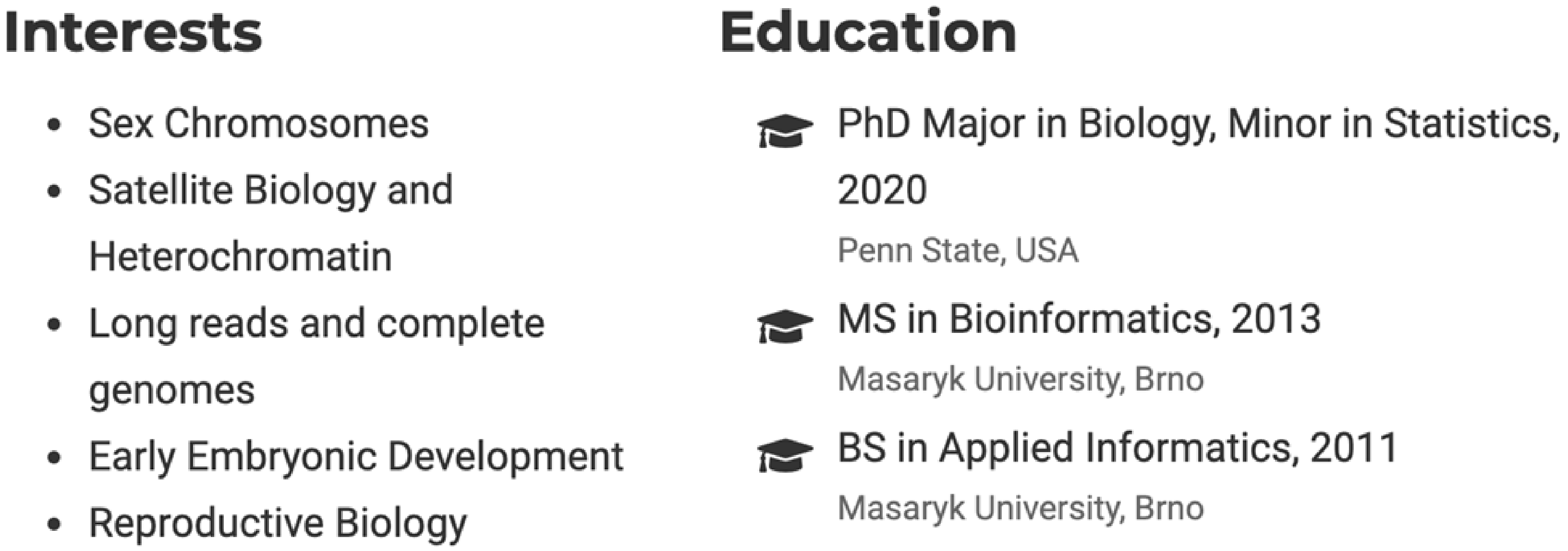
Repetitivní genom, metody NGS za využití dlouhých "readů"

Monika Čechová

@biomonika
biomonika.org



Interests

- Sex Chromosomes
- Satellite Biology and Heterochromatin
- Long reads and complete genomes
- Early Embryonic Development
- Reproductive Biology

Education

- PhD Major in Biology, Minor in Statistics, 2020

  Penn State, USA
- MS in Bioinformatics, 2013

  Masaryk University, Brno
- BS in Applied Informatics, 2011

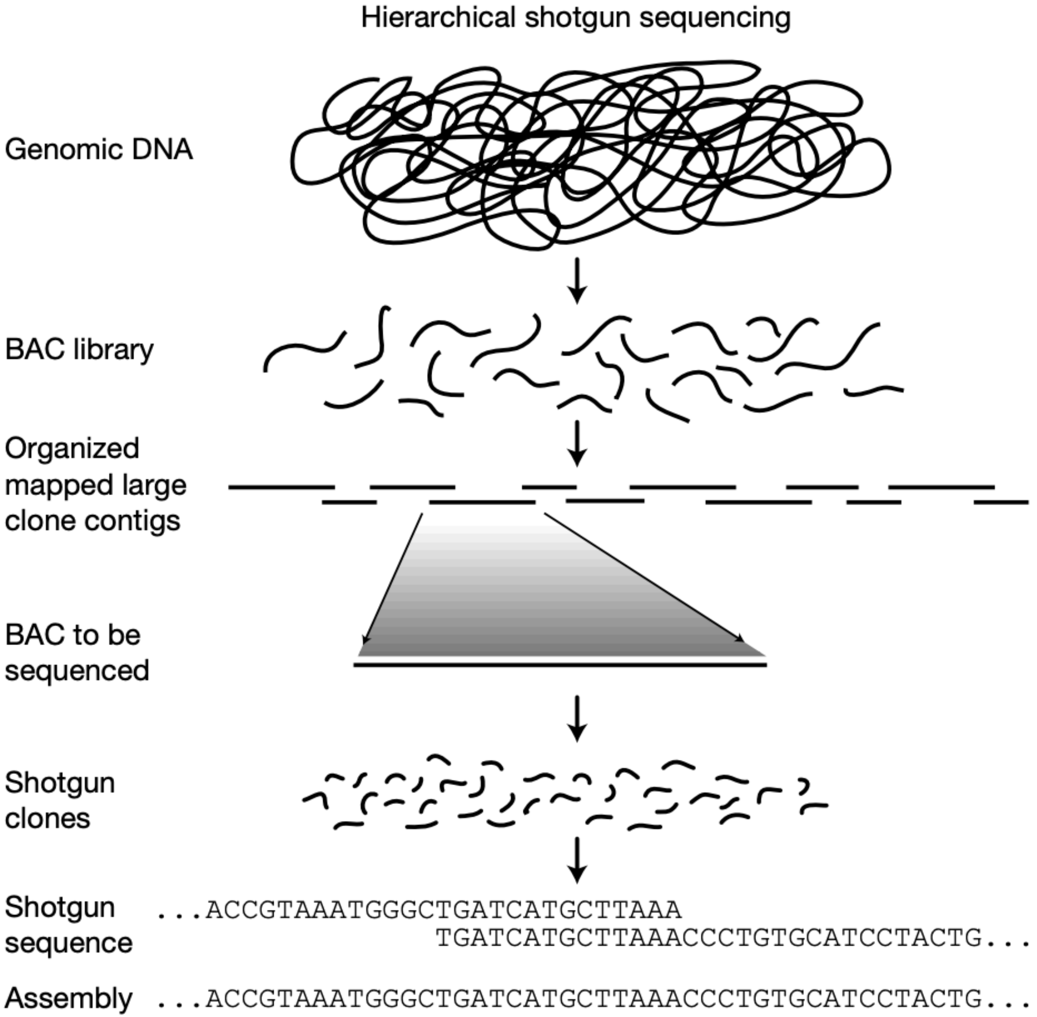
  Masaryk University, Brno

My motivation:
Answering biological questions using next-generation sequencing data

- Progress in the human genome assembly and analysis
- Overview of the sequencing technologies
- Long-read applications
- Error correction of long reads
- Structural variation detection: the basics

Progress in the human genome assembly and analysis

HUMAN GENOME: WHERE WE STARTED



Hierarchical shotgun sequencing

Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

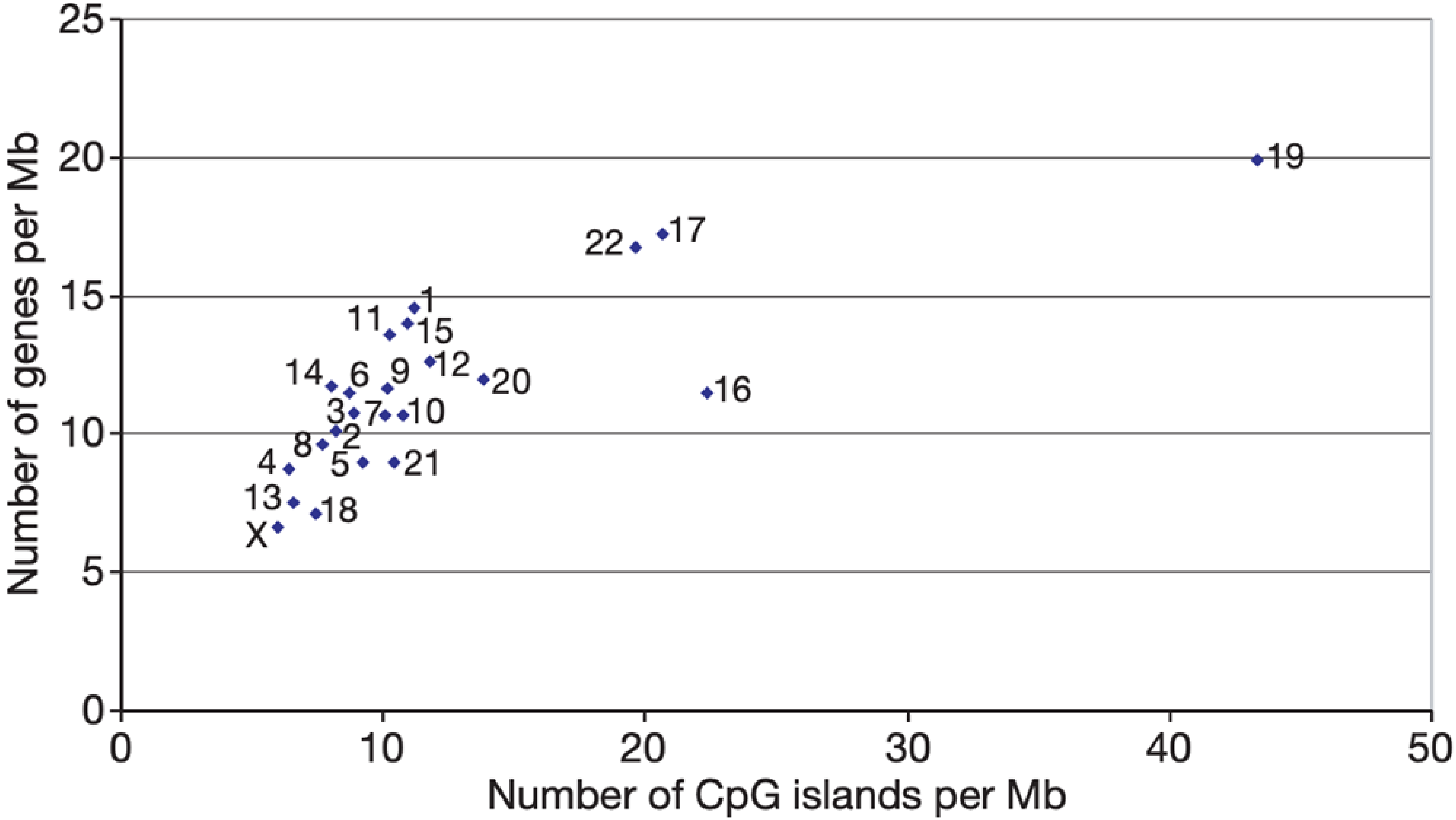
Shotgun sequence

```
...ACCGTAAATGGGCTGATCATGCTTAAA
                 TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

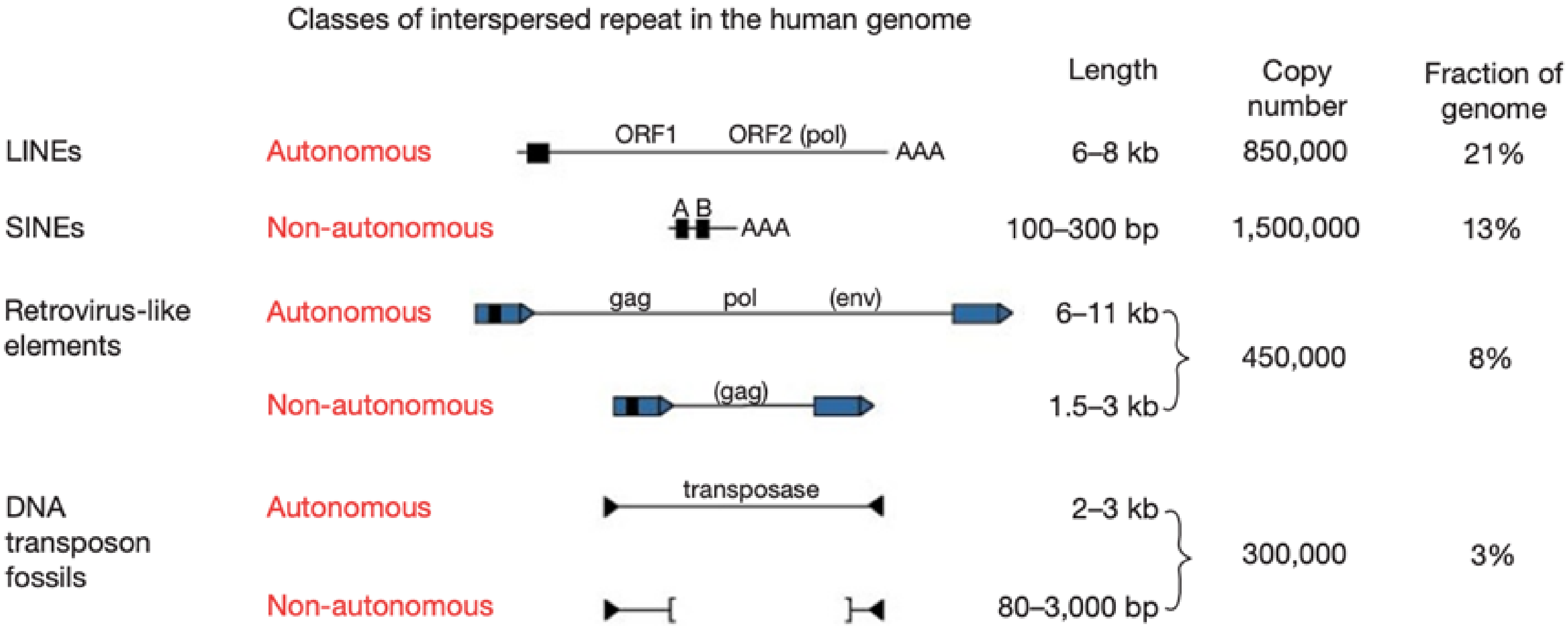
Assembly

