Long Read Sequencing

Dick McCombie
Davis Family Professor of Human Genetics
Cold Spring Harbor Laboratory

Advanced Sequencing Technologies and Applications course
Cold Spring Harbor Laboratory
2023

Significant advances in genome sequencing over last 16 years

DNA Sequencing Caught in Deluge of Data

Share full article

W. Richard McCombie, a professor of human genetics at the Cold Spring Harbor Laboratory, examining DNA samples. Kathy Kmonicek for The New York Times

Evolution of genome assemblies

- Initial references – very high quality – extremely expensive
- Period of lower quality Sanger assemblies (~2001-2007)
- Next gen assemblies (short read) – 2007- now
- Third generation – long read assemblies -2013/2014 –now – what can we do currently?
- T2T extremely complete genomes

Goodwin, McPherson and McCombie. Nat. Rev. Genetics. 2016

??

Short vs long reads

- Short read NGS has revolutionized resequencing
- *De novo* assembly is possible but not optimal with short reads
- Long reads improve the ability to do *de novo* assembly dramatically
- Even in organisms with a good reference, such as humans, resequencing misses many structural differences relative to the reference
- Plant genomes are very large in general
- There are significant structural differences between different strains of the same plant such as rice
- These structural differences contribute to salient biological differences

Advantages of Long Read length

Full scale of genetic variation
Repetitive regions
Structural variants
Enables higher quality alignments and assembly
Gapless genomes - T2T

The Telomere-to-Telomere Consortium

Long read sequencing of the hydatidaform mole CHM13 with multiple technologies

-PacBio HiFi

- ONT ultralong reads

- Illumina Arima Genomics Hi-C (Hi-C)

- BioNano optical maps

- single-cell DNA template strand sequencing (Strand-seq)

CHM13 reference

Vastly improves upon the previous "gold standard" reference genome GRCh38

- Introduces nearly 200 million base pairs of sequence
- 1956 new gene predictions, 99 of which are predicted to be protein coding
- Gapless assemblies for all chromosomes except Y
- Corrects errors in the prior reference
- Resolves highly repetitive/ complex regions

Each bar is a linear visualization of a chromosome, with the chromosome number shown at left. Red segments denote previously missing sequences that the T2T Consortium resolved.

GRAPHIC: V. ALTOUNIAN/*SCIENCE*; DATA: T2T CONSORTIUM

Filling the gaps

Laura M. Zahn

Science, 376 (6588), • DOI: 10.1126/science.abp8653

T2T phased diploid genomes

Improve variant detection compared to previous haploid references, simplify genotyping

Pan Genomes provide greater representation of diversity

Liao, WW., Asri, M., Ebler, J. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x

Fig. 1: Verkko assembly workflow.

Rautiainen, M., Nurk, S., Walenz, B.P. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* (2023). https://doi.org/10.1038/s41587-023-01662-6

Limitations of long reads

- Cost
- Throughput*
- Accuracy*
- DNA amount required*
- DNA quality required

*This is rapidly changing

Sample Input Requirements for Long Reads

Instrument	Initial DNA Input	Latest DNA Input
PacBio Sequel HiFi	10ug	1ug
PacBio Revio	5 ug	5 ug
Oxford Nanopore PromethION	5 ug	1ug

For smaller genomes <1Gb less DNA (300-400ng) may be sufficient

If PCR amplification is used, input may be as low as 5ng

Two “flavors” of long read sequencing

PACIFIC BIOSCIENCES®

Oxford NANOPORE Technologies

Significant advances in long read sequencing over last 9 years

Platform Comparisons

Instrument	Avg Readlength	Yield (Gb)	Cost per Gb
Illumina NovaSeqX	300 bp	8000	$2.00
PacBio Sequel II	15-20 kb	30	$18.00
PacBio Revio	15-20 kb	90	$10.00
Oxford Nanopore PromethION	40-80 kb	150	$5.00

PACIFIC
BIOSCIENCES®

PacBio

RSII

- ~85% single pass accuracy
- “short read” CCS accuracy >99.999%
- Up to 2Gb per SMRTcell
- Read lengths up to 60kb

