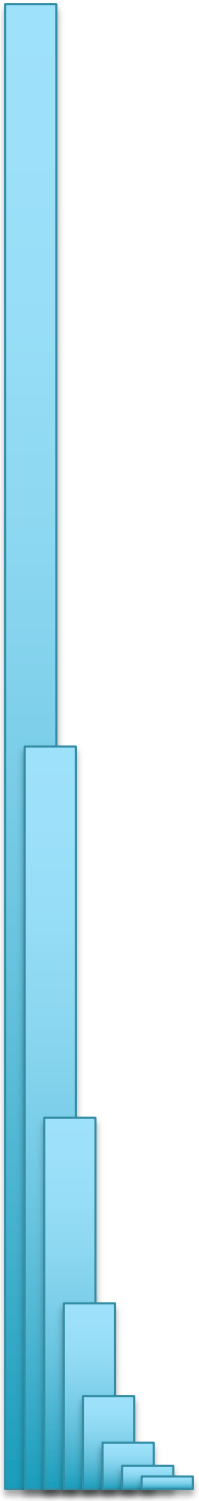
Read Mapping & Variant Calling

Michael Schatz

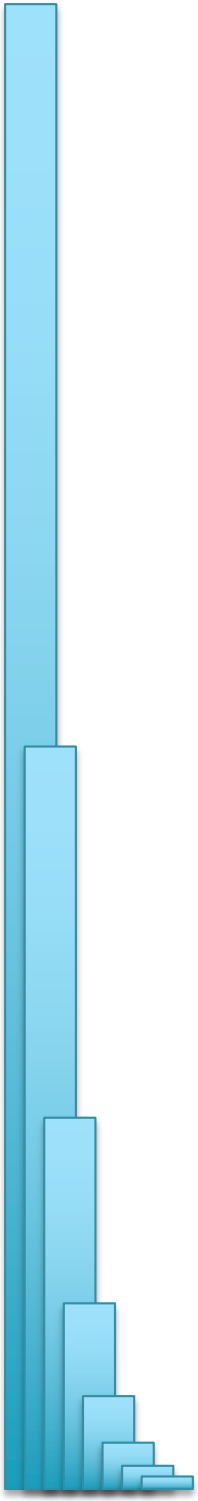
September 18, 2020

CMDB Quantitative Biology



Part 1: Recap
Part 2: The Human Genome
Part 3: Long Read Sequencing
Part 4: Read Alignment
Part 5: Variant Calling
Part 6: Structural Variant Calling

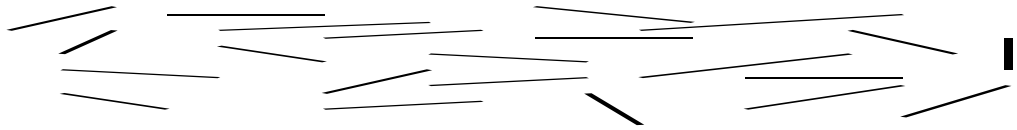
Assignment 2!



Part 1: Recap

Assembling a Genome

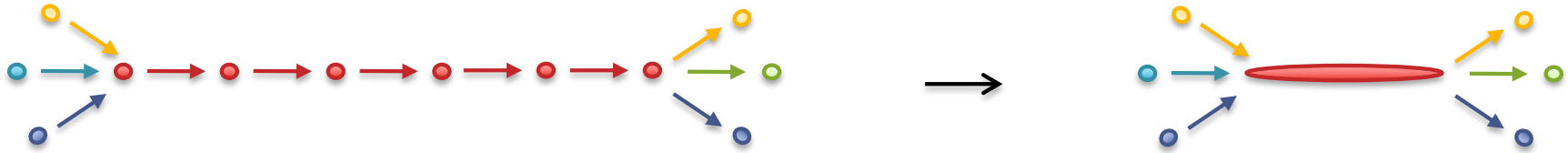
1. Shear & Sequence DNA



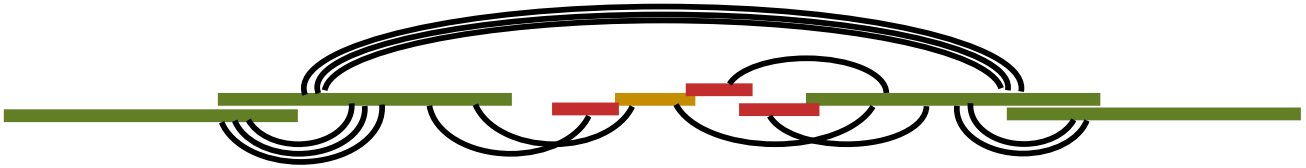
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

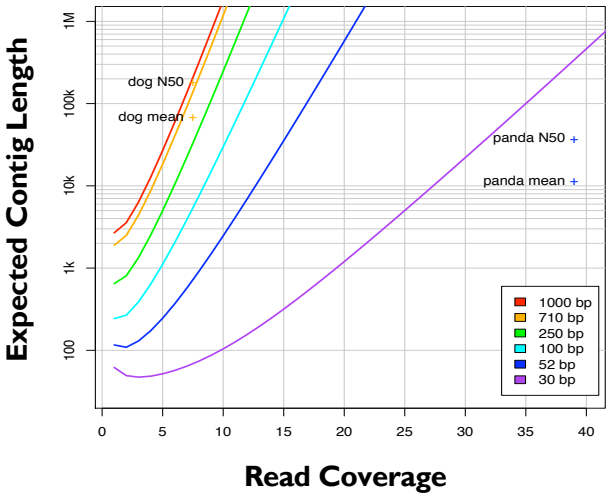


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

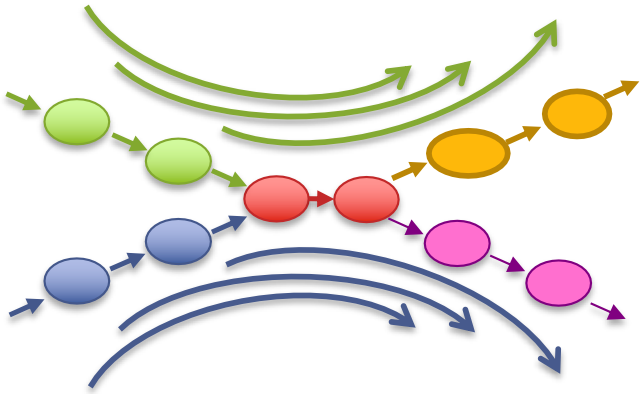
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

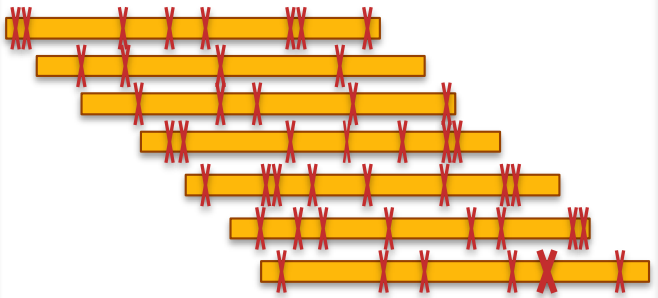
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



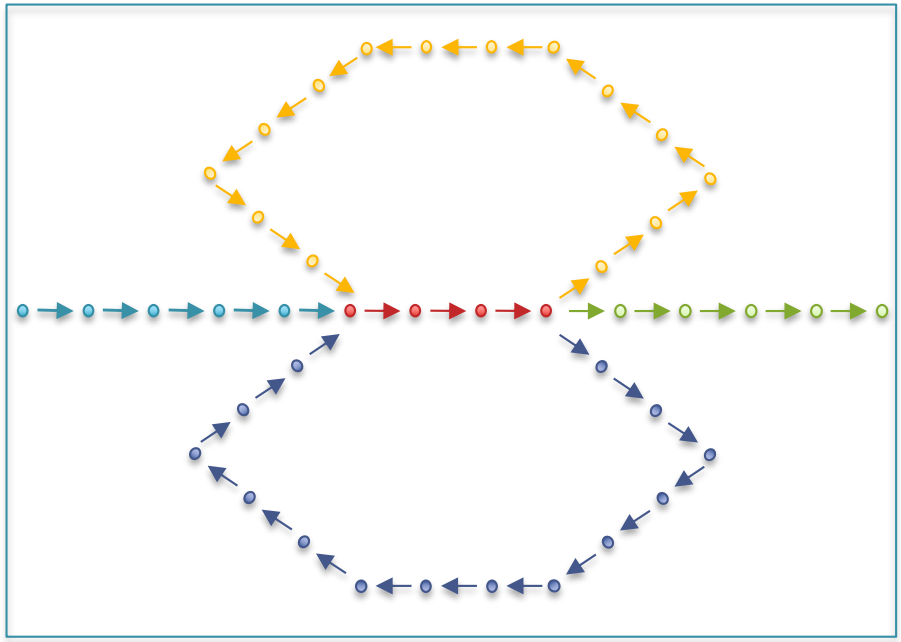
Errors obscure overlaps

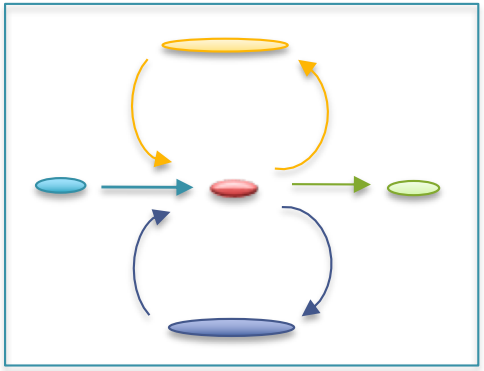
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"





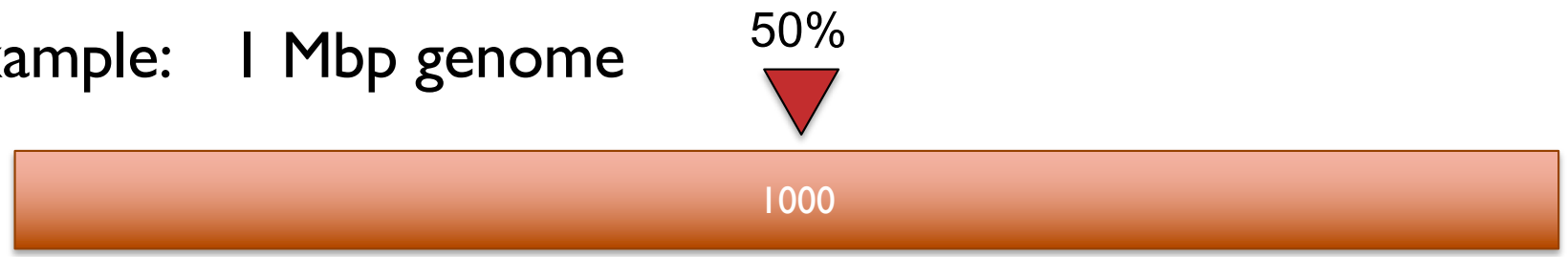
Why do contigs end?