```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. doi: 10.1038/35057062



Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. doi: 10.1038/35057062



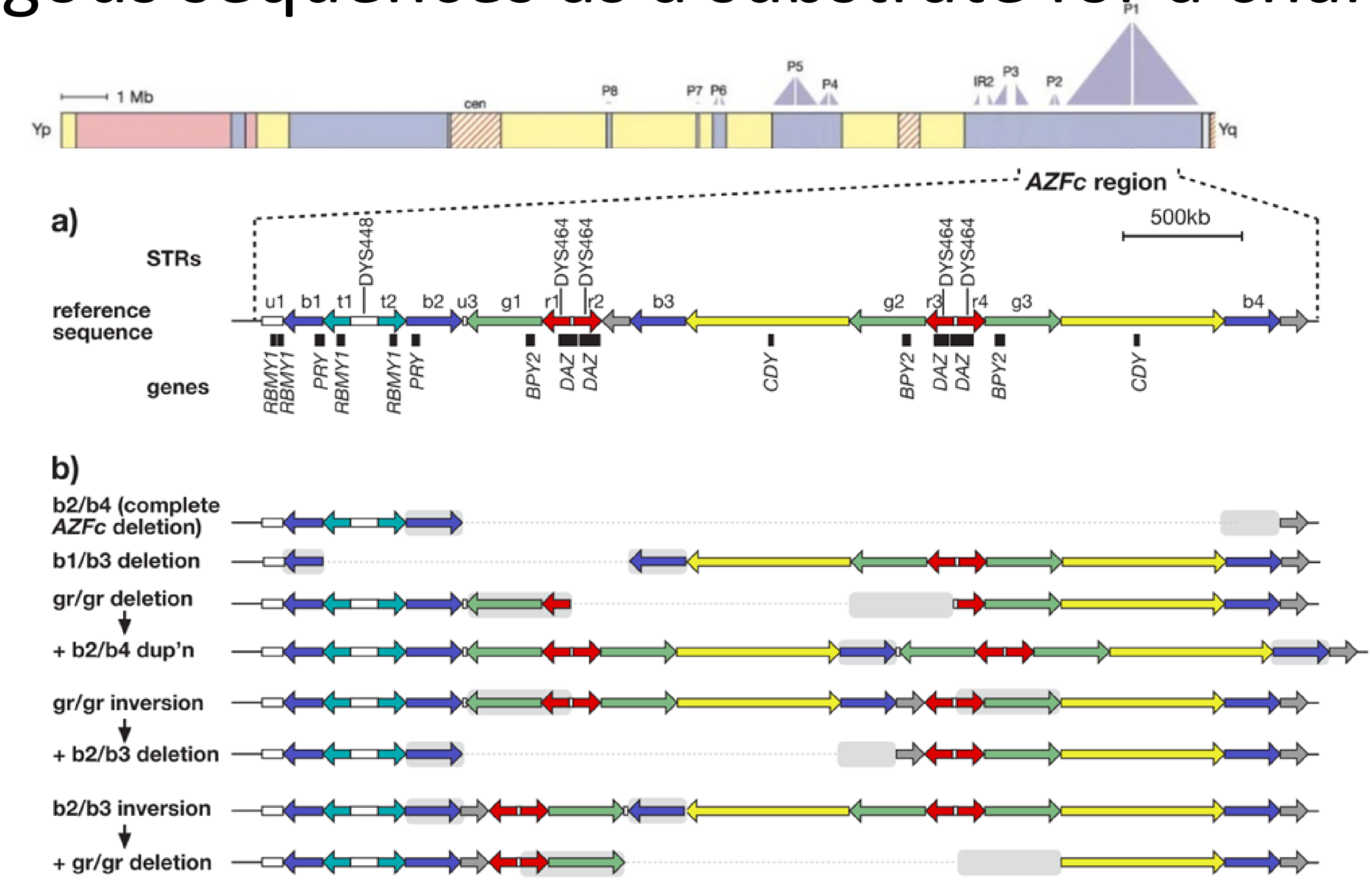
Classes of interspersed repeat in the human genome

		Length	Copy number	Fraction of genome
LINEs	Autonomous	6–8 kb	850,000	21%
SINEs	Non-autonomous	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	6–11 kb	450,000	8%
	Non-autonomous	1.5–3 kb		
DNA transposon fossils	Autonomous	2–3 kb	300,000	3%
	Non-autonomous	80–3,000 bp		

Figure 17 Almost all transposable elements in mammals fall into one of four classes. See text for details.

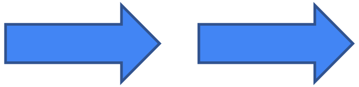
Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. doi: 10.1038/35057062

Homologous sequences as a substrate for a change

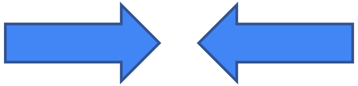


Jobling et al., 2008

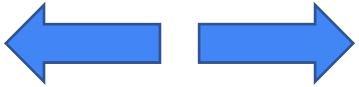
Homologous sequences as a substrate for a change



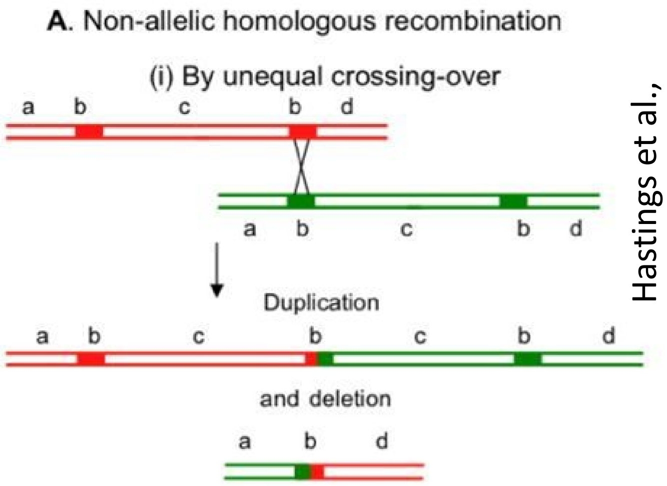
COPY NUMBER CHANGE



INVERSION



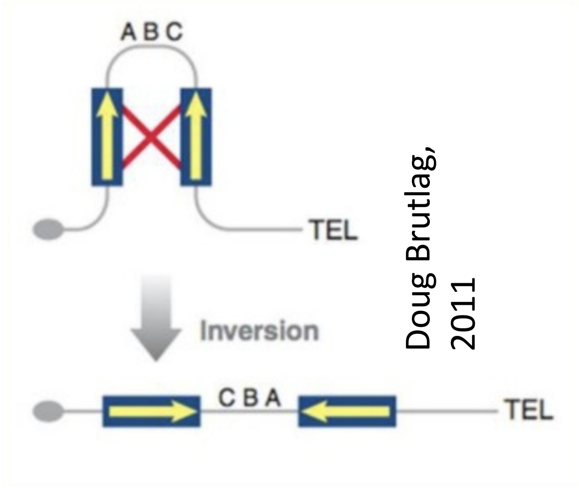
GENE CONVERSION



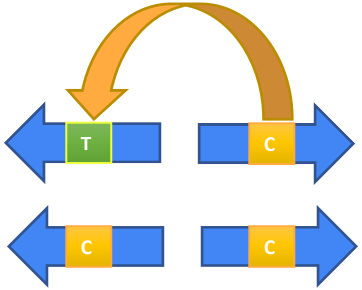
A. Non-allelic homologous recombination
(i) By unequal crossing-over
a b c b d
a b c b d
Duplication
a b c b c b d
and deletion
a b d

Hastings et al., 2009

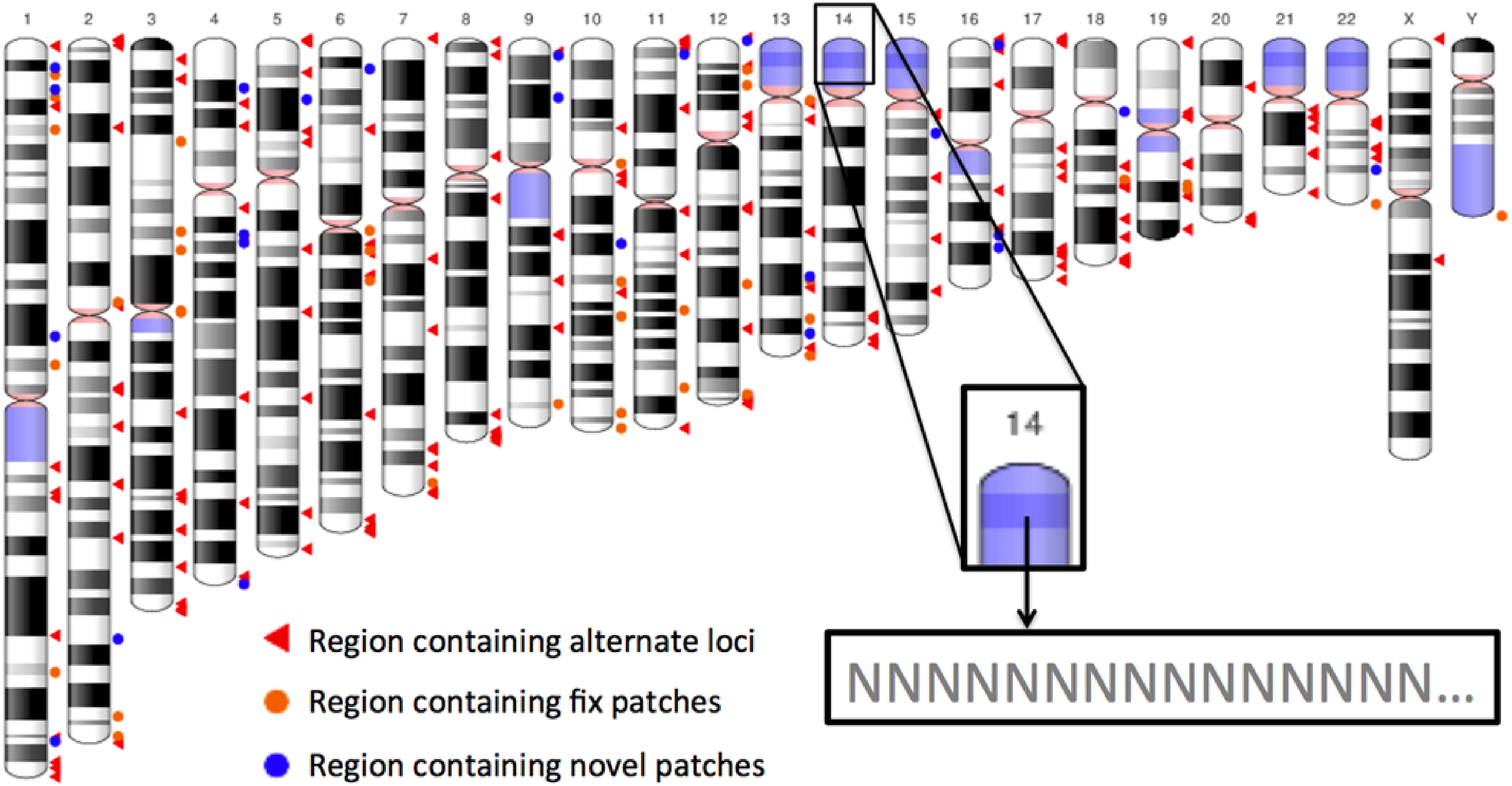
NAHR



Doug Brutlag, 2011

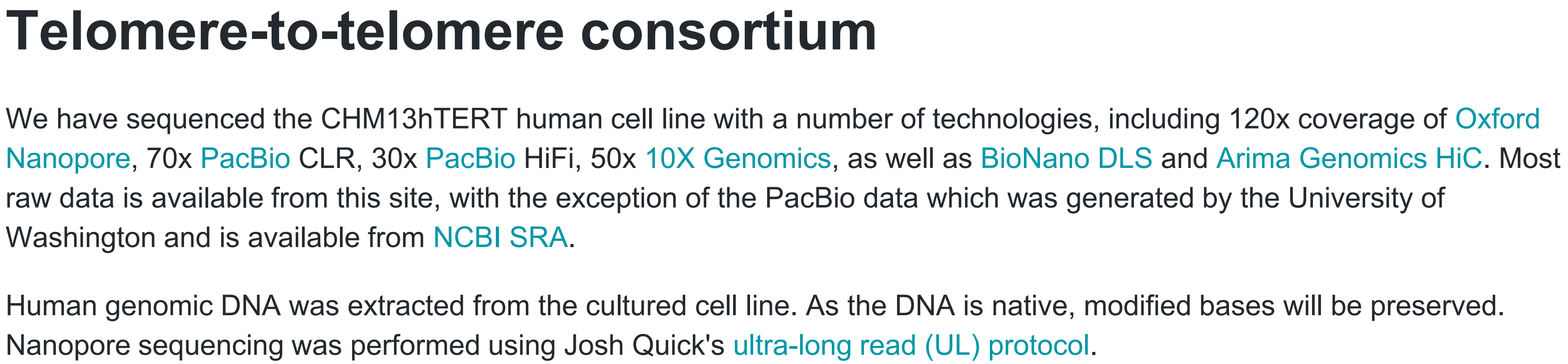


Human genome is incomplete


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
14
NNNNNNNNNNNNNNNNN...
◀ Region containing alternate loci
● Region containing fix patches
● Region containing novel patches

https://github.com/broadgsa/gatk/blob/master/doc_archive/dictionary/Reference_Genome_Components

HUMAN GENOME: WHERE WE ARE HEADING

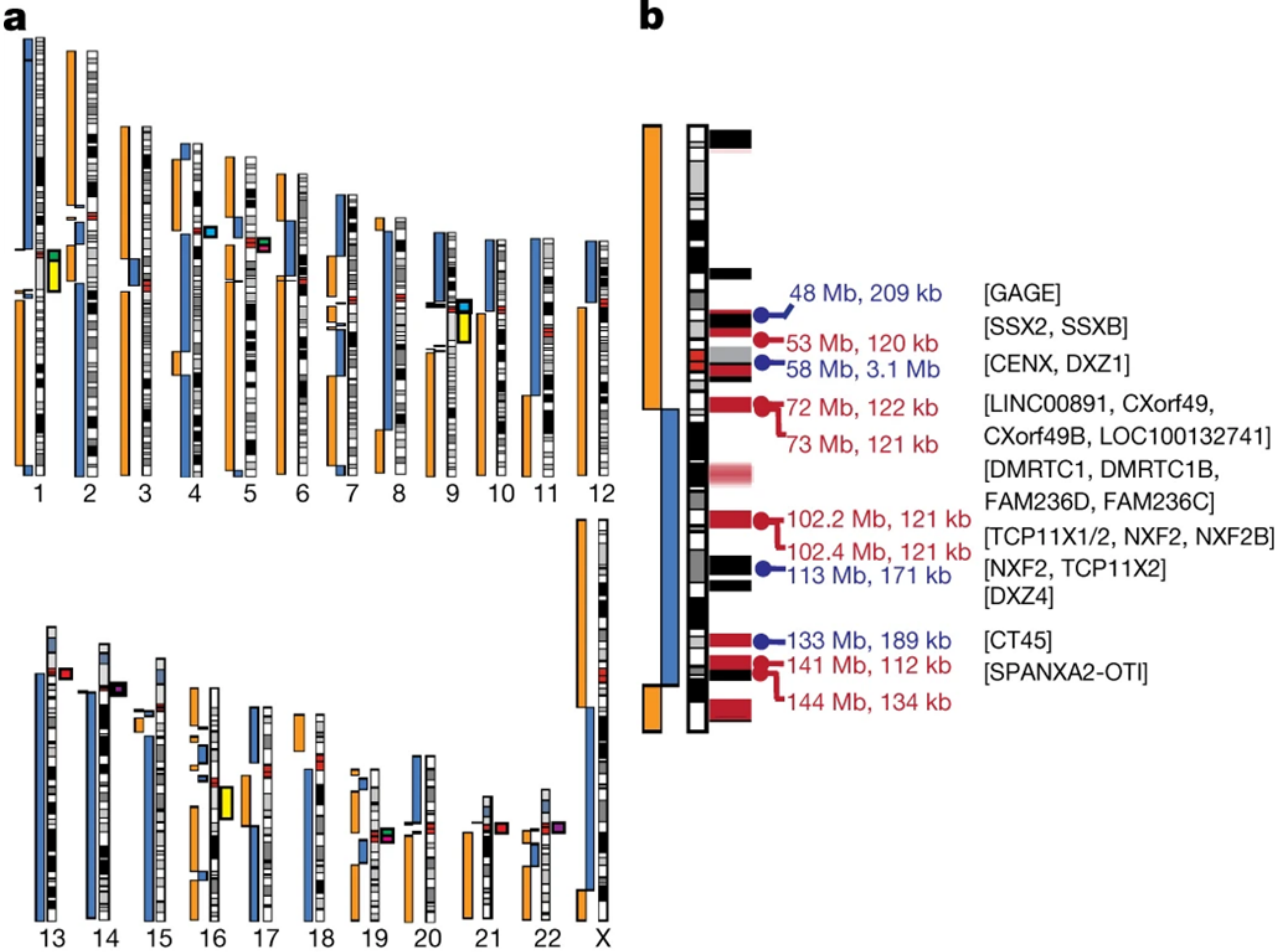


Telomere-to-telomere consortium

We have sequenced the CHM13hTERT human cell line with a number of technologies, including 120x coverage of Oxford Nanopore, 70x PacBio CLR, 30x PacBio HiFi, 50x 10X Genomics, as well as BioNano DLS and Arima Genomics HiC. Most raw data is available from this site, with the exception of the PacBio data which was generated by the University of Washington and is available from NCBI SRA.

