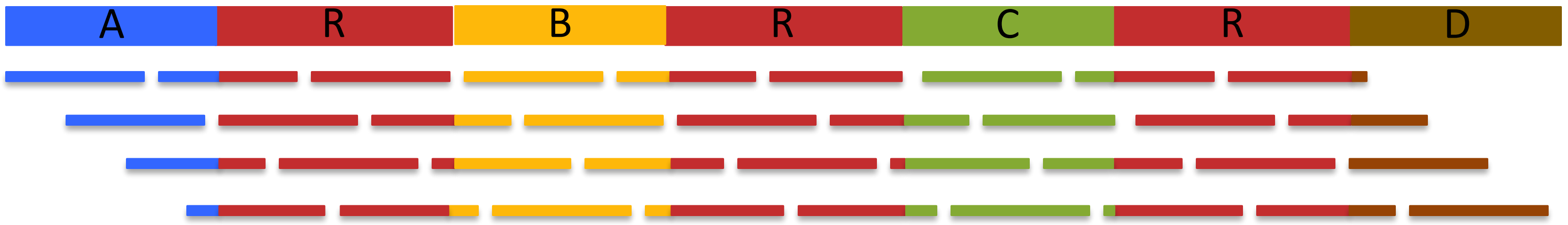
Home » Business, Policy & Funding » Business News » Bio-Rad Awarded $24M in 10x Genomics Pat

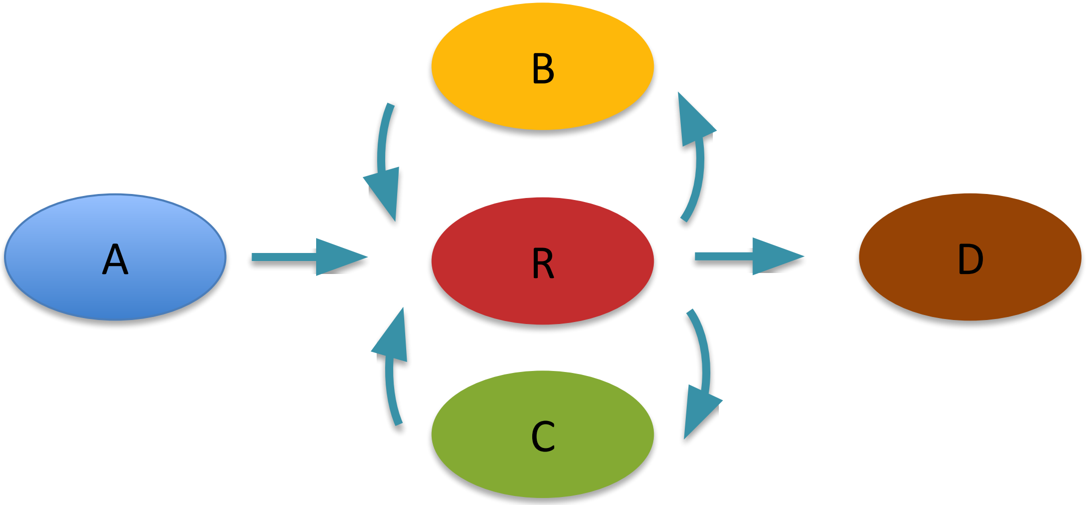
Bio-Rad Awarded $24M in 10x Genomics Patent Infringement Lawsuit

Nov 14, 2018

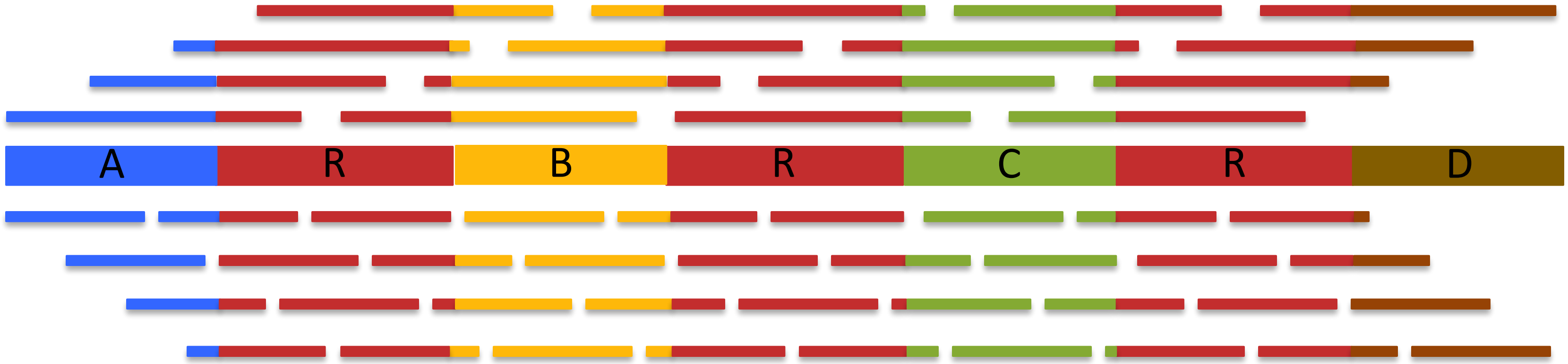
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

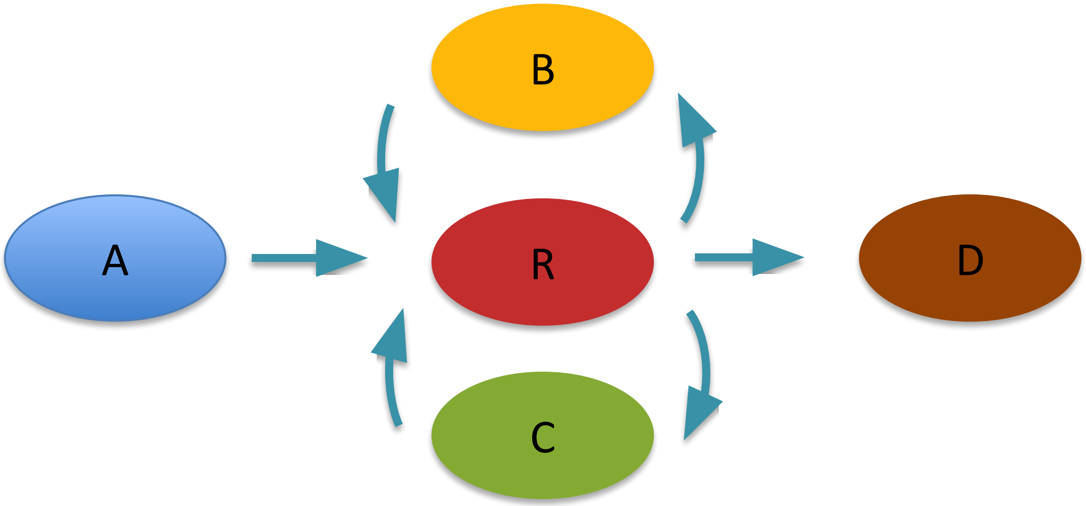
Assembly Complexity



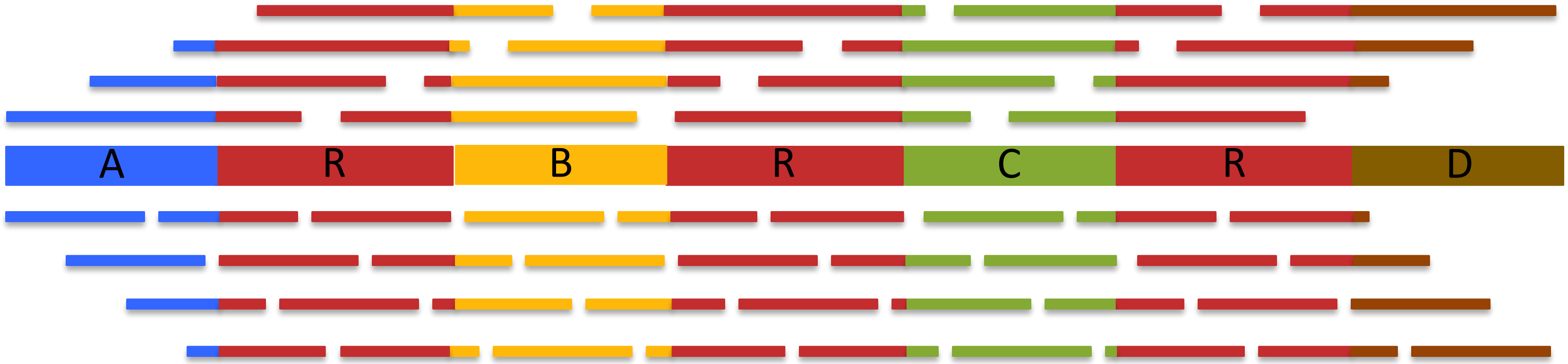


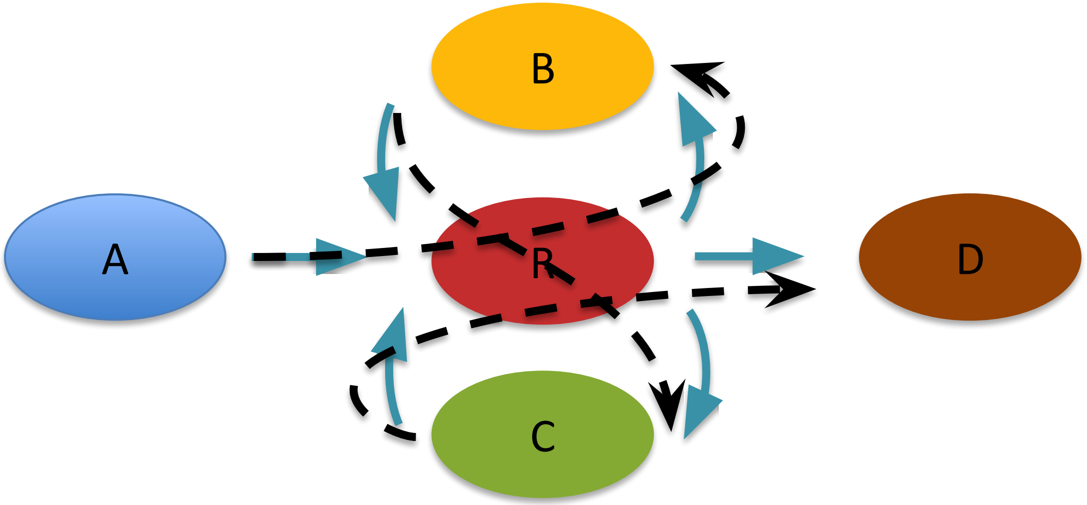
Assembly Complexity



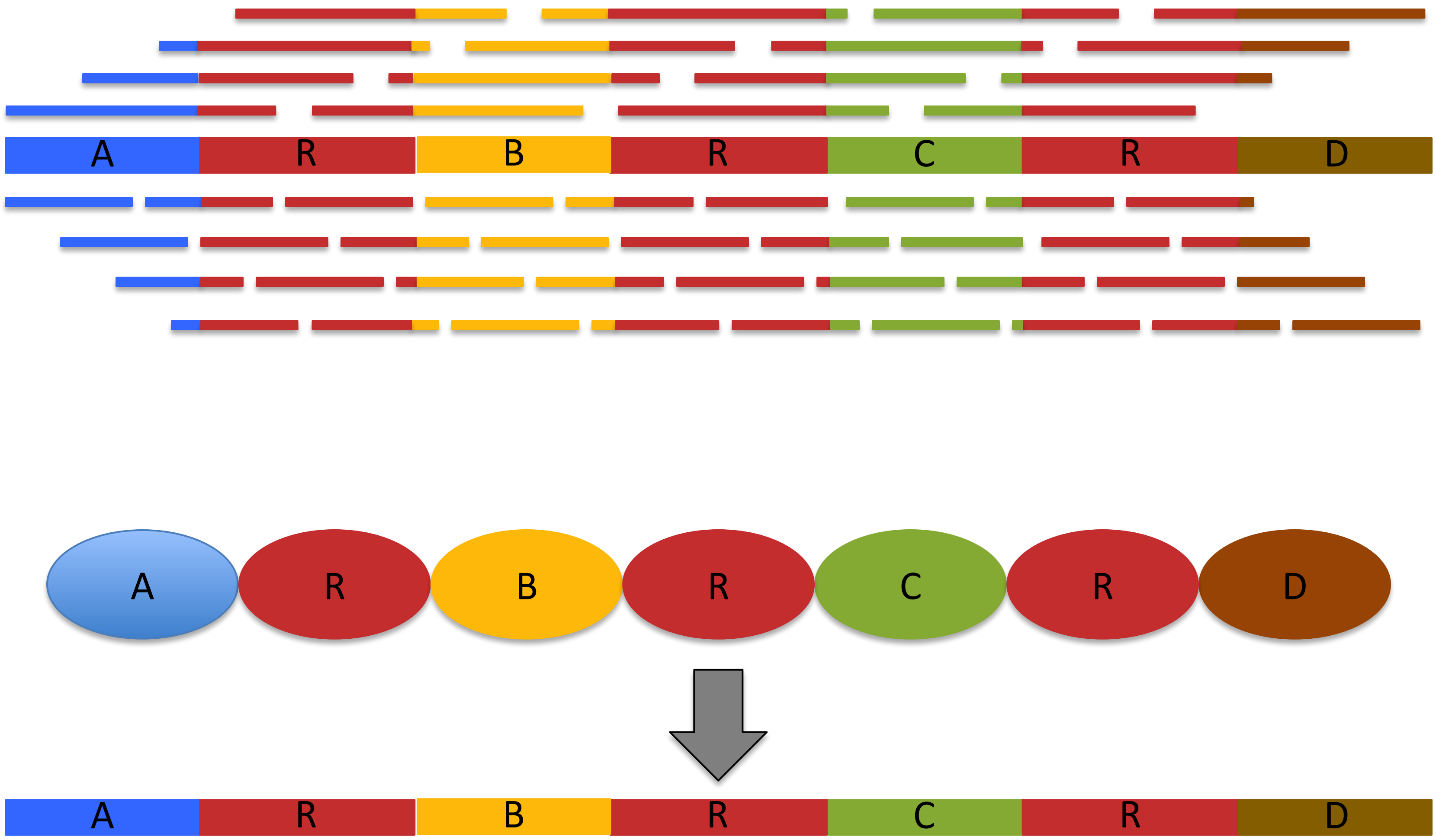


Assembly Complexity





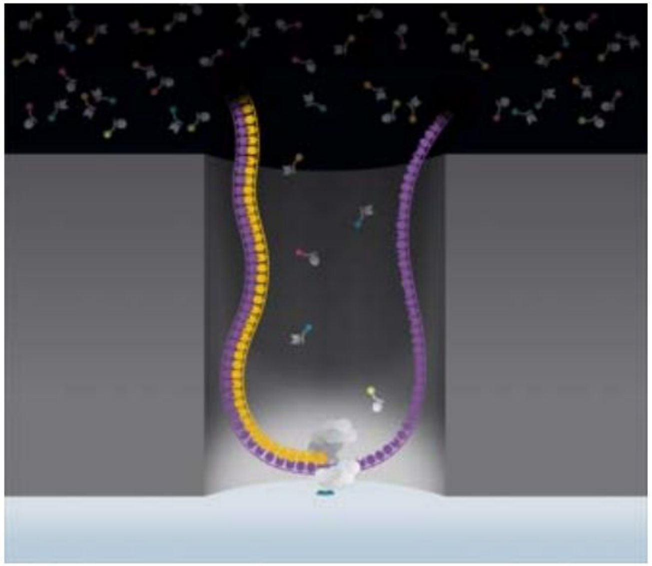
Assembly Complexity

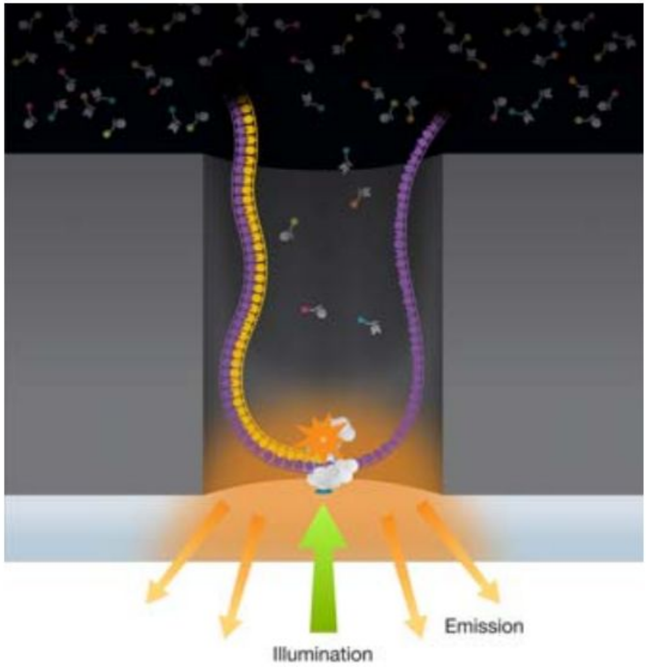


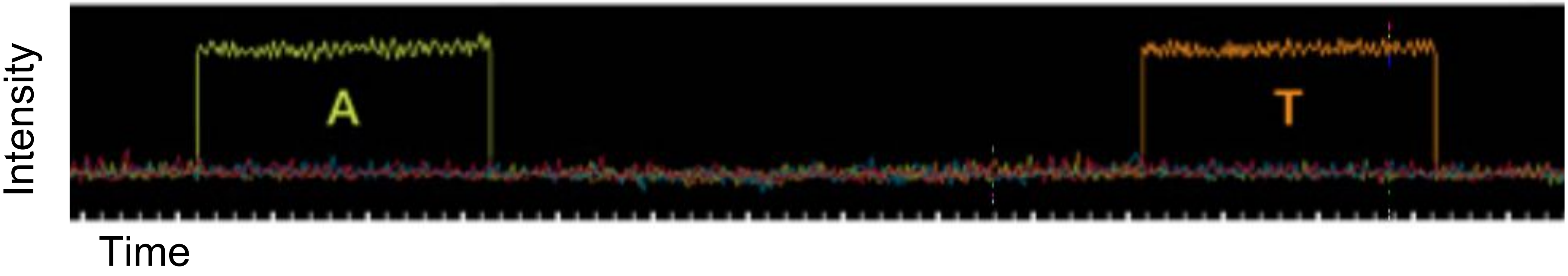
The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