(1) End of chromosome! ☺, (2) lack of coverage, (3) errors,
(4) heterozygosity and (5) repeats

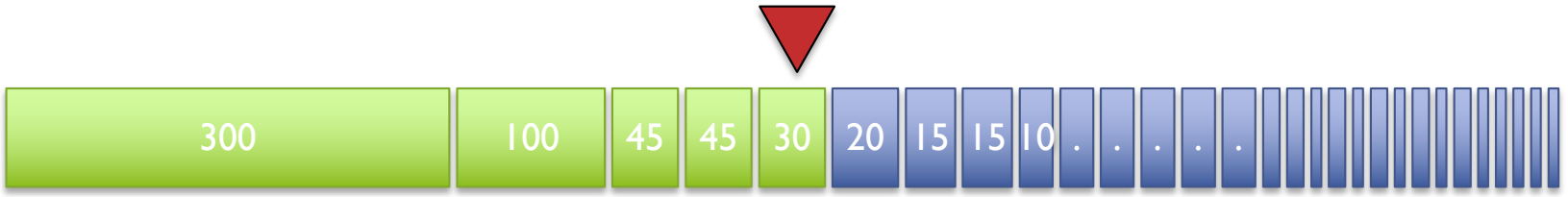
Contig N50

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome

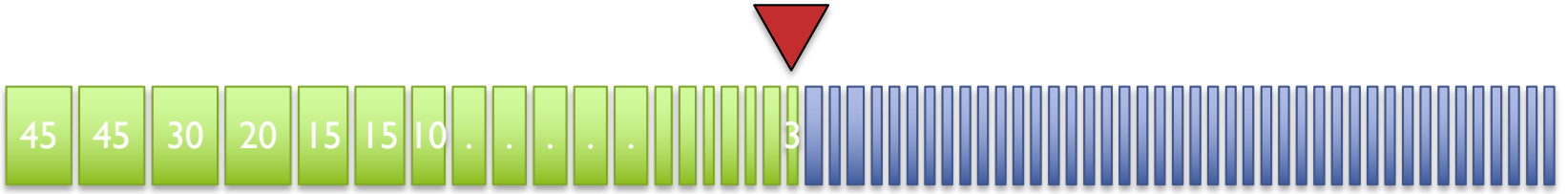


A



N50 size = 30 kbp

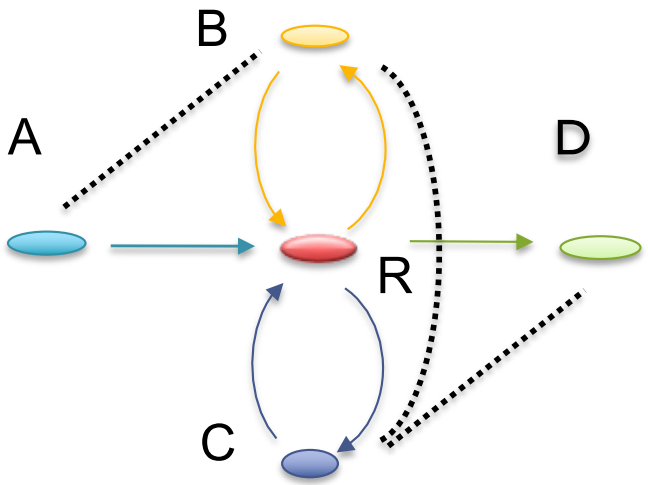
B



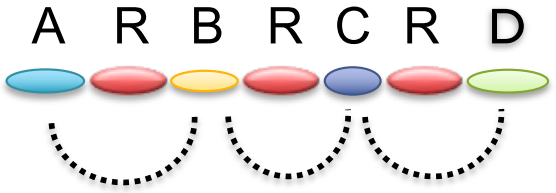
N50 size = 3 kbp

Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries



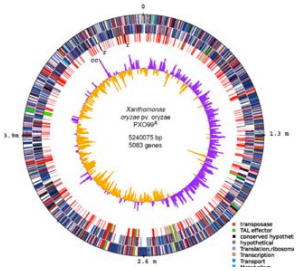
- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

Why do scaffolds end?

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

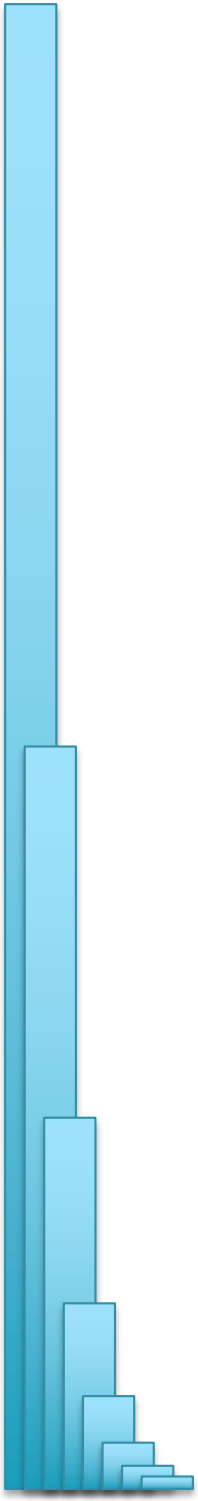
- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

Genome Assembly Lab
Due September 18 @ 11:59pm

1. ***Initialize Tools***
2. ***Download Reference Genome & Reads***
3. ***Decode the secret message***
    1. *Estimate coverage, check read quality*
    2. *Check kmer distribution*
    3. *Assemble the reads with spades*
    4. *Align to reference with MUMmer*
    5. *Extract foreign sequence*
    6. *dna-decode.py -d*

http://bxlab.github.io/cmdb-lab/



Part 2: The human genome

The scale of DNA in our body is staggering.

- **A typical human is comprised of roughly 40 trillion human cells**
  (excluding trillions of bacterial cells in our gut)

- If stretched out, each haploid genome would be **roughly 2 meters**.
- So, each cell has 4 meters of DNA.
- 40 trillion * 4 meters = 160 trillion meters.
- 160 trillion meters / 1609.34 = 99,750,623,441 miles
- 99,750,623,441 / 92,960,000 = 1,073.05 trips to the sun.

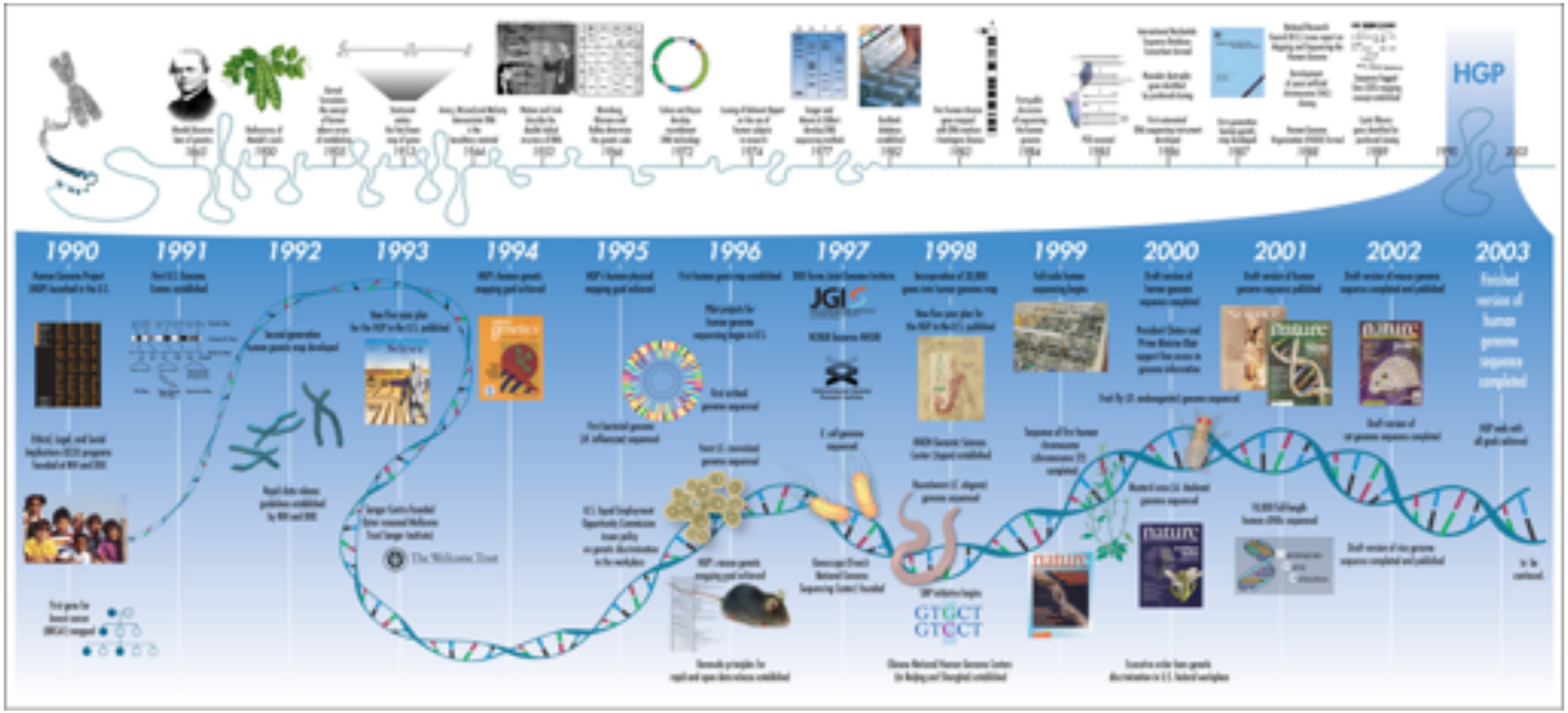
A typical cell replicates about 100 times

160 trillion meters x 100 =

1.69123746 light years

More info

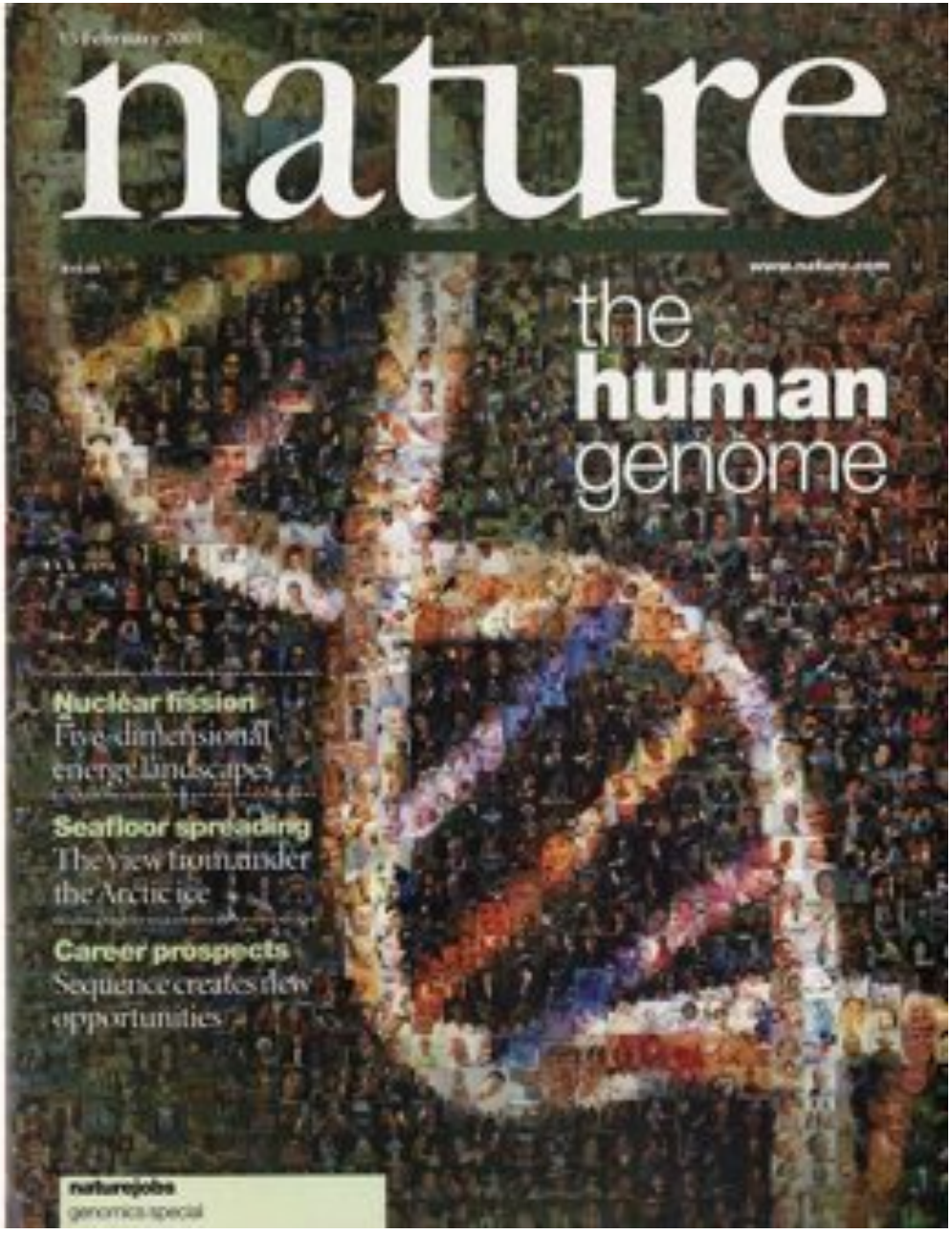
History of the Human Genome Project



The reference human genome

The reference human genome

The Sequence of the Human Genome
Venter et al.
Science 291. pp 1304-1351 (2001)



Initial sequencing and analysis of the human genome
International Human Genome Sequencing Consortium
Nature 409, pp 860–921 (2001)

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

opic. But the very nature of gov- creates a mind set that inspires increase their authority, always at cost of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master." o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the of the violent militia movement. n-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ıment abuse, civil disobedience

opic. But the very nature of gov- creates a mind set that inspires increase their authority, always at cost of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master."

o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the t of the violent militia movement. n-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Who is the reference human?