Human genomic DNA was extracted from the cultured cell line. As the DNA is native, modified bases will be preserved. Nanopore sequencing was performed using Josh Quick's ultra-long read (UL) protocol.

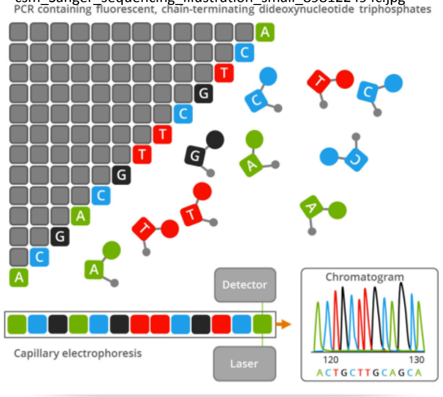
The X chromosome was selected for manual assembly, and was initially broken at three locations: the centromere (artificially collapsed in the assembly), a large segmental duplication (DMRTC1B, 120 kb), and a second segmental duplication with a paralogue on chromosome 2 (134 kb). Gaps in the GRCh38 reference (black) and known segmental duplications (red; paralogous to Y, pink) are annotated. Repeats larger than 100 kb are named with the expected size (kb) (blue, tandem repeats; red, segmental duplications).



Overview of the sequencing technologies

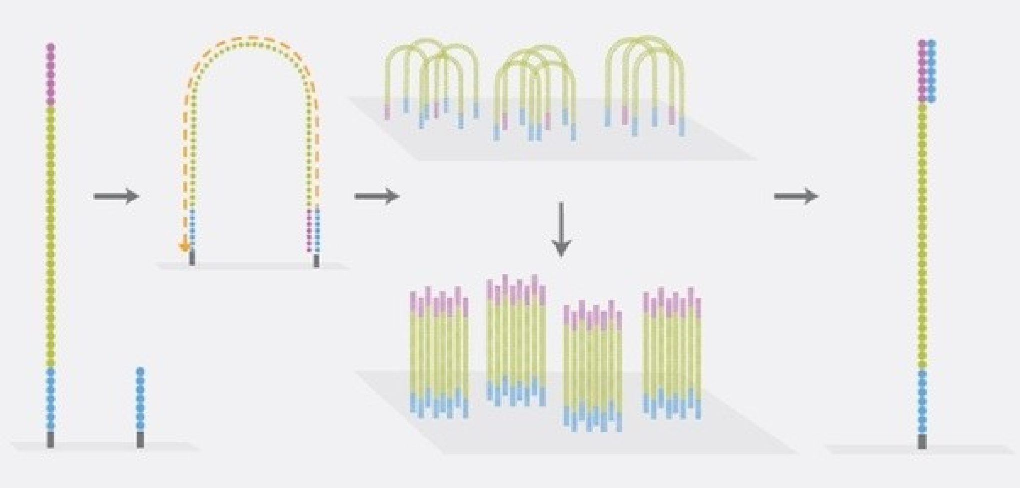
Outline

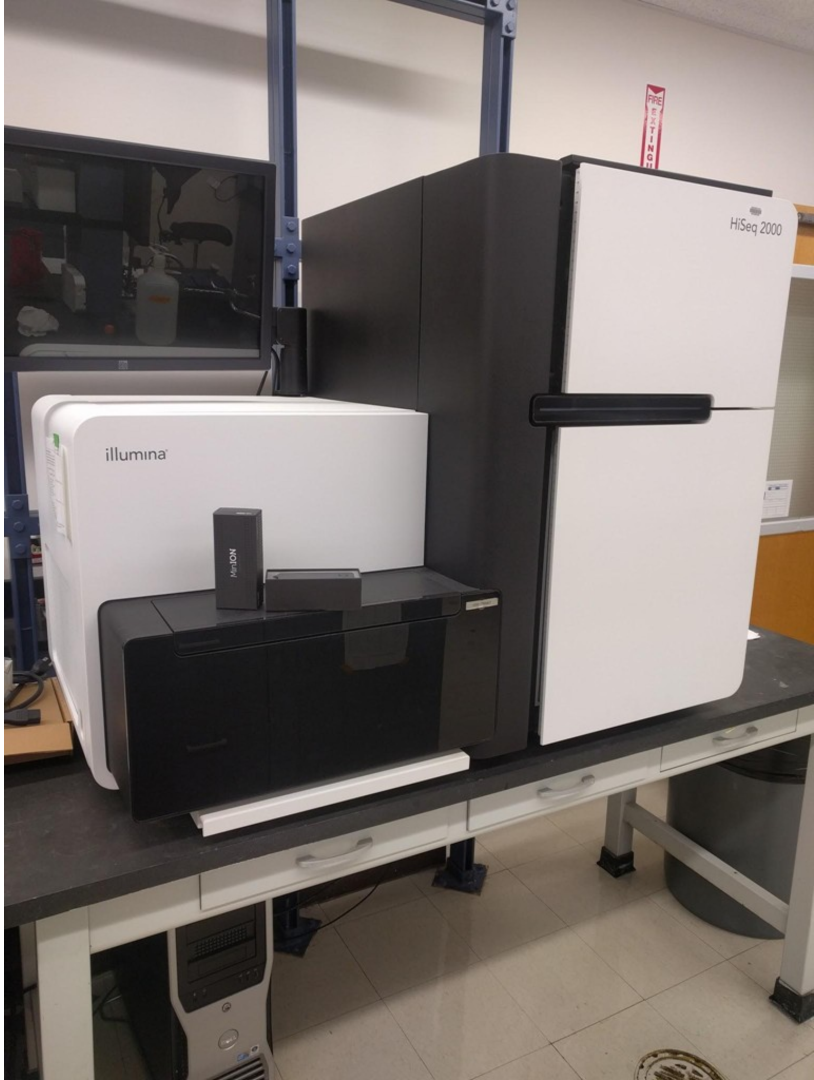
https://www.gatc-biotech.com/fileadmin/_processed_/csm_Sanger_sequencing_illustration_small_898122494c.jpg


PCR containing fluorescent, chain-terminating dideoxynucleotide triphosphates
Detector
Chromatogram
120
130
ACTGCTTGCAGCA
Capillary electrophoresis
Laser

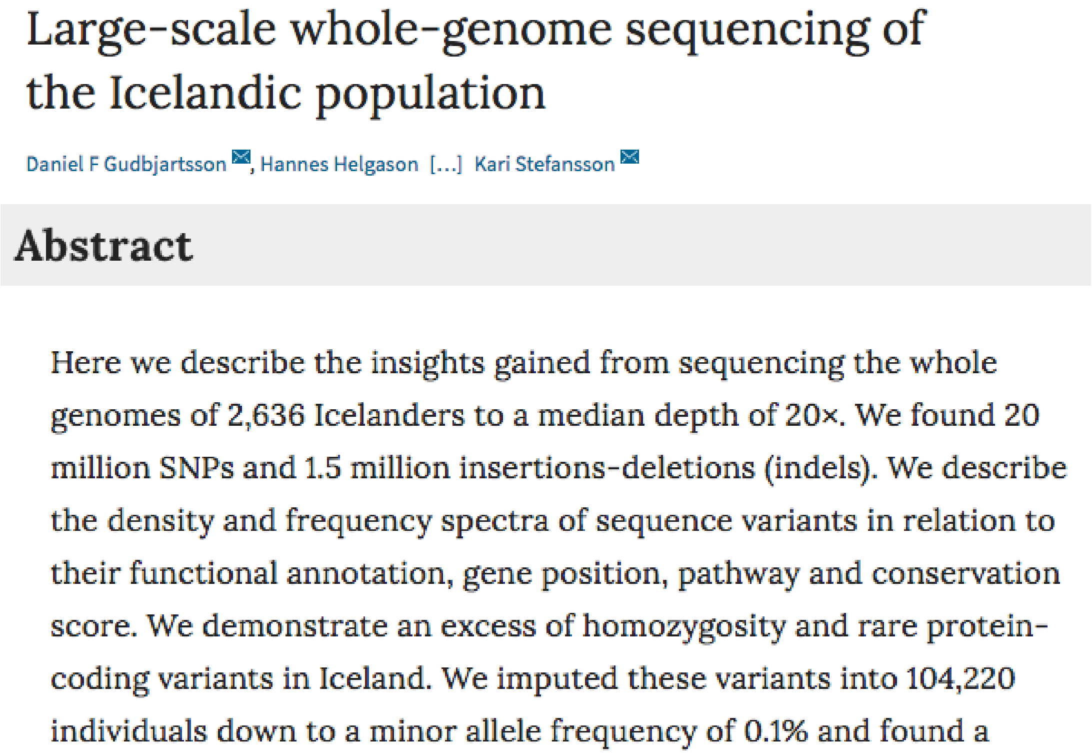
Illumina

- Introduction (5 mins)
  - https://www.youtube.com/watch?v=fCd6B5HRaZ8





Applications



Large-scale whole-genome sequencing of the Icelandic population

Daniel F Gudbjartsson, Hannes Helgason [...] Kari Stefansson

Abstract

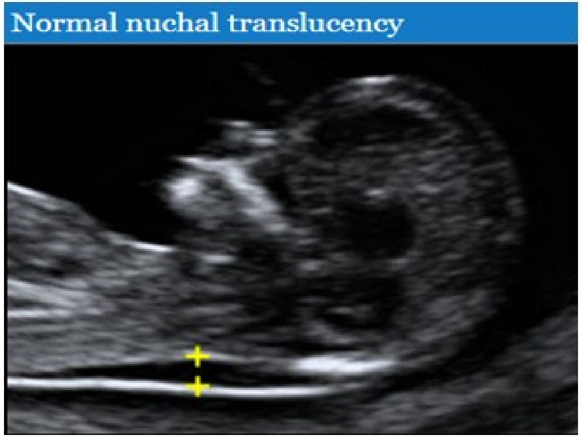
Here we describe the insights gained from sequencing the whole genomes of 2,636 Icelanders to a median depth of 20×. We found 20 million SNPs and 1.5 million insertions-deletions (indels). We describe the density and frequency spectra of sequence variants in relation to their functional annotation, gene position, pathway and conservation score. We demonstrate an excess of homozygosity and rare protein-coding variants in Iceland. We imputed these variants into 104,220 individuals down to a minor allele frequency of 0.1% and found a

23andMe

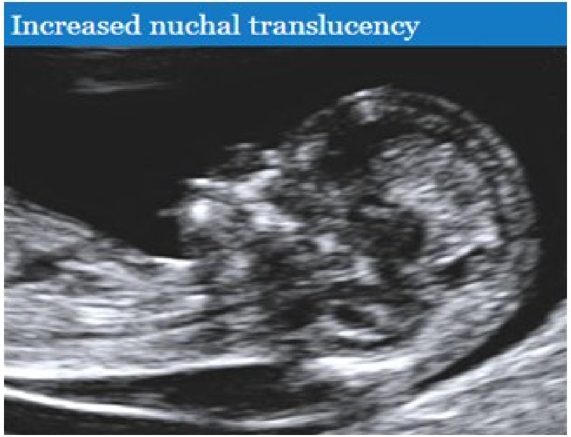
Sequencing versus genotyping?

Applications

illumina®



Normal nuchal translucency



Increased nuchal translucency

http://s19.postimage.org/nx6ifenv7/2012_10_11_215737.png

NIPT Is Noninvasive to the Mother and Baby

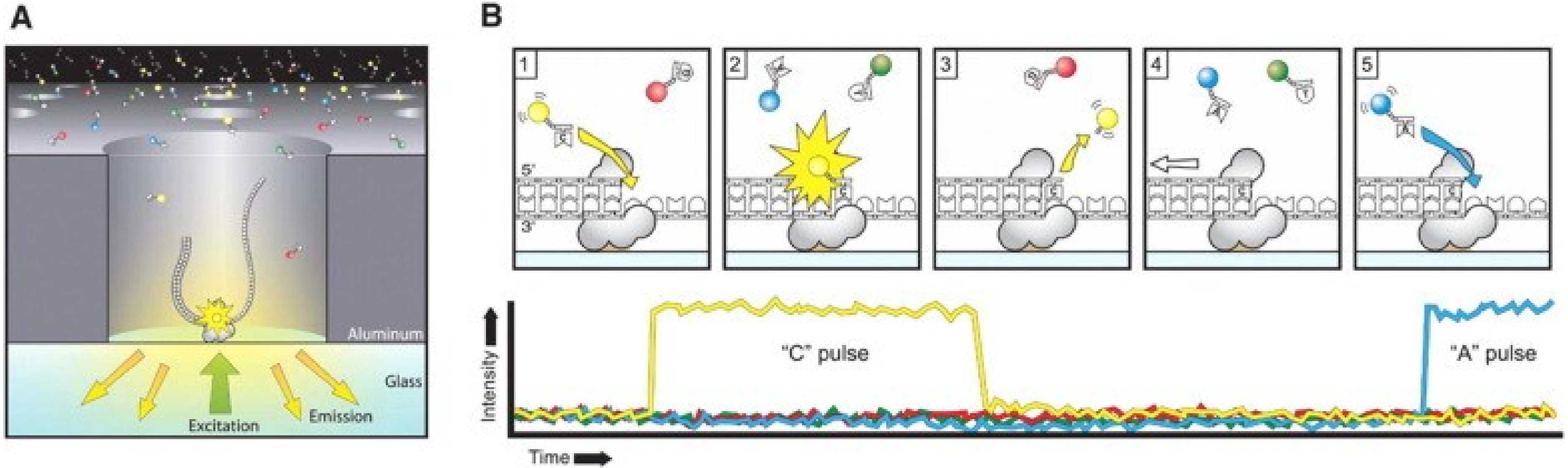
NIPT analyzes cell free DNA from a maternal blood sample (mixture of fetal and maternal DNA) to screen for common chromosomal conditions including trisomy 21 (Down syndrome), trisomy 18 (Edwards syndrome), and trisomy 13 (Patau syndrome).