Pacific Biosciences Sequel II

Released in 2018

Smaller, lower cost instrument

8 Million ZMW (155k RSII,1M Sequel I)

Early runs were rocky

Substantial recent improvement in performance up to 200Gb of CLR data or 30Gb of HiFi data

Upto 800Gb CLR or 120Gb HiFi in one week

Pacific Biosciences Revio (Available 2023)

Similar in size to Sequel

25M ZMW (1M Sequel I, 8M Sequel II)

Main focus is HiFi data

Runs 4 chips in parallel

Estimated up to 3Tb of HiFi data per week

Zero-Mode Waveguides Are the Observation Windows

DNA sequencing is performed on SMRT™ Cells, each containing tens of thousands of zero-mode waveguides (ZMWs)

A ZMW is a cylindrical hole, hundreds of nanometers in diameter, perforating a thin metal film supported by a transparent substrate

The ZMW provides a window for observing DNA polymerase as it performs sequencing by synthesis

Individual ZMW

Laser light illuminates the ZMW

DNA Polymerase as a Sequencing Engine

A single DNA polymerase molecule is attached to the bottom of the ZMW

A single incorporation event can be identified against the background of fluorescently labeled nucleotides

ZMW with DNA polymerase

ZMW with DNA polymerase and phospholinked nucleotides

Processive Synthesis with Phospholinked Nucleotides

Enzymatic incorporation of the labeled nucleotide creates a flash of light, which is captured by the optics system and converted into a base call with associated quality metrics using optimized algorithms

To generate consensus sequence from the data, an assembly process aligns the different fragments based on common sequences

LIGHTS ALL ASKEW IN THE HEAVENS; Men of Science More or Less Agog Over Results of Eclipse Observations. EINSTEIN THEORY TRIUMPHS Stars Not Where They Seemed or Were Calculated to be, but Nobody Need Worry. A BOOK FOR 12 WISE MEN No More in All the World Could Comprehend It, Said Einstein When His Daring Publishers Accepted It.

New York Times Nov. 9, 1919.

Yeast: S. cerevisiae W303

PacBio RS II sequencing at CSHL

Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science

Over 175x coverage in 16 SMRT cells using P5-C3

reads

0
5000
10000
15000
20000
25000
30000

Mean: 5910

83x over 10kbp

8.7x over 20kb

Max: 36,861bp

0
10000
20000
30000

S. cerevisiae W303

S288C Reference sequence

- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id

S. pombe dg21

ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id

O. sativa pv Indica (IR64)

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP + CA 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp

Structural Variations in SKBR3

SKRB3 cell line was derived by G. Trempe and L. J. Old in 1970 from pleural effusion cells of a patient, a white, Caucasian female

Most commonly used Her2-amplified breast cancer cell line

Often used for pre-clinical research on Her2-targeting therapeutics such as Herceptin (Trastuzumab) and resistance to these therapies.

(Davidson et al, 2000)

Nattestad, et al, Gen. Res. 2018

Assembly using PacBio yields far better contiguity

Number of sequences:
10,304

Total sequence length:
2.75 Gb

Mean: 266 kb

Max: 15 Mb

N50: 2.17 Mb

NG50: 1.86 Mb

PACIFIC BIOSCIENCES®

Number of sequences:
748,955

Total sequence length:
2.07 Gb

Mean: 2.8 kb

Max: 61 kb

N50: 3.3 kb

NG50: 1.9 kb

illumina®

Her2

PacBio
73X @ 10kb

split reads

Illumina
120X @ 100bp

split reads

8 Mb

Green arrow indicates an inverted duplication.

False positive and missing Illumina calls due to mis-mapped reads (especially low complexity).

Cancer lesion reconstruction from genomic threads

PacBio

chr17

By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome
2. Original translocation into chromosome 8
3. Duplication, inversion, and inverted duplication within chromosome 8
4. Final duplication from within chromosome 8

PacBio errors are randomly distributed

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTGTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTCGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCAGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCGGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGGT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTACTAGCTATCCGATCCTACTGACTTACTATGCT

Enough coverage makes error drop out

PacBio CCS
“HiFi” for longer
(~15-20kb)
fragments

99.99% Accurate

From Wenger et al (2019) Nature Biotechnology

Benefits of long read transcripts

Long read transcripts provide complete isoform information, enables identification of alternative splicing, fusion events, and allows for isoform level phasing

PacBio IsoSeq transcript sequencing

HiFi Reads	4.0 M
HiFi Base Yield	15.1 Gb
Mean HiFi Read Length	3,702 bp
Median HiFi Read Quality	Q39
HiFi Read Mean # of Passes	17

For SPK 3.0 human UHRR Iso-Seq libraries, per-SMRT Cell HiFi read counts typically ranged from ~3.3 Million to ~4.4 Million and HiFi base yields typically ranged from ~11 Gb to ~16 Gb.