The Buffalo News/Sunday, March 23, 1997

ment abuse, civil disobedience

opic. But the very nature of gov- creates a mind set that inspires increase their authority, always at cost of the people."

y, government has forgotten that rvant of the people," Parlato add- acting more like it's the master."

o and the Lapps share an abiding non-violent civil disobedience.

insist on being respectful in our of resistance," Barbara Lyn Lapp ut if we claim to care about our rs, we must protest government in-

violence has to be the watchword, said, calling civil disobedience the st of the violent militia movement. n-violence can serve as an anti- government oppression, he added. law is unjust or you're given an ithout moral or legal authority, you should refuse it," Parlato said. "And, if need be, you have to be brave enough to accept the consequences."

Rachel Lapp says she believes government can be good, when it controls the aggressors in society. Instead, it too often comes down on the side of the aggressors, who enforce child-protection laws, compulsory education, disclosure rules on tax forms and seat belt laws.

"We want people to see the correlation between what happened to us and what can happen to anyone when government gets out of hand," Rachel Lapp said.

The Lapps and Parlato will be joined by Samuel Radford III, a critic of public education who was arrested and pleaded guilty to reduced charges following a 1993 disturbance at the City Campus of Erie Community College.

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (human blueprint) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK

For more information please contact the
Clinical Genetics Service
843-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Pieter de Jong, RPCI

Appendix: Identifying the ancestry of segments of the human genome reference sequence

To compare Neandertal to present-day human haplotypes for the purpose of population genetic analysis, we needed to have long haploid sequences from present-day humans that were of known ancestry. To identify such segments, we took advantage of the fact that the human reference sequence is haploid over scales of tens of kilobases, because it is comprised of a tiling-path of Bacterial Artificial Chromosomes (BACs) or other clone types that are of typical size 50-150 kb (*S92*). We do not know of any other substantial source of high quality human haploid sequences of the requisite size.

Determining the ancestries of the libraries in the human genome reference sequence using HAPMIX

It is crucial to know the 'ancestry' of a clone to use it in a meaningful population genetic analysis. In what follows, we define 'ancestry' as the geographic region in which a clone's ancestor lived 1,000 years ago, inferred based on its genetic proximity to other individuals from that region today. This definition allows us to classify clones from Chinese Americans as "East Asian," from European Americans as "European", and from African Americans as either "West African" or "European".

To identify the ancestries of the libraries comprising most of the human genome reference sequence, we used a list of 26,558 clones tiling the great majority of the genome, most of which we were able to assign to a library of origin. Restricting to the autosomes, we identified 21,156 clones that seemed to fall into 9 libraries based on the naming scheme: CTA (n=199), CTB (n=356), CTC (n=452), CTD (n=1,426), RPCI-1 (n=740), RPCI-3 (n=456), RPCI-4 (n=716), RPCI-5 (n=802) and RPCI-11 (n=16,009). (In a subsequent re-examination, we identified additional clones that we likely could have classified into libraries, including 953 from RPCI-11, 632 from RPCI-1, and 490 from another library RPCI-13.) The median span of the 21,156 clones we analyzed was 112 kb, and 80% are >50kb in size. About 2/3 came from a single library, RPCI-11.

1. RPCI-11 is an African American: RPCI-11, the individual who contributed most of the human genome reference sequence, is consistent with having African American ancestry, with 42% of the clones of confident West African ancestry and 42% of the clones of confident European ancestry, and the ancestry of the remaining clones less confidently inferred. The finding of likely African American ancestry for RPCI-11 was previously reported in a study of the ancestry of RPCI-11 clones spanning the Duffy blood group locus (*S93*), and here we confirm this finding, and also expand the inference to the whole genome.
2. CTD is an East Asian: The majority of clones from CTD, the second largest library in its contribution to the human genome sequence, is likely an East Asian. In a HAPMIX analysis with CEU (European) – CHB+JPT (East Asian) as the proposed ancestral populations, the majority of clones are of confident East Asian origin, and there is no secondary mode of confident European ancestry, as might be expected from a Latino or South Asian individual.
3. The remaining 7 libraries are European: The remaining libraries (CTA, CTB, CTC, RPCI-1, RPCI-3, RPCI-4 and RPCI-5) are inferred to be of European ancestry, since they all have consistent distributions of inferred clone ancestries, with the majority of clones of confident European ancestry in both our HAPMIX analyses and no secondary modes.

A Draft Sequence of the Neandertal Genome

Green et al (2010) Science. DOI: 10.1126/science.1188021
Supplemental Note 16 (pg 145-146)

Who is the reference human?

nature methods
Techniques for life scientists and chemists

EDITORIAL

Nature Methods **7**, 331 (2010)
doi:10.1038/nmeth0510-331

E pluribus unum

If the human reference genome is to reflect more of the actual genomic diversity in humans, community participation is needed.

Please visit *methagora* to view and post comments on this article.

The human genome is ten years old. We acknowledge its reference assembly as an invaluable resource essential for many purposes such as the assembly of short reads from high-throughput sequencing platforms into chromosome context during resequencing projects. At the same time, we think necessary improvement of the reference genome depends on the willingness of the research community to provide data for the genome's less accessible regions.

First published in 2001, the human reference genome has, since 2007, been in the hands of the Genome Reference Consortium (GRC) a small group of fewer than 20 scientists from the European Bioinformatics Institute, the US National Center for Biotechnology Information, The Sanger Institute and The Genome Center at Washington University in St. Louis, who have committed to the improvement and completion of this reference, with very little financial support.

The reference genome is now in its 19th rendition, and probably the best measure of its improvement over the last ten years is the number of fragments it consists of. The very first version had ~150,000 gaps; the most recent build, GRCh37, has only around 250 gaps.

The only other publicly accessible *de novo* assembly of a human genome that contains chromosome sequences is HuRef. Obtained by traditional capillary sequencing, HuRef is the diploid genome of Craig Venter. It comes in 4,500 pieces and, like any individual genome, it contains many rare alleles.

GRCh37, in contrast, is a mosaic haploid genome derived from about 13 people. It still contains rare alleles, but the GRC recently decided to convert these to common haplotypes. Deciding which alleles are common and which are rare is proving challenging, and the GRC members are collaborating with members of the 1000 Genomes project to collect enough data to make these decisions.

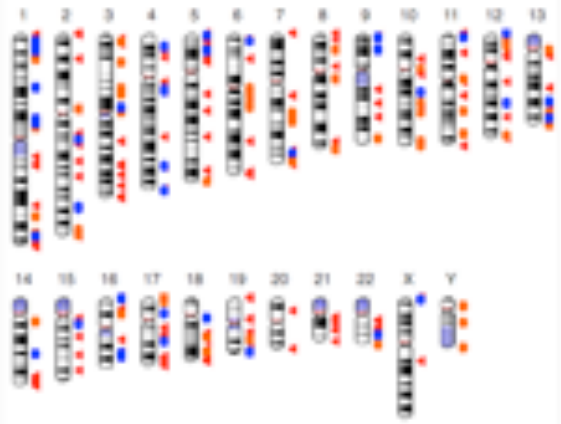
Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human Overview | Human Genome Issues | Human Assembly Data

Human Genome Overview

Information about the continuing improvement of the human genome



Region containing alternate loci
Region containing fix patches
Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p11

The GRC is working hard to provide the best possib
by both generating multiple representations (alterna
represented by a single path. Additionally, we are re
allows users who are interested in a specific locus t
affecting users who need chromosome coordinate s

Download data:

- GRCh38.p11 (latest minor release) FTP
- GRCh38 (latest major release) FTP
- Genomic regions under review FTP
- Current Tiling Path Files (TPFs)

Transitioning to GRCh38? Try the NCBI Remapp
assembly alignments used by the GRC.

Next assembly update
The next assembly update (GRCh38.p12) will be

GRCh38.p11 | GRCh37.p13 | GRCh37

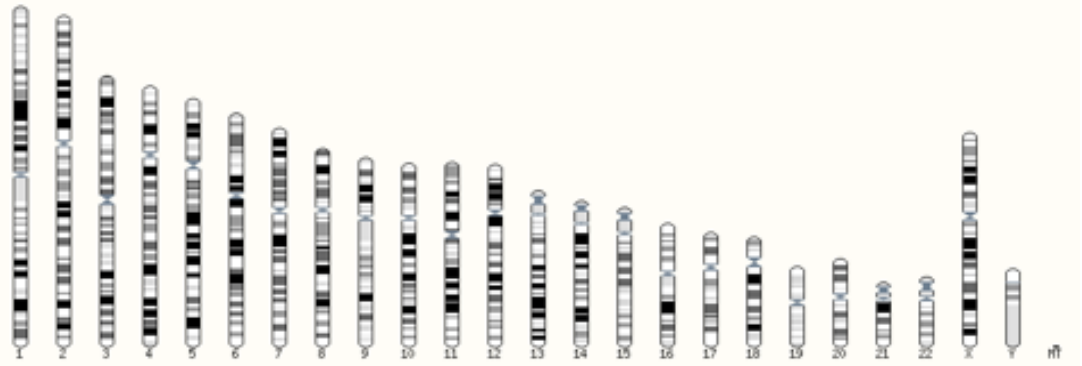
GRCh38.p11

Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordi
of patch scaffolds is now: 64 FIX and 59 NOVEL.
Assembly accessions: GenBank: GCA_000001405.26 , RefSeq: GCF_000001405.37

Pseudoautosomal regions

Name	Chr	Start	Stop
PAR#1	X	10,001	2,781,479
PAR#2	X	155,701,383	156,030,895
PAR#1	Y	10,001	2,781,479
PAR#2	Y	56,887,903	57,217,415

The human genome - basic stats



- 3.096 billion base pairs (haploid)
- 20,454 protein coding genes
- 226,950 coding transcripts (isoforms of a gene that each encode a distinct protein product)

Assembly	GRCh38.p12 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.27, Dec 2013
Base Pairs	3,609,003,417
Golden Path Length	3,096,649,726
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Mar 2019
Database version	97.38
Gencode version	GENCODE 31

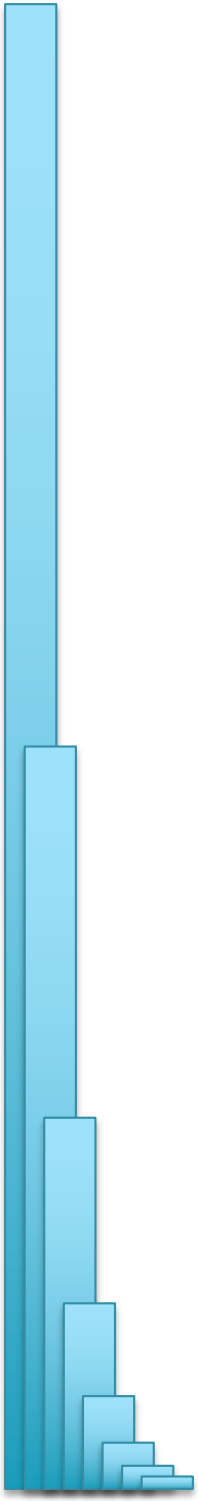
Gene counts (Primary assembly)

Coding genes	20,454 (incl 660 readthrough)
Non coding genes	23,940
Small non coding genes	4,871
Long non coding genes	16,848 (incl 302 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,204 (incl 8 readthrough)
Gene transcripts	226,950

http://uswest.ensembl.org/Homo_sapiens/Location/Genome

The human reference genome continues to change.

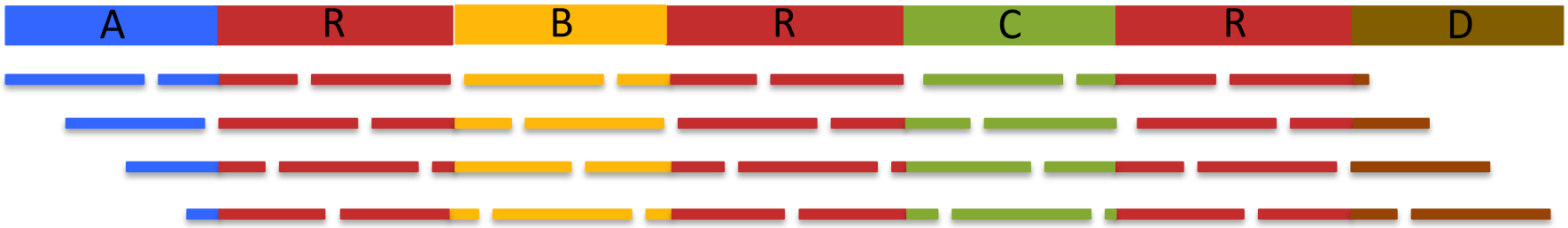
- Ongoing efforts to fill "gaps" and properly/thoroughly represent complex structures and loci in the genome (e.g., Major Histocompatibility Complex)
- Each improvement leads to a new genome "build". Currently on build 38.
- Experimental and computational methods provide new genome annotations
  - New gene models, transcription factor binding sites, and loci where human individuals differ (i.e., polymorphisms)
- Therefore, the human reference genome is by no means "complete"!
- How does the same genome yield such phenotypic diversity across tissue types?
- How does the genome evolve within an individual (tissues) and among a population?