PacBio

- Introduction to SMRT sequencing (2 mins)
    - https://www.youtube.com/watch?NHCJ8PtYCFc
- Advanced overview of the technology
    - https://vimeo.com/121267426 (9 mins)

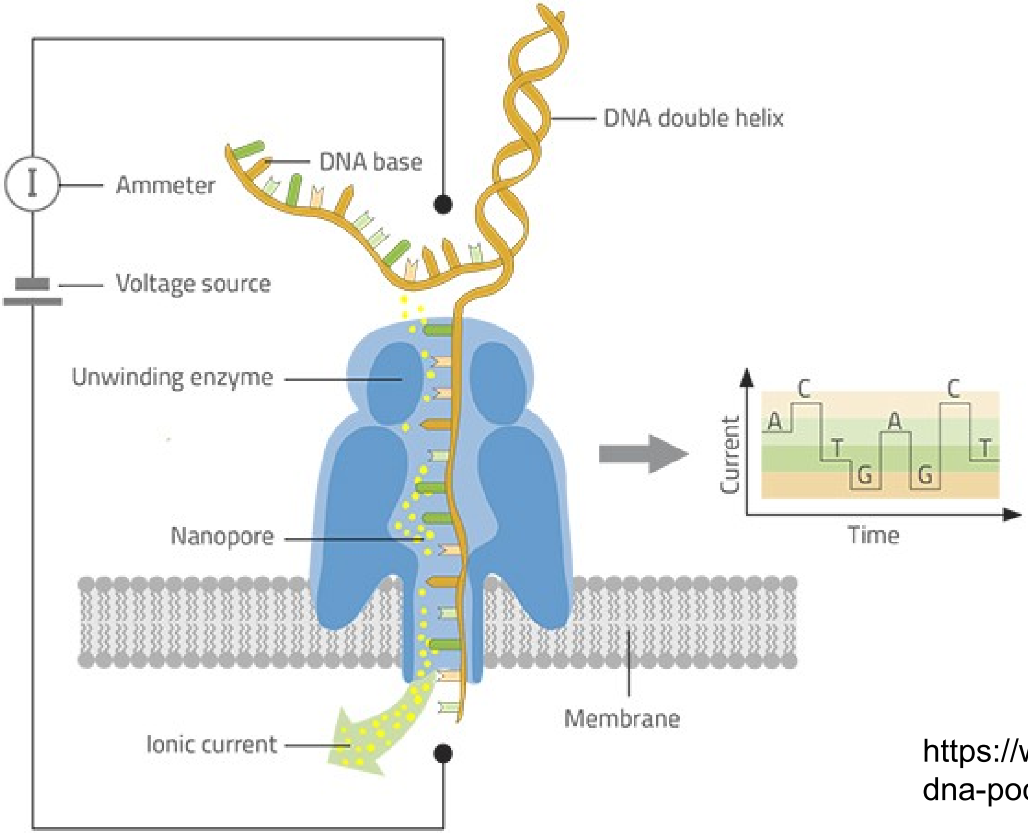
https://www.pacb.com/wp-content/uploads/sequel-2Bimage-300x295.jpg

PacBio



https://doi.org/10.1016/j.gpb.2015.08.002

Nanopore

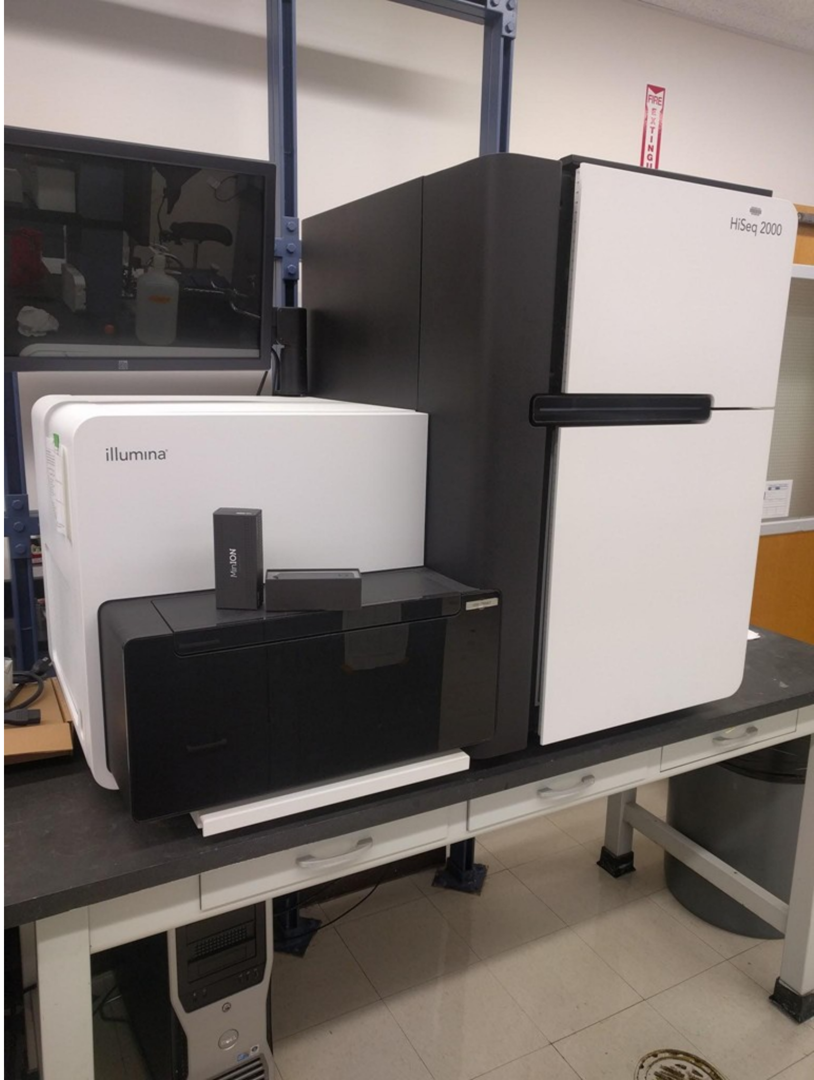


https://www.scienceinschool.org/content/decoding-dna-pocket-sized-sequencer

Nanopore

- Introduction (3 mins)
  - https://www.youtube.com/watch?v=GUb1TZvMWsw
- Advanced slides on the technology
  - https://github.com/lmmx/talk-transcripts/blob/master/Nanopore/NoThanksIveAlreadyGotOne.md

https://twitter.com/MakovaLab

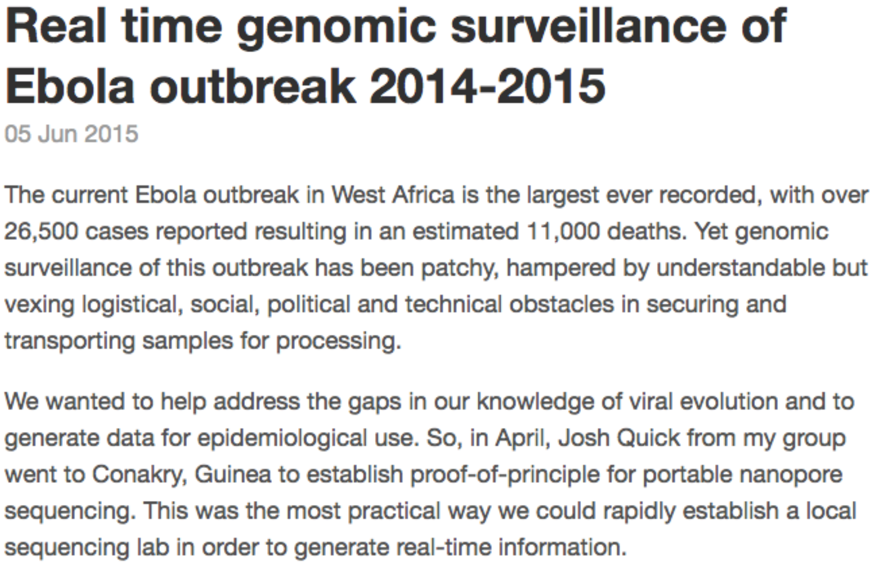


Nanopore

- Introduction (3 mins)
  - https://www.youtube.com/watch?v=GUb1TZvMWsw
- Advanced slides on the technology
  - https://github.com/lmmx/talk-transcripts/blob/master/Nanopore/NoThanksIveAlreadyGotOne.md

https://twitter.com/MakovaLab

Real-time surveillance

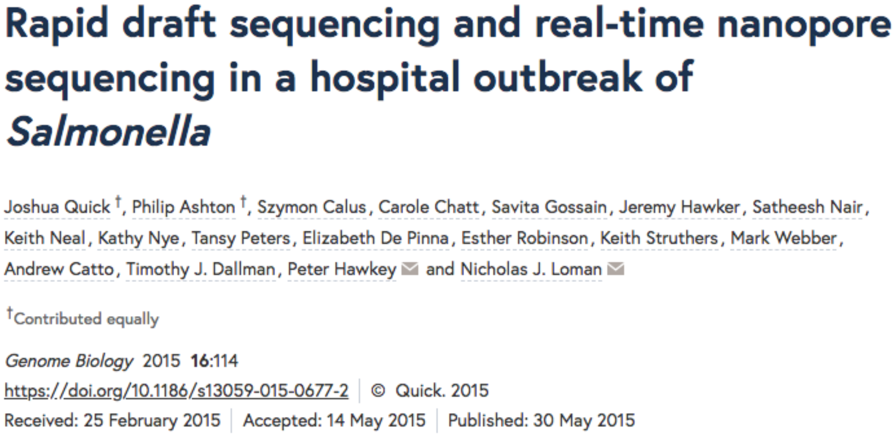


Real time genomic surveillance of Ebola outbreak 2014-2015

05 Jun 2015

The current Ebola outbreak in West Africa is the largest ever recorded, with over 26,500 cases reported resulting in an estimated 11,000 deaths. Yet genomic surveillance of this outbreak has been patchy, hampered by understandable but vexing logistical, social, political and technical obstacles in securing and transporting samples for processing.

We wanted to help address the gaps in our knowledge of viral evolution and to generate data for epidemiological use. So, in April, Josh Quick from my group went to Conakry, Guinea to establish proof-of-principle for portable nanopore sequencing. This was the most practical way we could rapidly establish a local sequencing lab in order to generate real-time information.



Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*

Joshua Quick [†], Philip Ashton [†], Szymon Calus, Carole Chatt, Savita Gossain, Jeremy Hawker, Satheesh Nair, Keith Neal, Kathy Nye, Tansy Peters, Elizabeth De Pinna, Esther Robinson, Keith Struthers, Mark Webber, Andrew Catto, Timothy J. Dallman, Peter Hawkey ✉ and Nicholas J. Loman ✉

[†] Contributed equally

Genome Biology 2015 **16**:114

https://doi.org/10.1186/s13059-015-0677-2 | © Quick. 2015

Received: 25 February 2015 | Accepted: 14 May 2015 | Published: 30 May 2015

Sequencing the station: Investigation aims to identify unknown microbes in space

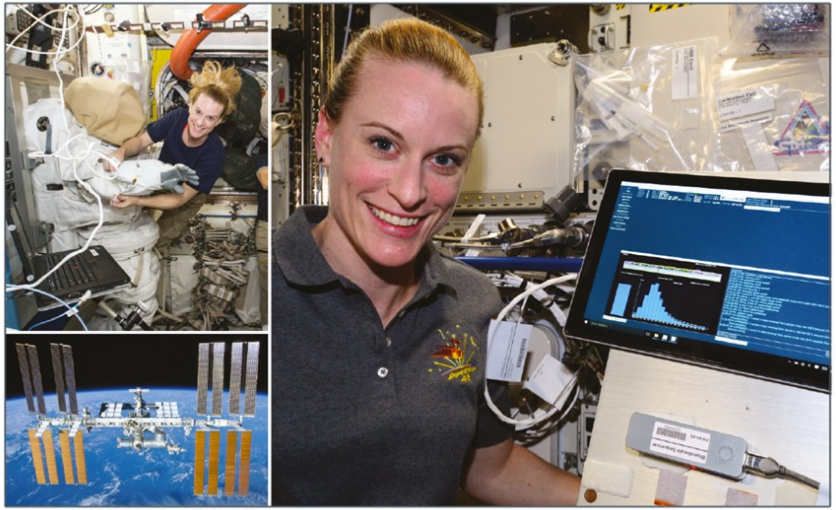


Fig. 1 Astronaut Kate Rubins on the ISS

https://phys.org/news/2017-04-sequencing-station-aims-unknown-microbes.html

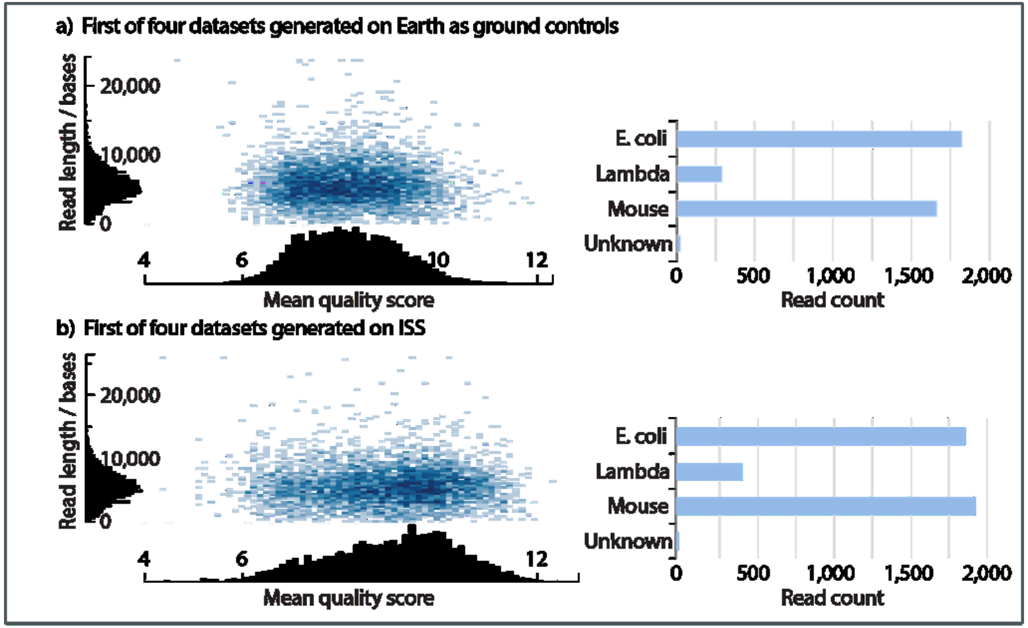
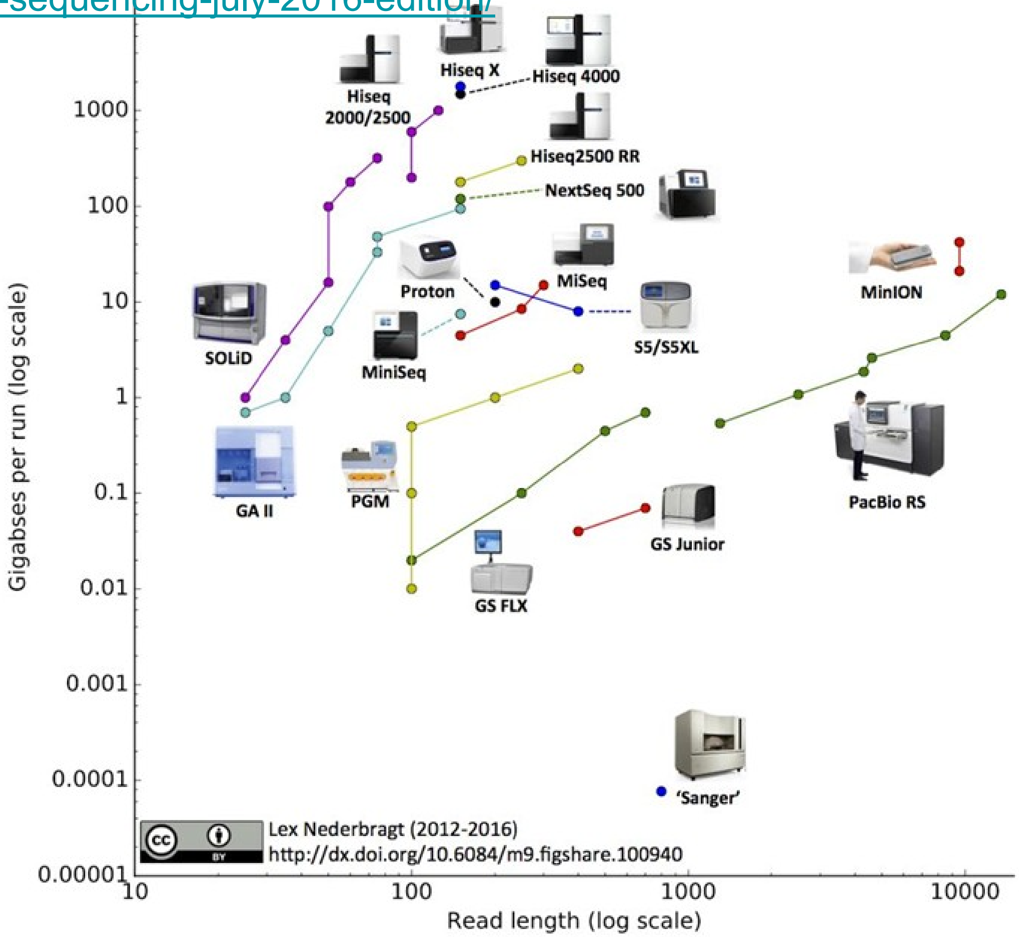


Fig. 2 Analysis workflow showing read quality for runs a) on Earth b) on the ISS

https://nanoporetech.com/resource-centre/posters/dna-sequencing-microgravity-international-space-station-iss-using-minion https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

Long-read sequencing

- Sample quality
- Library preparation (size selection, repair)
- Protocols
- Sequencing throughput
- Multiplexing

Budgeting is important: get yourself familiar with the cost of reagents and sequencing!