High accuracy, low throughput

>300ng total RNA input

Value	Analysis Metric
3,701,876	HiFi Reads
7,241,572,252	HiFi Yield (bp)
1,956	HiFi Read Length (mean, bp)
Q40	HiFi Read Quality (median)
19	HiFi Number of Passes (mean)
386,924	<Q20 Reads
887,074,413	<Q20 Yield (bp)
2,292	<Q20 Read Length (mean, bp)
Q15	<Q20 Read Quality (median)

CSHL run

Oxford NANOPORE Technologies

PromethION

24 independent flowcells

500bp/s sequencing speed

3000 pores per flowcells = 144,000 pores (fully loaded) (MinION cells 512 pores)

On board single or duplex basecalling

>140Gb in CSHL hands

>100M cDNA reads

Up to ~5 Tb fully loaded in one week

Sequencing "flavors" include:

Ligation based - standard methods for gDNA
Q20 - enables higher accuracy including duplex
Barcoding - allows multiplexing up to 96 sample
16S - enables 16S metagenome sequencing
PCR sequencing - long-range PCR for low mass samples
Cas9 - enables Cas9 mediated target enrichment

Rapid - enables library prep <2hrs w/o mechanical shearing
Ultra-long - enables N50s up to 100kb
Native barcoding - PCR-free barcoding to preserve epigenetic marks
Field kit - enables sequencing in the field w/o cold chain
Short fragment - enables sequencing of fragments <1000bp

Oxford Nanopore relies on CsgG and a non-destructive motor protein

Cis side voltage drives DNA through pore

Motor protein mediates DNA unwinding and translocation speed

Ions flow through the pore to change membrane potential

Small changes in measured voltage are translated into k-mers

Nature Biotechnology 30, 326–328 (2012) doi:10.1038/nbt.2181Published online 10 April 2012

Nanopore Sensing Summary

Nanopore = ‘very small hole’

Ionic current flows through the pore Introduce analyte of interest into the pore

Identify target analyte by the characteristic disruption or block to the electrical current

Block or ‘State’, Dwell, Noise

Raw Data and Data Reduction

Nanopore errors are (mostly) randomly distributed

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTTTTCCGATCCTACTGACTTACTATGCT

ATGCTGTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTCGCTAGCTAGCTTTTTTTTT CCGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTTTTCAGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTT CGGATCCTACTGACTTACTATGCT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTTT CCGATCCTACTGACTTACTATGGT

ATGCTCTTCGATCGATGCTGCTAGCTAGCTAGCTTTTTTT CCGATCCTACTGACTTACTATGCT

Enough coverage makes error (mostly) drop out

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. “Detecting DNA cytosine methylation using nanopore sequencing.” Nat Methods. 2017;14(4):407-410. doi:10.1038/nmeth.4184

Figure 1: Differences in event distribution between methylated and unmethylated 6-mers in nanopore sequencing data.

(**a**–**c**) Event distribution for the three sample 6-mers AGGTAG (**a**), AGGTMG (**b**) and TMGAGT (**c**) using R7.3 pore data from PCR-amplified *E. coli* DNA. 'M' represents 5-mC. (**d**–**f**) Data from the R9 pore for the AGGTAG (**d**), AGGTMG (**e**) and TMGAGT (**f**) 6-mers from PCR-amplified *E. coli* DNA. Data from M.SssI-methylated DNA are in green, and data from untreated DNA are in blue. Solid lines indicate Gaussian fits. Events are from the template strand and have been transformed to account for per-read differences with respect to the model (Supplementary Note).

Nanopores are sensitive to DNA modifications which alter the electrical current as the nucleotide transduces the pore. Simpson et al trained a HMM to distinguish signal differences between 5methylC and unmethylated cytosines. They used synthetically modified 6 mers from bacteria to test the shift of methylation at different positions.