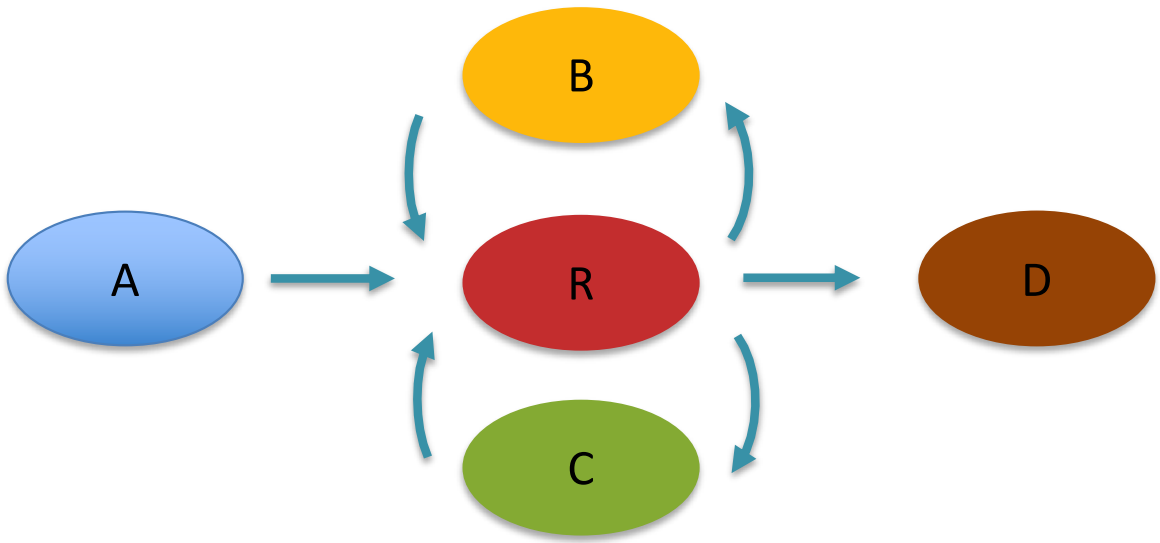


Part 3: Long Reads

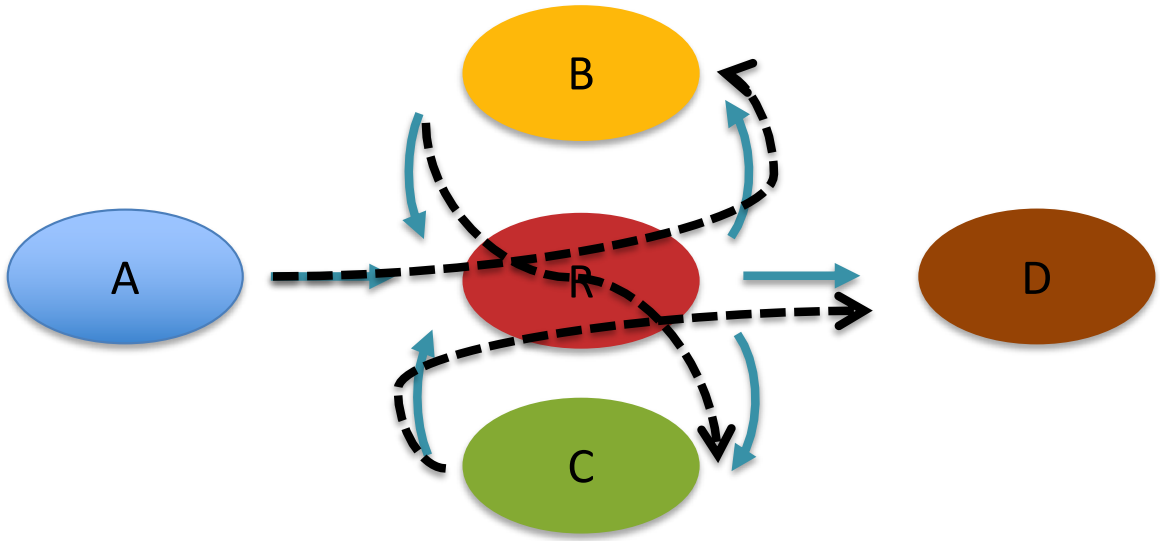
Genomics Arsenal in the Year 2020

Assembly Complexity

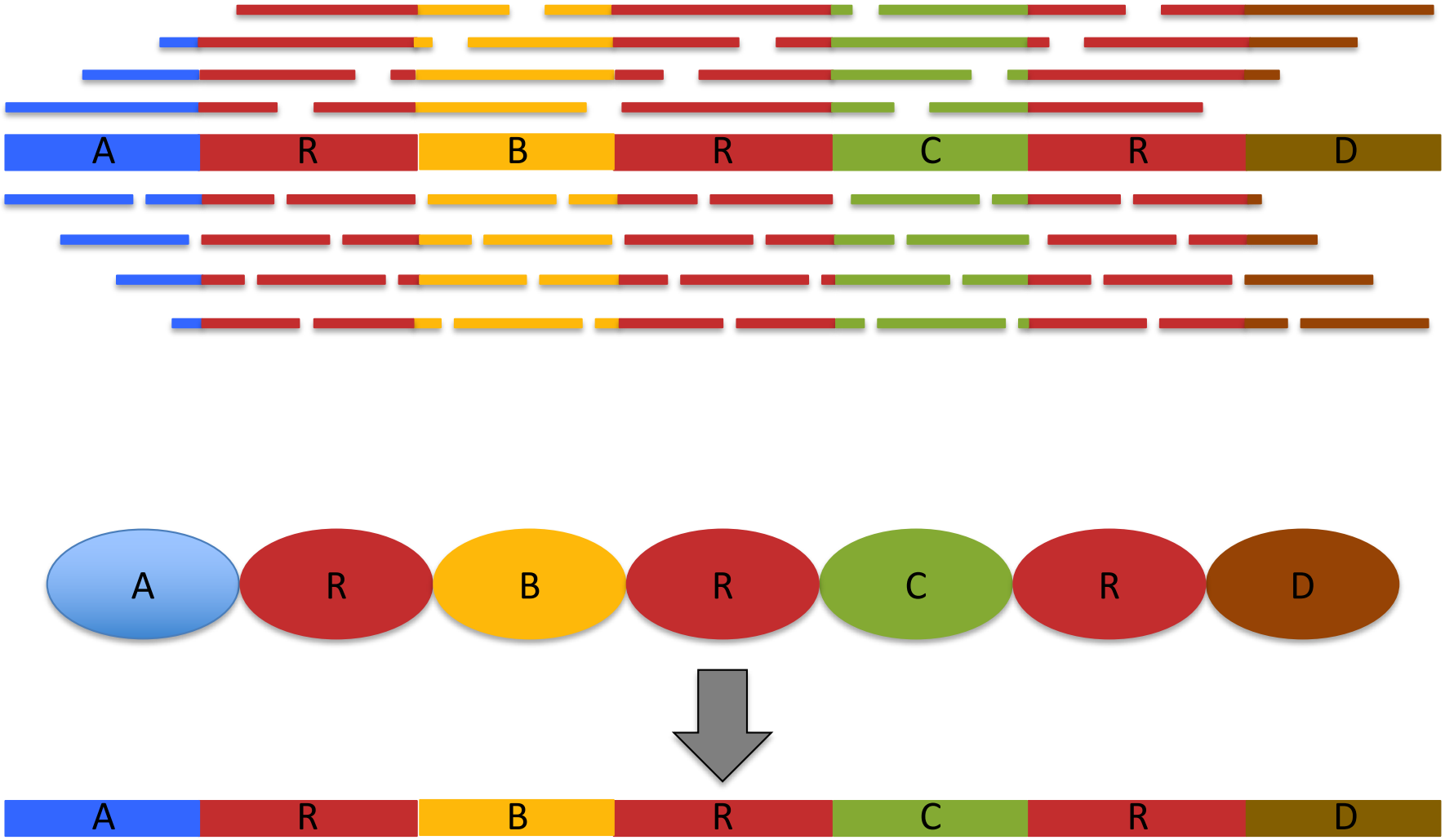




Assembly Complexity



Assembly Complexity



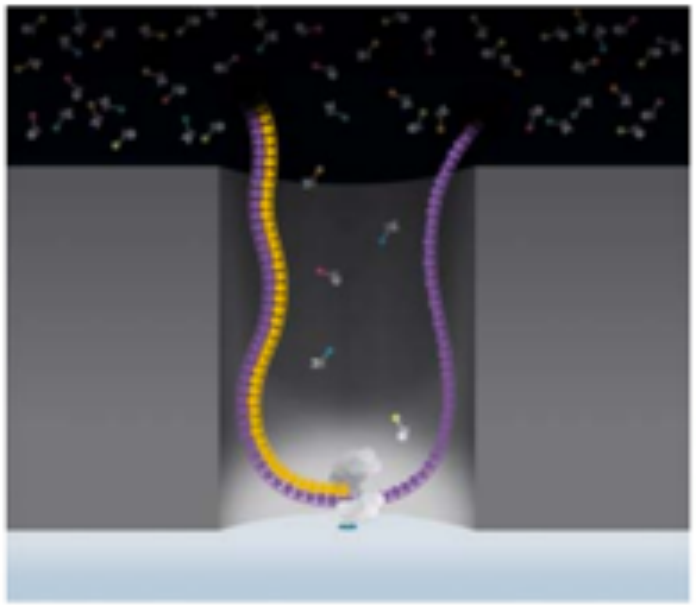
The advantages of SMRT sequencing

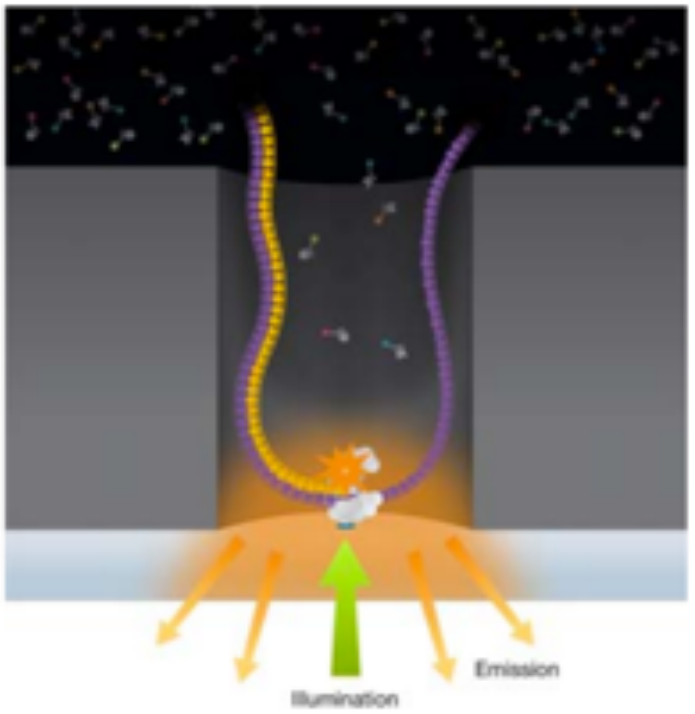
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

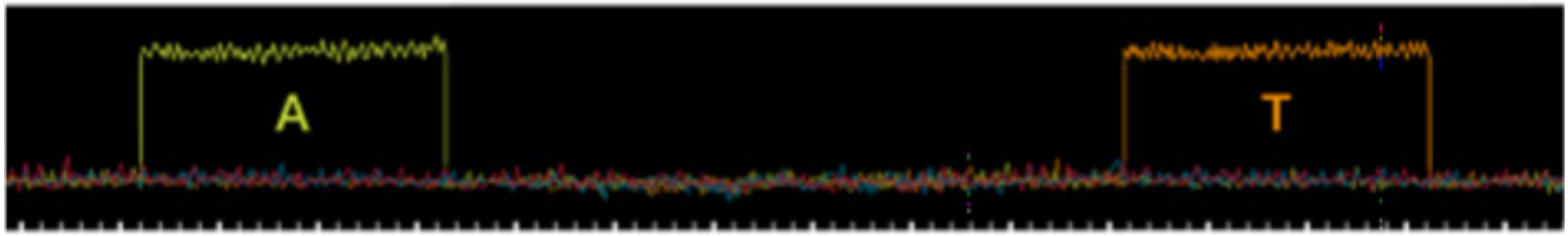
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