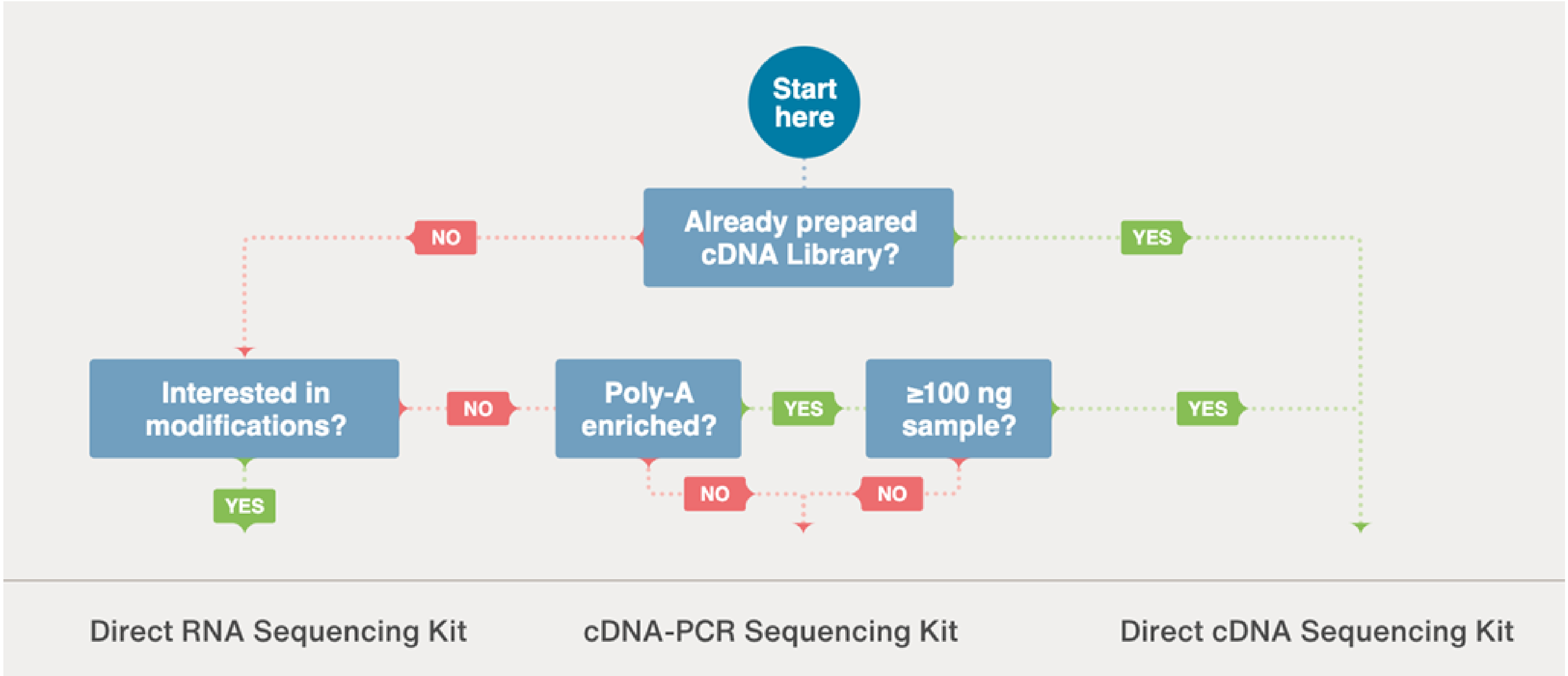
There’s always a trade-off.

Long-read applications

Long reads

1. **Transcriptome (isoforms)**
2. **Assembly (PacBio HiFi + ultra-long Nanopore reads)**
3. **Epigenetic modifications**

1. RNA sequencing with Nanopore



Start here

Already prepared cDNA Library?

NO

YES

Interested in modifications?

NO

Poly-A enriched?

YES

≥100 ng sample?

YES

YES

NO

NO

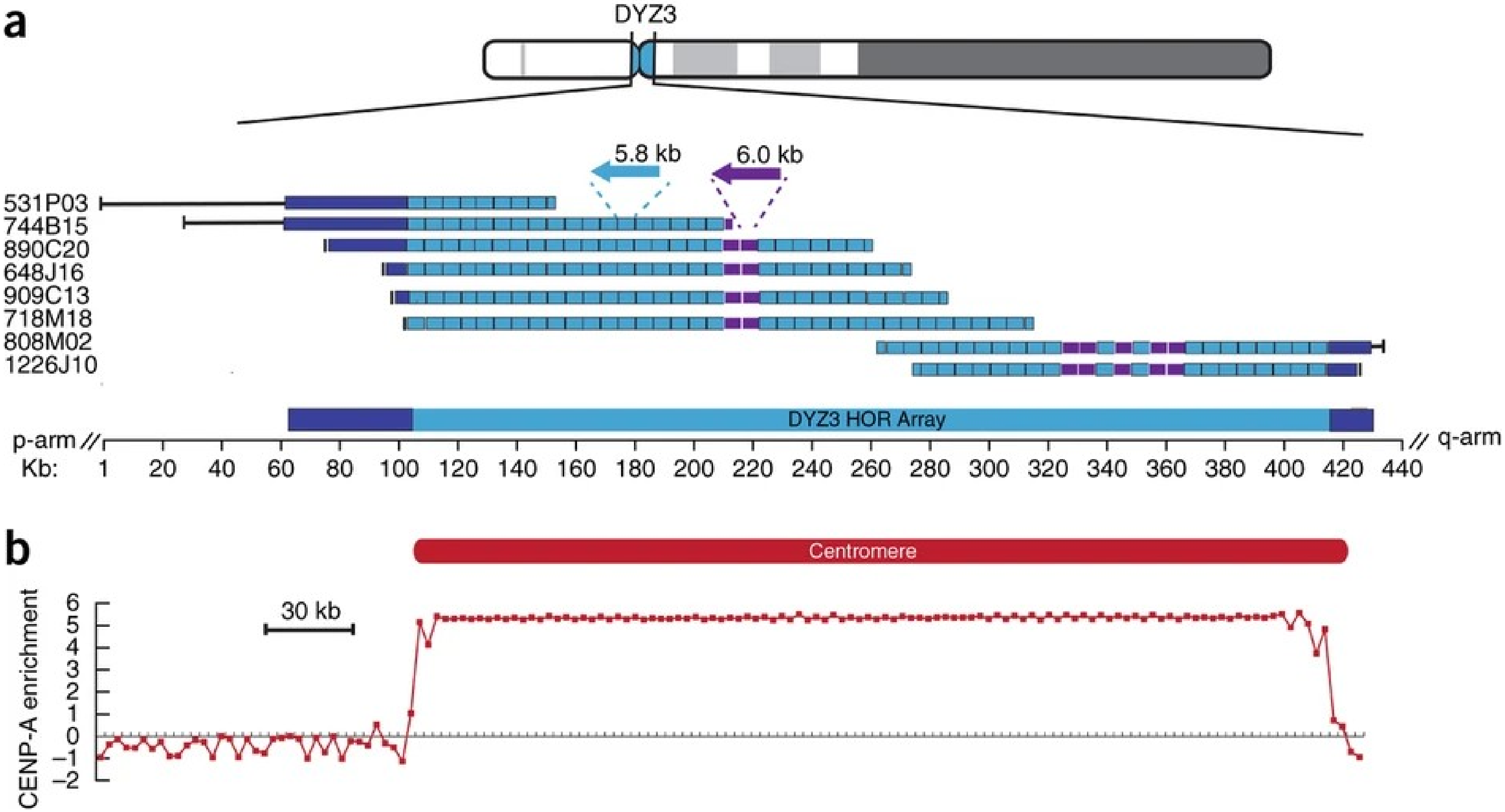
Direct RNA Sequencing Kit

cDNA-PCR Sequencing Kit

Direct cDNA Sequencing Kit

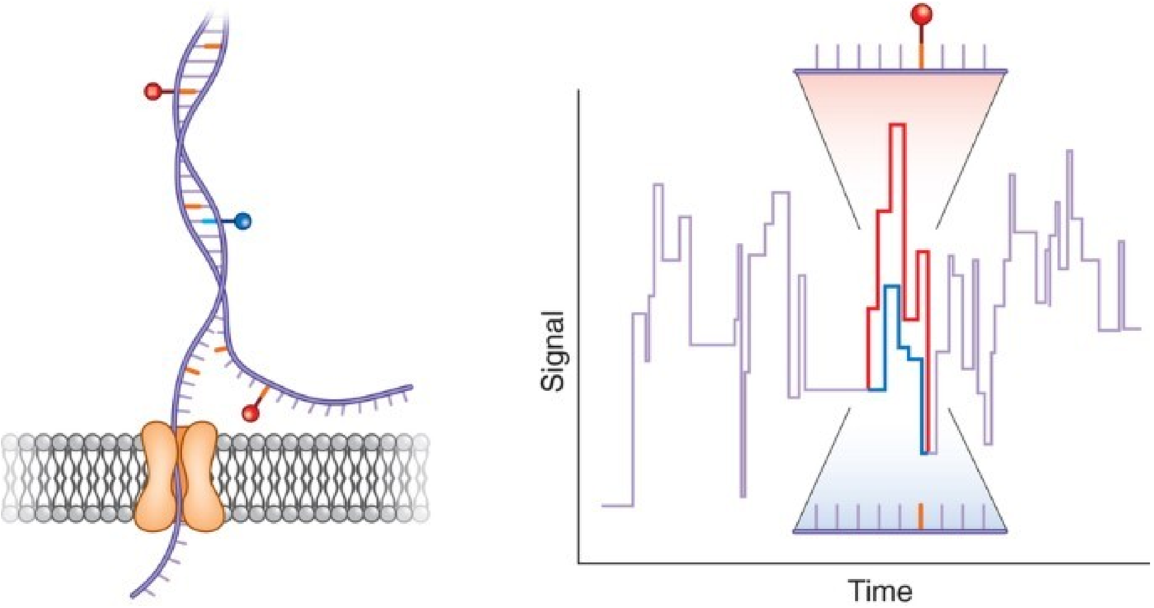
https://nanoporetech.com/applications/rna-sequencing

2. Linear assembly of the RP11 Y centromere



3. Nanopore sequencing meets epigenetics

Figure 1 | DNA methylation can be read out directly by nanopore sequencing. Single-stranded DNA alters ionic current in a sequence-dependent manner as it passes through a pore (left), with methylated bases highlighted with red and blue tags. The raw current signal (right) indicates small changes due to methylation that a new set of algorithms can robustly interpret.



Schatz, M. Nanopore sequencing meets epigenetics. *Nat Methods* 14, 347–348 (2017). https://doi.org/10.1038/nmeth.4240

Error correction of long reads

Error profiles

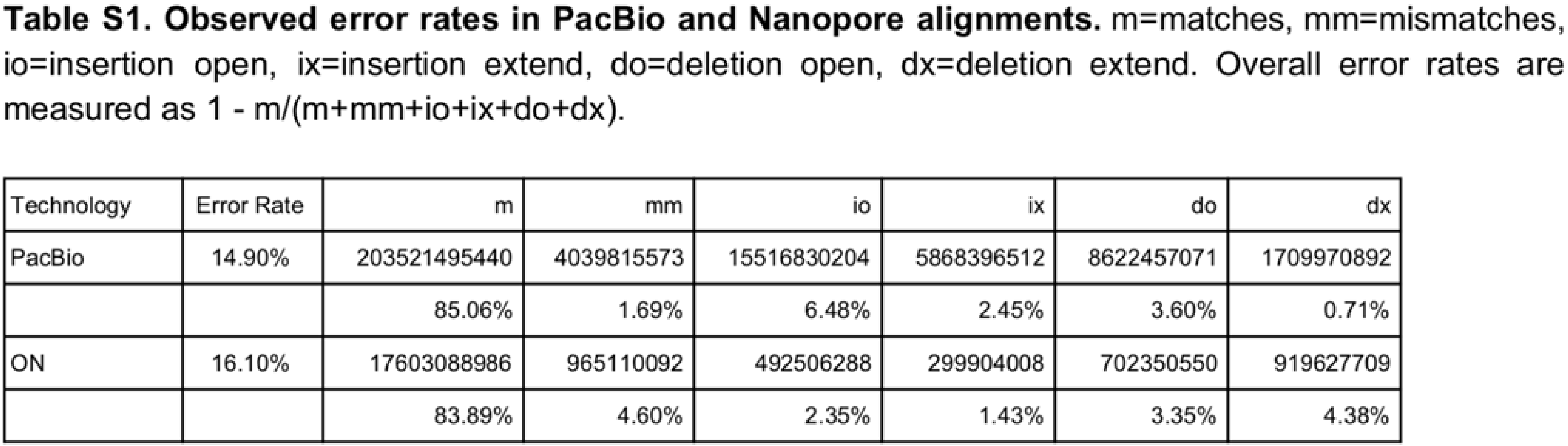


Table S1. Observed error rates in PacBio and Nanopore alignments. m=matches, mm=mismatches, io=insertion open, ix=insertion extend, do=deletion open, dx=deletion extend. Overall error rates are measured as 1 - m/(m+mm+io+ix+do+dx).

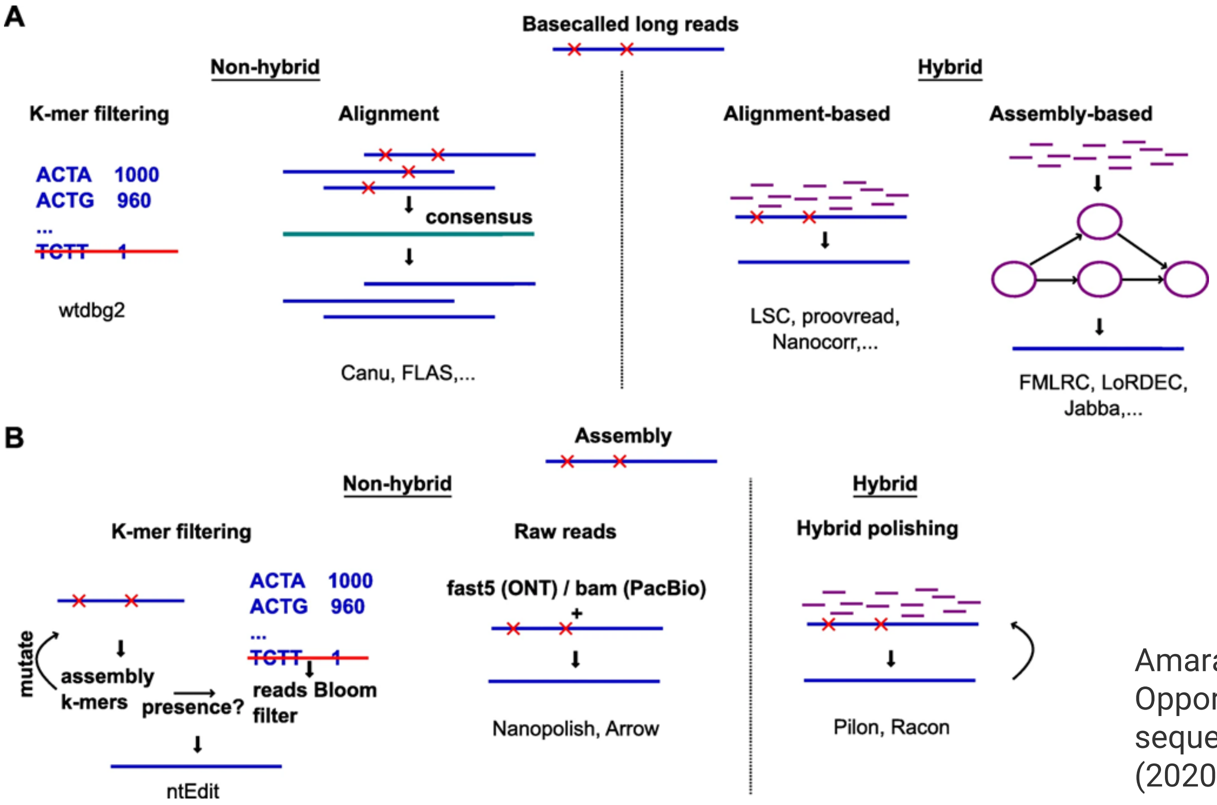
Technology	Error Rate	m	mm	io	ix	do	dx
PacBio	14.90%	203521495440	4039815573	15516830204	5868396512	8622457071	1709970892
		85.06%	1.69%	6.48%	2.45%	3.60%	0.71%
ON	16.10%	17603088986	965110092	492506288	299904008	702350550	919627709
		83.89%	4.60%	2.35%	1.43%	3.35%	4.38%

RS Harris, **M Cechova**, KD Makova. *Noise-Cancelling Repeat Finder: Uncovering tandem repeats in error-prone long-read sequencing data. BIOINFORMATICS*, 2019.