PacBio: SMRT Sequencing

Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



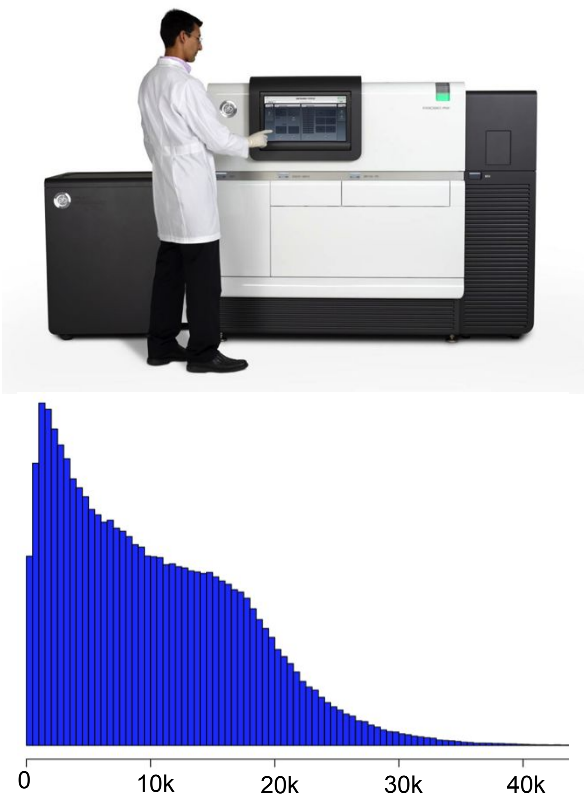




http://www.youtube.com/watch?v=v8p4ph2MAvI

SMRT Sequencing Data

PacBio RS II

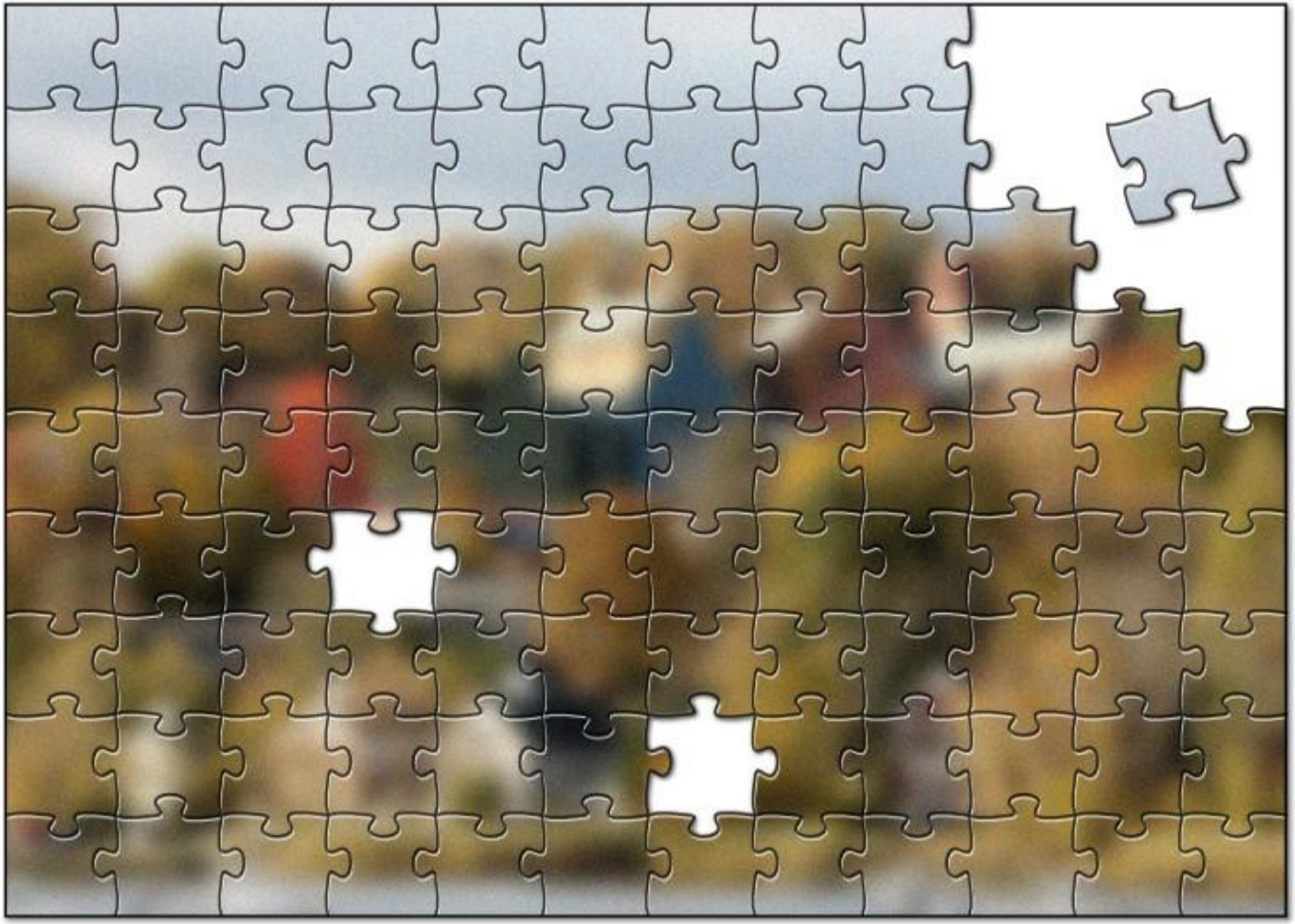


CSHL/PacBio

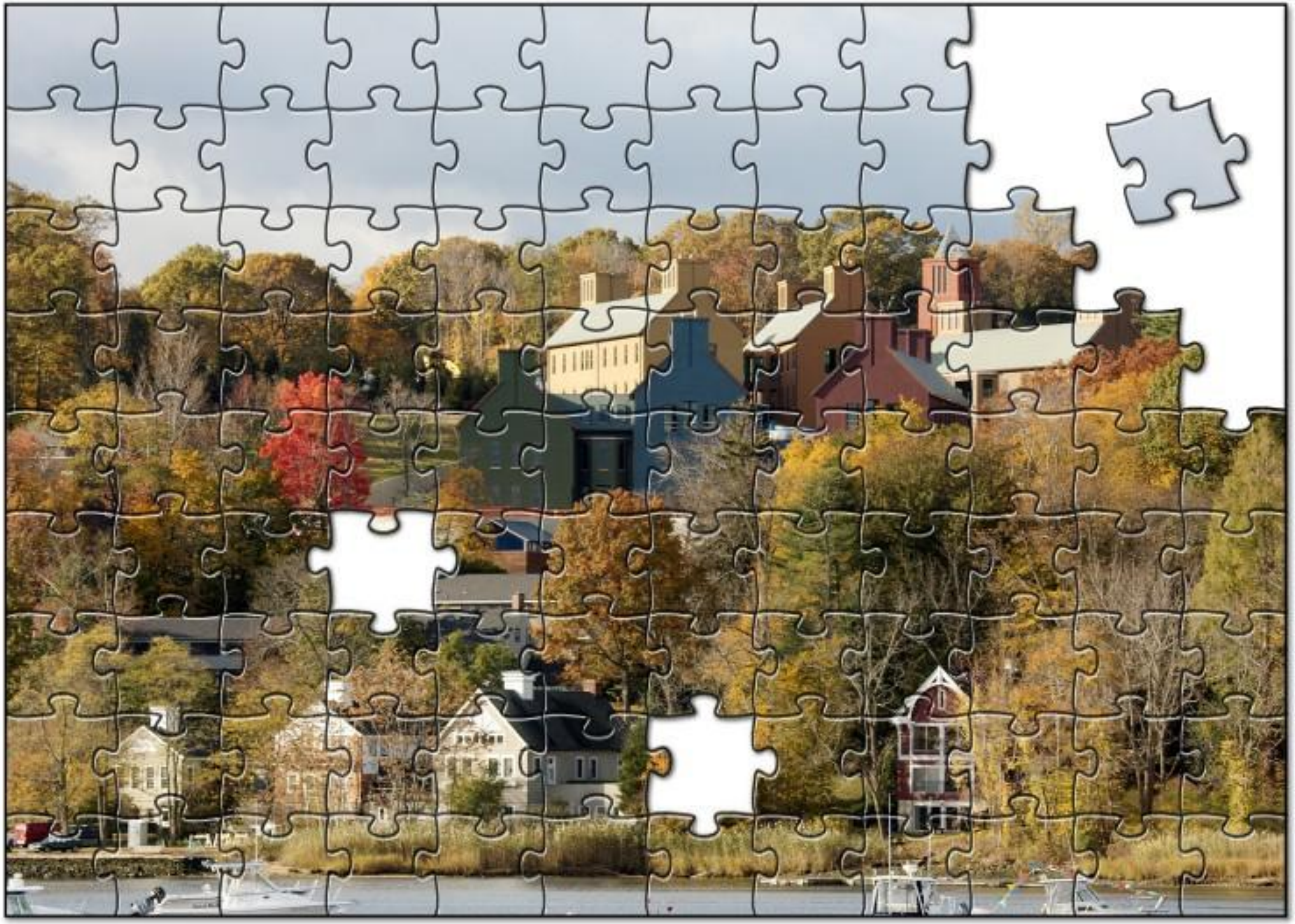
```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| |||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4%

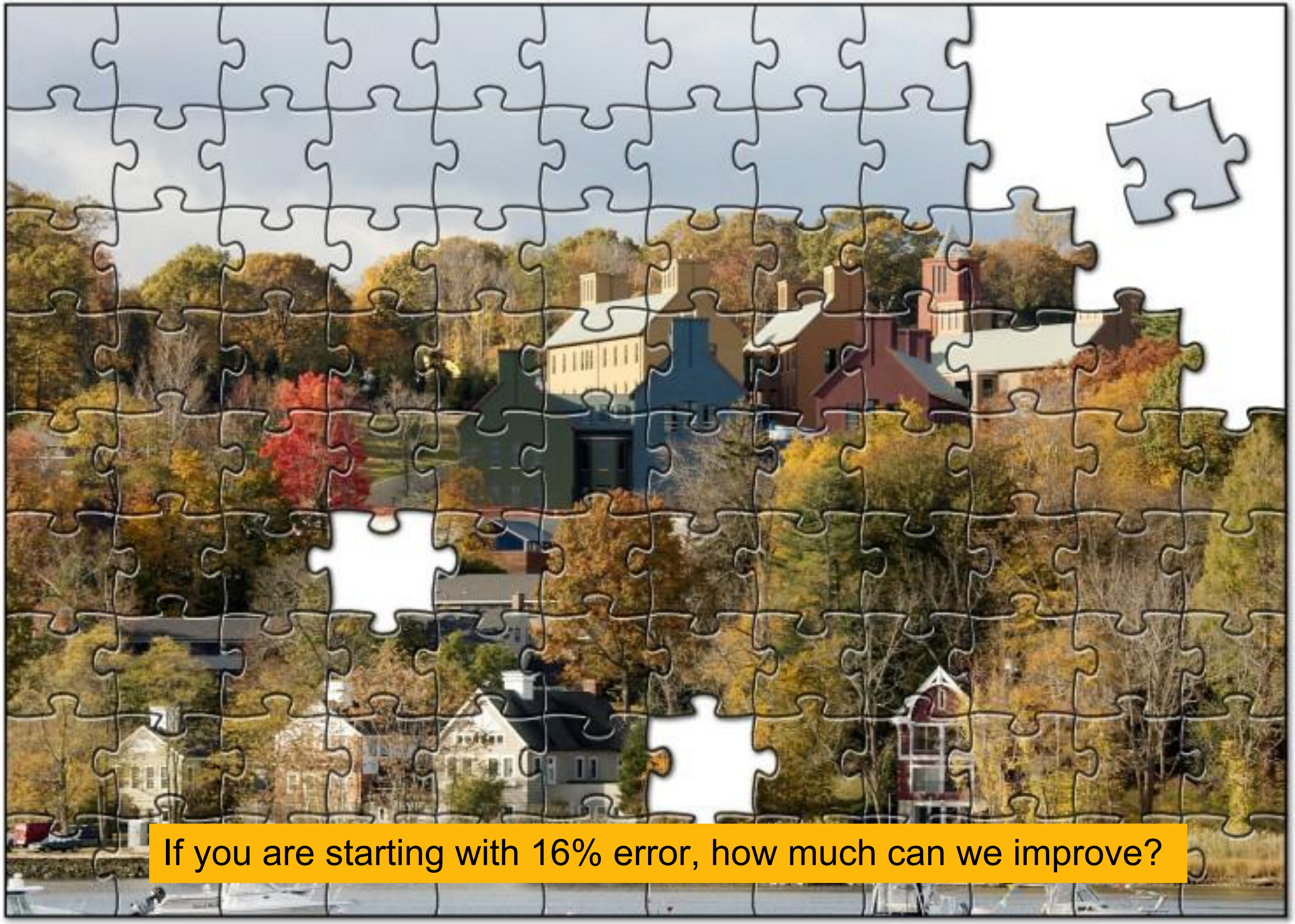
Single Molecule Sequences



“Corrective Lens” for Sequencing

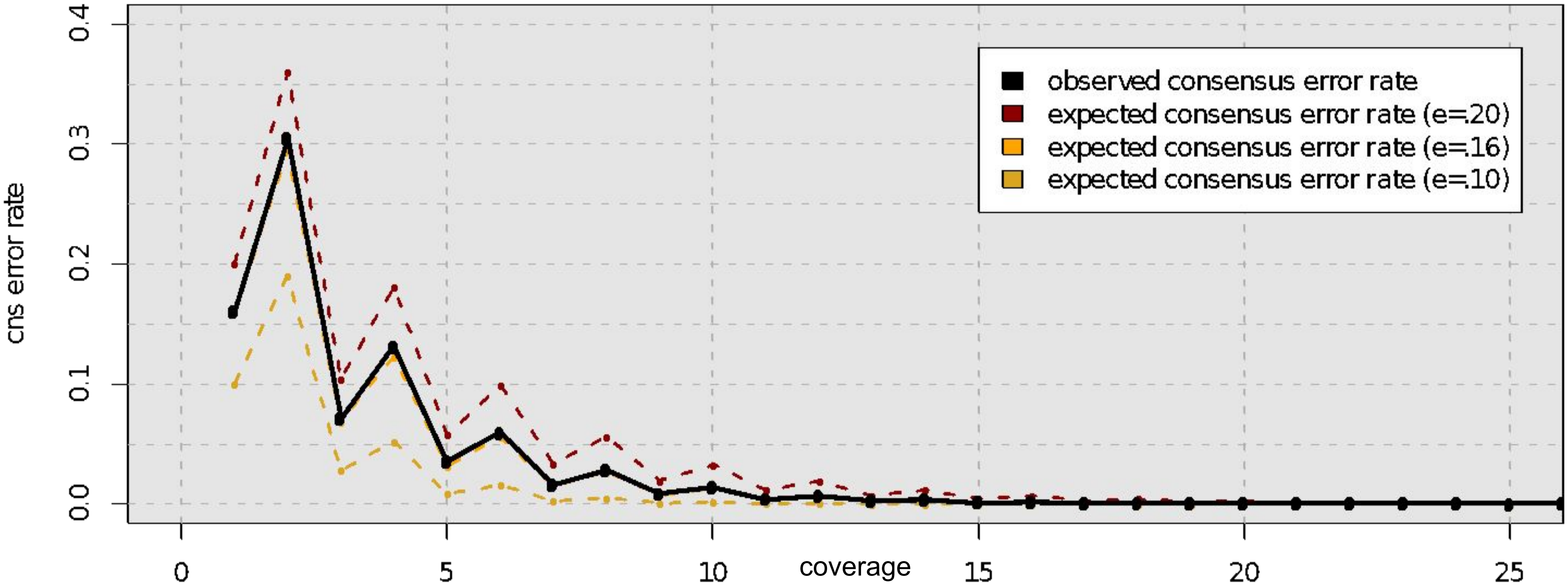


“Corrective Lens” for Sequencing



If you are starting with 16% error, how much can we improve?

Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

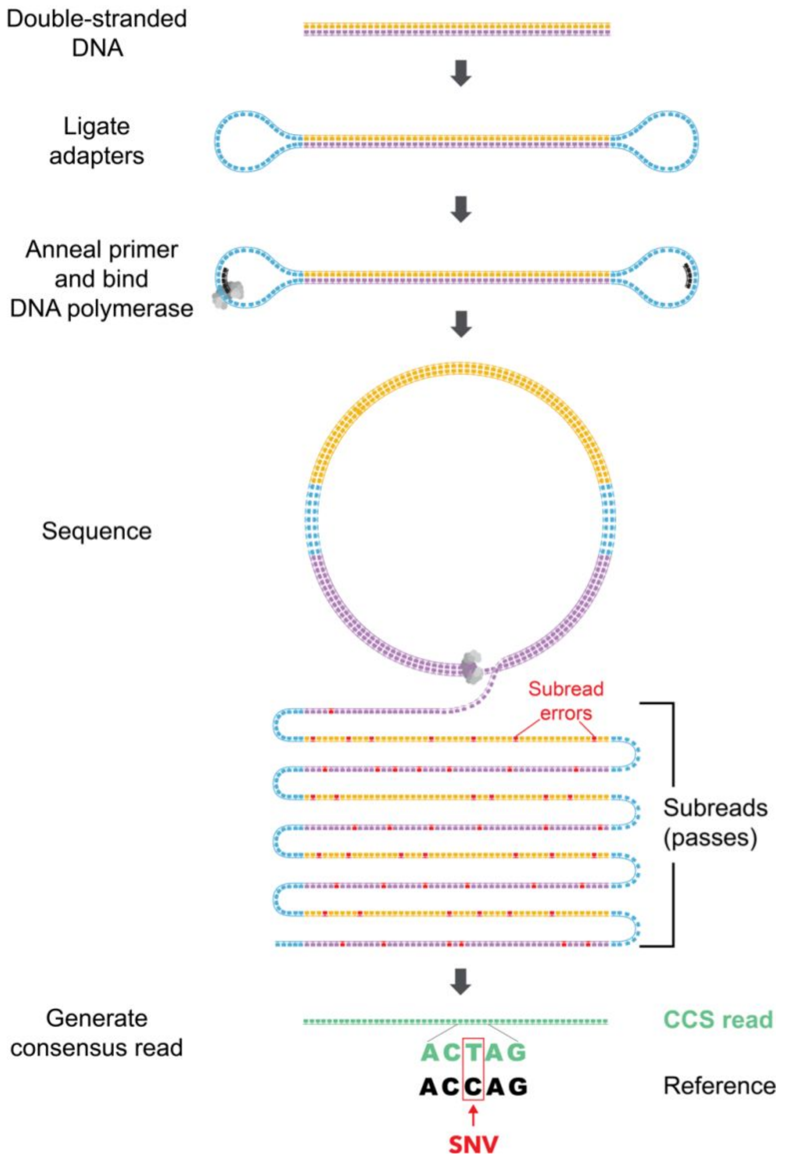
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Circular Consensus Reads

High-quality reads produced by sequencing the same molecule multiple times

Higher accuracy for low-coverage sequences like somatic variants or lowly expressed transcripts in RNA-seq, more interpretable alignments, faster assembly

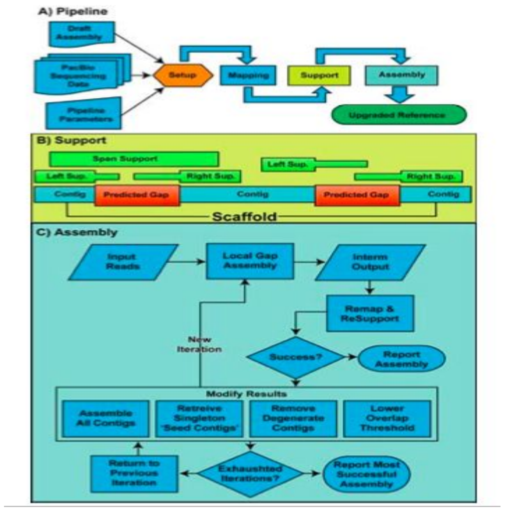
Limits read length, very expensive, calling consensus is currently slow



https://www.biorxiv.org/content/10.1101/519025v2

PacBio Assembly Algorithms

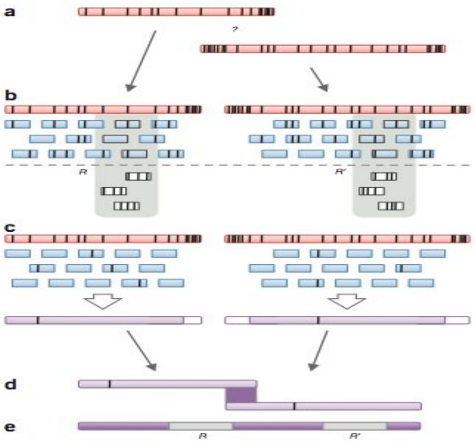
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

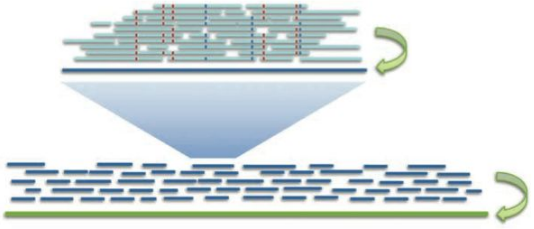
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

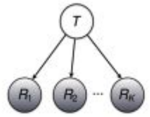
Koren, Schatz, *et al* (2012)
Nature Biotechnology.
30:693–700

Canu/FALCON & Quiver/Arrow



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$



Quiver Performance Results
Comparison to Reference Genome
(*M. ruber* ; 3.1 MB ; SMRT® Cells)

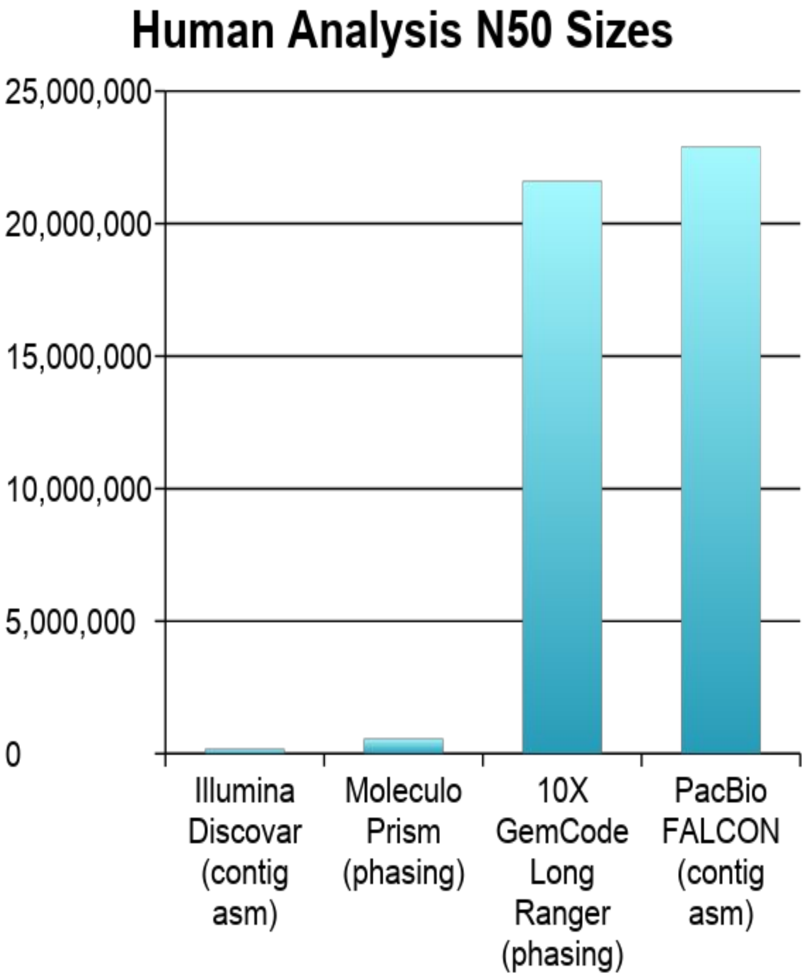
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

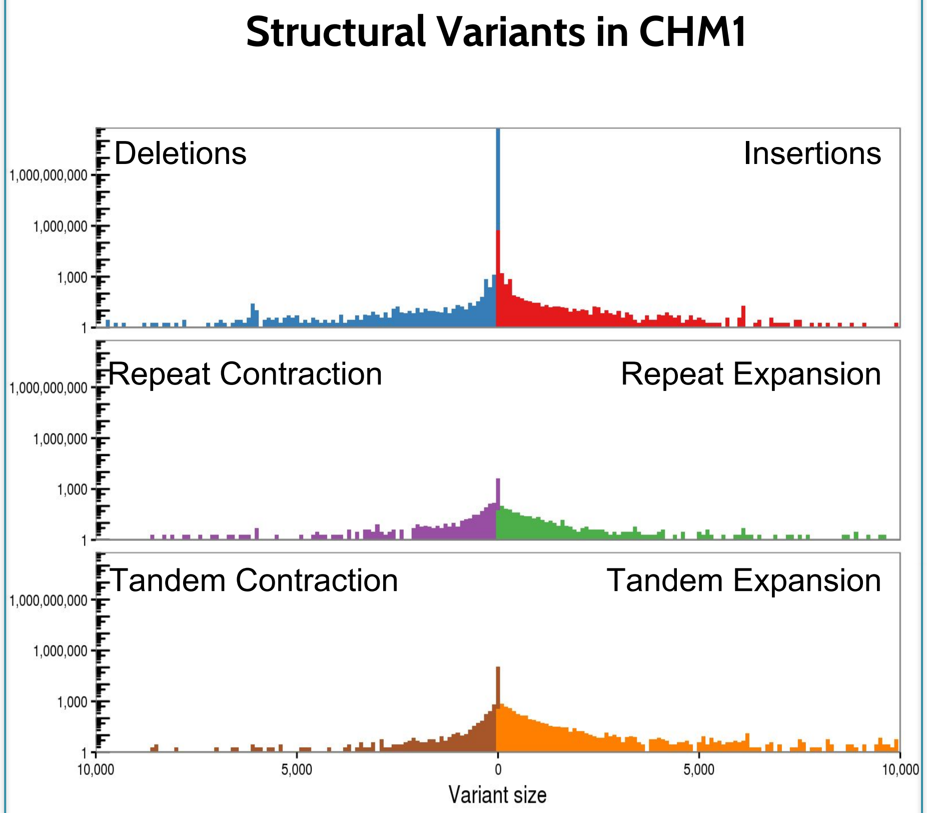
Chin *et al* (2016)
Nature Methods. 13:1050–1054

< 5x PacBio Coverage > 50x

Recent Long Read Assemblies



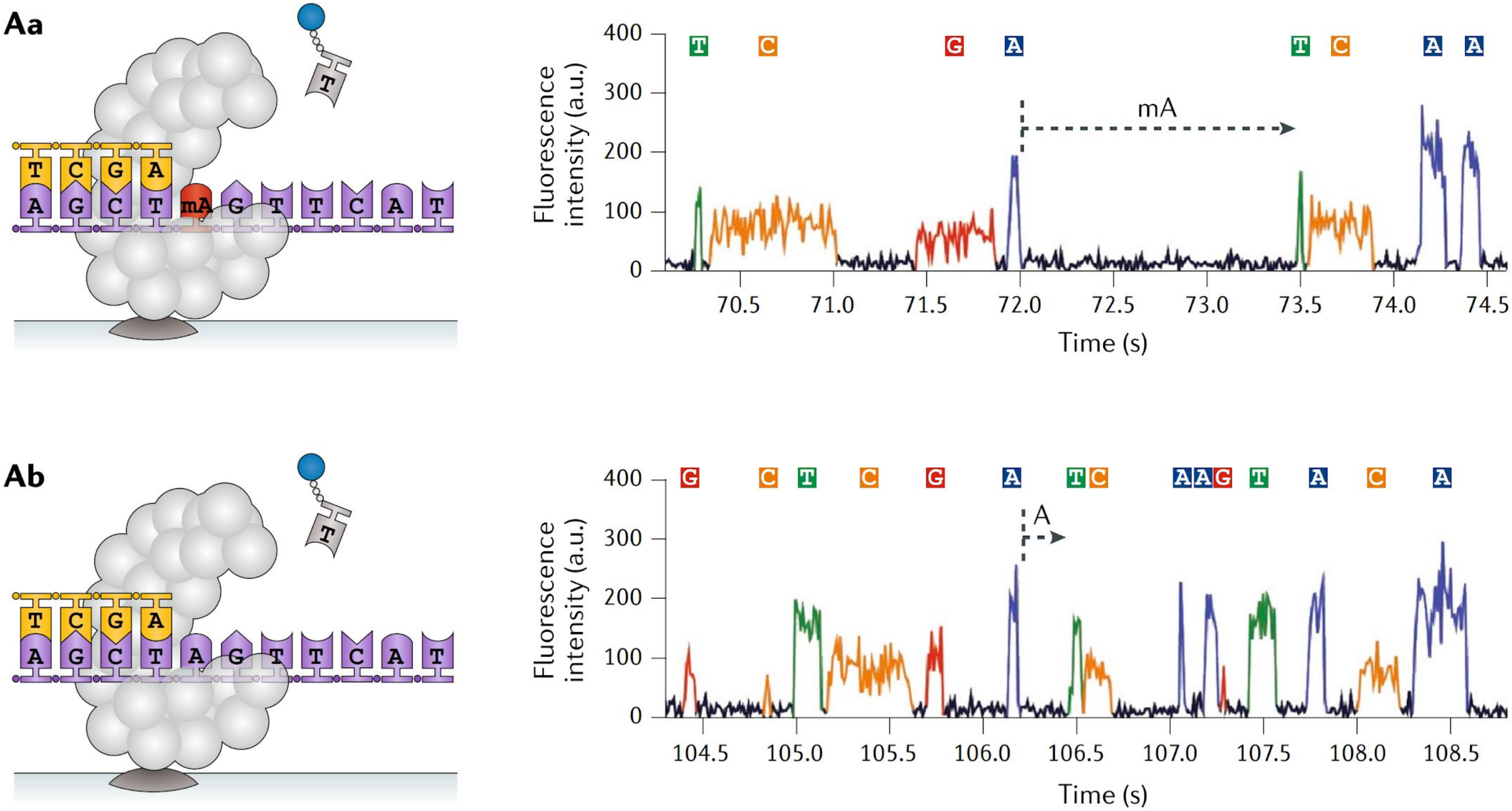
Third-generation sequencing and the future of genomics
Lee et al (2016) *bioRxiv*
doi: http://dx.doi.org/10.1101/048603



Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad & Schatz (2016) *Bioinformatics*.
doi: 10.1093/bioinformatics/btw369

Methylation Detection

- **Methylation -** an epigenetic modification that can have a variety of effects, such as gene repression
- Can detect methylation from raw PacBio signal



PacBio Roadmap

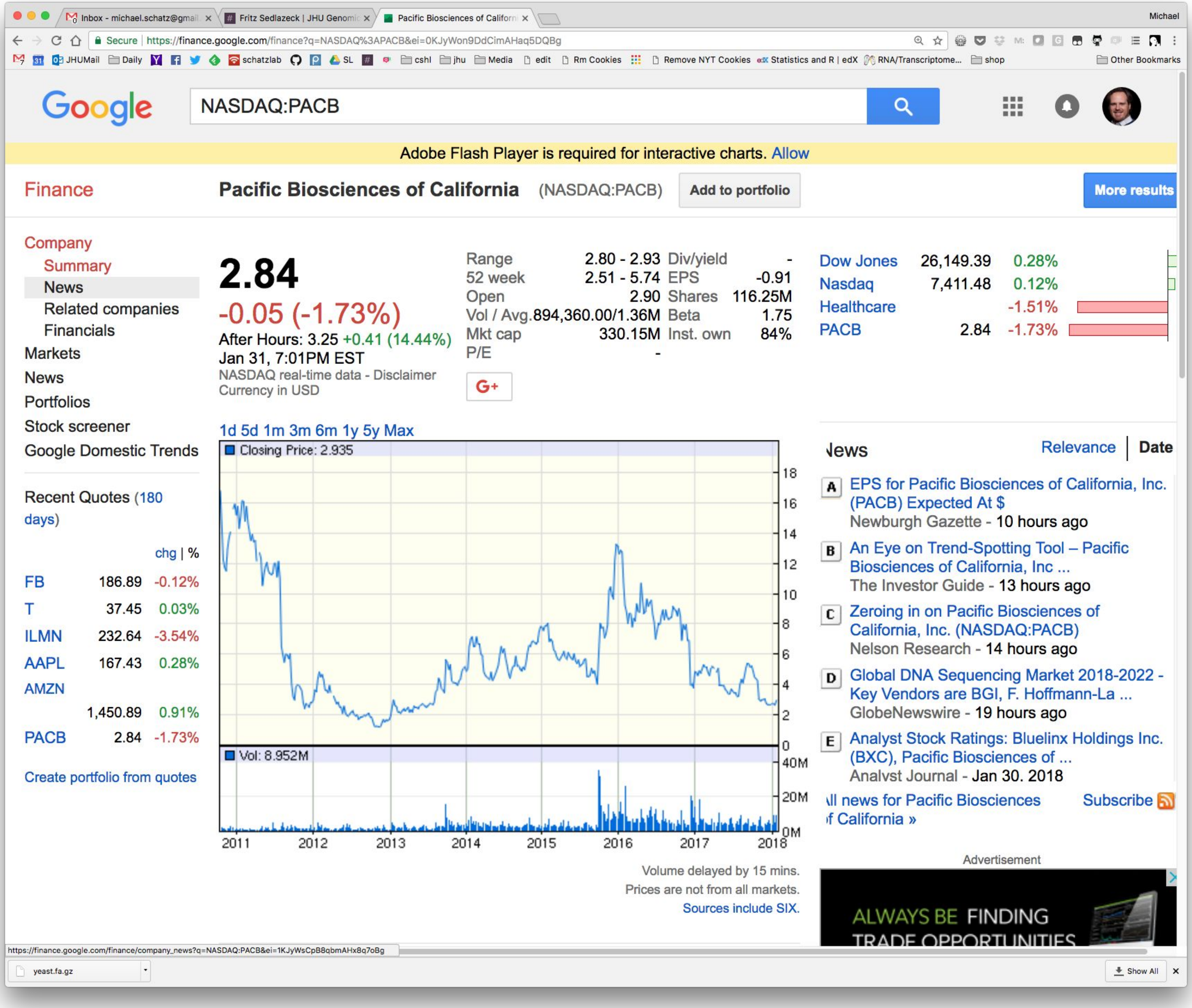
PacBio Sequel II

$350k instrument cost
841 lbs

~$6k / human @ 50x

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~10 GB / SMRTcell
~$1000 / SMRTcell



Inbox - michael.schatz@gmail
Fritz Sedlazeck | JHU Genomic
Pacific Biosciences of Californ
Michael
Secure | https://finance.google.com/finance?q=NASDAQ%3APACB&ei=0KJyWon9DdCimAHaq5DQBg
JHUMail
Daily
schatzlab
SL
cshl
jhu
Media
edit
Rm Cookies
Remove NYT Cookies
Statistics and R | edX
RNA/Transcriptome...
shop
Other Bookmarks
Google

NASDAQ:PACB

Adobe Flash Player is required for interactive charts. Allow

Finance

Pacific Biosciences of California (NASDAQ:PACB) Add to portfolio

More results

Company
- Summary
- News
- Related companies
- Financials

Markets

News

Portfolios

Stock screener

Google Domestic Trends

Recent Quotes (180 days)

		chg | %
FB	186.89	-0.12%
T	37.45	0.03%
ILMN	232.64	-3.54%
AAPL	167.43	0.28%
AMZN	1,450.89	0.91%
PACB	2.84	-1.73%

Create portfolio from quotes

2.84

-0.05 (-1.73%)

After Hours: 3.25 +0.41 (14.44%)

Jan 31, 7:01PM EST

NASDAQ real-time data - Disclaimer

Currency in USD

Range	2.80 - 2.93	Div/yield	-
52 week	2.51 - 5.74	EPS	-0.91
Open	2.90	Shares	116.25M
Vol / Avg.	894,360.00/1.36M	Beta	1.75
Mkt cap	330.15M	Inst. own	84%
P/E	-		

G+

Dow Jones	26,149.39	0.28%
Nasdaq	7,411.48	0.12%
Healthcare		-1.51%
PACB	2.84	-1.73%

Volume delayed by 15 mins.
Prices are not from all markets.
Sources include SIX.

\News Relevance | Date

A EPS for Pacific Biosciences of California, Inc. (PACB) Expected At $
Newburgh Gazette - 10 hours ago

B An Eye on Trend-Spotting Tool – Pacific Biosciences of California, Inc ...
The Investor Guide - 13 hours ago

C Zeroing in on Pacific Biosciences of California, Inc. (NASDAQ:PACB)
Nelson Research - 14 hours ago

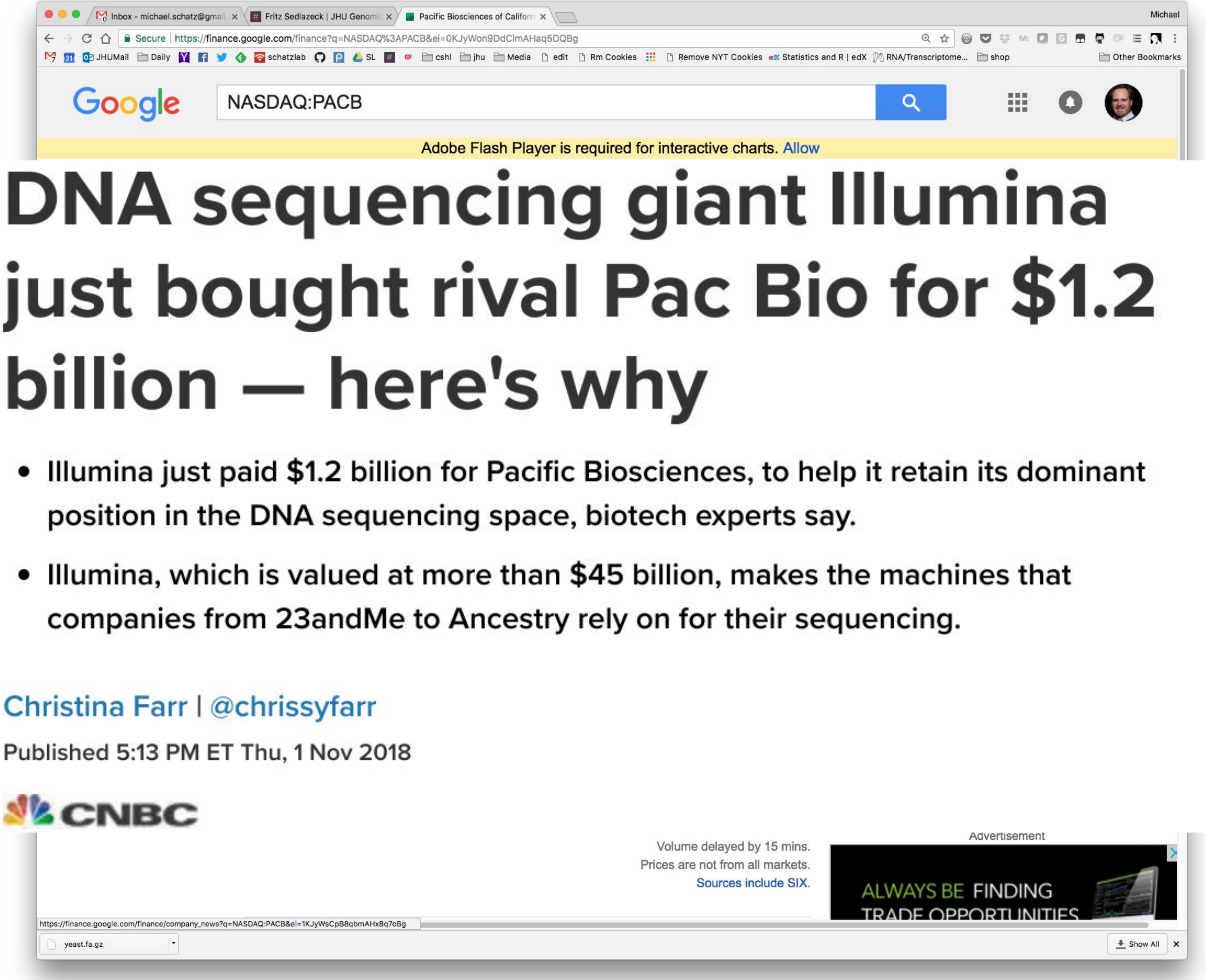