PacBio: SMRT Sequencing

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

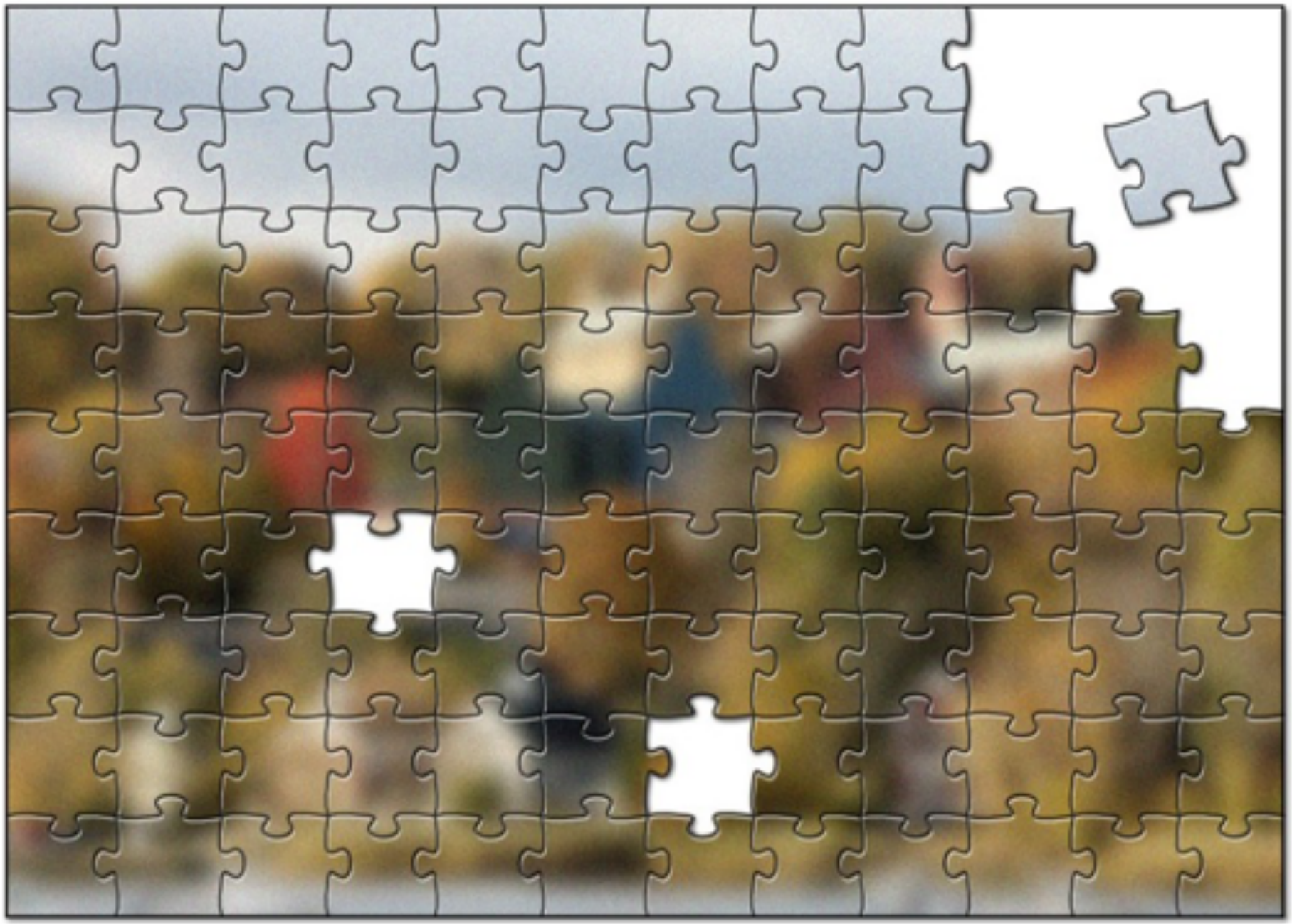




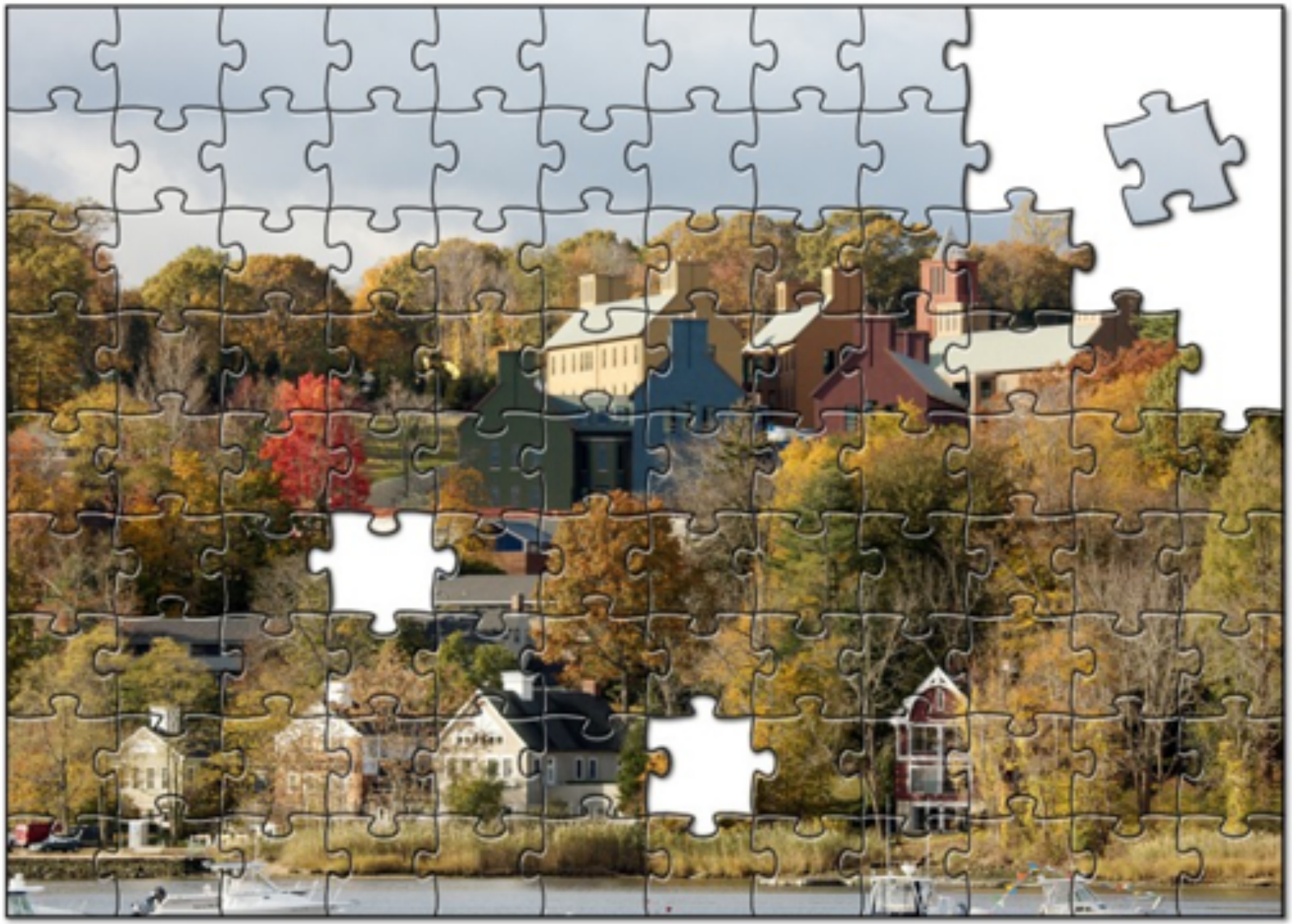


http://www.youtube.com/watch?v=v8p4ph2MAvI

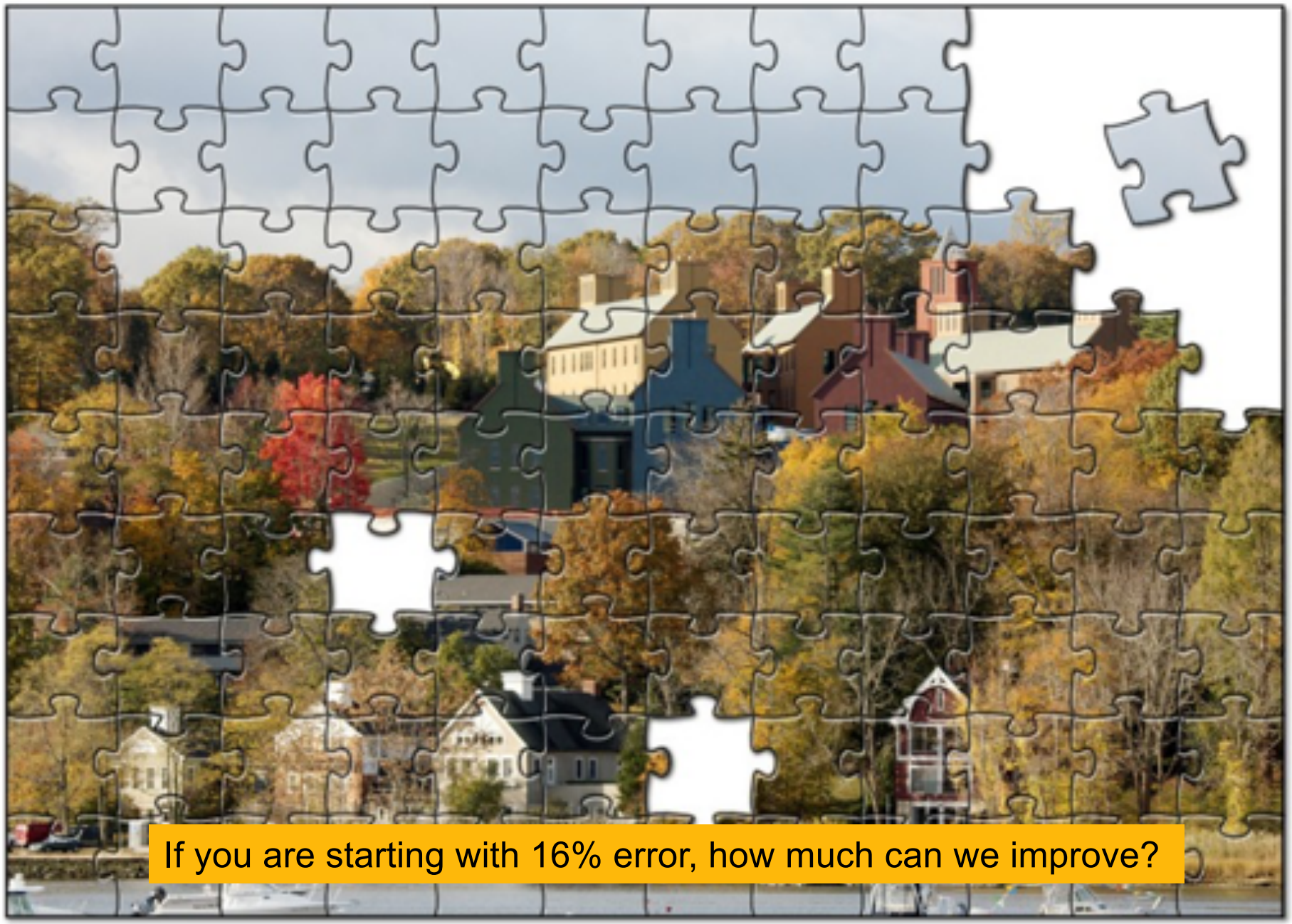
Single Molecule Sequences



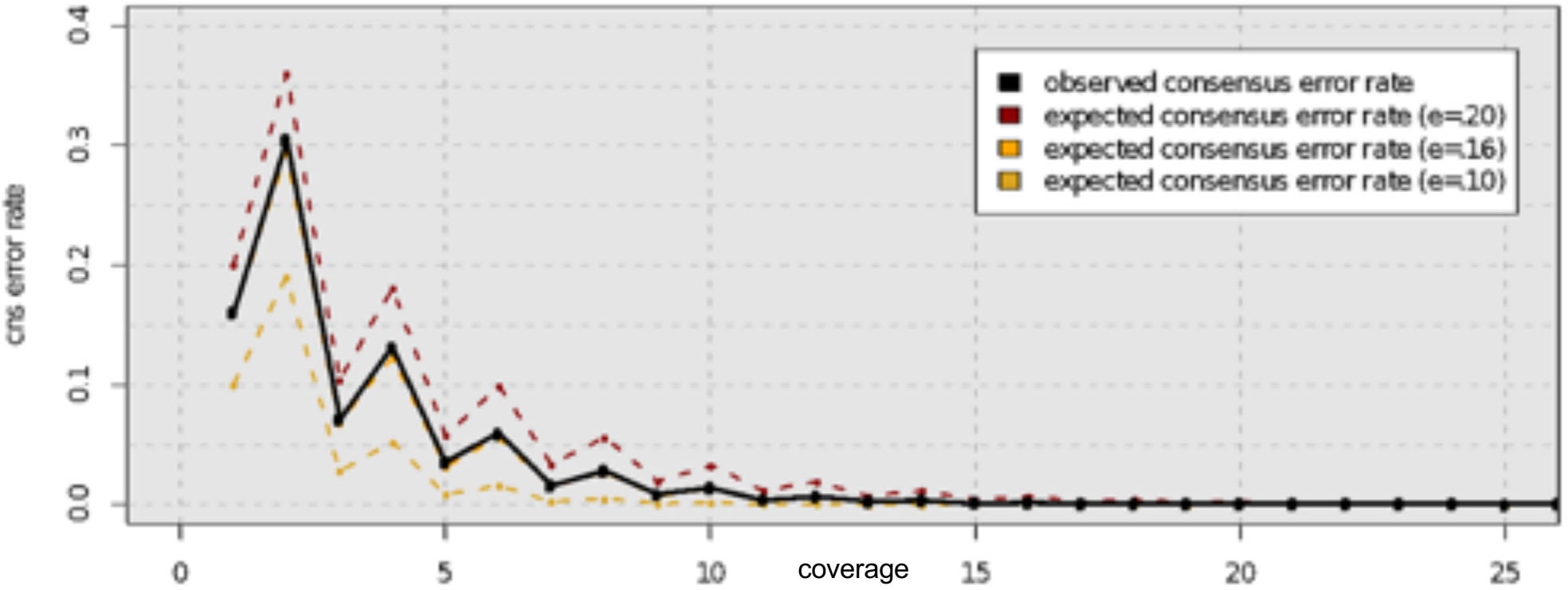
“Corrective Lens” for Sequencing



“Corrective Lens” for Sequencing



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

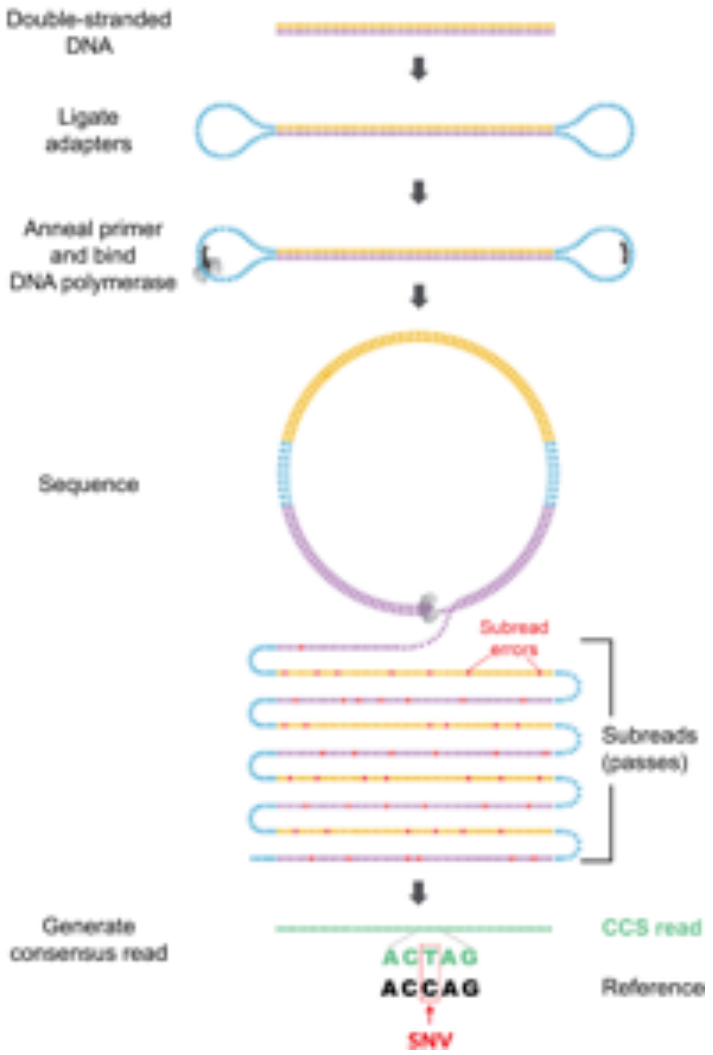
$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

“HiFi” Circular Consensus Reads

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

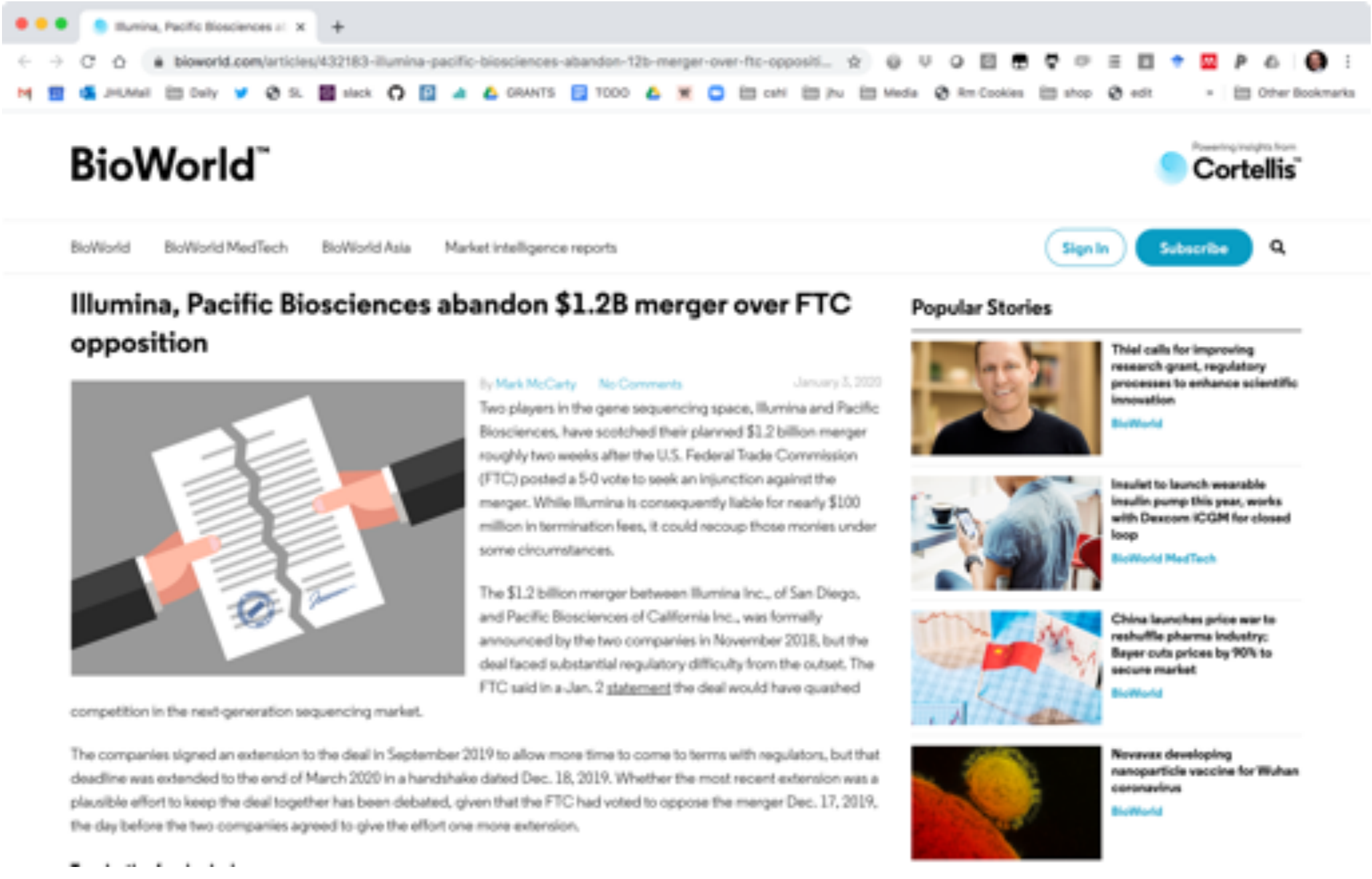
Higher accuracy for low-coverage sequences like somatic variants or lowly expressed transcripts in RNA-seq, more interpretable alignments, faster assembly

Limits read length, used to be very expensive but more manageable now



Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome
Wenger et al (2019) Nature Biotechnology doi:10.1038/s41587-019-0217-9

PACB



BioWorld™

Powering insights from Cortellis™

BioWorld | BioWorld MedTech | BioWorld Asia | Market intelligence reports | Sign In | Subscribe

Illumina, Pacific Biosciences abandon $1.2B merger over FTC opposition

By Mark McCarty | No Comments | January 3, 2020

Two players in the gene sequencing space, Illumina and Pacific Biosciences, have scotched their planned $1.2 billion merger roughly two weeks after the U.S. Federal Trade Commission (FTC) posted a 5-0 vote to seek an injunction against the merger. While Illumina is consequently liable for nearly $100 million in termination fees, it could recoup those monies under some circumstances.

The $1.2 billion merger between Illumina Inc., of San Diego, and Pacific Biosciences of California Inc., was formally announced by the two companies in November 2018, but the deal faced substantial regulatory difficulty from the outset. The FTC said in a Jan. 2 statement the deal would have quashed competition in the next-generation sequencing market.