Error correction

- Error correction with short reads
- Error correction with long reads
- Assembly merging

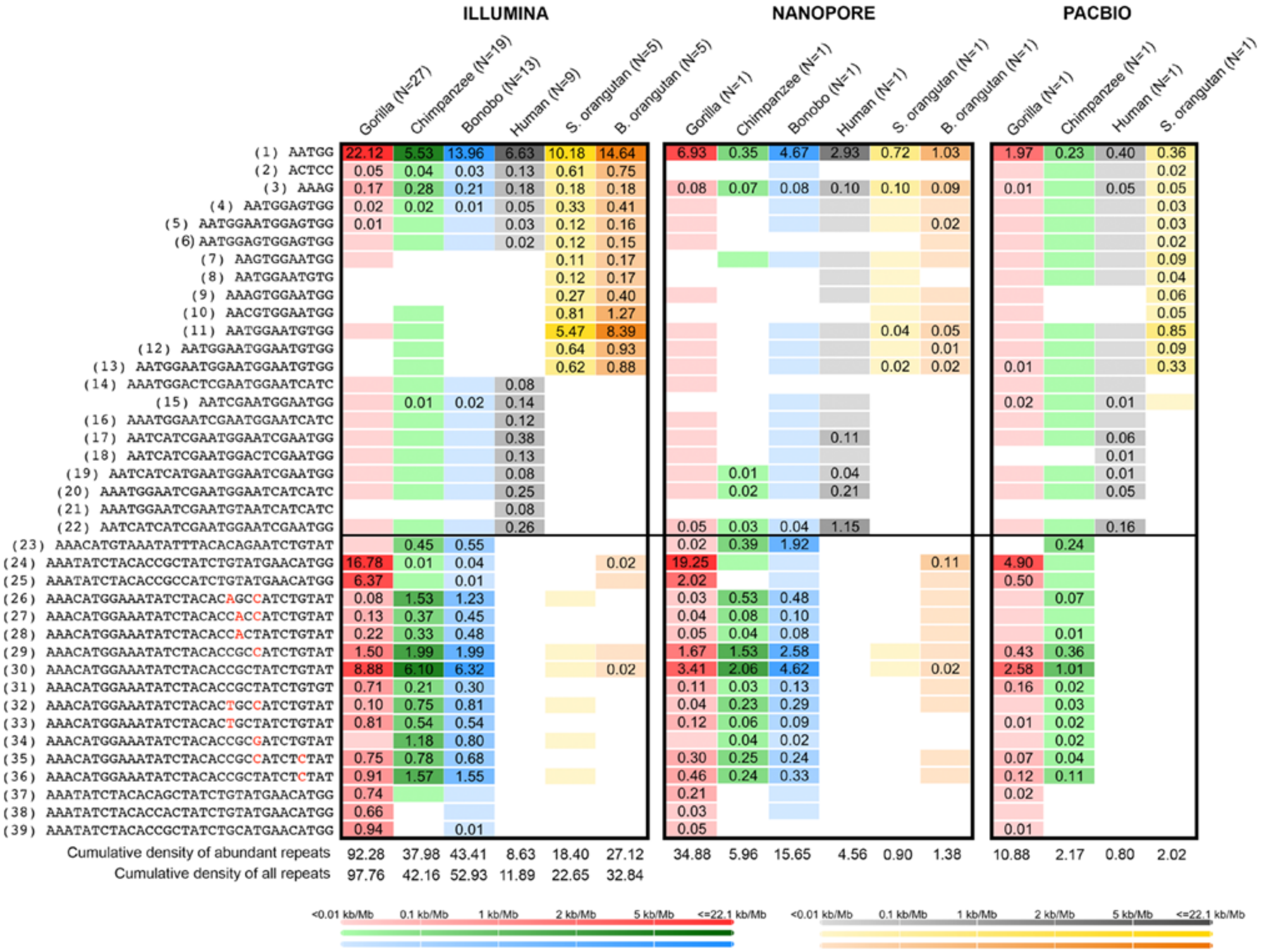
Error correction



- Error correction with short reads
- Error correction with long reads
- Assembly merging

Amarasinghe, S.L., Su, S., Dong, X. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 21, 30 (2020). https://doi.org/10.1186/s13059-020-1935-5

We demonstrated that orthogonal technologies (such as Illumina, Nanopore, and PacBio) are generally concordant in distinguishing between highly and lowly abundant repeated motifs.



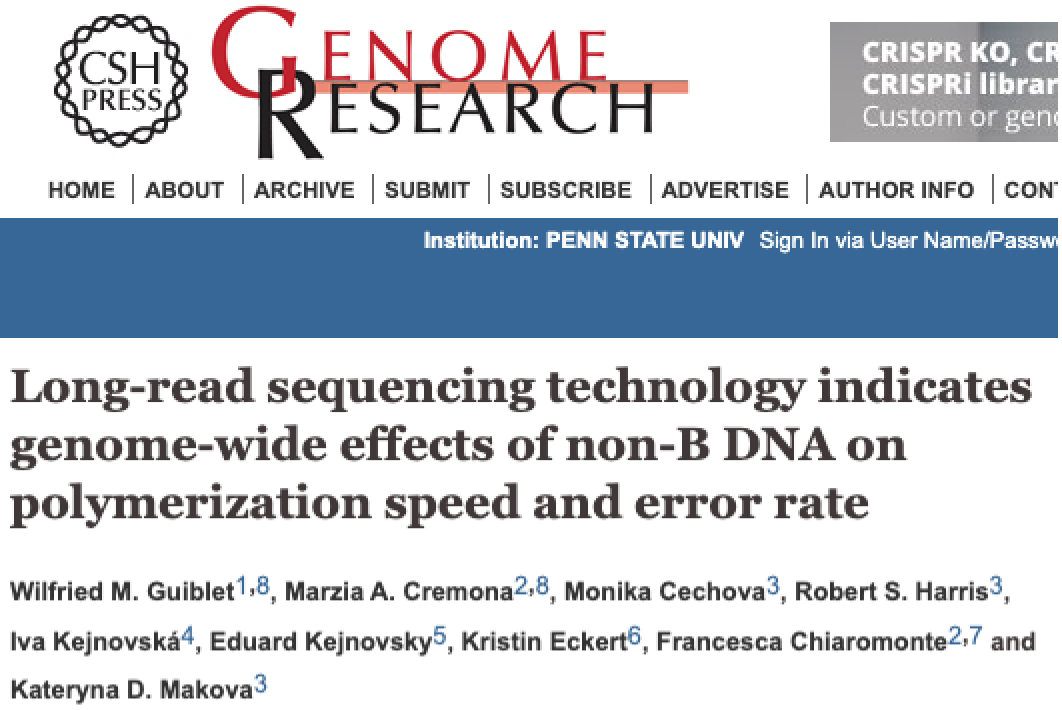
	ILLUMINA						NANOPORE						PACBIO			
	Gorilla (N=27)	Chimpanzee (N=19)	Bonobo (N=13)	Human (N=9)	S. orangutan (N=5)	B. orangutan (N=5)	Gorilla (N=1)	Chimpanzee (N=1)	Bonobo (N=1)	Human (N=1)	S. orangutan (N=1)	B. orangutan (N=1)	Gorilla (N=1)	Chimpanzee (N=1)	Human (N=1)	S. orangutan (N=1)
(1) AATGG	22.12	5.53	13.96	6.63	10.18	14.64	6.93	0.35	4.67	2.93	0.72	1.03	1.97	0.23	0.40	0.36
(2) ACTCC	0.05	0.04	0.03	0.13	0.61	0.75										0.02
(3) AAAG	0.17	0.28	0.21	0.18	0.18	0.18	0.08	0.07	0.08	0.10	0.10	0.09	0.01		0.05	0.05
(4) AATGGAGTGG	0.02	0.02	0.01	0.05	0.33	0.41										0.03
(5) AATGGAATGGAGTGG	0.01			0.03	0.12	0.16						0.02				0.03
(6) AATGGAGTGGAGTGG				0.02	0.12	0.15										0.02
(7) AAGTGGAATGG					0.11	0.17										0.09
(8) AATGGAATGTG					0.12	0.17										0.04
(9) AAAGTGGAATGG					0.27	0.40										0.06
(10) AACGTGGAATGG					0.81	1.27										0.05
(11) AATGGAATGTGG					5.47	8.39					0.04	0.05				0.85
(12) AATGGAATGGAATGTGG					0.64	0.93						0.01				0.09
(13) AATGGAATGGAATGGAATGTGG					0.62	0.88					0.02	0.02	0.01			0.33
(14) AAATGGACTCGAATGGAATCATC				0.08												
(15) AATCGAATGGAATGG		0.01	0.02	0.14									0.02		0.01	
(16) AAATGGAATCGAATGGAATCATC				0.12												
(17) AATCATCGAATGGAATCGAATGG				0.38						0.11					0.06	
(18) AATCATCGAATGGACTCGAATGG				0.13											0.01	
(19) AATCATCATGAATGGAATCGAATGG				0.08				0.01		0.04					0.01	
(20) AAATGGAATCGAATGGAATCATCATC				0.25				0.02		0.21					0.05	
(21) AAATGGAATCGAATGTAATCATCATC				0.08												
(22) AATCATCATCGAATGGAATCGAATGG				0.26			0.05	0.03	0.04	1.15					0.16	
(23) AAACATGTAAATATTTACACAGAATCTGTAT		0.45	0.55				0.02	0.39	1.92					0.24		
(24) AAATATCTACACCGCTATCTGTATGAACATGG	16.78	0.01	0.04			0.02	19.25					0.11	4.90			
(25) AAATATCTACACCGCCATCTGTATGAACATGG	6.37		0.01				2.02						0.50			
(26) AAACATGGAAATATCTACACAGCCATCTGTAT	0.08	1.53	1.23				0.03	0.53	0.48					0.07		
(27) AAACATGGAAATATCTACACCACCATCTGTAT	0.13	0.37	0.45				0.04	0.08	0.10							
(28) AAACATGGAAATATCTACACCACTATCTGTAT	0.22	0.33	0.48				0.05	0.04	0.08					0.01		
(29) AAACATGGAAATATCTACACCGCCATCTGTAT	1.50	1.99	1.99				1.67	1.53	2.58				0.43	0.36		
(30) AAACATGGAAATATCTACACCGCTATCTGTAT	8.88	6.10	6.32			0.02	3.41	2.06	4.62			0.02	2.58	1.01		
(31) AAACATGGAAATATCTACACCGCTATCTGTGT	0.71	0.21	0.30				0.11	0.03	0.13				0.16	0.02		
(32) AAACATGGAAATATCTACACTGCCATCTGTAT	0.10	0.75	0.81				0.04	0.23	0.29					0.03		
(33) AAACATGGAAATATCTACACTGCTATCTGTAT	0.81	0.54	0.54				0.12	0.06	0.09				0.01	0.02		
(34) AAACATGGAAATATCTACACCGCGATCTGTAT		1.18	0.80					0.04	0.02					0.02		
(35) AAACATGGAAATATCTACACCGCCATCTCTAT	0.75	0.78	0.68				0.30	0.25	0.24				0.07	0.04		
(36) AAACATGGAAATATCTACACCGCTATCTCTAT	0.91	1.57	1.55				0.46	0.24	0.33				0.12	0.11		
(37) AAATATCTACACAGCTATCTGTATGAACATGG	0.74						0.21						0.02			
(38) AAATATCTACACCACTATCTGTATGAACATGG	0.66						0.03									
(39) AAATATCTACACCGCTATCTGCATGAACATGG	0.94		0.01				0.05						0.01			
Cumulative density of abundant repeats	92.28	37.98	43.41	8.63	18.40	27.12	34.88	5.96	15.65	4.56	0.90	1.38	10.88	2.17	0.80	2.02
Cumulative density of all repeats	97.76	42.16	52.93	11.89	22.65	32.84										

<0.01 kb/Mb 0.1 kb/Mb 1 kb/Mb 2 kb/Mb 5 kb/Mb <=22.1 kb/Mb

<0.01 kb/Mb 0.1 kb/Mb 1 kb/Mb 2 kb/Mb 5 kb/Mb <=22.1 kb/Mb

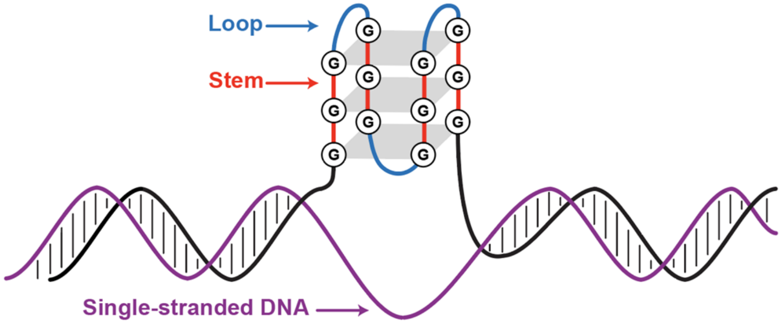
Cechova et al., 2019

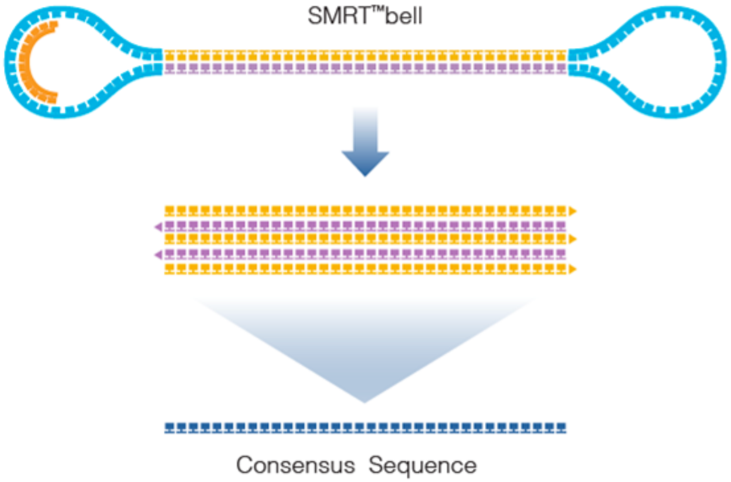
Harris et al., 2019


CSH PRESS
GENOME RESEARCH
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CON
Institution: PENN STATE UNIV Sign In via User Name/Passw