D Global DNA Sequencing Market 2018-2022 - Key Vendors are BGI, F. Hoffmann-La ...
GlobeNewswire - 19 hours ago

E Analyst Stock Ratings: Bluelinx Holdings Inc. (BXC), Pacific Biosciences of ...
Analyst Journal - Jan 30, 2018

ll news for Pacific Biosciences f California »

Subscribe

https://finance.google.com/finance/company_news?q=NASDAQ:PACB&ei=1KJyWsCpB8qbmAHx8q7oBg
yeast.fa.gz
Show All



Inbox - michael.schatz@gmail
Fritz Sedlazeck | JHU Genomic
Pacific Biosciences of Californ
Michael
Secure | https://finance.google.com/finance?q=NASDAQ%3APACB&ei=0KJyWon9DdCimAHaq5DQBg
JHUMail
Daily
schatzlab
SL
cshl
jhu
Media
edit
Rm Cookies
Remove NYT Cookies
Statistics and R | edX
RNA/Transcriptome...
shop
Other Bookmarks
Google
NASDAQ:PACB

Adobe Flash Player is required for interactive charts. Allow

DNA sequencing giant Illumina just bought rival Pac Bio for $1.2 billion — here's why

- Illumina just paid $1.2 billion for Pacific Biosciences, to help it retain its dominant position in the DNA sequencing space, biotech experts say.
- Illumina, which is valued at more than $45 billion, makes the machines that companies from 23andMe to Ancestry rely on for their sequencing.

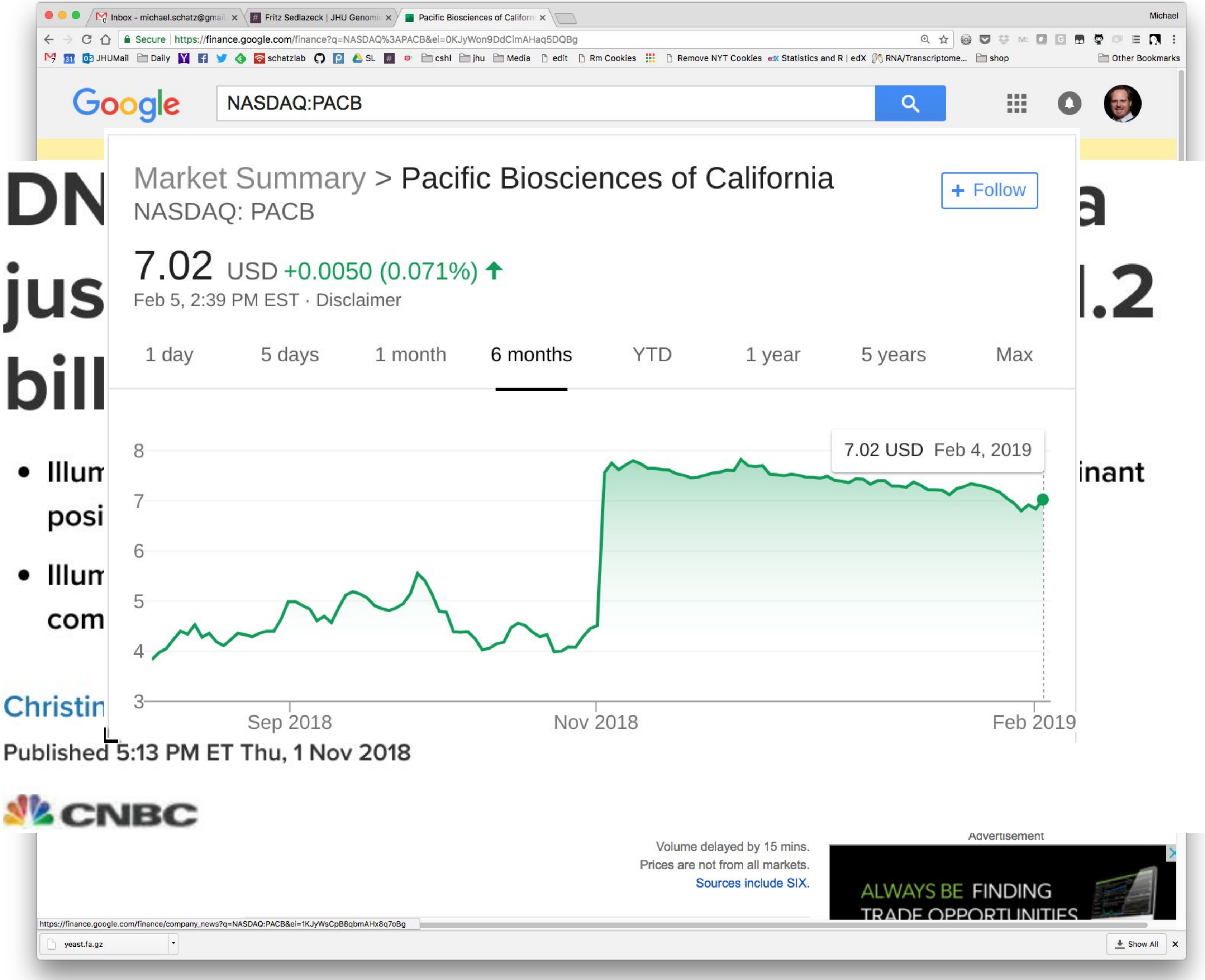
Christina Farr | @chrissyfarr

Published 5:13 PM ET Thu, 1 Nov 2018

CNBC

Volume delayed by 15 mins.
Prices are not from all markets.
Sources include SIX.

https://finance.google.com/finance/company_news?q=NASDAQ:PACB&ei=1KJyWsCpB8qbmAHx8q7oBg
yeast.fa.gz
Show All



Inbox - michael.schatz@gmail
Fritz Sedlazeck | JHU Genomic
Pacific Biosciences of Californ
Michael
Secure | https://finance.google.com/finance?q=NASDAQ%3APACB&ei=0KJyWon9DdCimAHaq5DQBg
JHUMail
Daily
schatzlab
SL
cshl
jhu
Media
edit
Rm Cookies
Remove NYT Cookies
Statistics and R | edX
RNA/Transcriptome...
shop
Other Bookmarks

NASDAQ:PACB

Market Summary > Pacific Biosciences of California

NASDAQ: PACB

+ Follow

7.02 USD +0.0050 (0.071%) ↑

Feb 5, 2:39 PM EST · Disclaimer

1 day
5 days
1 month
6 months
YTD
1 year
5 years
Max
8
7
6
5
4
3
7.02 USD Feb 4, 2019
Sep 2018
Nov 2018
Feb 2019

Volume delayed by 15 mins.
Prices are not from all markets.
Sources include SIX.

https://finance.google.com/finance/company_news?q=NASDAQ:PACB&ei=1KJyWsCpB8qbmAHx8q7oBg
yeast.fa.gz
Show All

DN a
jus l.2
bill

- Illum inant
  posi
- Illum
  com

Christin

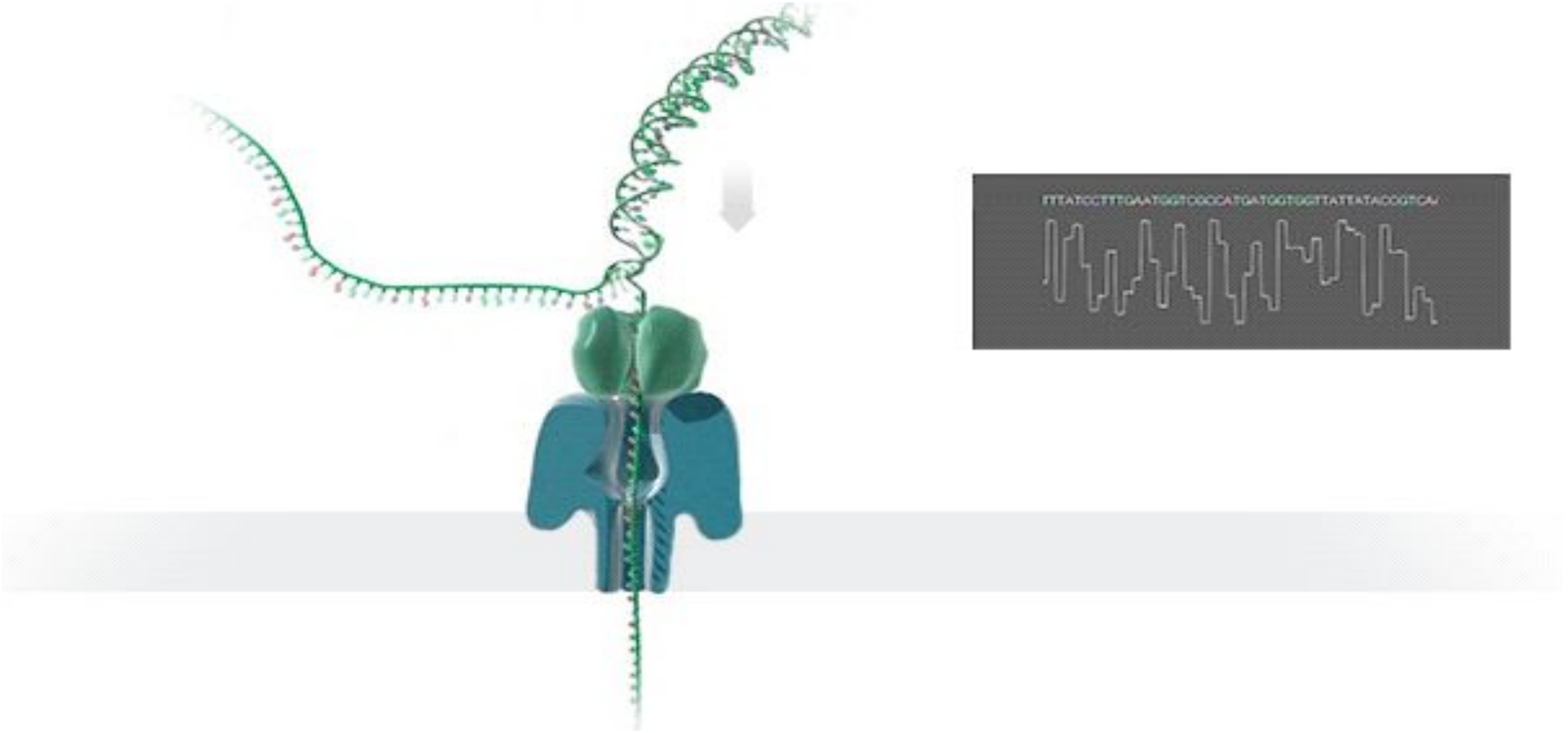
Published 5:13 PM ET Thu, 1 Nov 2018

CNBC

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore

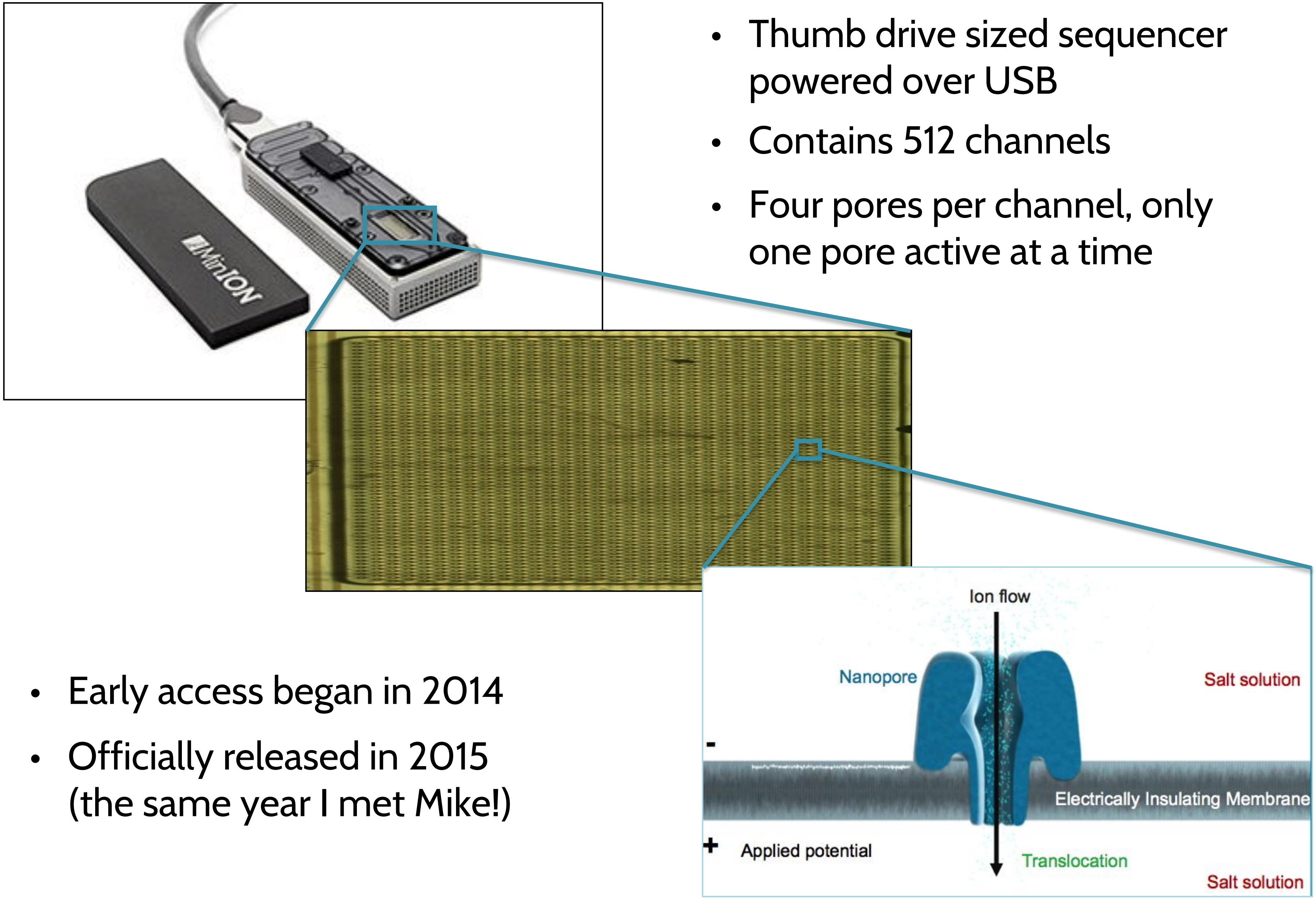


https://www.youtube.com/watch?v=CE4dW64x3Ts

nanoporetech.com/applications/dna-nanopore-sequencing

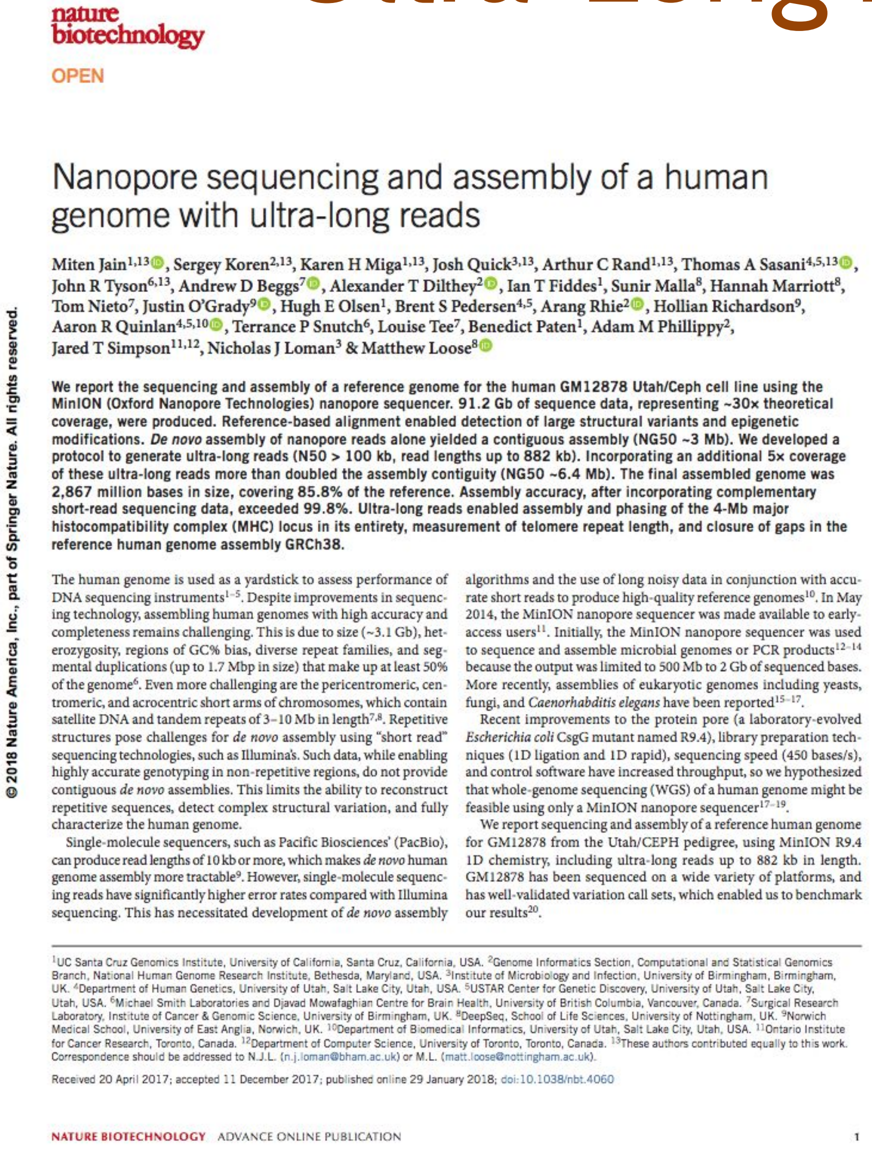
Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Contains 512 channels
- Four pores per channel, only one pore active at a time



- Early access began in 2014
- Officially released in 2015 (the same year I met Mike!)

“Ultra-Long Read” Assembly



nature
biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O’Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). We developed a protocol to generate ultra-long reads (N50 > 100 kb, read lengths up to 882 kb). Incorporating an additional 5× coverage of these ultra-long reads more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4-Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length, and closure of gaps in the reference human genome assembly GRCh38.

The human genome is used as a yardstick to assess performance of DNA sequencing instruments[1–5]. Despite improvements in sequencing technology, assembling human genomes with high accuracy and completeness remains challenging. This is due to size (~3.1 Gb), heterozygosity, regions of GC% bias, diverse repeat families, and segmental duplications (up to 1.7 Mbp in size) that make up at least 50% of the genome[6]. Even more challenging are the pericentromeric, centromeric, and acrocentric short arms of chromosomes, which contain satellite DNA and tandem repeats of 3–10 Mb in length[7,8]. Repetitive structures pose challenges for *de novo* assembly using “short read” sequencing technologies, such as Illumina's. Such data, while enabling highly accurate genotyping in non-repetitive regions, do not provide contiguous *de novo* assemblies. This limits the ability to reconstruct repetitive sequences, detect complex structural variation, and fully characterize the human genome.