Significance of Structural Variants

- Often missed by short read approaches
- May uncover causative mutations underlying disease, especially where typical screening has failed
- SVs account for much more variation in the genome than SNPs
- SVs can alter methylation and gene expression
- However, the functional impact of most SVs remains unknown

Alkan, C., Coe, B. & Eichler, E. Genome structural variation discovery and genotyping. Nat Rev Genet 12, 363–376 (2011) doi:10.1038/nrg2958

Structural Variant Comparison of SKBR3

(Hicks et al, 2006,

- Strong concordance between long read platforms
- Substantially more variants than detected by short reads

Structural Variant Comparison of SKBR3

- PCR validation shows most Illumina-only calls are false positives
- Especially translocations or inversions caused by smaller insertions or deletions

Preliminary Structural Variations Analysis

62bp repeat expansion in BRCA1 detected in normal tissue that is undetectable using a panel or short read sequencing

3,662 SVs specific to the tumor, most are undetectable using short read sequencing

	Total	Deletions	Duplications	Insertions	Inversions	Translocations
All SVs in normal	9816	5225	578	3727	130	156
All SVs in tumor	13737	7020	988	5292	202	235
SVs only in tumor (Also exclude NA12878)	3662	1805	420	1250	98	89

SVs in sample 51 not detected by short reads.

Insertions found in BRCA1 and CHEK2. Insertions and duplications found in NOTCH1.

Living Fossils Project

- Sequence, assemble and annotate reference quality genomes of living fossils and their radiated sister species
- Perform phylogenetic, gene family and comparative genomic analyses
- Interrogate epigenetic mechanisms with smRNA sequencing and transposon analysis

Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

The genome of the Wollemi pine, a critically endangered "living fossil" unchanged since the Cretaceous, reveals extensive ancient transposon activity

Dennis Wm. Stevenson, Srividya Ramakrishnan, Cristiane de Santis Alves, Laís Araujo Coelho, Melissa Kramer, Sara Goodwin, Olivia Mendevil Ramos, Gil Eshel, Veronica M. Sondervan, Samantha Frangos, Cecilia Zumajo-Cardona, Katherine Jenike, Shujun Ou, Xiaojin Wang, Yin Peng Lee, Stella Loke, Maurizio Rossetto, Hannah McPherson, Sebastiano Nigris, Silvia Moschin, Damon P. Little, Manpreet S. Katari, Kranthi Varala, Sergios-Orestis Kolokotronis, Barbara Ambrose, Larry J. Croft, Gloria M. Coruzzi, Michael Schatz, W. Richard McCombie, Robert A. Martienssen

doi: https://doi.org/10.1101/2023.08.24.554647

Wollemi nobilis is a critically endangered gymnosperm recently threatened by wildfires in Australia

Wollemia nobilis Genome Assembly

Previous Assembly with GuppyV3 and wtbg2 assembler	
Genome size	15.6 Gbp
No of Contigs	223,812
N50 Contig-size	312 Kbp
Max Contig-size	7 Mbp
Assembly Quality	Q20 (99%)

Current Assembly with GuppyV4 and Flye assembler	
Genome size	11.66 Gbp
No of Contigs	23,120
N50 Contig-size	8.83 Mbp
Max Contig-size	54.83 Mbp
Assembly Quality	Q31 (99.9%)

Recently published 25Gb Chinese pine genome:
contig N50 of 2.6 Mb

Niu et al 2022 Cell
https://doi.org/10.1016/j.cell.2021.12.006

BioNano Saphyr - Optical Mapping

Total processing time: as few as 3 days*

BioNano Scaffolding Increases Contiguity of Long Read Assemblies

Inclusion of ~100X BioNano data has increased the contiguity of large, complex gymnosperm assemblies

Wollemia ONT only: N50 **8.8Mb**
Wollemia + BioNano: N50 **38Mb**

Gnetum ONT only: N50 **1.8Mb**
Gnetum + BioNano: N50 **20Mb**

GBE
Genome Biology and Evolution

JOURNAL ARTICLE

Long-Read Sequencing Reveals Rapid Evolution of Immunity- and Cancer-Related Genes in Bats