The companies signed an extension to the deal in September 2019 to allow more time to come to terms with regulators, but that deadline was extended to the end of March 2020 in a handshake dated Dec. 18, 2019. Whether the most recent extension was a plausible effort to keep the deal together has been debated, given that the FTC had voted to oppose the merger Dec. 17, 2019, the day before the two companies agreed to give the effort one more extension.

Popular Stories

Thiel calls for improving research grant, regulatory processes to enhance scientific innovation
BioWorld

Insulet to launch wearable insulin pump this year, works with Dexcom iCGM for closed loop
BioWorld MedTech

China launches price war to reshuffle pharma industry; Bayer cuts prices by 90% to secure market
BioWorld

Novavax developing nanoparticle vaccine for Wuhan coronavirus
BioWorld

pacbio stock

Google

pacbio stock

Sign in

All Finance News Shopping Videos More Settings Tools

About 537,000 results (0.37 seconds)

Market Summary > **Pacific Biosciences of California Inc**
NASDAQ: PACB

7.81 USD −0.060 (0.76%)
Closed: Sep 17, 7:56 PM EDT · Disclaimer
After hours 8.00 +0.19 (2.43%)

1 day 5 days 1 month 6 months YTD 1 year 5 years **Max**

Open	7.89	Div yield	-
High	7.99	Prev close	7.87
Low	7.56	52-wk high	8.10
Mkt cap	1.36B	52-wk low	2.20
PE ratio	-		

More about Pacific Biosciences ...

finance.yahoo.com › ... ▾

Pacific Biosciences of Californ (PACB) Stock Price, News ...

Find the latest Pacific Biosciences of Californ (PACB) stock quote, history, news and other

Pacific Biosciences
Biotechnology company

PACBIO

Pacific Biosciences of California, Inc. is an American biotechnology company founded in 2004 that develops and manufactures systems for gene sequencing and some novel real time biological observation. Wikipedia

Headquarters: Menlo Park, CA

Founded: 2004

Ceo: Michael Hunkapiller pacb.com

Verify these facts to help others.