Single-molecule sequencers, such as Pacific Biosciences’ (PacBio), can produce read lengths of 10 kb or more, which makes *de novo* human genome assembly more tractable[9]. However, single-molecule sequencing reads have significantly higher error rates compared with Illumina sequencing. This has necessitated development of *de novo* assembly algorithms and the use of long noisy data in conjunction with accurate short reads to produce high-quality reference genomes[10]. In May 2014, the MinION nanopore sequencer was made available to early-access users[11]. Initially, the MinION nanopore sequencer was used to sequence and assemble microbial genomes or PCR products[12–14] because the output was limited to 500 Mb to 2 Gb of sequenced bases. More recently, assemblies of eukaryotic genomes including yeasts, fungi, and *Caenorhabditis elegans* have been reported[15–17].

Recent improvements to the protein pore (a laboratory-evolved *Escherichia coli* CsgG mutant named R9.4), library preparation techniques (1D ligation and 1D rapid), sequencing speed (450 bases/s), and control software have increased throughput, so we hypothesized that whole-genome sequencing (WGS) of a human genome might be feasible using only a MinION nanopore sequencer[17–19].

We report sequencing and assembly of a reference human genome for GM12878 from the Utah/CEPH pedigree, using MinION R9.4 1D chemistry, including ultra-long reads up to 882 kb in length. GM12878 has been sequenced on a wide variety of platforms, and has well-validated variation call sets, which enabled us to benchmark our results[20].

[1]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California, USA. [2]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA. [3]Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA. [5]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah, USA. [6]Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada. [7]Surgical Research Laboratory, Institute of Cancer & Genomic Science, University of Birmingham, UK. [8]DeepSeq, School of Life Sciences, University of Nottingham, UK. [9]Norwich Medical School, University of East Anglia, Norwich, UK. [10]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA. [11]Ontario Institute for Cancer Research, Toronto, Canada. [12]Department of Computer Science, University of Toronto, Toronto, Canada. [13]These authors contributed equally to this work. Correspondence should be addressed to N.J.L. (n.j.loman@bham.ac.uk) or M.L. (matt.loose@nottingham.ac.uk).

Received 20 April 2017; accepted 11 December 2017; published online 29 January 2018; doi:10.1038/nbt.4060

Current Nanopore Assembly

nature
biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

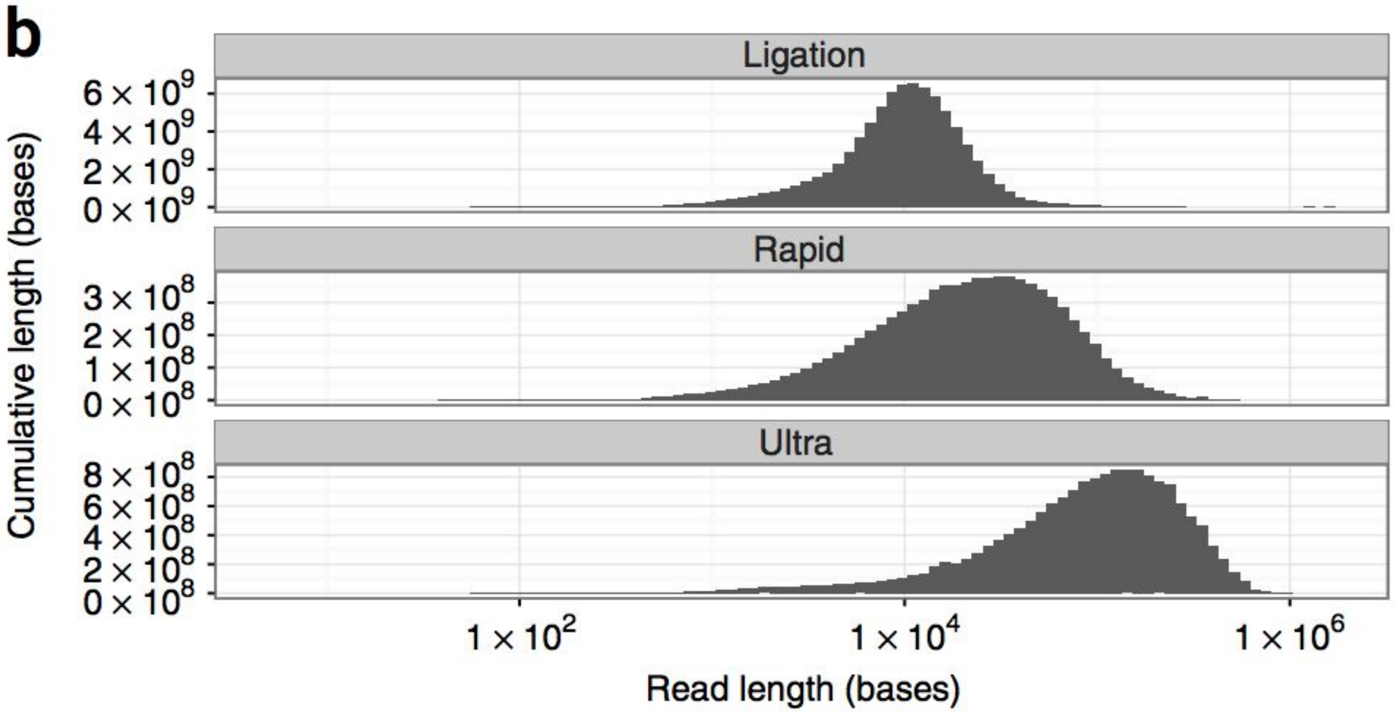
**We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Referenc
modifications. *De novo* assembly o
protocol to generate ultra-long reac
of these ultra-long reads more than
2,867 million bases in size, coveri
short-read sequencing data, exceed
histocompatibility complex (MHC)
reference human genome assembl**

The human genome is used as a yardst
DNA sequencing instruments[1–5]. Despi
ing technology, assembling human gen
completeness remains challenging. Thi
erozygosity, regions of GC% bias, dive
mental duplications (up to 1.7 Mbp in s
of the genome[6]. Even more challenging
tromeric, and acrocentric short arms of
satellite DNA and tandem repeats of 3–
structures pose challenges for *de novo*
sequencing technologies, such as Illumi
highly accurate genotyping in non-repe
contiguous *de novo* assemblies. This lir
repetitive sequences, detect complex s
characterize the human genome.

Single-molecule sequencers, such as
can produce read lengths of 10 kb or mor
genome assembly more tractable[9]. Howe
ing reads have significantly higher error
sequencing. This has necessitated deve

[1]UC Santa Cruz Genomics Institute, University
Branch, National Human Genome Research Ins
UK. [4]Department of Human Genetics, Universi
Utah, USA. [6]Michael Smith Laboratories and D
Laboratory, Institute of Cancer & Genomic Scie
Medical School, University of East Anglia, Nor
for Cancer Research, Toronto, Canada. [12]Depar
Correspondence should be addressed to N.J.L.

Received 20 April 2017; accepted 11 Decemb



Current Nanopore Assembly

nature
biotechnology

OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