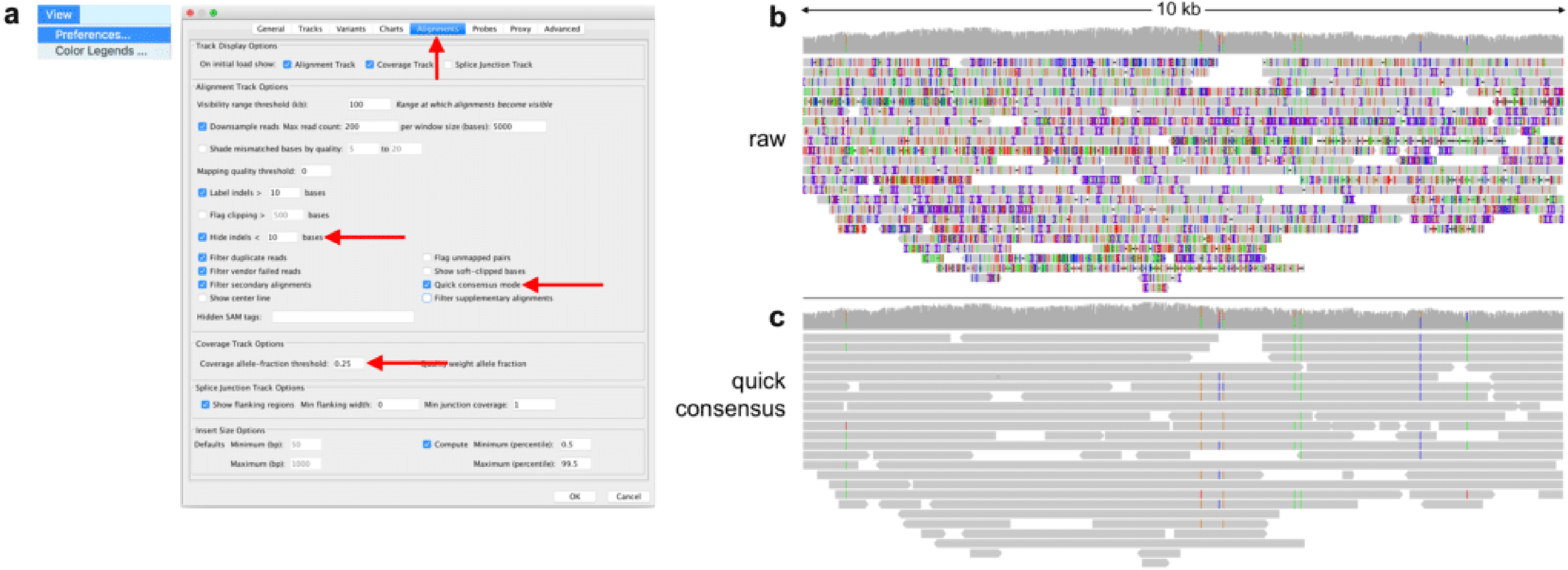
Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate

Wilfried M. Guiblet[1,8], Marzia A. Cremona[2,8], Monika Cechova[3], Robert S. Harris[3], Iva Kejnovská[4], Eduard Kejnovsky[5], Kristin Eckert[6], Francesca Chiaromonte[2,7] and Kateryna D. Makova[3]





The visualization of sequencing errors



Track Display Options

On initial load show: ☑ Alignment Track ☑ Coverage Track □ Splice Junction Track

Alignment Track Options

Visibility range threshold (kb): 100 Range at which alignments become visible

☑ Downsample reads Max read count: 200 per window size (bases): 5000

□ Shade mismatched bases by quality: 5 to 20

Mapping quality threshold: 0

☑ Label indels > 10 bases

□ Flag clipping > 500 bases

☑ Hide indels < 10 bases

☑ Filter duplicate reads

☑ Filter vendor failed reads

☑ Filter secondary alignments

□ Show center line

□ Flag unmapped pairs

□ Show soft-clipped bases

☑ Quick consensus mode

□ Filter supplementary alignments

Hidden SAM tags:

Coverage Track Options

Coverage allele-fraction threshold: 0.25 □ Quality weight allele fraction

Splice Junction Track Options

☑ Show flanking regions Min flanking width: 0 Min junction coverage: 1

Insert Size Options

Defaults Minimum (bp): 50

Maximum (bp): 1000

☑ Compute Minimum (percentile): 0.5

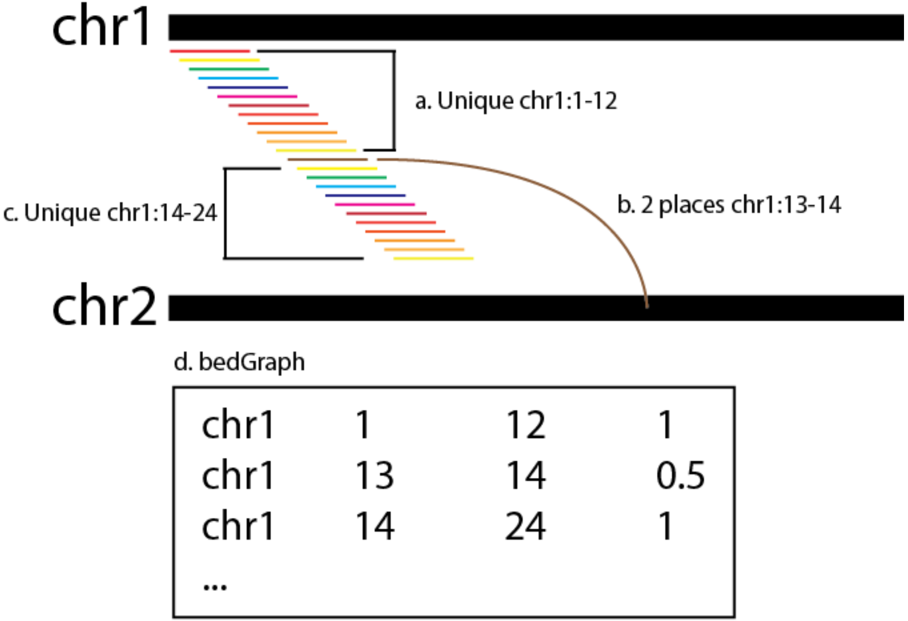
Maximum (percentile): 99.5

https://www.pacb.com/blog/igv-3-improves-support-pacbio-long-reads/

Mappability profiles

Alignability *– These tracks provide a measure of how often the sequence found at the particular location will align within the whole genome. Unlike measures of uniqueness, alignability will tolerate up to 2 mismatches. These tracks are in the form of signals ranging from 0 to 1 and have several configuration options.*

Uniqueness *– These tracks are a direct measure of sequence uniqueness throughout the reference genome. These tracks are in the form of signals ranging from 0 to 1 and have several configuration options*



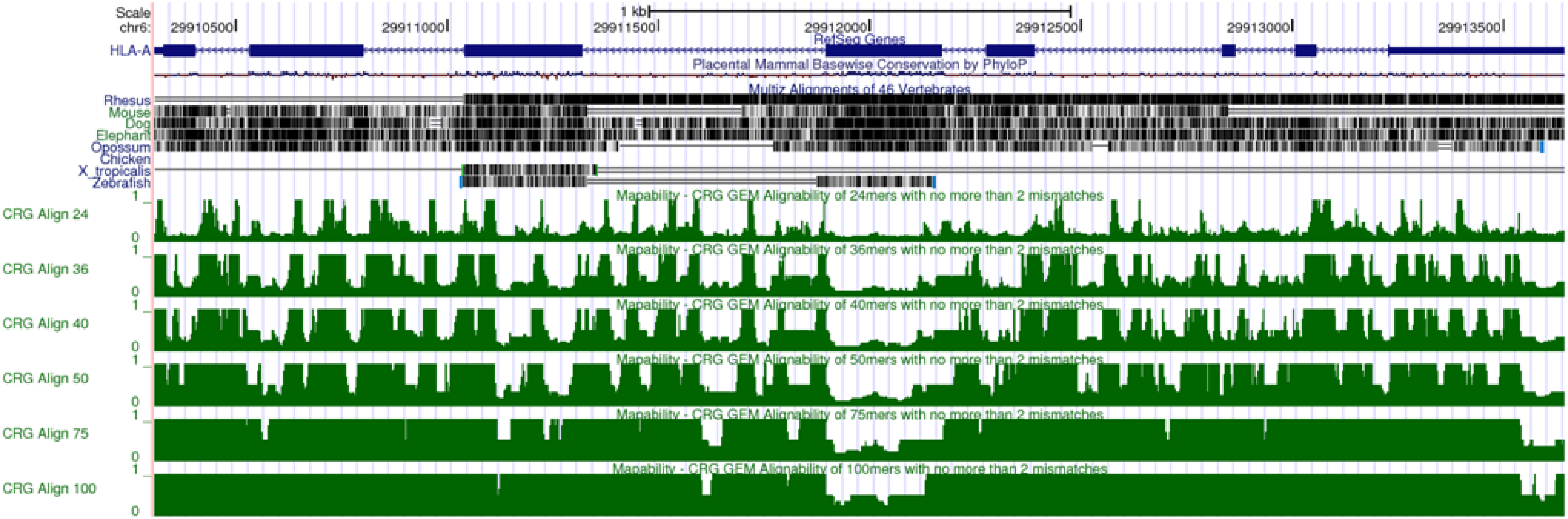
chr1	1	12	1
chr1	13	14	0.5
chr1	14	24	1
...			

https://davetang.org/muse/2013/07/08/encode-mappability/

Mappability profiles

Influence of paralogous genes on the mappability scores: the example of the HLA-A gene.

The HLA-A gene is part of the Major Histocompatibility Complex (MHC) involving a large gene family with numerous paralogs. This screenshot of the UCSC genome browser (with the six mappability tracks in green) illustrates the low uniqueness of the HLA-A gene (especially, its exon 4) which could render its targeting by RNASeq difficult (if only uniquely mapping reads are considered).

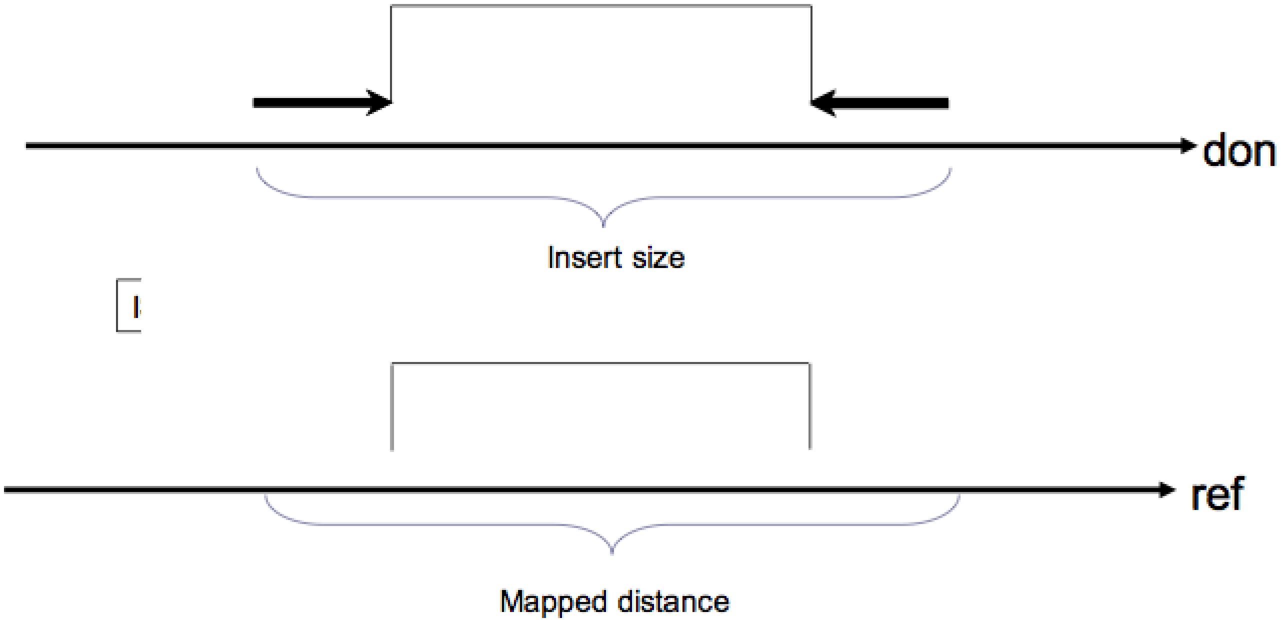


Scale
chr6:
1 kb
29910500
29911000
29911500
29912000
29912500
29913000
29913500
RefSeq Genes
HLA-A
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
CRG Align 24
Mapability - CRG GEM Alignability of 24mers with no more than 2 mismatches
CRG Align 36
Mapability - CRG GEM Alignability of 36mers with no more than 2 mismatches
CRG Align 40
Mapability - CRG GEM Alignability of 40mers with no more than 2 mismatches
CRG Align 50
Mapability - CRG GEM Alignability of 50mers with no more than 2 mismatches
CRG Align 75
Mapability - CRG GEM Alignability of 75mers with no more than 2 mismatches
CRG Align 100
Mapability - CRG GEM Alignability of 100mers with no more than 2 mismatches
1
0

doi: https://doi.org/10.1371/journal.pone.0030377.g009

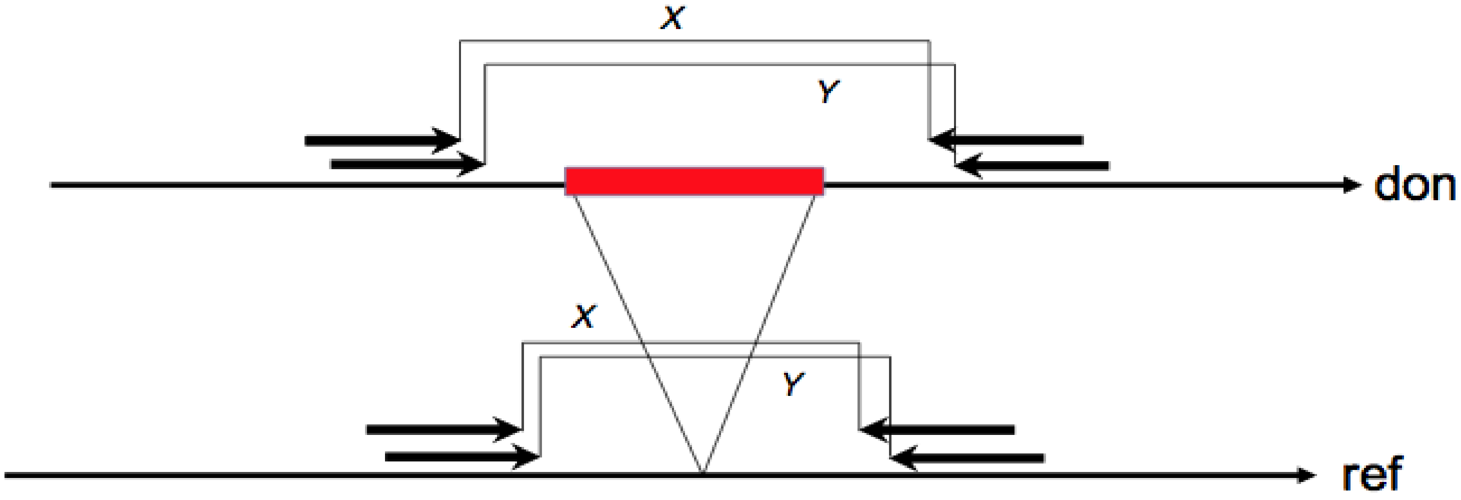
Structural variation detection: the basics