Armin Scheben, Olivia Mendivil Ramos, Melissa Kramer, Sara Goodwin, Sara Oppenheim, Daniel J Becker, Michael C Schatz, Nancy B Simmons, Adam Siepel, W Richard McCombie Author Notes

Genome Biology and Evolution, Volume 15, Issue 9, September 2023, evad148,
https://doi.org/10.1093/gbe/evad148
Published: 20 September 2023 **Article history** ▾

PDF Split View Cite Permissions Share ▾

Abstract

Bats are exceptional among mammals for their powered flight, extended lifespans, and robust immune systems and therefore have been of particular interest in comparative genomics. Using the Oxford Nanopore Technologies long-read platform, we sequenced the genomes of two bat species with key phylogenetic positions, the Jamaican fruit bat (*Artibeus jamaicensis*) and the Mesoamerican mustached bat (*Pteronotus mesoamericanus*), and carried out a comprehensive comparative genomic analysis with a diverse collection of bats and other mammals. The high-quality, long-read genome assemblies revealed a contraction of interferon (IFN)-α at the immunity-related type I IFN locus in bats, resulting in a shift in relative IFN-ω and IFN-α copy numbers. Contradicting previous hypotheses of constitutive expression of IFN-α being a feature of the bat immune system, three bat species lost all IFN-α genes. This shift to IFN-ω could contribute to the increased viral tolerance that has made bats a common reservoir for viruses that can be transmitted to humans. Antiviral genes stimulated by type I IFNs also showed evidence of rapid evolution, including a lineage-specific duplication of IFN-induced transmembrane genes and positive selection in *IFIT2*. In addition, 33 tumor suppressors and 6 DNA-repair genes showed signs of positive selection, perhaps contributing to increased longevity and reduced cancer rates in bats. The robust immune systems of bats rely on both bat-wide and lineage-specific evolution in the immune gene repertoire, suggesting diverse immune

The type I Interferon (IFN) locus is contracted in bats compared to other mammals

Positive selection on the bat ancestral branch suggests strong enrichment of innate immunity genes

Eight of the ten most significant GO terms are related to immunity

DNA repair genes and tumor suppressors are positively selected in bats

Symbol	Name	aBSREL *P*-Value	Sites under Selection	Bat Branch ω	Outgroup Branch ω
CDH1	*Cadherin-1*	2.0e−6	3	0.17	0.29
CAT	*Catalase*	7.0e−6	1	0.38	>10
BIK	*BCL2 interacting killer*	3.7e−5	1	1.92	0.89
PALB2	*Partner and localizer of BRCA2*	1.8e−4	3	0.67	0.43
LATS2	*Large tumor suppressor kinase 2*	2.5e−4	10	0.23	0.11
SLC39A4	*Solute carrier family 39 member 4*	3.8e−4	3	0.55	0.42
SPARCL1	*SPARC like 1*	4.0e−4	4	0.88	0.95
PLA2G7	*Phospholipase A2 group VII*	8.5e−4	5	>10	1.70
GALR1	*Galanin receptor 1*	8.7e−4	1	0.10	0.08
CD79A	*CD79a molecule*	9.0e−4	1	>10	0.37
MYO1A	*Myosin IA*	9.8e−4	7	0.94	0.14

HEALTH

Bats could play vital role in preventing and treating cancer: 'first step' discovery

By Rob Bailey-Millado

Published Sep. 22, 2023, 12:28 p.m. ET

Bats get a deadly bad rap — but they could play a lifesaving role in cancer treatment, according to a "fascinating new study" that examines their remarkable immune systems.

As documented in peer-reviewed research in Genome Biology and Evolution, published Thursday by Oxford University Press, scientists from New York hope that by better understanding the winged creatures' extraordinary ability to both host and survive infections, they can develop ways to treat — and prevent — cancer in humans.