Key people

Profiles

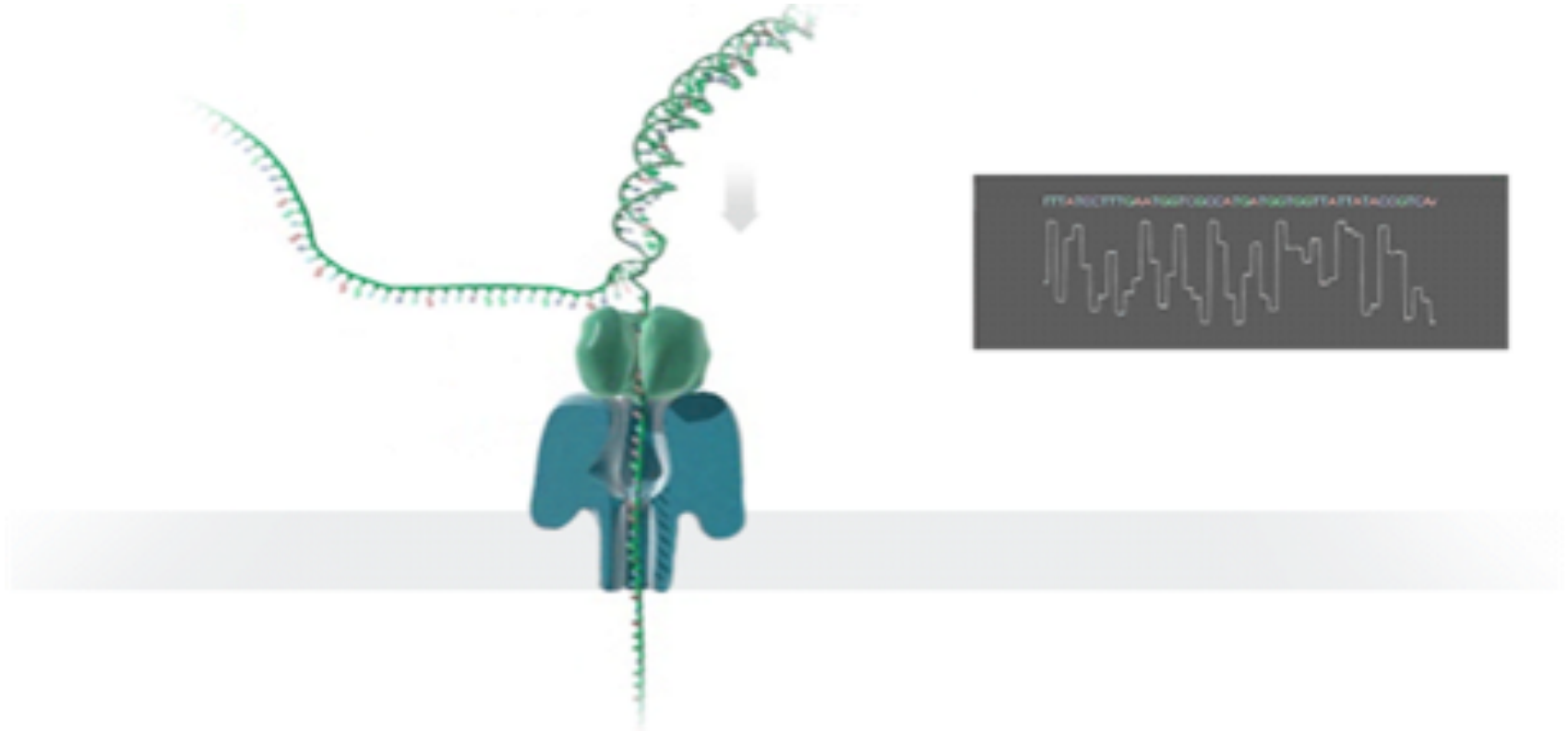
LinkedIn YouTube Twitter

People also search for View 10+ more

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

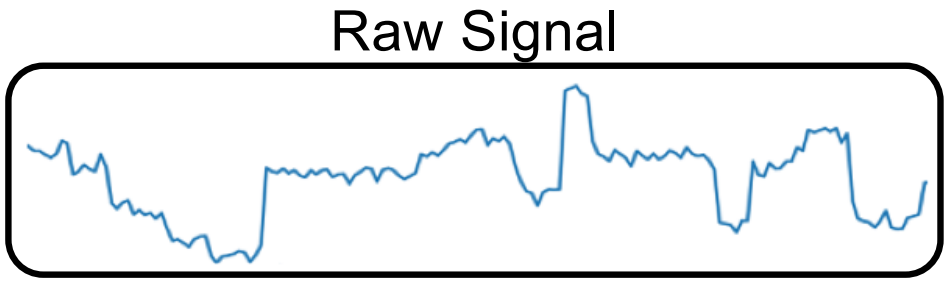
Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore



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

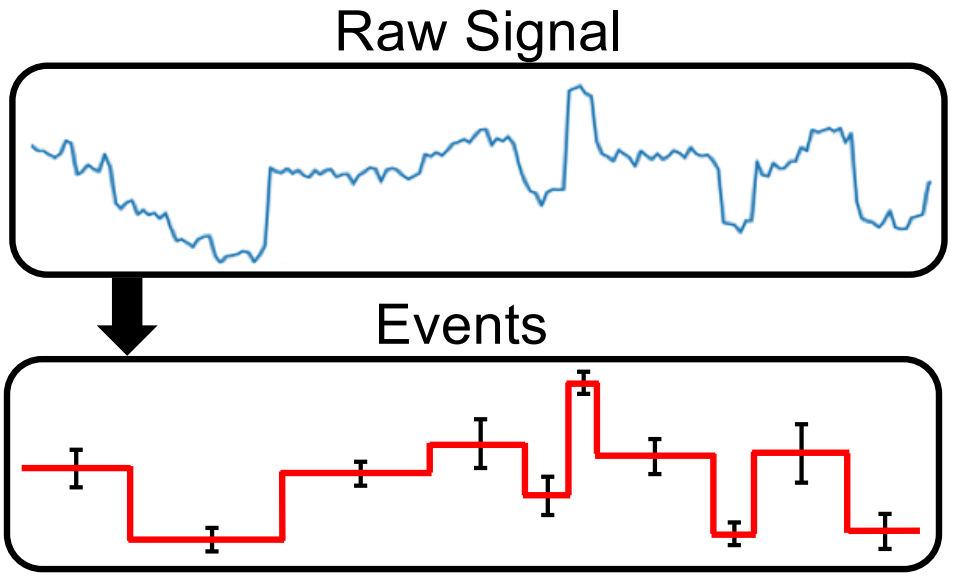
nanoporetech.com/applications/dna-nanopore-sequencing

Nanopore Basecalling



Translation of raw signal into basepairs

Nanopore Basecalling

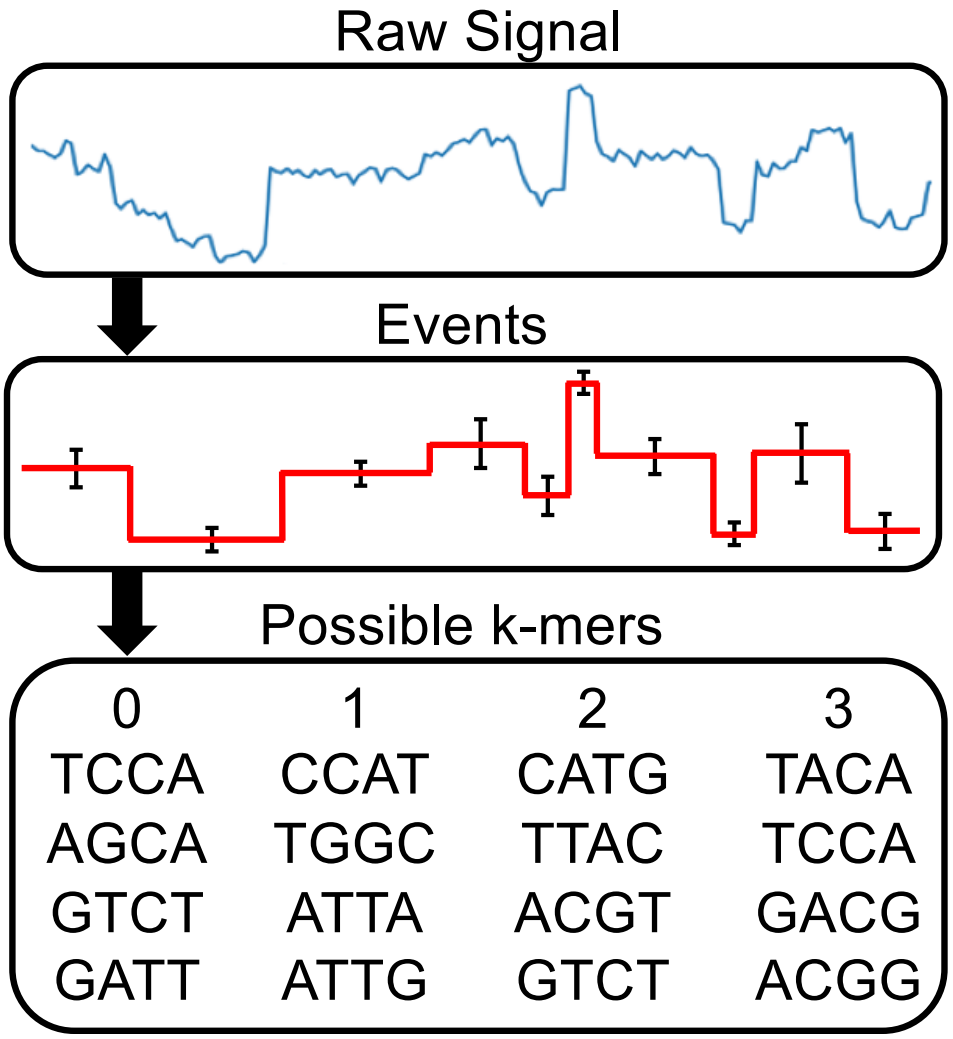


Translation of raw signal into basepairs

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

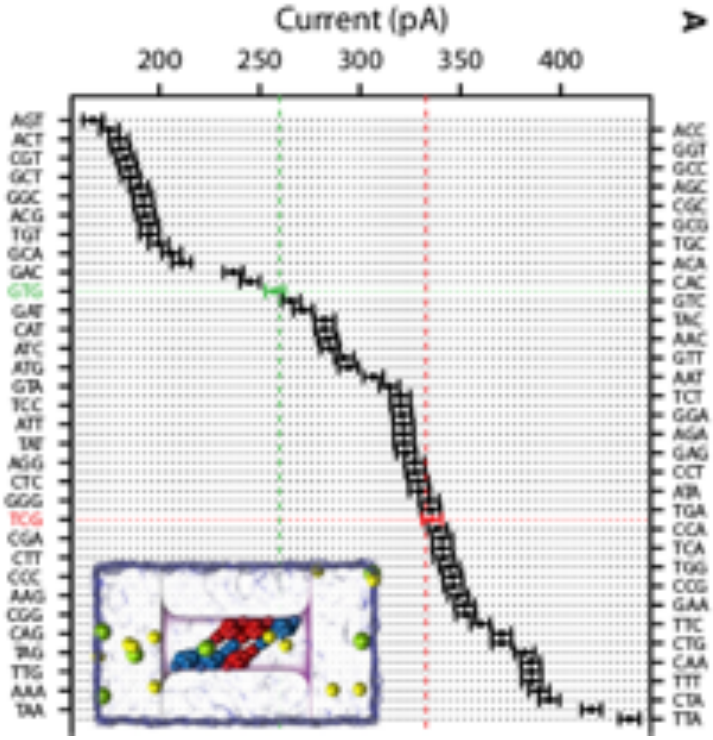
Modern basecalers use neural networks directly on raw signal

Nanopore Basecalling



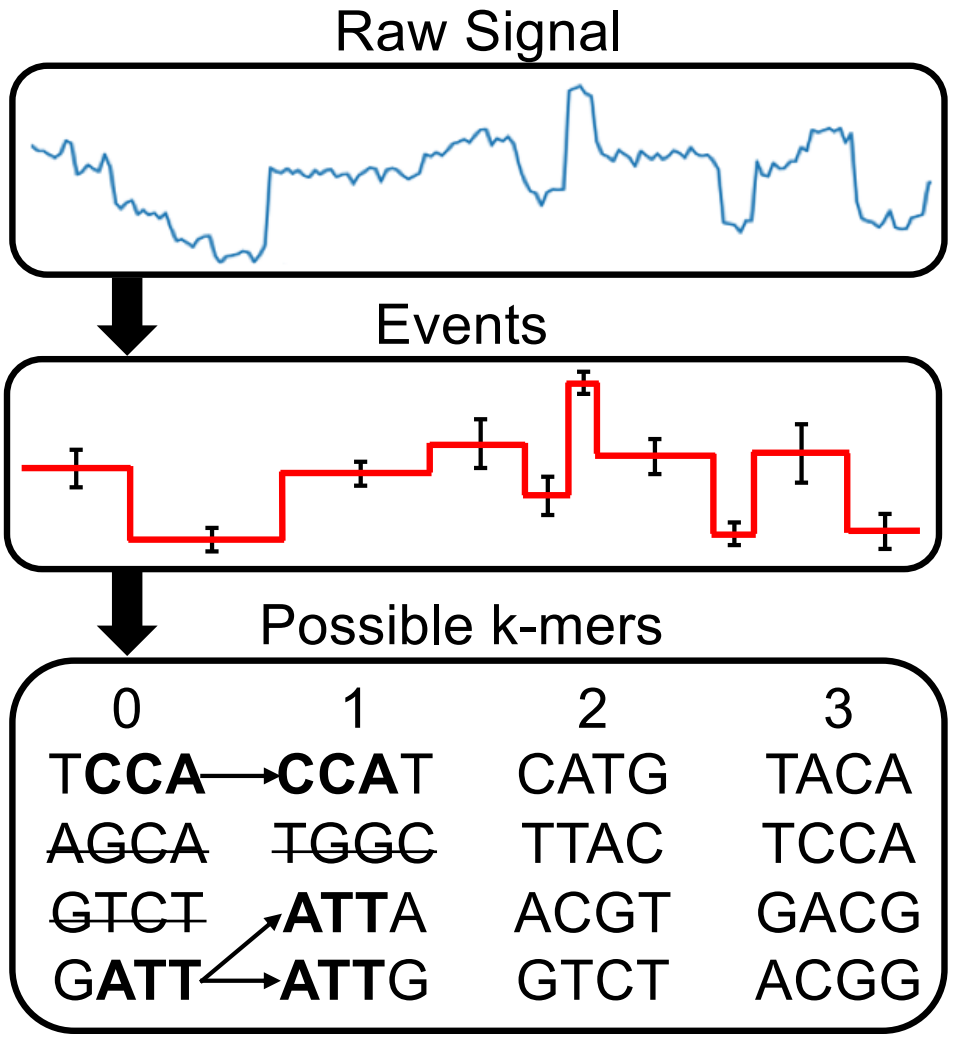
(Based on probability of event matches)