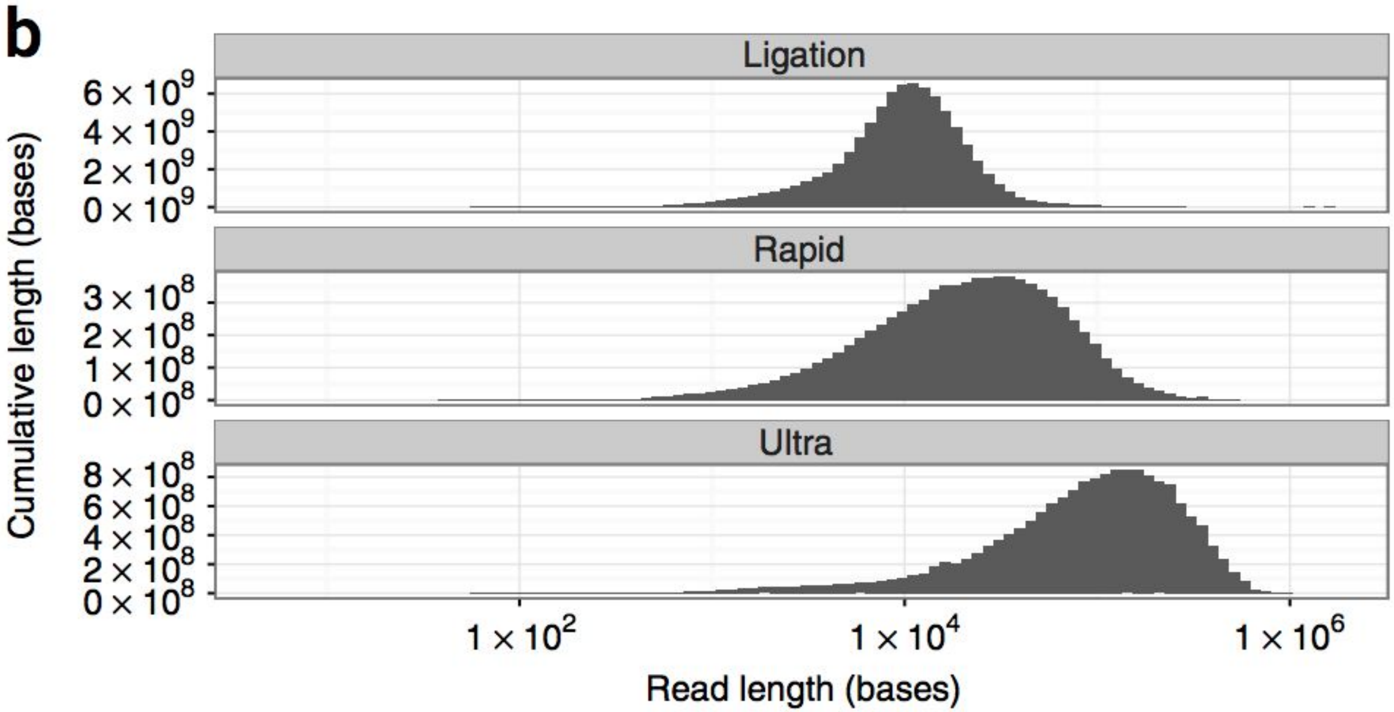
We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical
coverage, were produced. Referenc
modifications. *De novo* assembly o
protocol to generate ultra-long reac
of these ultra-long reads more thar
2,867 million bases in size, coveri
short-read sequencing data, exceec
histocompatibility complex (MHC)
reference human genome assembly

The human genome is used as a yardst
DNA sequencing instruments[1–5]. Despi
ing technology, assembling human gen
completeness remains challenging. Thi
erozygosity, regions of GC% bias, dive
mental duplications (up to 1.7 Mbp in s
of the genome[6]. Even more challenging
tromeric, and acrocentric short arms of
satellite DNA and tandem repeats of 3–
structures pose challenges for *de novo*
sequencing technologies, such as Illumi
highly accurate genotyping in non-repe
contiguous *de novo* assemblies. This lir
repetitive sequences, detect complex s
characterize the human genome.

Single-molecule sequencers, such as
can produce read lengths of 10 kb or mor
genome assembly more tractable[9]. Howe
ing reads have significantly higher error
sequencing. This has necessitated deve

[1]UC Santa Cruz Genomics Institute, University
Branch, National Human Genome Research Ins
UK. [4]Department of Human Genetics, Universi
Utah, USA. [6]Michael Smith Laboratories and D
Laboratory, Institute of Cancer & Genomic Scie
Medical School, University of East Anglia, Nor
for Cancer Research, Toronto, Canada. [12]Depar
Correspondence should be addressed to N.J.L.

Received 20 April 2017; accepted 11 Decemb

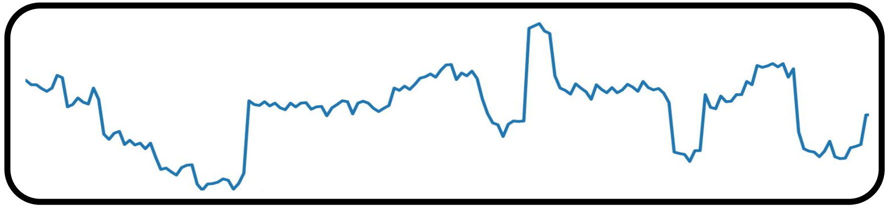


Same group recently reported a read 2.3 million bases long!

No theoretical upper limit

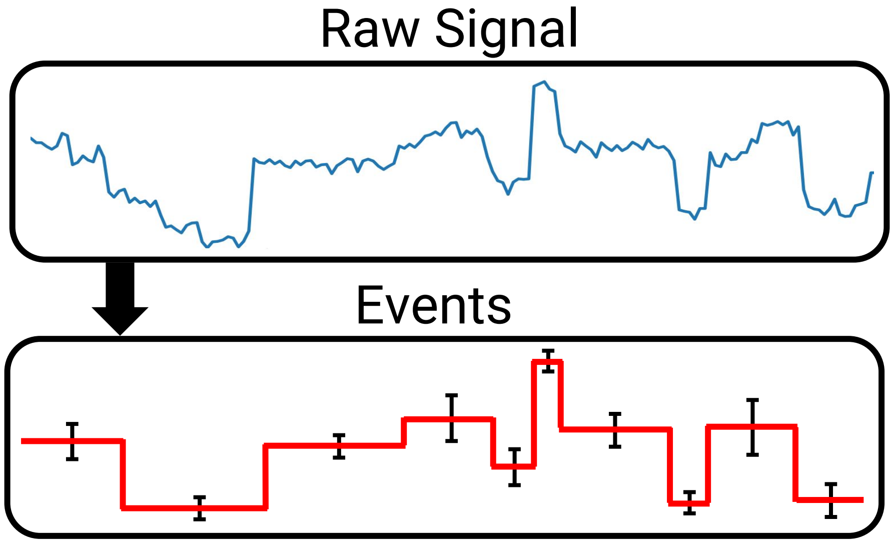
Nanopore Basecalling

Raw Signal



Translation of raw signal into basepairs

Nanopore Basecalling

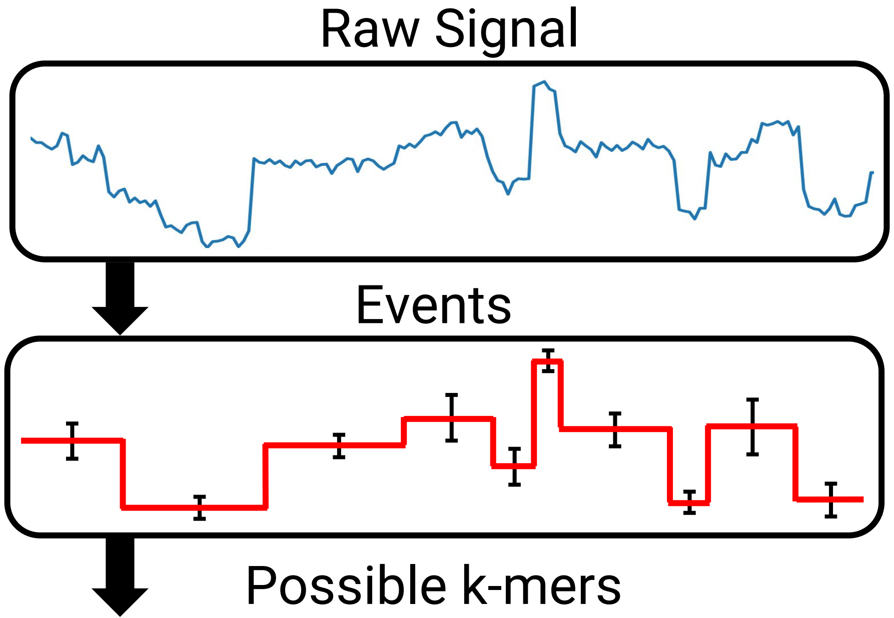


Translation of raw signal into basepairs

Early basecallers began by estimating k-mer boundaries using “events”, which were then input to an HMM

Modern basecalers use neural networks directly on raw signal

Nanopore Basecalling

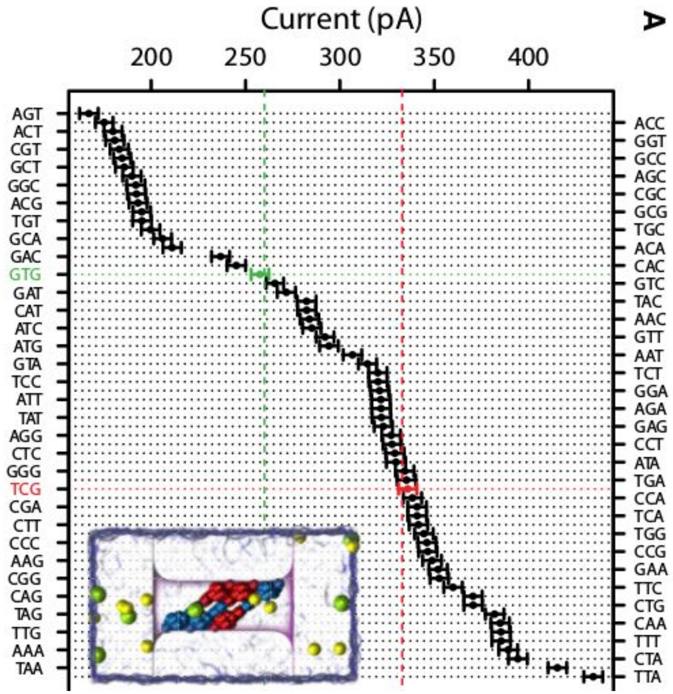


Possible k-mers

0	1	2	3
TCCA	CCAT	CATG	TACA
AGCA	TGGC	TTAC	TCCA
GTCT	ATTA	ACGT	GACG
GATT	ATTG	GTCT	ACGG

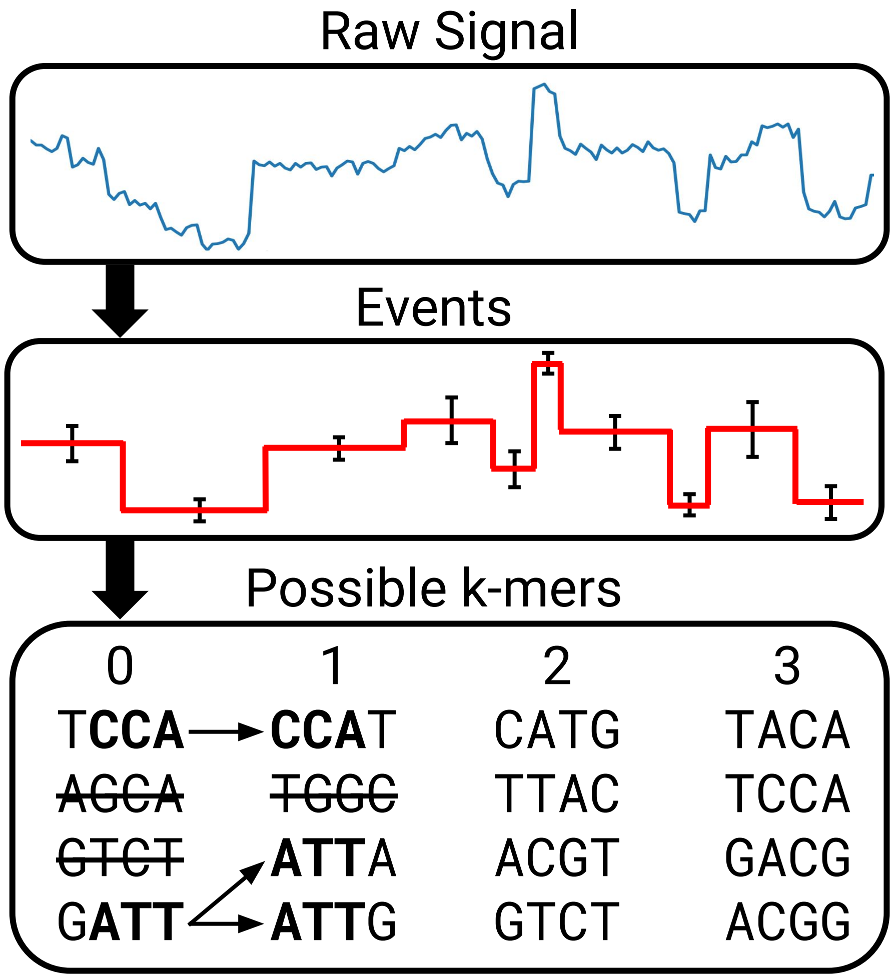
(Based on probability of event matches)

ONT releases k-mer models with expected current distribution of every k-mer



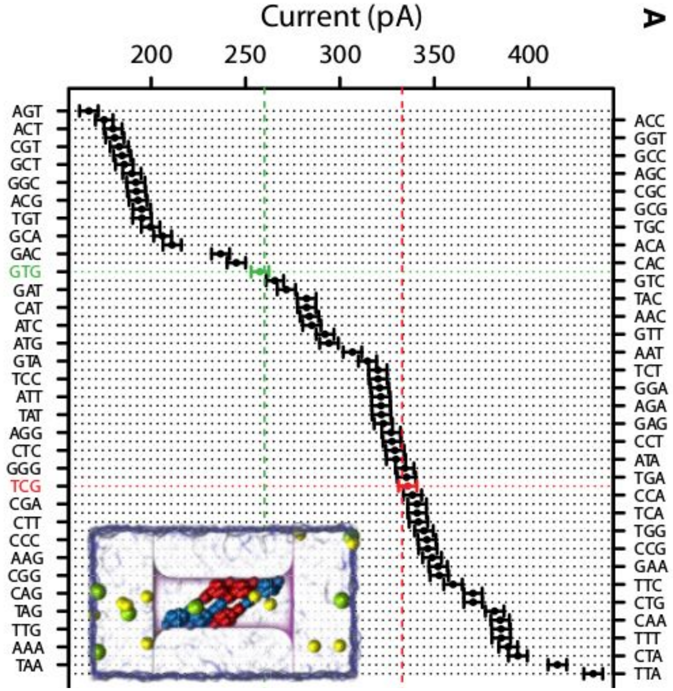
DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling



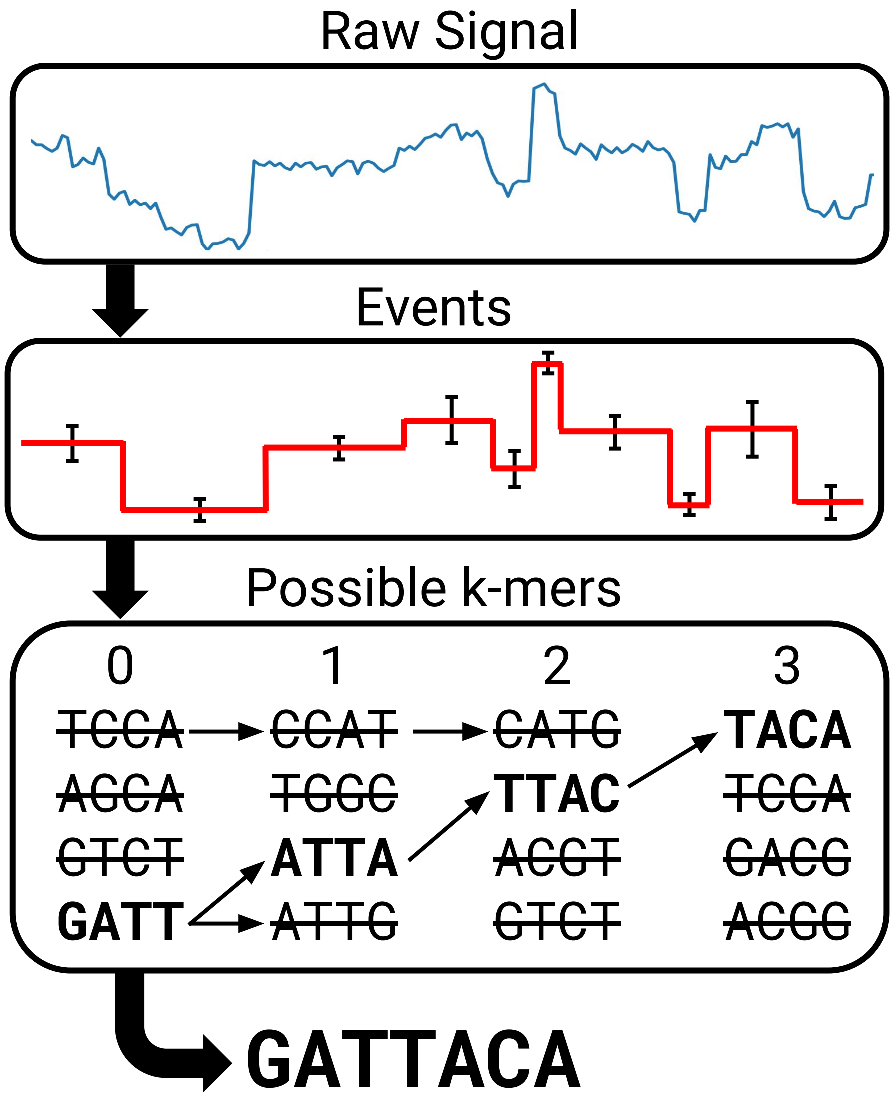
0	1	2	3
T**CCA** →	**CCA**T	CATG	TACA
~~AGCA~~	~~TGGC~~	TTAC	TCCA
~~GTCT~~	**ATT**A	ACGT	GACG
G**ATT** →	**ATT**G	GTCT	ACGG

Certain k-mers can be eliminated based on possible transitions



DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

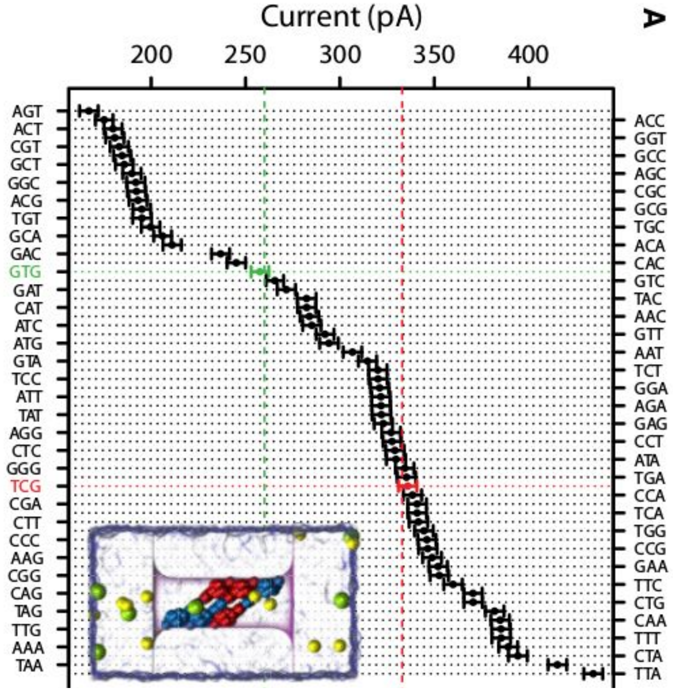
Nanopore Basecalling



0	1	2	3
~~TCCA~~	~~CCAT~~	~~CATG~~	**TACA**
~~AGCA~~	~~TGGC~~	**TTAC**	~~TCCA~~
~~GTCT~~	**ATTA**	~~ACGT~~	~~GACG~~
GATT	~~ATTG~~	~~GTCT~~	~~ACGG~~

GATTACA

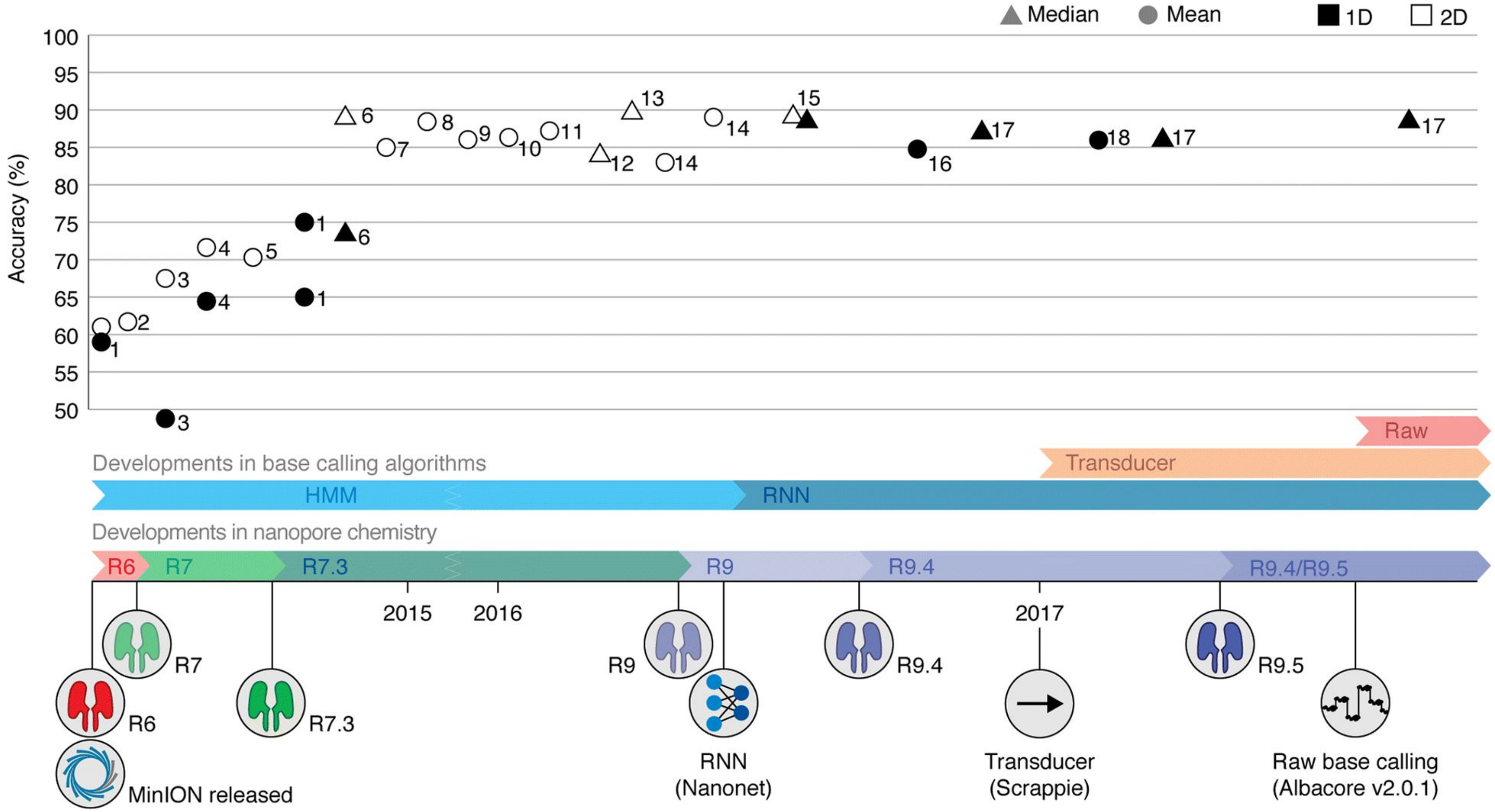
Final sequence determined by most probable k-mers



"DNA Base-Calling from a Nanopore Using a Viterbi Algorithm"
Timp et al. (2012) *Biophysical Journal*

Basecaller/Pore Timeline

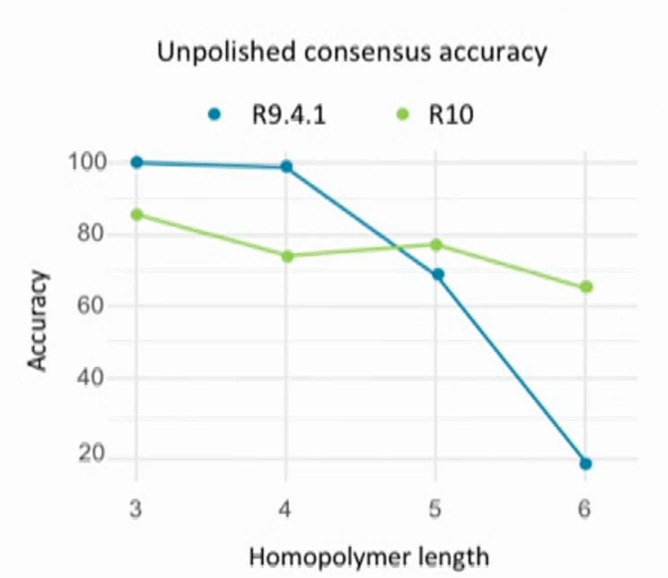
Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy

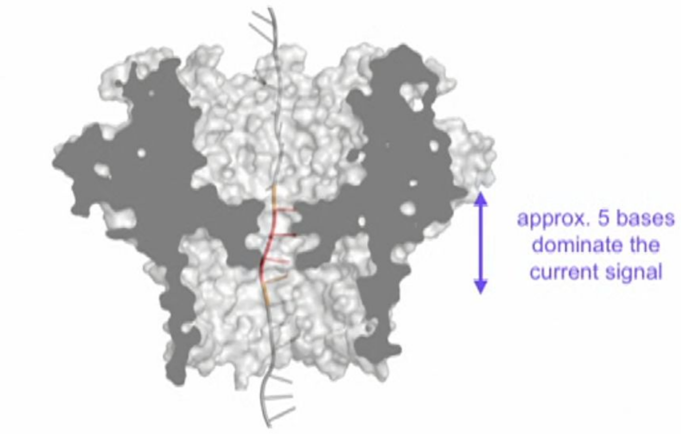


From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology.* https://doi.org/10.1186/s13059-018-1462-9

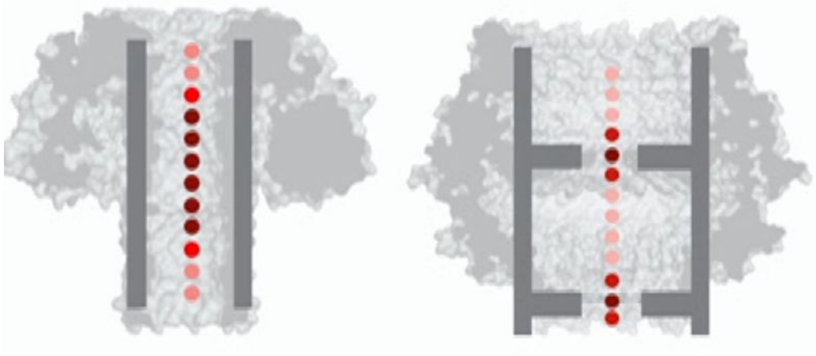
New Pore Chemistries

ONT is developing alternate pore chemistries to improve accuracy, particularly for homopolymers





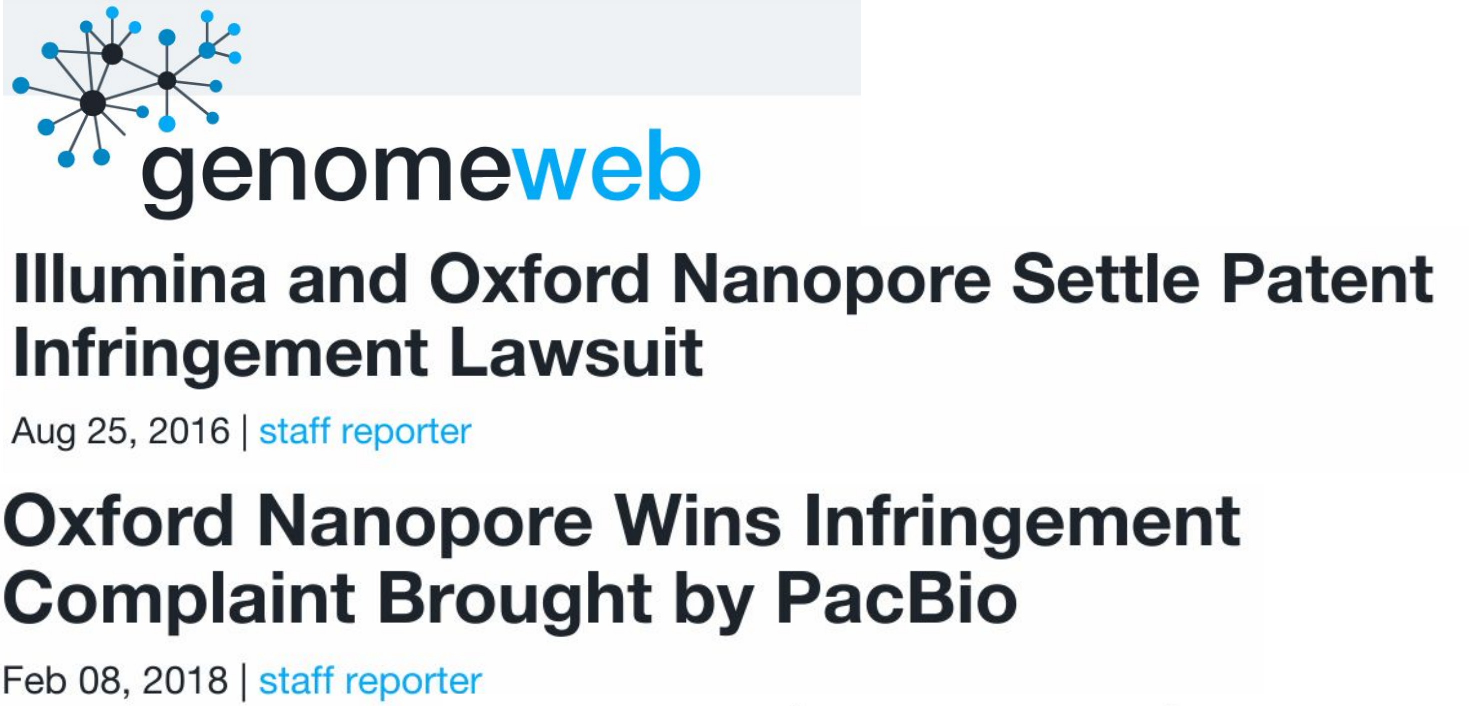
Standard pore chemistry
"R9"



Pore with long reader head Lysenin –
"R8"

Multiple points of contribution
"R10"

From 2018 London Calling Keynote
https://vimeo.com/272526835



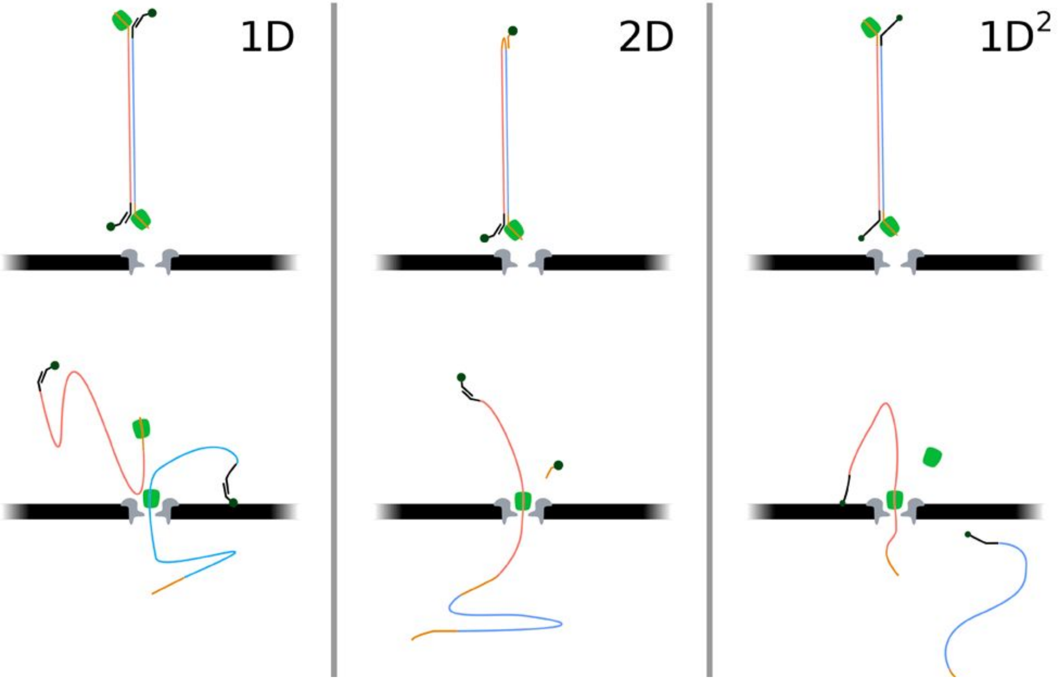
genomeweb

Illumina and Oxford Nanopore Settle Patent Infringement Lawsuit

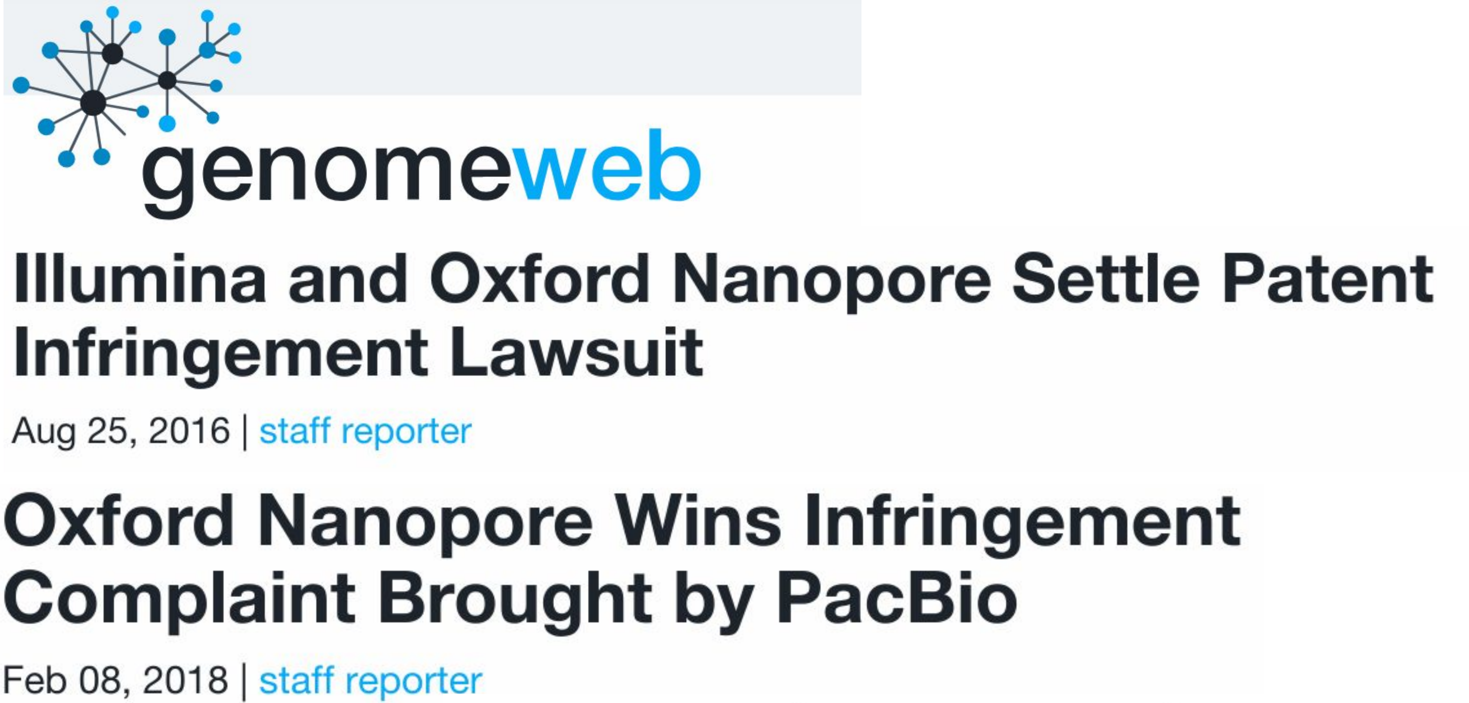
Aug 25, 2016 | staff reporter

Oxford Nanopore Wins Infringement Complaint Brought by PacBio

Feb 08, 2018 | staff reporter



https://f1000research.com/articles/6-1083/v1



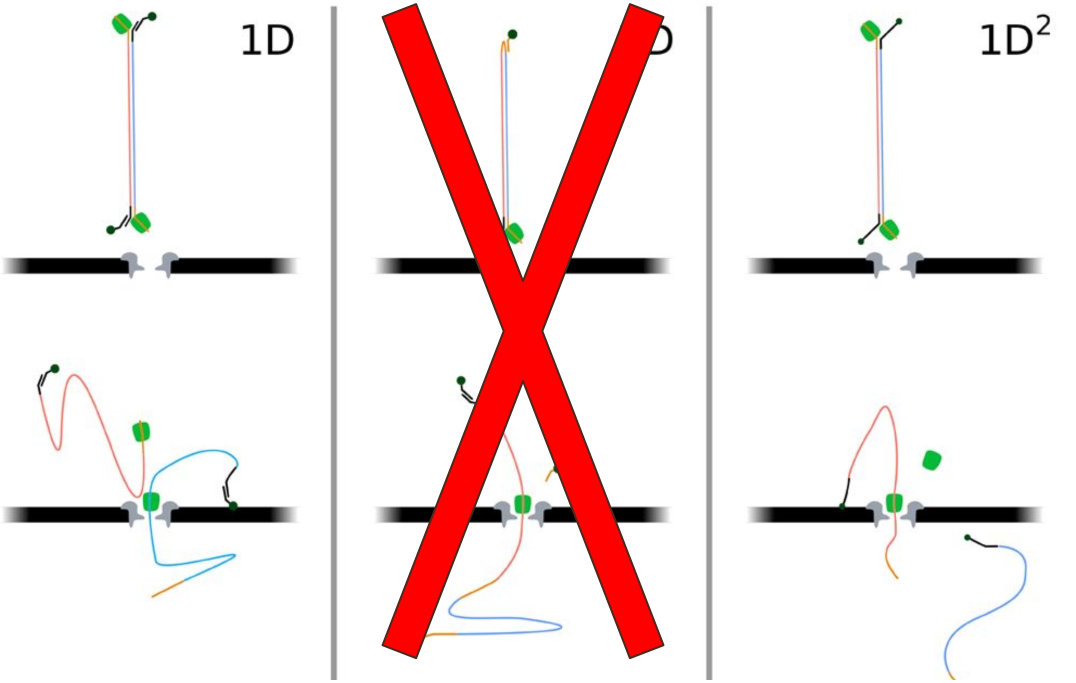
genomeweb

Illumina and Oxford Nanopore Settle Patent Infringement Lawsuit