It might sound bat crazy, but these disease-riddled cave-dwelling creatures of the night might hold the key to curing CANCER

- **Bats are recognized for their ability to tolerate viruses and low rates of cancer**
- **READ MORE: Now the CDC wants to monitor your POOP to track flu outbreaks**

By BRITNEY NGUYEN FOR DAILYMAIL.COM

PUBLISHED: 07:31 EDT, 21 September 2023 | **UPDATED:** 08:47 EDT, 21 September 2023

Long read sequencing uncovers DNMs in Autism quad

Long reads, increased detection of De Novo variants including SNPs, InDels and SVs

Further improvement when mapping to the more complete CHM13 reference

Michelle D. Noyes, William T. Harvey, David Porubsky, Arvis Sulovari, Ruiyang Li, Nicholas R. Rose, Peter A. Audano, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Tuomo Mantere, Tina A. Graves-Lindsay, Ashley D. Sanders, Sara Goodwin, Melissa Kramer, Younes Mokrab, Michael C. Zody, Alexander Hoischen, Jan O. Korbel, W. Richard McCombie, Evan E. Eichler

Familial long-read sequencing increases yield of de novo mutations

The American Journal of Human Genetics, Volume 109, Issue 4

Long Read Sequencing of Early Onset Cancer Pedigrees

SV Filtering Workflow

- Each individual will have ~25000 SV calls per genome
- Merge and genotype all calls across family members (~34000)
- Filter by family structure (~1400)
- Pull variants in/near genes (~450)
- Filter common events (~300)
- Filter false positives / ambiguous events/ select likely genes (~<100)

- No family history of cancer
- Standard IMPACT panel did not detect drivers

Homozygous intronic gene deletion in proband, heterozygous in healthy parents

Collaboration with Zsofia Stadler MSKCC

Assess regulatory impact of intronic events

Many events are intronic, so prioritize those that overlap potential regulatory regions

BCL2 suppresses apoptosis, increased expression leads to multiple types of cancer

Structural Variant overlaps an ENCODE enhancer

Long Read Sequencing of Early Onset Cancer Pedigrees

ONT signal data allows for direct detection of methylation state

Hypermethylation of promoter region of tumor suppressor in proband compared to healthy parents

Collaboration with Zsofia Stadler MSKCC

Phasing Variants and Methylation

Long reads allow phasing of large blocks (~1-10Mb) without use of trio information. (Read length and coverage)

Jyun-Hong Lin, Liang-Chi Chen, Shu-Qi Yu and Yao-Ting Huang, LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants, Bioinformatics, 2022.

Phasing methylation provides allele specific context

Proband

Unaffected parent

Unaffected parent

IGV showing increased methylation in proband (colored bases are “unprotected” converted bases)

PEG3 downregulated in colon cancer

Overlaps Enhancer region

Improved T2T reference genome uncovers new variants

ZNF793 deletion overlaps last exon
Has been associated with lung cancer and Barrett's esophagus

Break in the alignment of CHM13 compared to hg38 in this region

Colon cancer trio 2

Improved T2T reference genome uncovers new variants

Colon cancer trio 1

Intronic deletion in RPH3AL gene (homozygous in proband)

Rabphillin-3A-like is a likely tumor suppressor gene, involved in medulloblastoma and down regulated in bladder cancer

Region unique to CHM13 compared to hg38

Nanopore error rates are improving

Chemistry and software improvements are increasing accuracy

ONT simplex reads (single read):
~1-3% error (Q18-Q20)

ONT duplex reads (sequence template and its complement):
approaching Q30, lower yield but improving

ONT software improvements are increasing base quality

"Guppy V5" sup accuracy model

"Guppy V4" high accuracy model

Nanopore Quality Distribution
Read Density
0.150
0.100
0.050
0.000
10
20
30
40
Qualscore
run
GuppyV6_duplex
GuppyV6_simplex

Yield vs N50
160,000,000,000
120,000,000,000
80,000,000,000
40,000,000,000
0
90,000
67,500
45,000
22,500
0
2020
2021
UltraLong Kit
Flowcell Pass Yield (bp)
N50 readlength

Transcriptome Sequencing on Oxford Nanopore

ONT PCR cDNA

Low input (~1ng poly A+)

PCR may introduce biases

Enriched for full length cDNA (template switching)

Multiplex up to 24 samples

Much higher throughput (>60 million reads per PromethION cell, up to ~180M)

Lengths ~700bp

Recent paper shows 40 fold fewer long reads (8 fold fewer bases) are required to cover 6000 genes across 95%

Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for Transcript Profiling Using Long-Read Technologies. Front Genet. 2020 Jul 7;11:606.