Matepair signature: no SVs



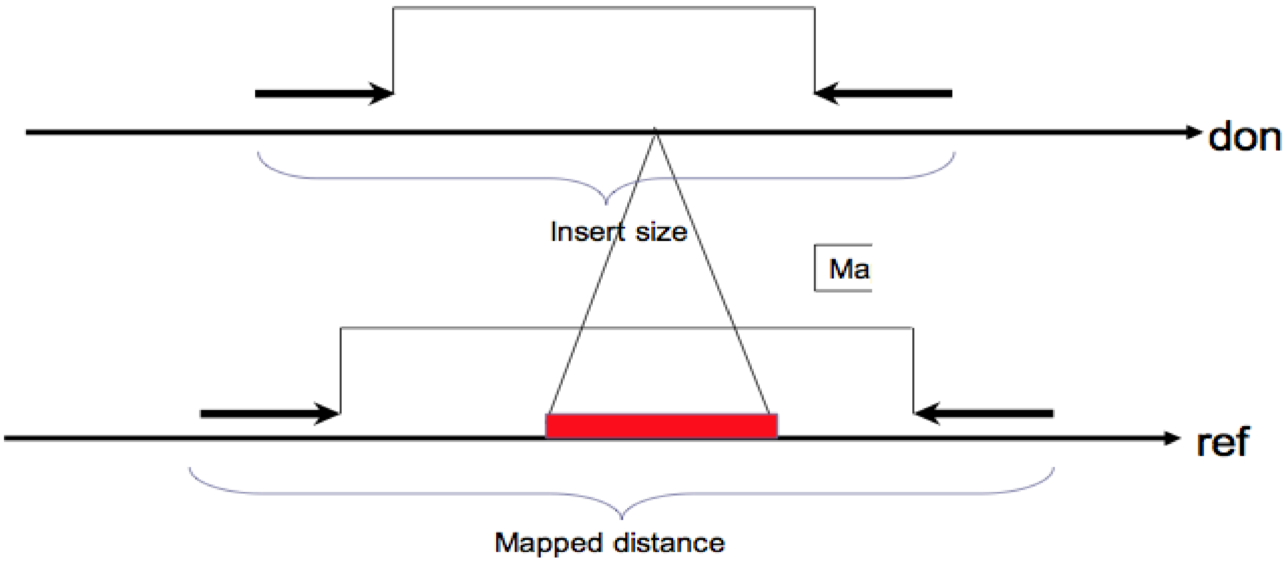
adapted from Prof. Paul Medvedev, Algorithms & Data Structures in Bioinformatics

Insertion: consistency



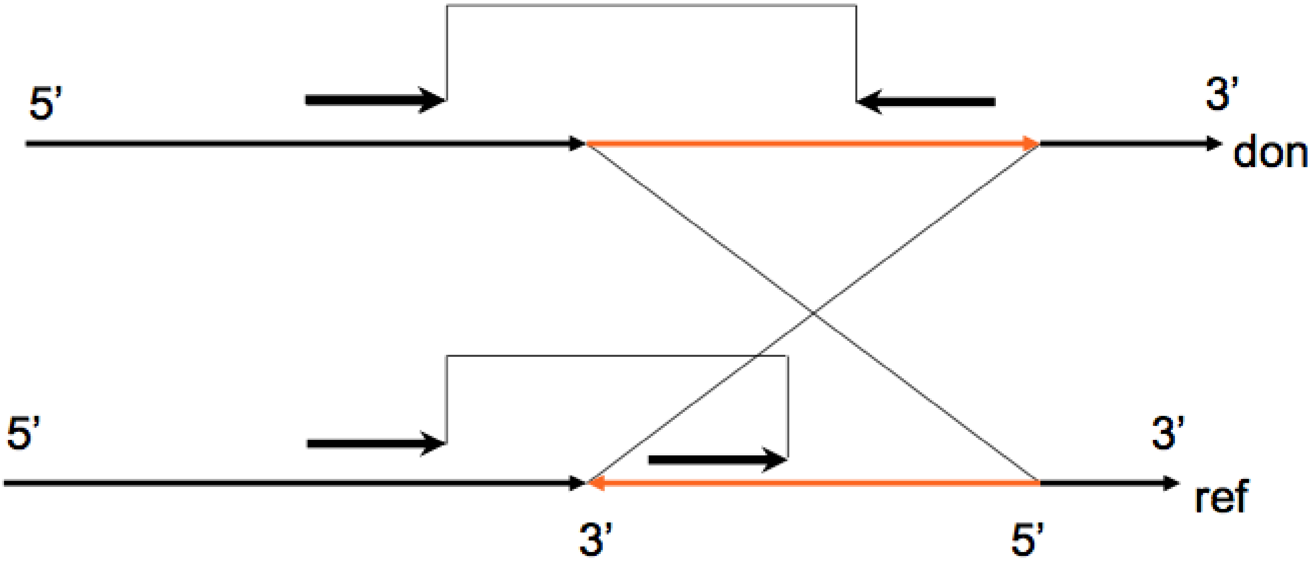
adapted from Prof. Paul Medvedev, Algorithms & Data Structures in Bioinformatics

Deletion: signature



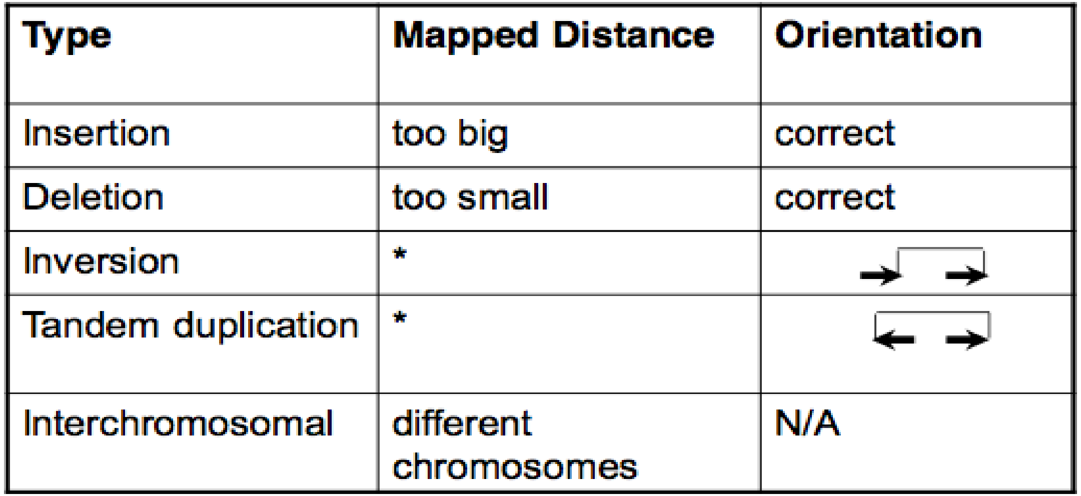
adapted from Prof. Paul Medvedev, Algorithms & Data Structures in Bioinformatics

Inversion: signature



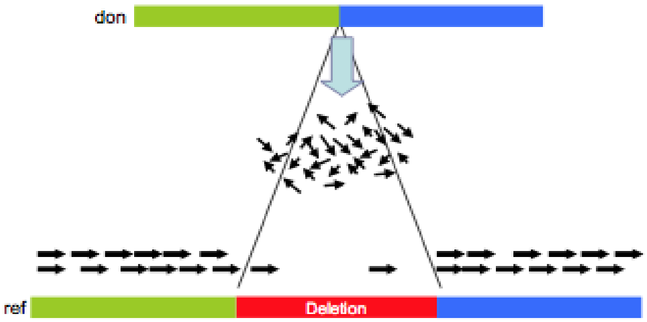
Prof. Paul Medvedev, Algorithms & Data Structures in Bioinformatics

SV summary



Type	Mapped Distance	Orientation
Insertion	too big	correct
Deletion	too small	correct
Inversion	*	→ →
Tandem duplication	*	← →
Interchromosomal	different chromosomes	N/A

Depth-of-coverage

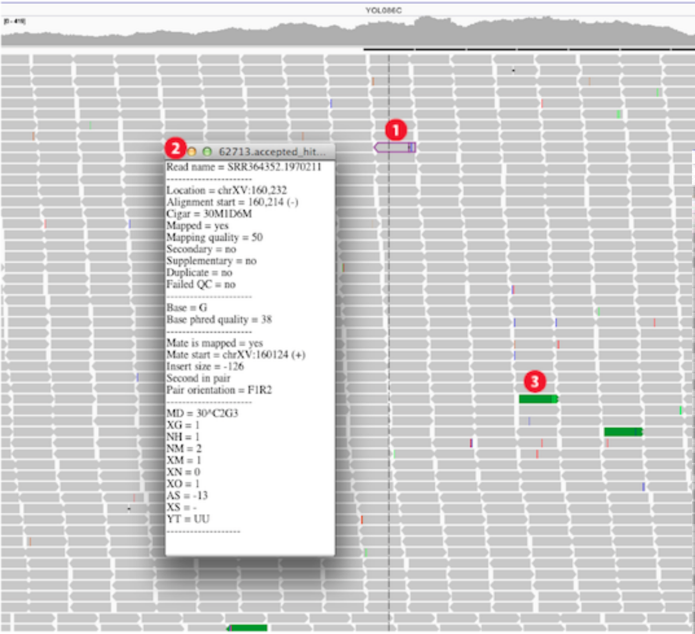


Depth-of-coverage can help detect SVs

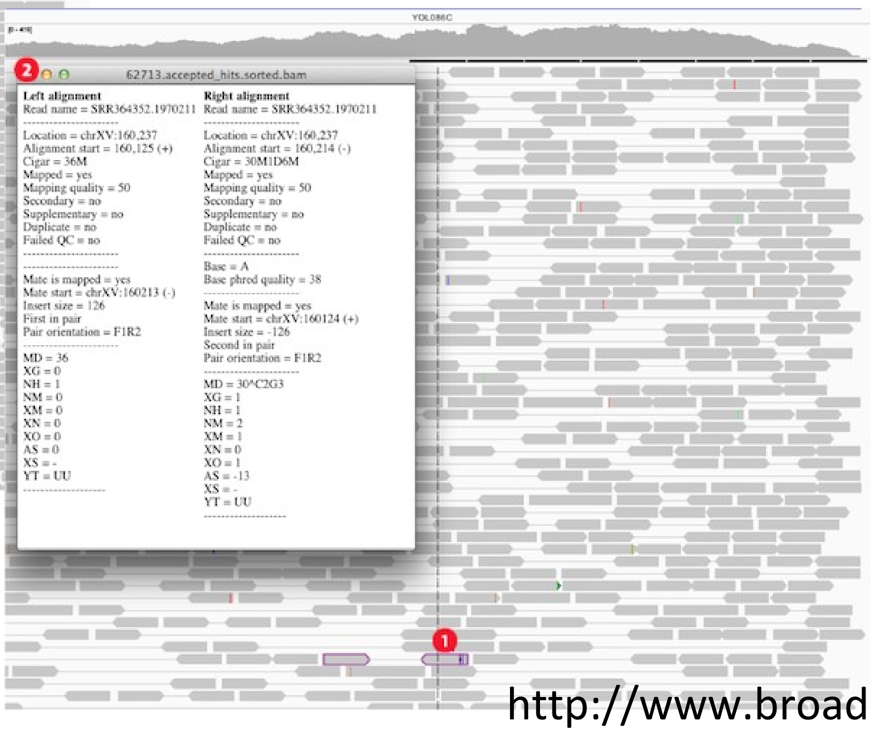
adapted from Prof. Paul Medvedev, Algorithms & Data Structures in Bioinformatics

IGV

- view as pairs



```
Read name = SRR364352.1970211
----------------------
Location = chrXV:160,232
Alignment start = 160,214 (-)
Cigar = 30M1D6M
Mapped = yes
Mapping quality = 50
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no
----------------------
Base = G
Base phred quality = 38
----------------------
Mate is mapped = yes
Mate start = chrXV:160124 (+)
Insert size = -126
Second in pair
Pair orientation = F1R2
----------------------
MD = 30^C2G3
XG = 1
NH = 1
NM = 2
XM = 1
XN = 0
XO = 1
AS = -13
XS = -
YT = UU
-------------------
```

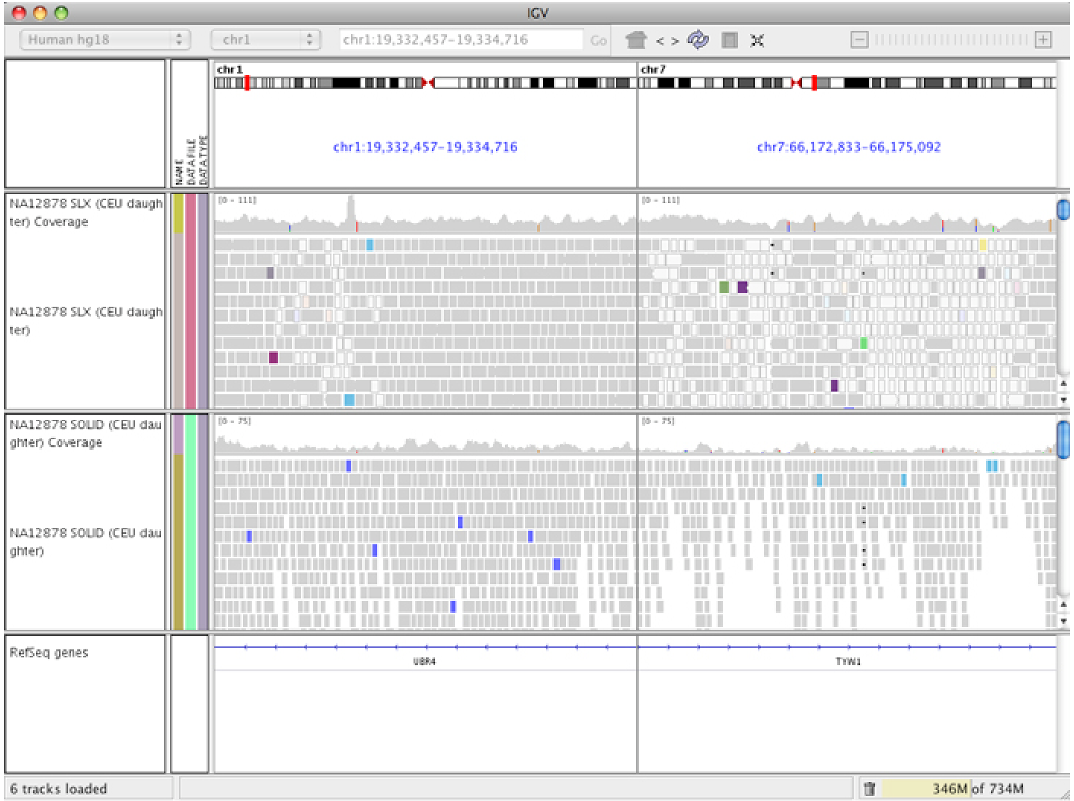


```
Left alignment                    Right alignment
Read name = SRR364352.1970211     Read name = SRR364352.1970211
----------------------            ----------------------
Location = chrXV:160,237          Location = chrXV:160,237
Alignment start = 160,125 (+)     Alignment start = 160,214 (-)
Cigar = 36M                       Cigar = 30M1D6M
Mapped = yes                      Mapped = yes
Mapping quality = 50              Mapping quality = 50
Secondary = no                    Secondary = no
Supplementary = no                Supplementary = no
Duplicate = no                    Duplicate = no
Failed QC = no                    Failed QC = no
----------------------            ----------------------
----------------------            Base = A
Mate is mapped = yes              Base phred quality = 38
Mate start = chrXV:160213 (-)     ----------------------
Insert size = 126                 Mate is mapped = yes
First in pair                     Mate start = chrXV:160124 (+)
Pair orientation = F1R2           Insert size = -126
----------------------            Second in pair
MD = 36                           Pair orientation = F1R2
XG = 0                            ----------------------
NH = 1                            MD = 30^C2G3
NM = 0                            XG = 1
XM = 0                            NH = 1
XN = 0                            NM = 2
XO = 0                            XM = 1
AS = 0                            XN = 0
XS = -                            XO = 1
YT = UU                           AS = -13
-------------------               XS = -
                                  YT = UU
                                  -------------------
```

http://www.broadinstitute.org/igv

IGV

- split screen view



IGV

Human hg18 | chr1 | chr1:19,332,457-19,334,716 Go

chr1 — chr1:19,332,457-19,334,716

chr7 — chr7:66,172,833-66,175,092

NA12878 SLX (CEU daughter) Coverage

NA12878 SLX (CEU daughter)

NA12878 SOLID (CEU daughter) Coverage

NA12878 SOLID (CEU daughter)

RefSeq genes

6 tracks loaded

346M of 734M

http://www.broadinstitute.org/igv

Hands-on session (not graded)

- run TRF and NCRF and compare the results for the GGAAT repeat
- visualize the alpha satellites in the centromeric portion of the human Y chromosome
- solve the Rosalind problem "Error correction in reads"

Thank you for your attention