Aug 25, 2016 | staff reporter

Oxford Nanopore Wins Infringement Complaint Brought by PacBio

Feb 08, 2018 | staff reporter



https://f1000research.com/articles/6-1083/v1

More Throughput

MinION

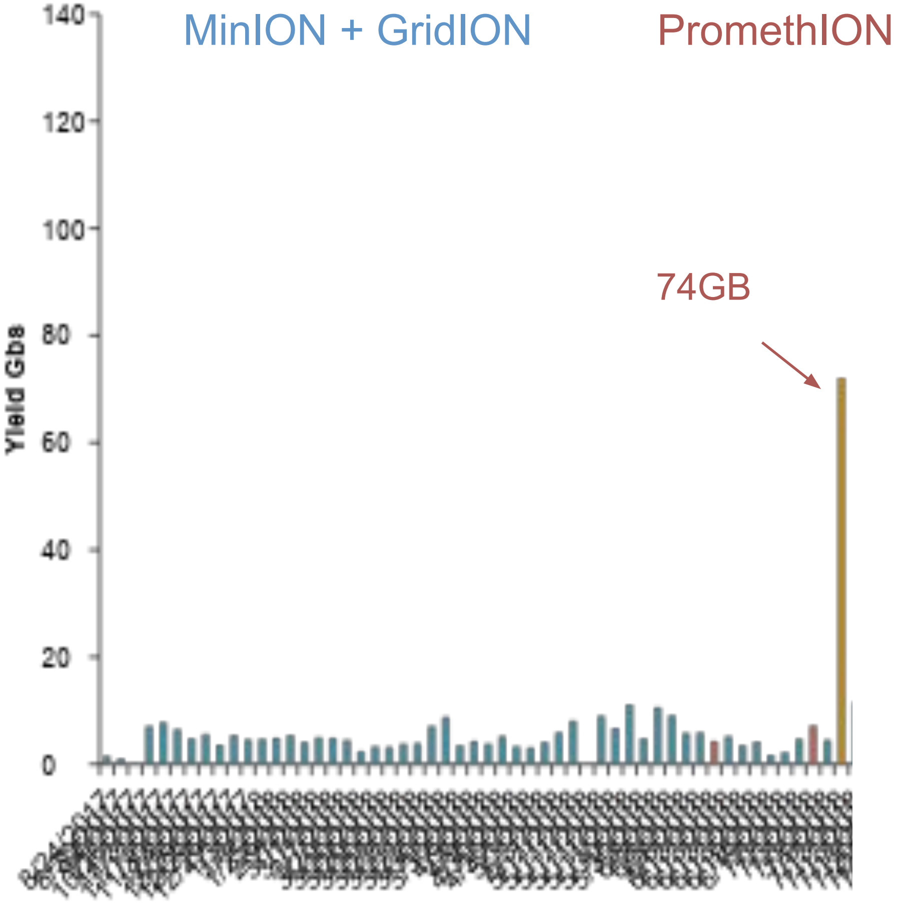
Quick Mobile
Sequencing
$1k / instrument
5-6 GB / day

PromethION

High Throughput Desktop
Sequencer
$75k / instrument
>>1000GB / day

Nanopore Performance at CSHL

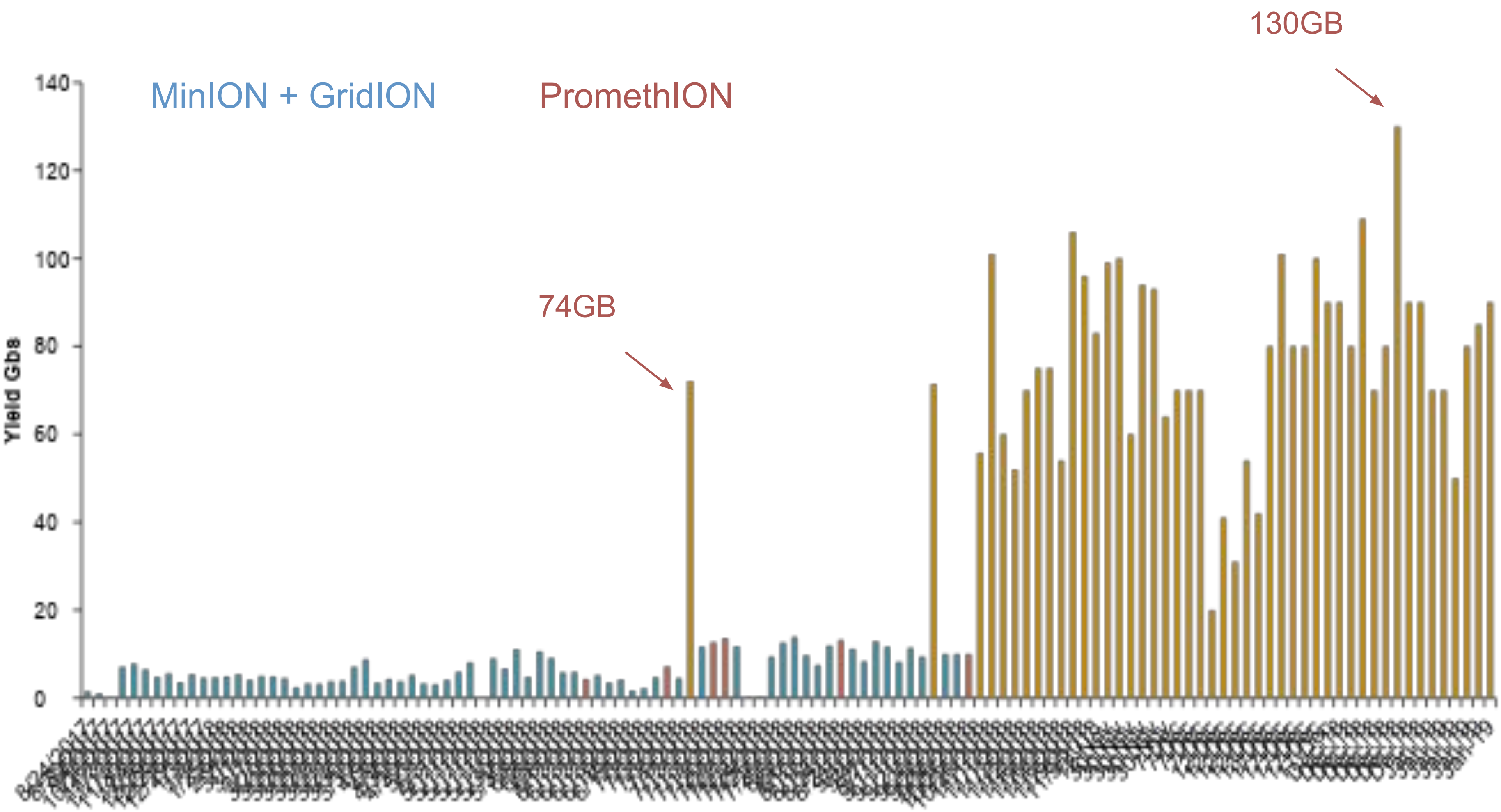
Sara Goodwin



Part of collaboration between JHU and CSHL to sequence 100 tomato genomes in 100 days

Nanopore Performance at CSHL

Sara Goodwin



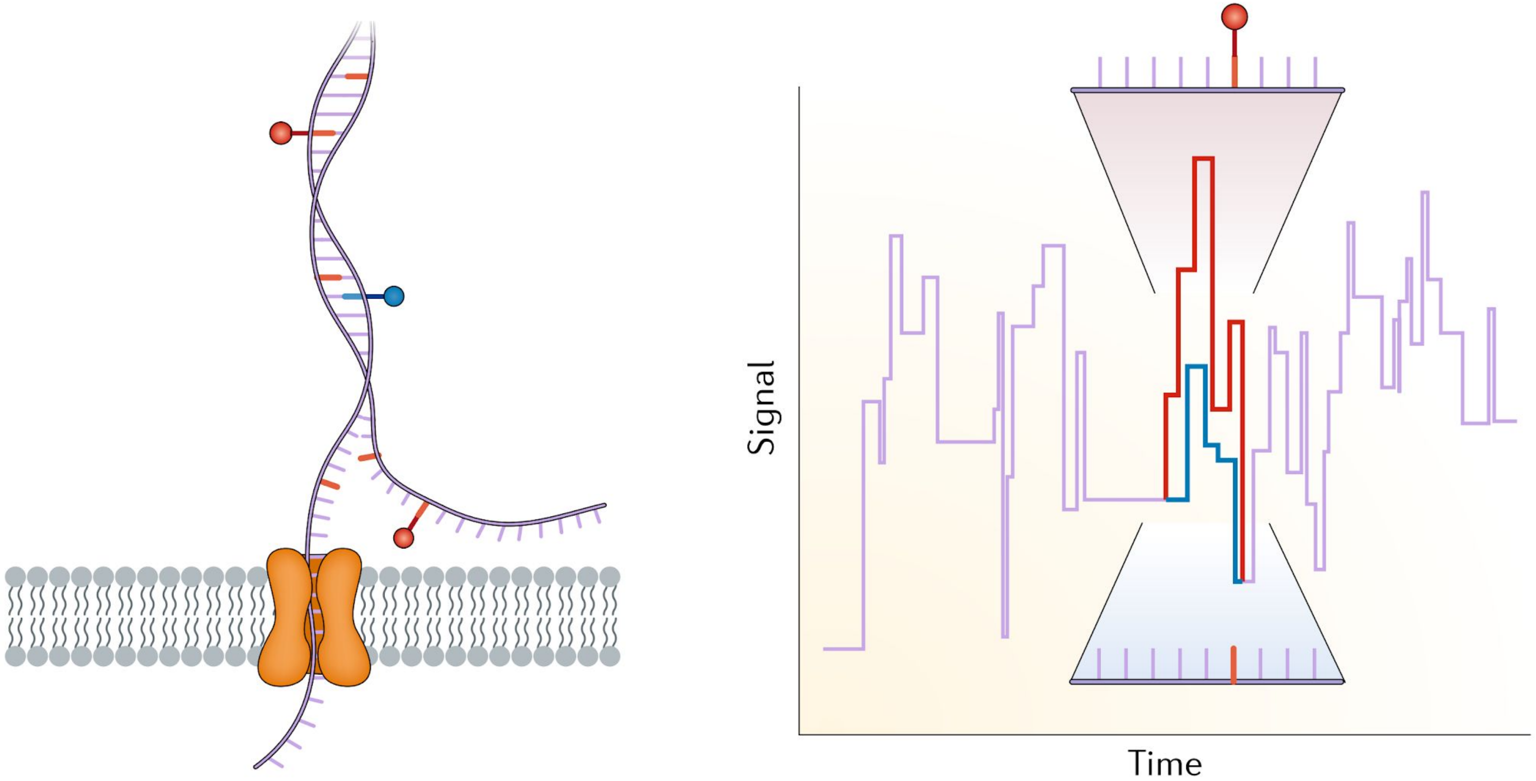
74GB

130GB

DNA Modification Detection

Like PacBio, ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



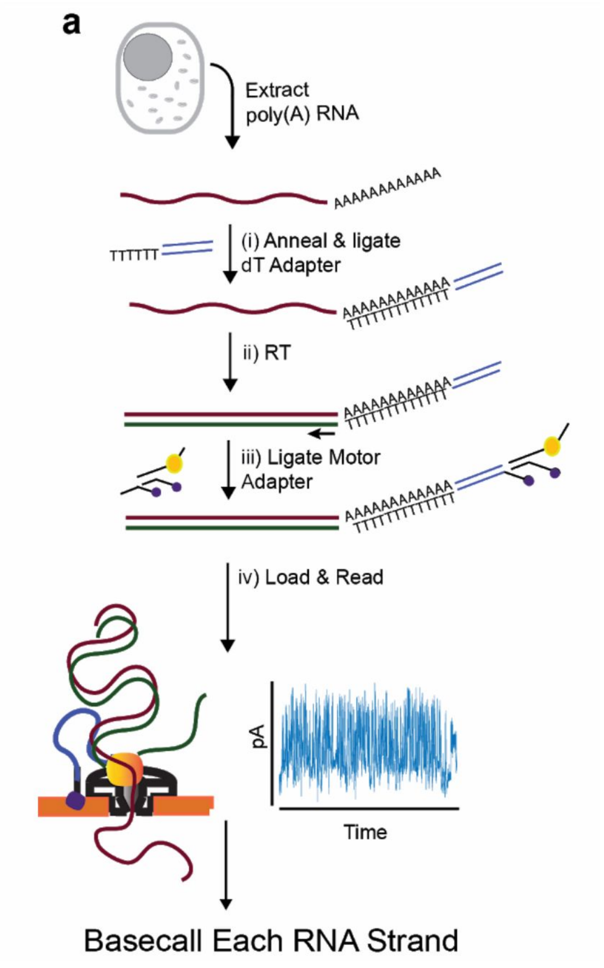
Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

Direct RNA-seq

Standard RNA sequencing (RNA-seq) requires creation of complementary DNA (cDNA)

ONT recently introduced direct RNA sequencing

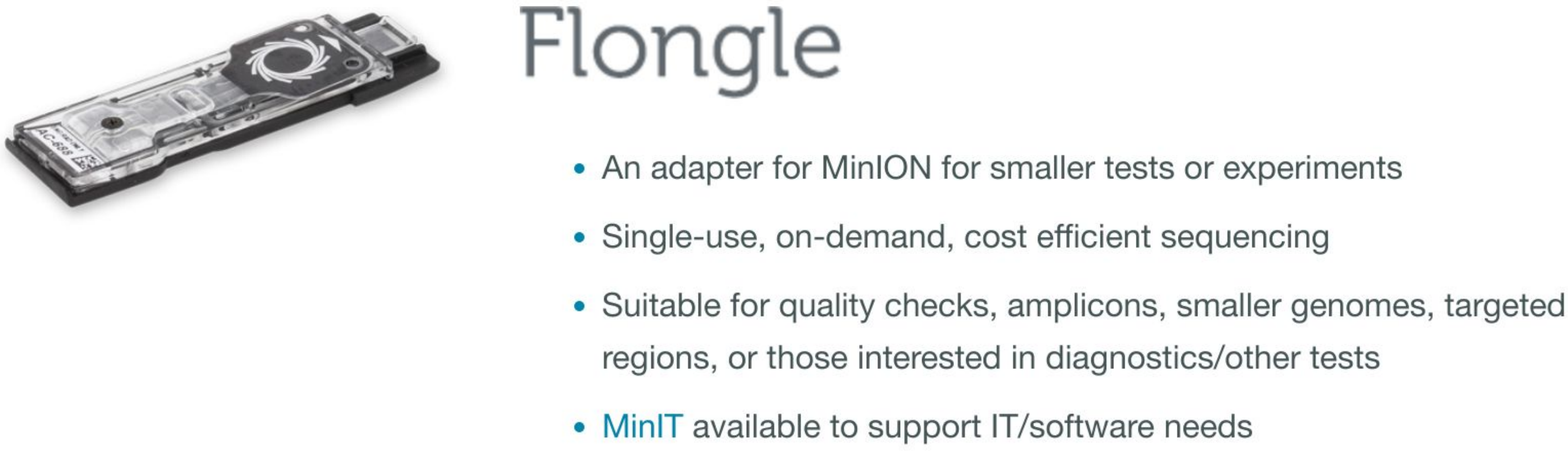
Allows detection of RNA modifications, and potentially secondary structure



Nanopore native RNA sequencing of a human poly(A) transcriptome

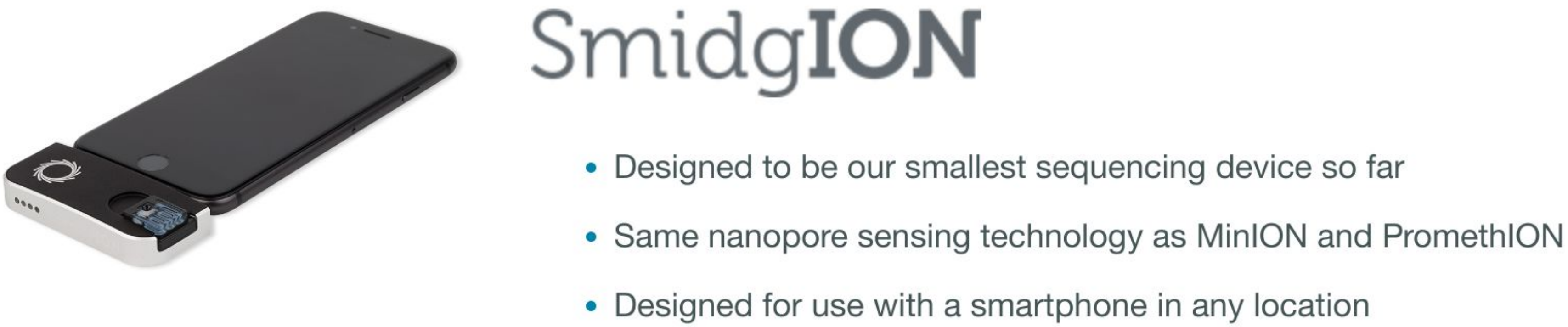
Workman et al. *BioRxiv* (https://www.biorxiv.org/content/10.1101/459529v1)

Less Throughput (coming soon)



Flongle

- An adapter for MinION for smaller tests or experiments
- Single-use, on-demand, cost efficient sequencing
- Suitable for quality checks, amplicons, smaller genomes, targeted regions, or those interested in diagnostics/other tests
- MinIT available to support IT/software needs

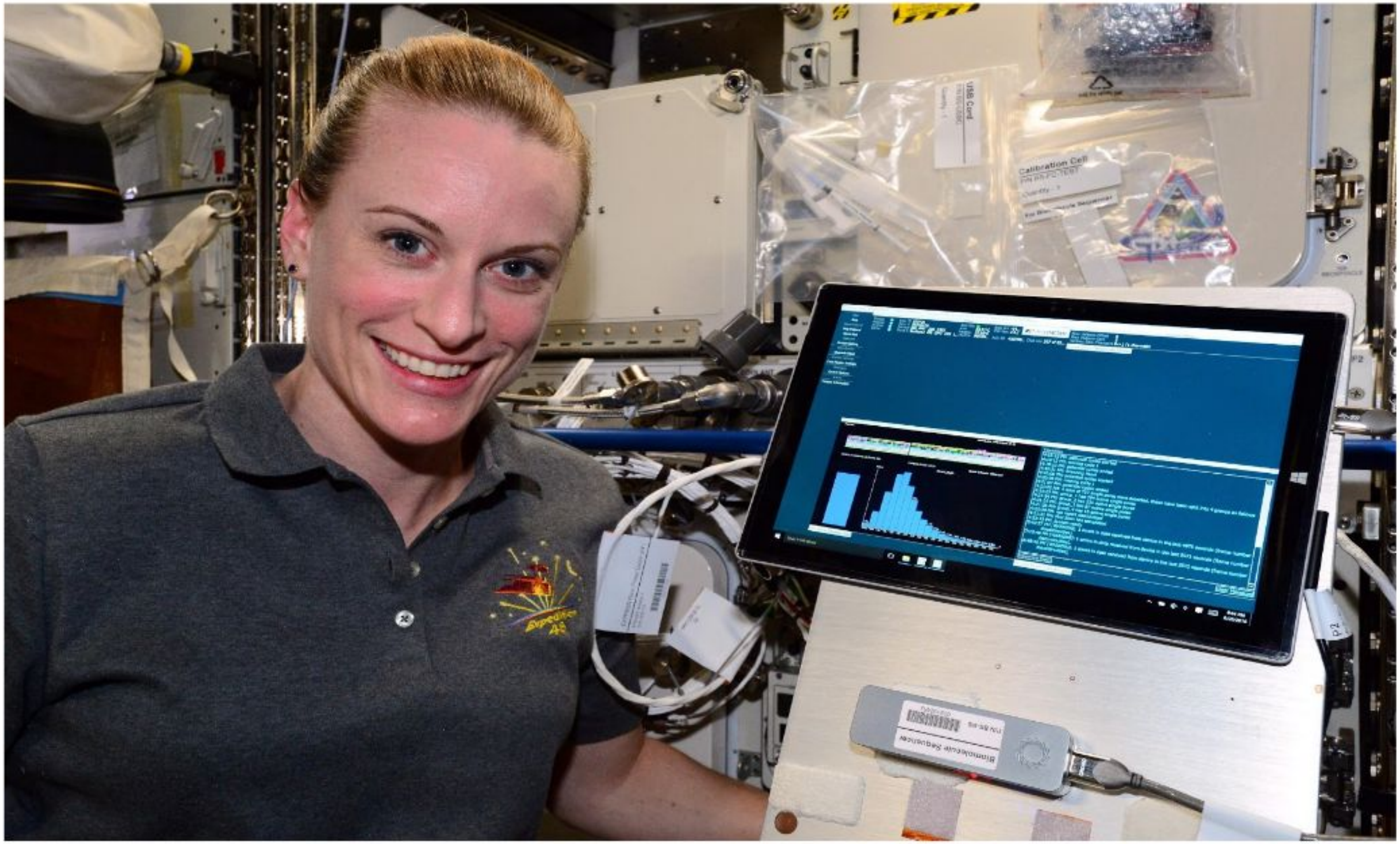


SmidgION

- Designed to be our smallest sequencing device so far
- Same nanopore sensing technology as MinION and PromethION
- Designed for use with a smartphone in any location

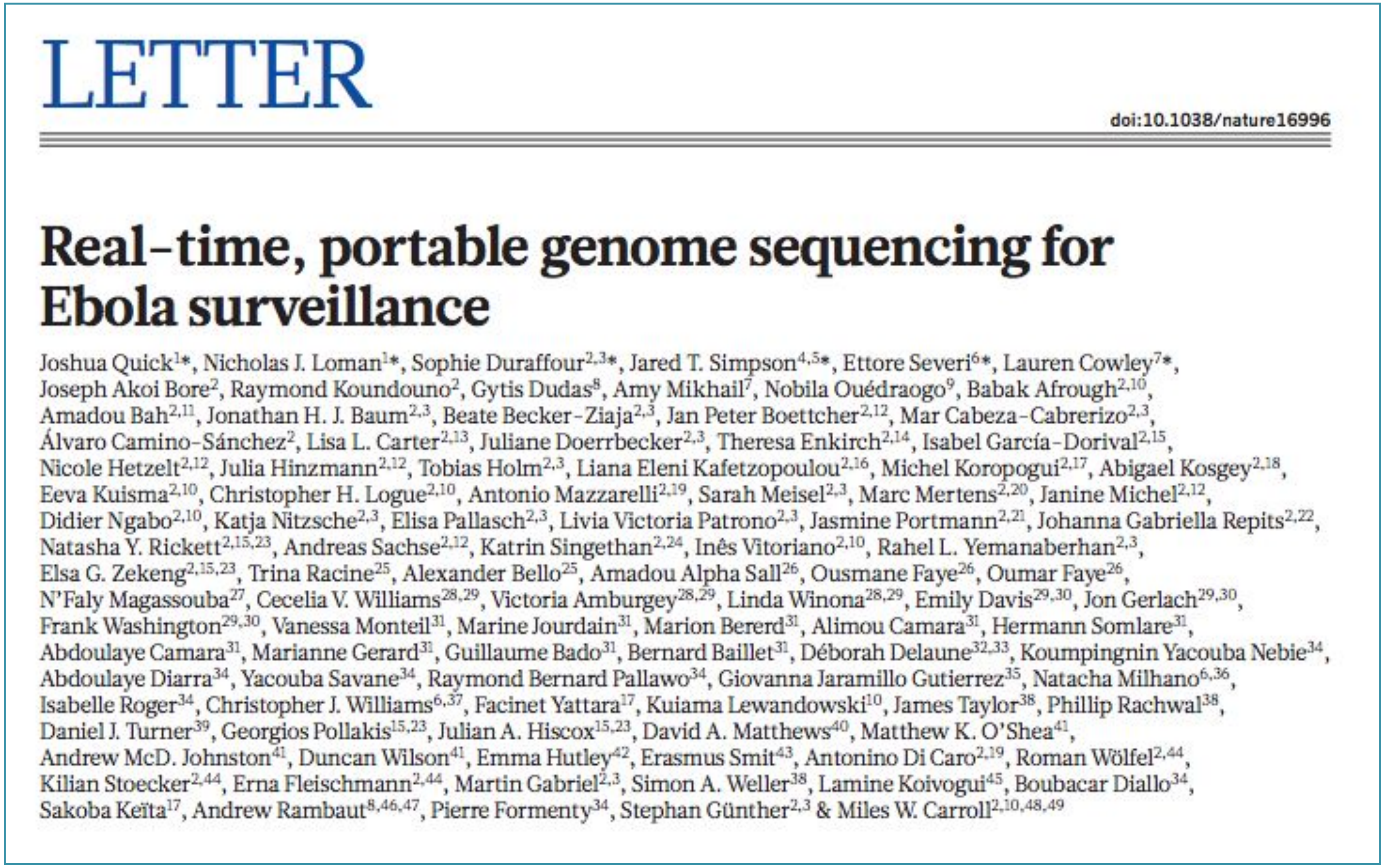
https://nanoporetech.com/products

Extremely Portable Sequencing!



Kate Rubins sequencing DNA on the ISS

Ebola Surveillance



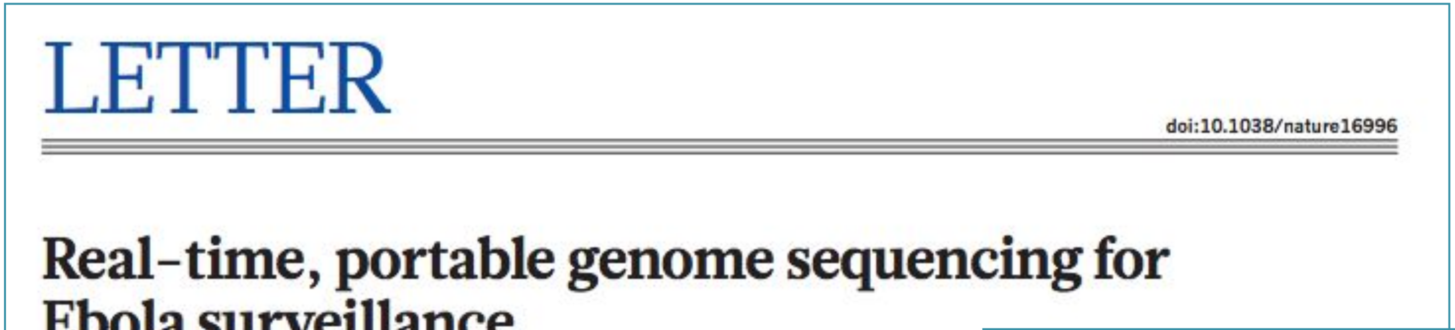
LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Severi[6]*, Lauren Cowley[7]*, Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraogo[9], Babak Afrough[2,10], Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar Cabeza-Cabrerizo[2,3], Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Isabel García-Dorival[2,15], Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Michel Koropogui[2,17], Abigael Kosgey[2,18], Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc Mertens[2,20], Janine Michel[2,12], Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Portmann[2,21], Johanna Gabriella Repits[2,22], Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel L. Yemanaberhan[2,3], Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Faye[26], Oumar Faye[26], N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Emily Davis[29,30], Jon Gerlach[29,30], Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Camara[31], Hermann Somlare[31], Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Delaune[32,33], Koumpingnin Yacouba Nebie[34], Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gutierrez[35], Natacha Milhano[6,36], Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], James Taylor[38], Phillip Rachwal[38], Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthew K. O'Shea[41], Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di Caro[2,19], Roman Wölfel[2,44], Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koivogui[45], Boubacar Diallo[34], Sakoba Keïta[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. Carroll[2,10,48,49]

Ebola Surveillance



LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Se
Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraog
Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Ma
Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Is
Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Miche
Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc M
Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Port
Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel
Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Fa
N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Er
Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Berard[31], Alimou Cam
Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Dela
Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gu
Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], Jam
Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthe
Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di
Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koi
Sakoba Keïta[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. C

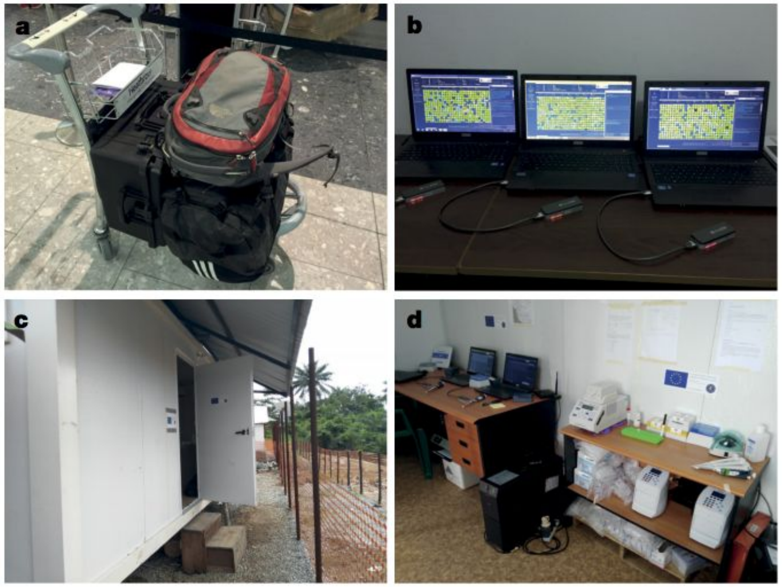


Figure 1 | Deployment of the portable genome surveillance system in Guinea. a, We were able to pack all instruments, reagents and disposable consumables within aircraft baggage. **b,** We initially established the genomic surveillance laboratory in Donka Hospital, Conakry, Guinea. **c,** Later we moved the laboratory to a dedicated sequencing laboratory in Coyah prefecture. **d,** Within this laboratory we separated the sequencing instruments (on the left) from the PCR bench (to the right). An uninterruptable power supply can be seen in the middle that provides power to the thermocycler. (Photographs taken by J.Q. and S.D.)

Ebola Surveillance

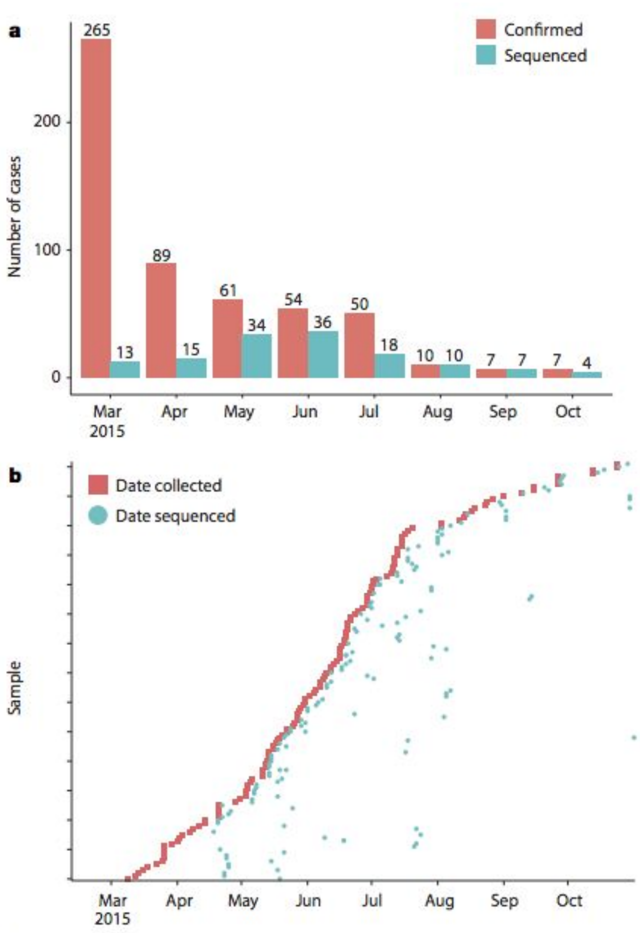


Figure 2 | Real-time genomics surveillance in context of the Guinea Ebola virus disease epidemic. a, Here we show the number of reported cases of Ebola virus disease in Guinea (red) in relation to the number of EBOV new patient samples ($n = 137$, in blue) generated during this study. **b,** For each of the 142 sequenced samples, we show the relationship between sample collection date (red) and the date of sequencing (blue). Twenty-eight samples were sequenced within three days of the sample being taken, and sixty-eight samples within a week. Larger gaps represent retrospective sequencing of cases to provide additional epidemiological context.

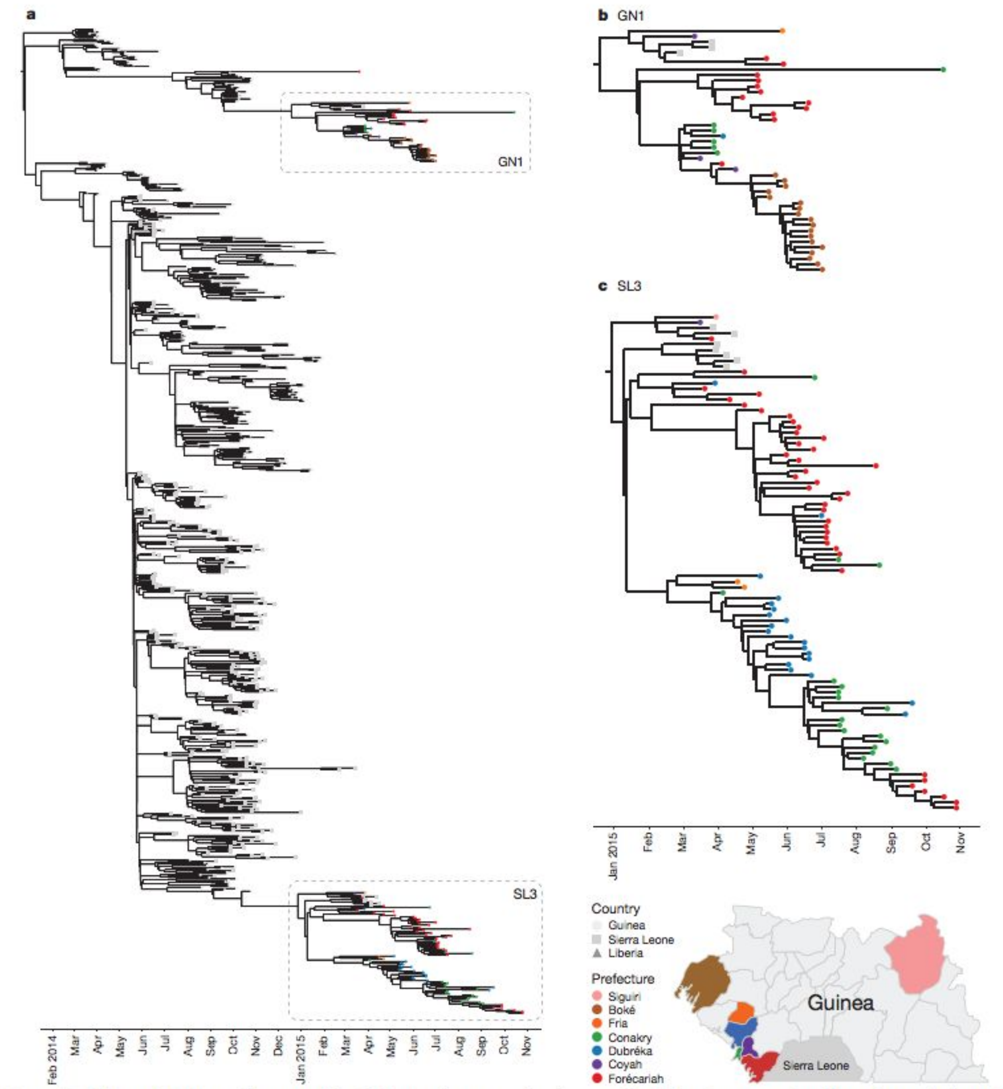


Figure 3 | Evolution of EBOV over the course of the Ebola virus disease epidemic. a, Time-scaled phylogeny of 603 published sequences with 125 high quality sequences from this study. The shape of nodes on the tree demonstrates country of origin. Our results show Guinean samples (coloured circles) belong to two previously identified lineages, GN1 and SL3. **b,** GN1 is deeply branching with early epidemic samples. **c,** SL3 is related to cases identified in Sierra Leone. Samples are frequently clustered by geography (indicated by colour of circle) and this provides information as to origins of new introductions, such as in the Boké epidemic in May 2015. Map figure adapted from SimpleMaps website (http://simplemaps.com/resources/svg-gn).

ReadUntil Sequencing

ONT machines can stop sequencing a read and immediately start on another in real-time

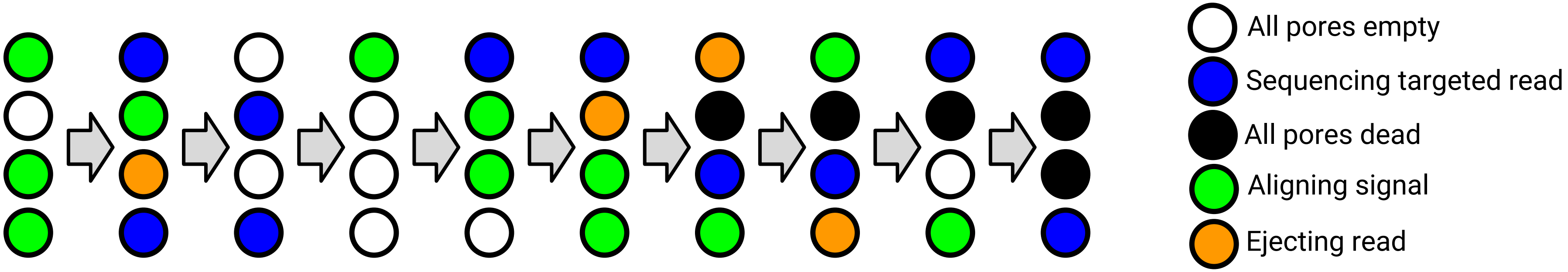
- Each channel has four pores, non-active pores have reads docked

Can potentially avoid sequencing unwanted reads

- For example: reads that align to the human genome, reads that *do not* align to a database of pathogens, reads that align to a region already sequenced to a desired depth

MinION has up to 512 active channels, each reading 450 bp/sec

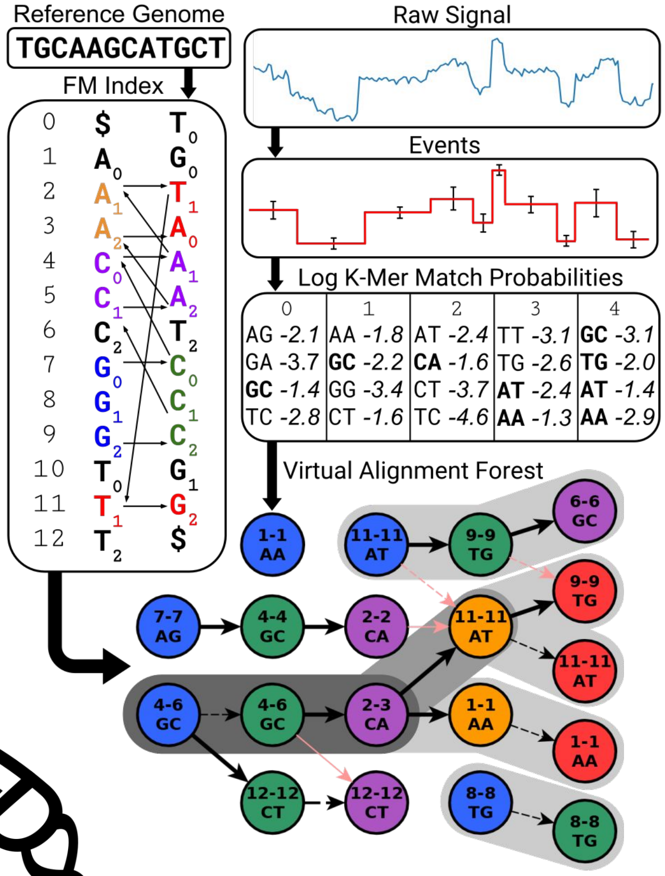
- Actual number of active channels is variable



Utility for **N**anopore **C**urrent **AL**ignment to **L**arge **E**xpanses of **D**NA

AKA UNCALLED

- Probabilistically considers all possible k-mers that the streaming signal could represent
- Finds seeds in the reference consistent with those k-mers using an FM index
- Clusters seeds into potential alignments until one or more locations has sufficient support



Reference Genome

TGCAAGCATGCT

FM Index

0	\$	T_0
1	A_0	G_0
2	A_1	T_1
3	A_2	A_0
4	C_0	A_1
5	C_1	A_2
6	C_2	T_2
7	G_0	C_0
8	G_1	C_1
9	G_2	C_2
10	T_0	G_1
11	T_1	G_2
12	T_2	\$

Raw Signal

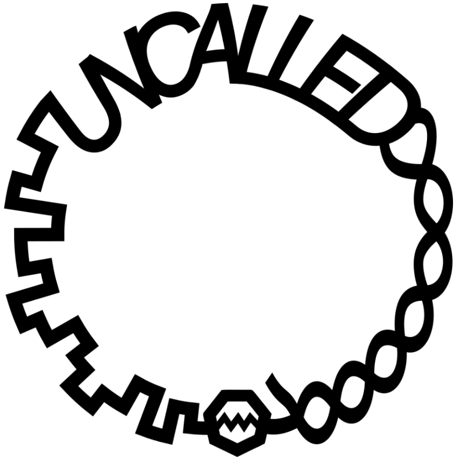
Events

Log K-Mer Match Probabilities

0	1	2	3	4
AG *-2.1*	AA *-1.8*	AT *-2.4*	TT *-3.1*	**GC** *-3.1*
GA *-3.7*	**GC** *-2.2*	**CA** *-1.6*	TG *-2.6*	**TG** *-2.0*
GC *-1.4*	GG *-3.4*	CT *-3.7*	**AT** *-2.4*	**AT** *-1.4*
TC *-2.8*	CT *-1.6*	TC *-4.6*	**AA** *-1.3*	AA *-2.9*

Virtual Alignment Forest

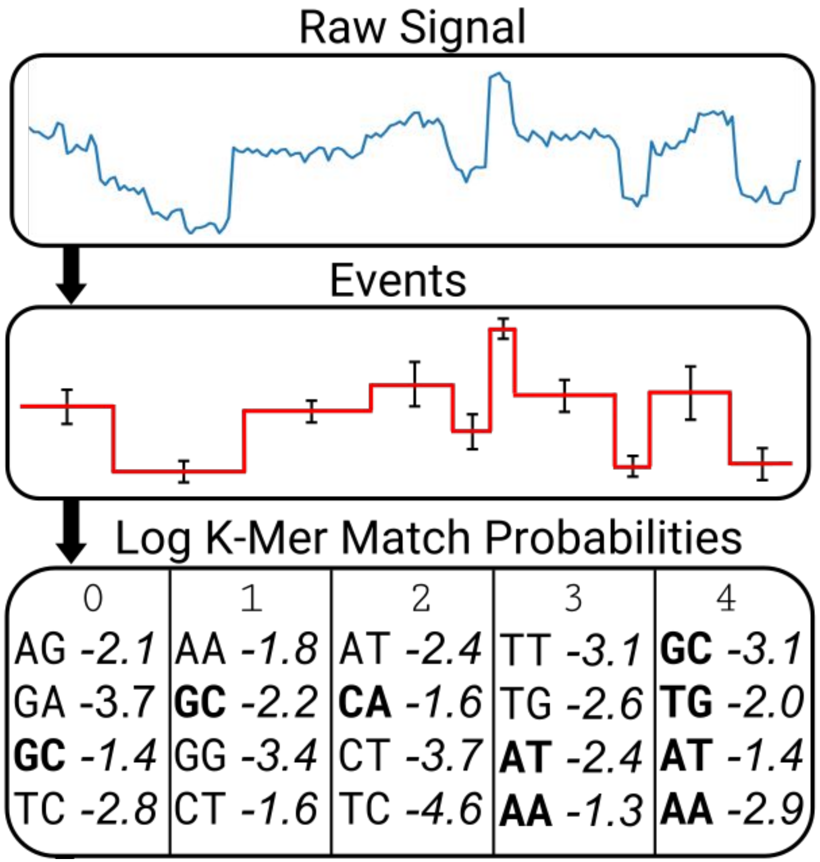
Goal: align reads as fast as ONT machines can sequence them



Started in 2017 as my final project for this class!
(with Taher Mun and Yunfan Fan)

UNCALLED Signal Processing

- Stretches of similar signal are collapsed into **events**
  - Averages out noise and reduces amount of signal to process