ONT direct cDNA

Requires more input (~100ng poly A+)

Does not use PCR

Enriched for full length cDNA (template switching)

Can multiplex with native barcoding kit

Less throughput (20-40 million reads per PromethION cell), lengths a bit longer ~1.5kb

ONT direct RNA

Many tools have been/are being developed to use the raw ONT signal data to detect modifications

Tombo (Stoiber et al 2017)
Nanocompore (Leger et al 2021)
xPore (Pratanwanich et al 2021)
nanoRMS (Begik et al 2021)

Input requirement (500 ng total RNA or 50 ng poly-A_+ RNA)

RNA length preserved

No PCR, RT is optional

Can detect base modifications (6mA, 5mC)

Output is much lower, 6-10 million reads on PromethION

Lengths 1.5-2kb

Ribosomal RNA depletion is an issue

Recent PromethION sequencing metrics

PCR cDNA cell

169 million reads
N50 readlength 899bp
108Gb total yield

Average human transcript ~2kb

Long reads span junctions and provide connection

Transcript Coverage

StringTie2 (Kovaka 2019) used to assemble transcripts and detect genes where transcripts were fully covered end to end by reads

Illumina data - 16,476 genes

ONT full data set - 25,478 genes

ONT 50pct downsample - 21,322 genes

ONT 75pct downsample - 19,088 genes

Transcriptome Sequencing Cost Comparisons Across Platforms

Illumina cost per million reads
$6

ONT PCR cDNA cost per million reads
$10

PacBio IsoSeq cost per million reads
$600

Tradeoffs on accuracy and length, so it is key to assess the method that will address critical questions of your experiment

Targeting Sequencing Methods for ONT

Molecular Method

- Enrich for regions of interest from each sample prior to sequencing
- ACME capture, nCATS, PCR

Computational Method

- Utilize “adaptive sampling” capability to selectively eject molecules from the pore
- Reads are mapped to a reference in real time
- Enrich for targets or deplete unwanted regions

These methods may be used separately or combined for further enrichment

Conclusions

- Substantial variation is missed by short reads
- Phasing can reveal more detailed context of variation
- Interpretation of SVs must be optimized to understand regulatory effects of widespread variation

Summary

Long read platforms have matured significantly in the last few years
PacBio and Oxford Nanopore producing similar length distributions
Overcome high error sequencing with improved informatics
Oxford Nanopore exciting for methylation & direct RNA capabilities

Long reads are crucial for accurate SV calling
Finding thousands to tens of thousands of additional SVs over short reads
Resolves the false positives observed with short reads
Detecting potential cancer risk factors that would otherwise go unnoticed

Sample & DNA requirements one of the largest barriers for clinical application
Continue to advance protocols for extracting, preparing samples
Organoids (as opposed to primary tumors) enable large DNA amounts for long read sequencing, though it remains much more difficult then cell culture
Organoids also enable application and profiling of other molecular and pharmaceutical assays

Future goals

Reduce sample DNA input - tumors, single cell, targeting - Shruti Iyer
Analyze data from projects for relevant genome properties
Improve long read sequencing efficiency - read length, yield, combination of input data types
Optimum cost benefit analyses of different long read approaches and coverage
Optimize long read transcriptome sequencing

Acknowledgements

McCombie Lab

Sara Goodwin
Melissa Kramer
Olivia Mendivil Ramos
Stephanie Muller
Robert Wappel
Senem Mavruk
Elena Ghiban
Shruti Iyer

Siepel Lab

Armin Scheben

Spector Lab

Sonam Bhatia
Gayatri Arun

Schatz Lab

Sam Kovaka
Melanie Kirsche
Rachel Sherman
Katie Jenike
Sergey Aganeov
Srividya Ramakrishnan

Timp Lab

Isac Lee

AMNH

Nancy Simmons
Sara Oppenheim

Fritz Sedlazeck

Medhat Helmy

Mayo Clinic

Mark Ebbert

Living Fossils Consortium

Karen Kostroff

Memorial Sloan Kettering Cancer Center

Zsofia Stadler
Matilde Borio
Zalak Patel
Sami Belhadj

Funding

NCI
NSF
NHGRI
Northwell Health