- Ideally each event represents a single k-mer, but many errors occur
  - ~50% of events are **stays**
  - ~1% of events are **skips**
  - A read's event length is usually ~2x greater than its basepair length
- Probability of each event matching every k-mer is then computed
  - Expected current for each k-mer modeled by normal distribution
  - ONT releases 6-mer models (I use 5-mers)

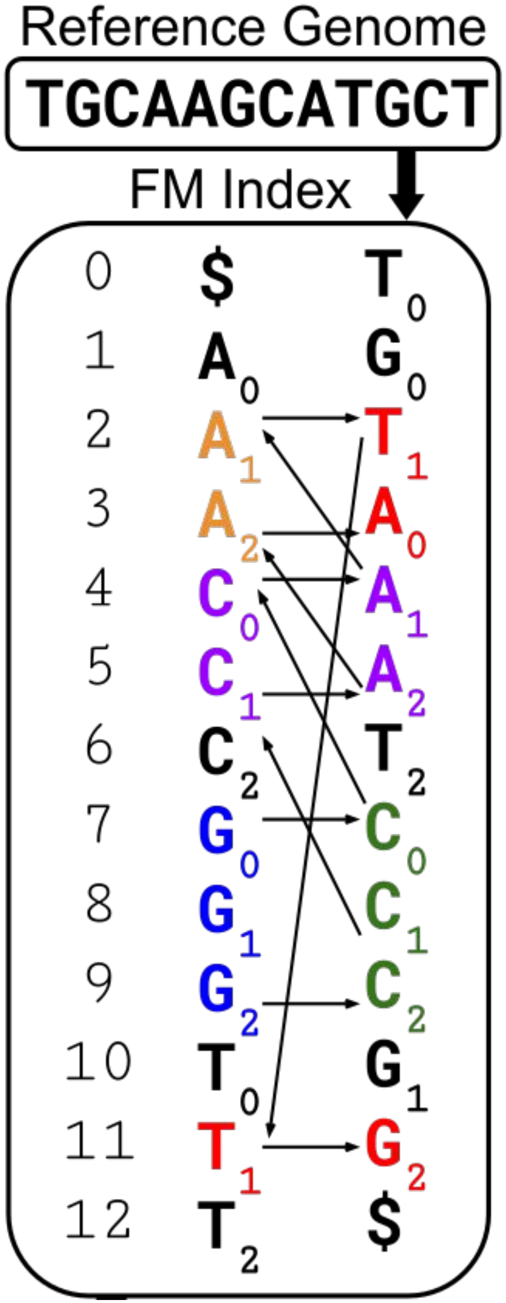


0	1	2	3	4
AG -2.1	AA -1.8	AT -2.4	TT -3.1	**GC** -3.1
GA -3.7	**GC** -2.2	**CA** -1.6	TG -2.6	**TG** -2.0
GC -1.4	GG -3.4	CT -3.7	**AT** -2.4	**AT** -1.4
TC -2.8	CT -1.6	TC -4.6	**AA** -1.3	**AA** -2.9

FM Index

- Used by many aligners such as BWA, Bowtie, and HISAT
- Finds exact string matches of arbitrary length
- Time to align is constant with respect to reference size
- Very small memory footprint
- UNCALLED uses BWA's FM index
  - Interchangeable - started with my own implementation

You will learn more about this soon!



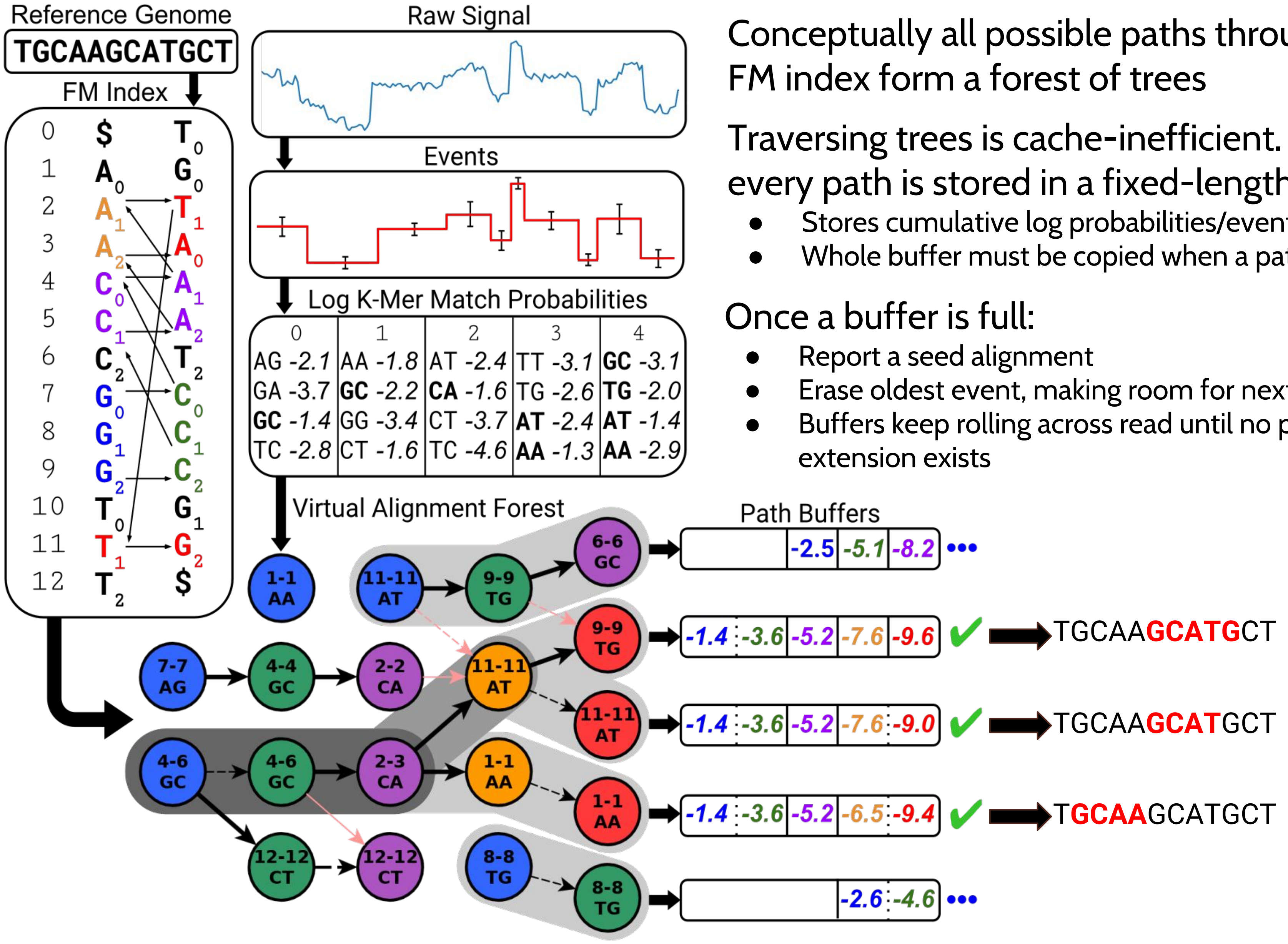
Reference Genome

TGCAAGCATGCT

FM Index

0	$	T_0
1	A_0	G_0
2	A_1	T_1
3	A_2	A_0
4	C_0	A_1
5	C_1	A_2
6	C_2	T_2
7	G_0	C_0
8	G_1	C_1
9	G_2	C_2
10	T_0	G_1
11	T_1	G_2
12	T_2	$

UNCALLED Algorithm



0	$	T_0
1	A_0	G_0
2	A_1	T_1
3	A_2	A_0
4	C_0	A_1
5	C_1	A_2
6	C_2	T_2
7	G_0	C_0
8	G_1	C_1
9	G_2	C_2
10	T_0	G_1
11	T_1	G_2
12	T_2	$

0	1	2	3	4
AG *-2.1*	AA *-1.8*	AT *-2.4*	TT *-3.1*	**GC** *-3.1*
GA -3.7	**GC** *-2.2*	**CA** *-1.6*	TG *-2.6*	**TG** *-2.0*
GC *-1.4*	GG *-3.4*	CT *-3.7*	**AT** *-2.4*	**AT** *-1.4*
TC *-2.8*	CT *-1.6*	TC *-4.6*	**AA** *-1.3*	**AA** *-2.9*

Conceptually all possible paths through the FM index form a forest of trees

Traversing trees is cache-inefficient. Instead every path is stored in a fixed-length buffer

- Stores cumulative log probabilities/event types
- Whole buffer must be copied when a path splits

Once a buffer is full:

- Report a seed alignment
- Erase oldest event, making room for next
- Buffers keep rolling across read until no possible extension exists

E. coli Alignments

Aligned 21K *E. coli* reads to the *E. coli* reference genome

- Reads provided by Winston Timp's lab

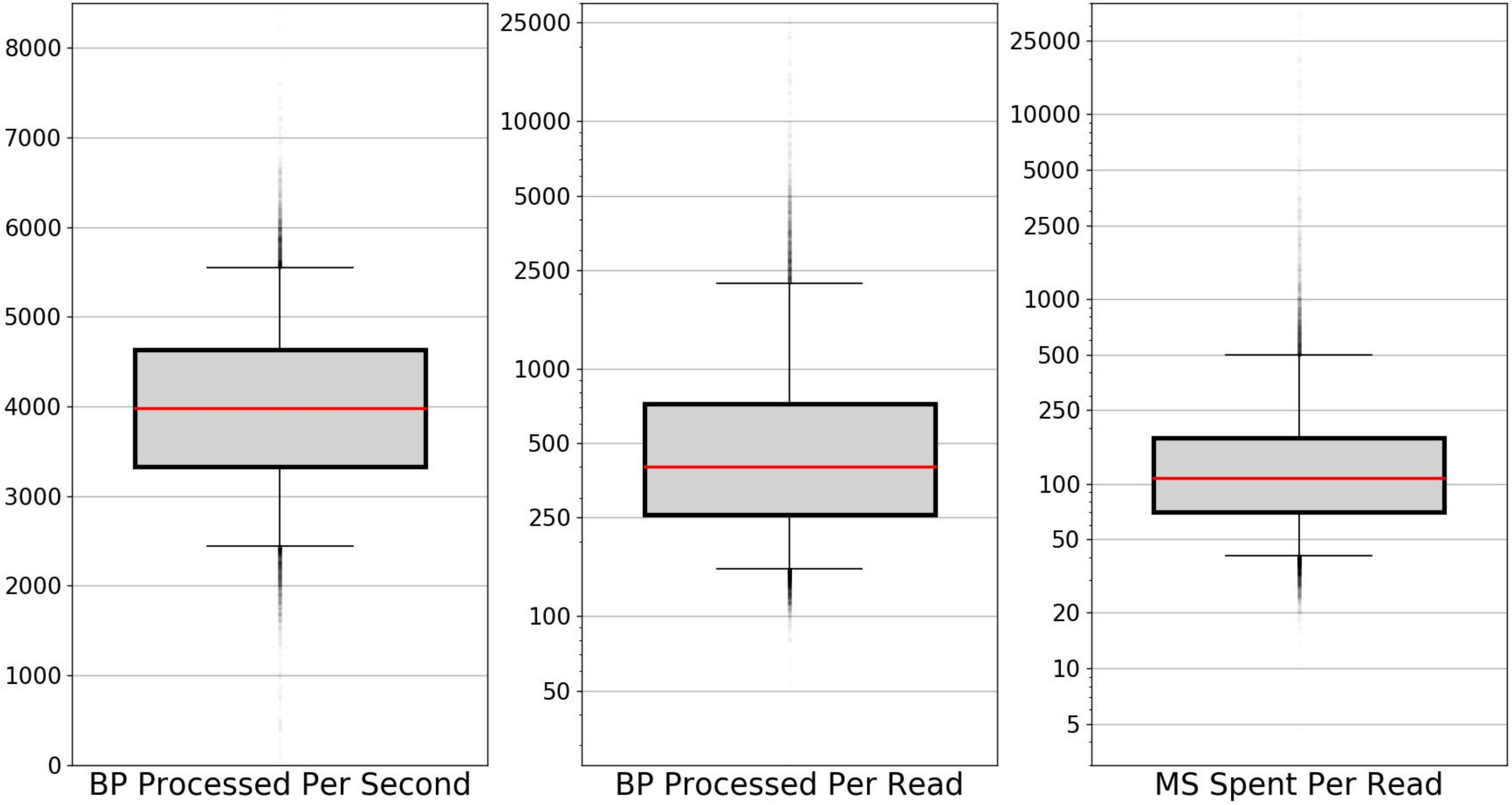
Used minimap2 alignments as ground truth

- FPs and TNs include reads that failed to be basecalled or were unaligned according to minimap2
- Some "false positives" could be alignments found by UNCALLED that minimap2 couldn't find

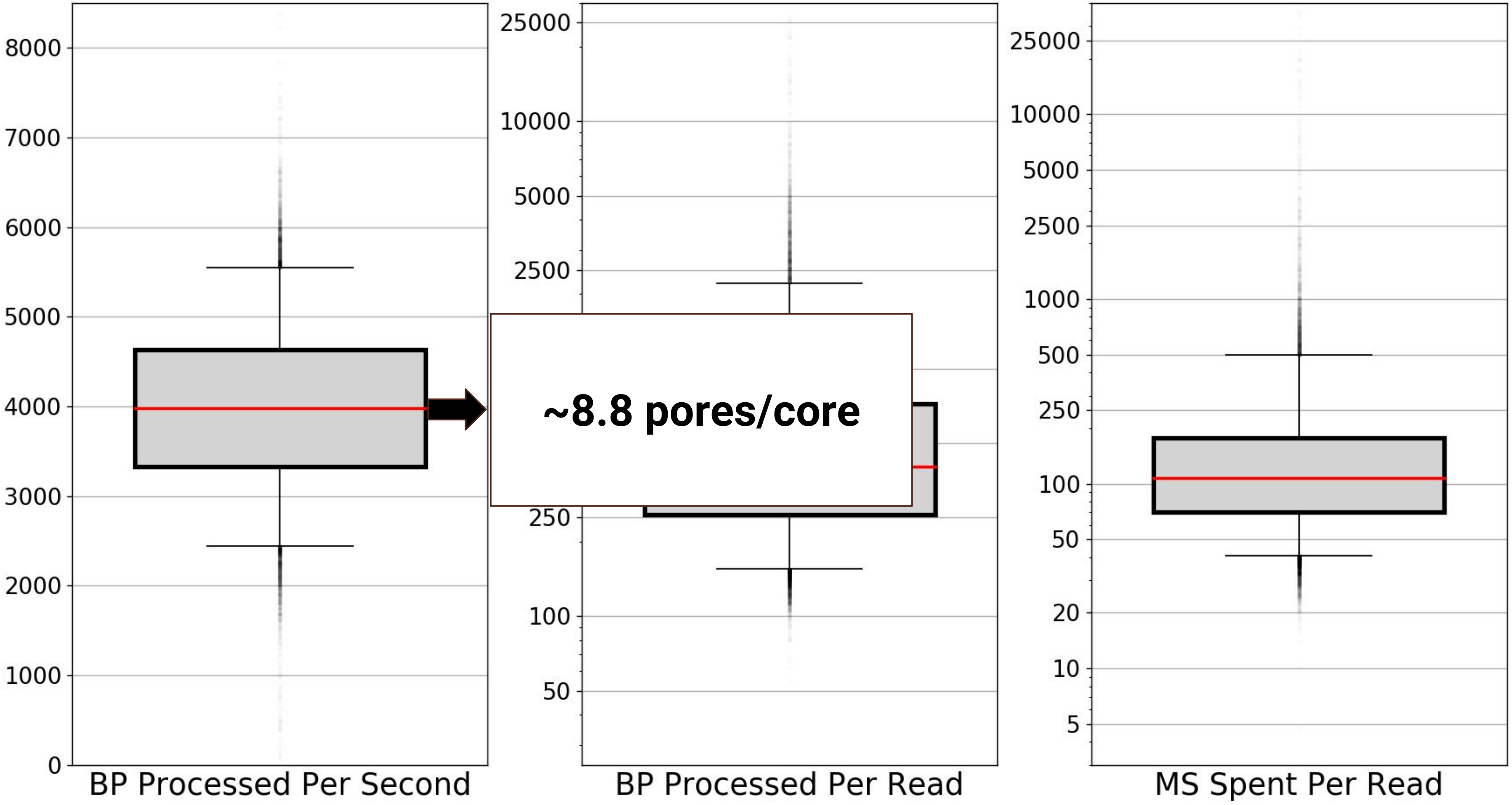
	P	N
T	81.82%	7.99%
F	1.15%	9.04%

50% of "FPs" were
unaligned by m.m.2
or not basecalled

E. coli Timing

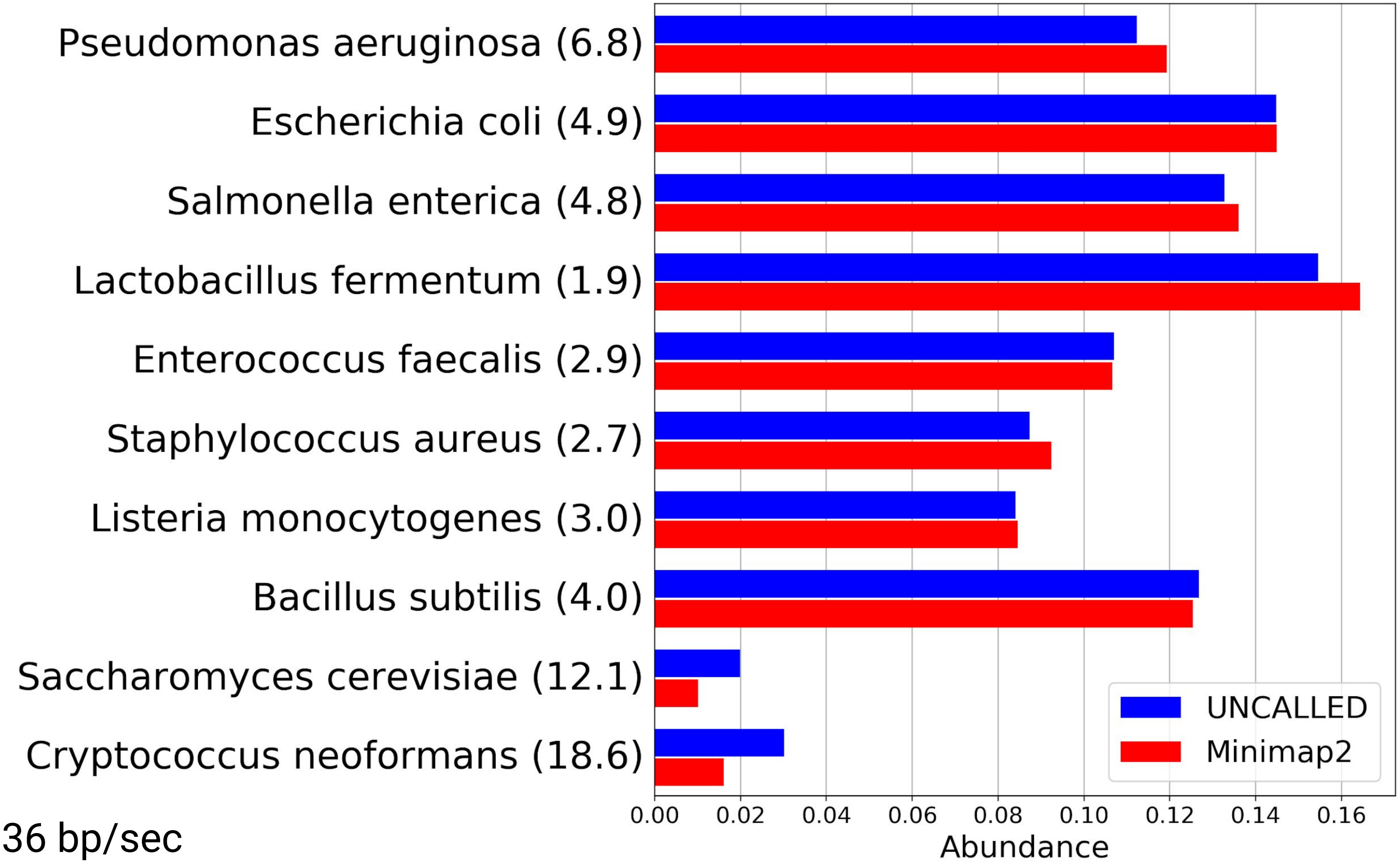


E. coli Timing



~8.8 pores/core

Mock Community Abundance Estimates

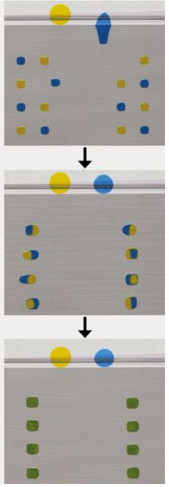


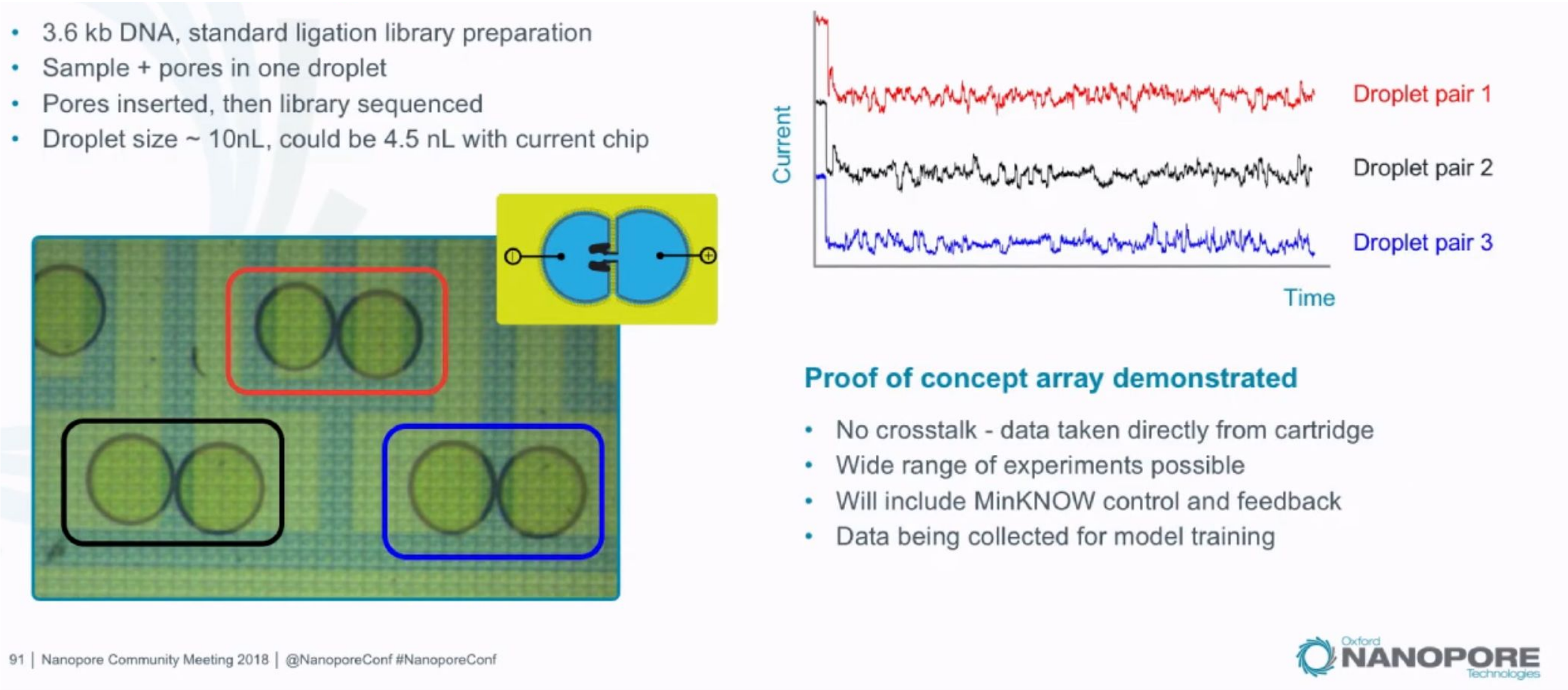
1,136 bp/sec
~2.5 pores/core
3.5x slower than E. coli
reference 12.9x larger

Class Project

- UNCALLED has come a long way since the class project
- How we split the work between the three of us:
  - Collecting/parsing raw nanopore signal data
  - Signal processing/k-mer matching
  - FM Index construction/basic search algorithm
- All of us brainstormed how the algorithm should work
- We did not have a functional aligner in the end
  - Created a signal-based FM index (later turned out to be unnecessary)
  - Figured out how to compute event/k-mer match probabilities (but messed up signal normalization)
  - Could produce seed alignments based on a very simple algorithm (but had no way to filter the many many false positives)
- Despite the incompleteness it was a successful project!

VolTRAX - Library Prep (+ sequencing?)





- 3.6 kb DNA, standard ligation library preparation
- Sample + pores in one droplet
- Pores inserted, then library sequenced
- Droplet size ~ 10nL, could be 4.5 nL with current chip

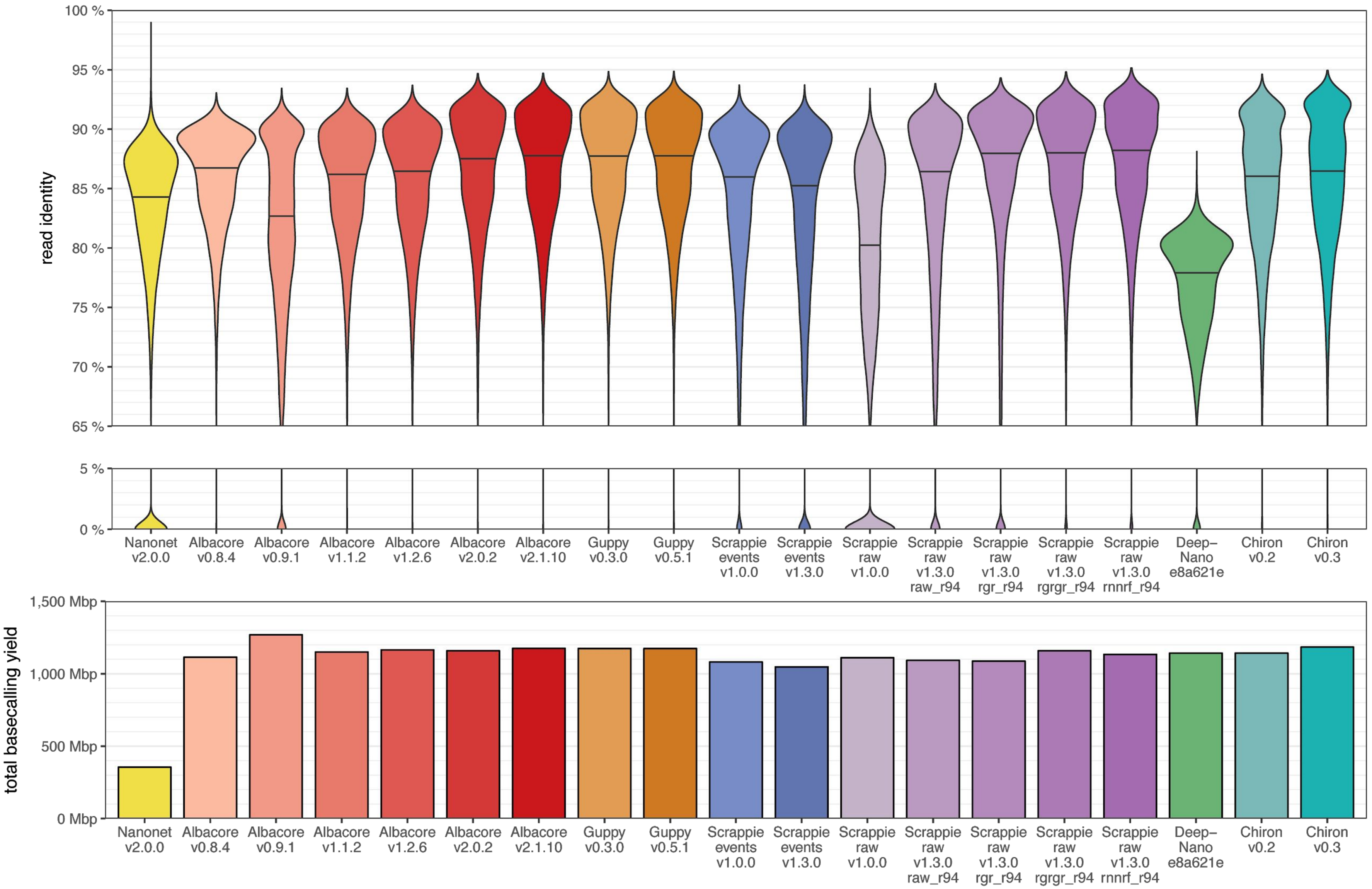
Current
Time
Droplet pair 1
Droplet pair 2
Droplet pair 3

Proof of concept array demonstrated

- No crosstalk - data taken directly from cartridge
- Wide range of experiments possible
- Will include MinKNOW control and feedback
- Data being collected for model training

91 | Nanopore Community Meeting 2018 | @NanoporeConf #NanoporeConf

Basecaller Comparison



https://github.com/rrwick/Basecalling-comparison

Welcome to Applied Comparative Genomics

https://github.com/schatzlab/appliedgenomics2
019

Questions?