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

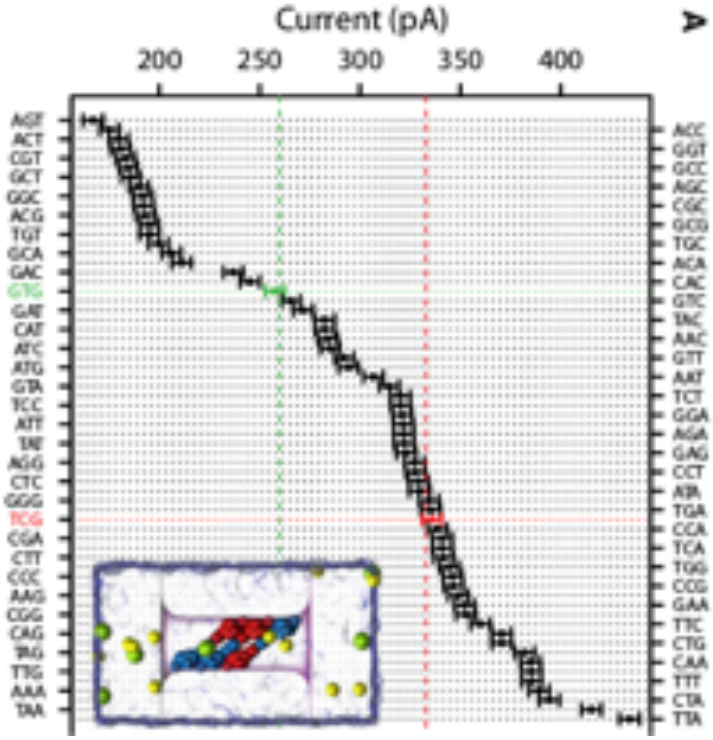


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling

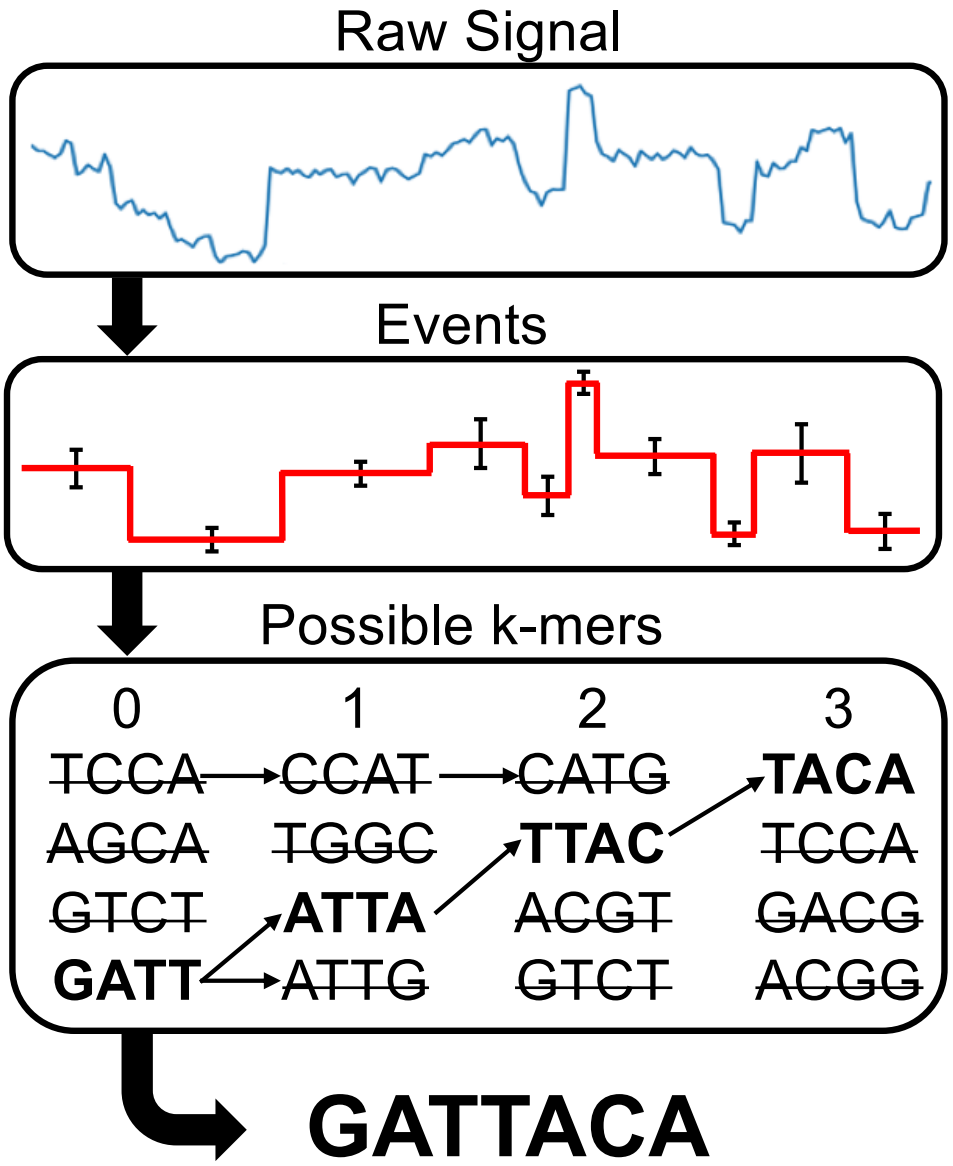


Certain k-mers can be eliminated based on possible transitions

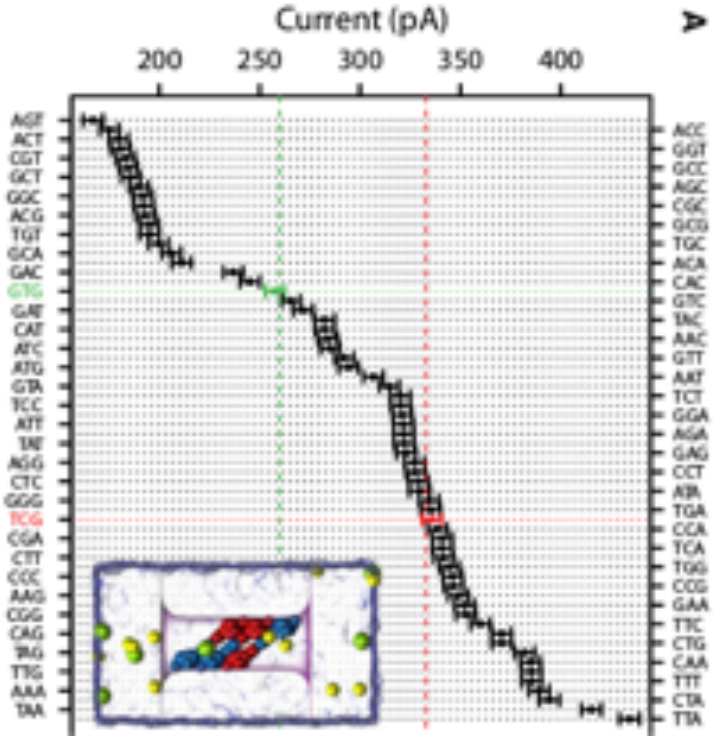


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling



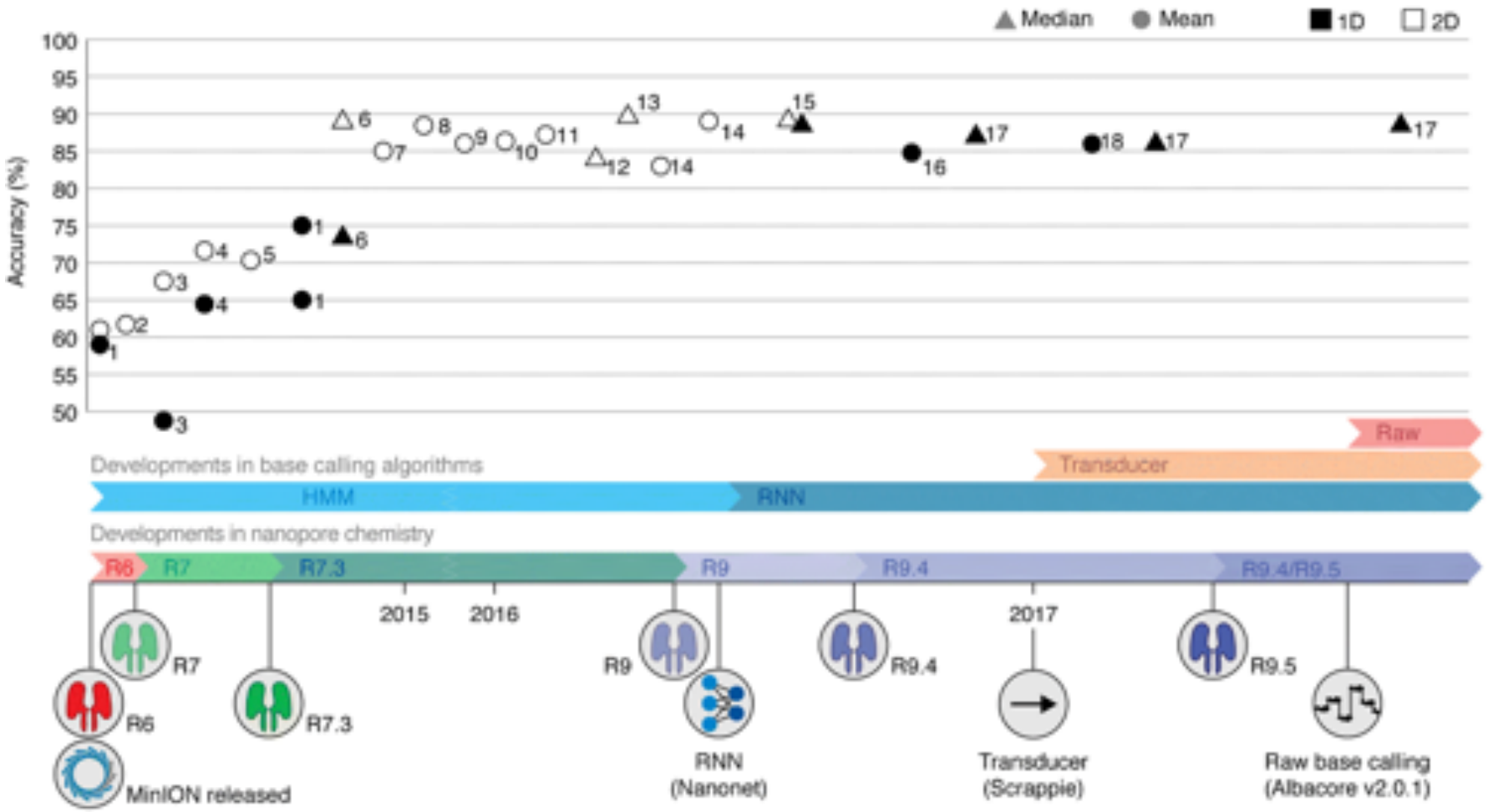
Final sequence determined by most probable k-mers



"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

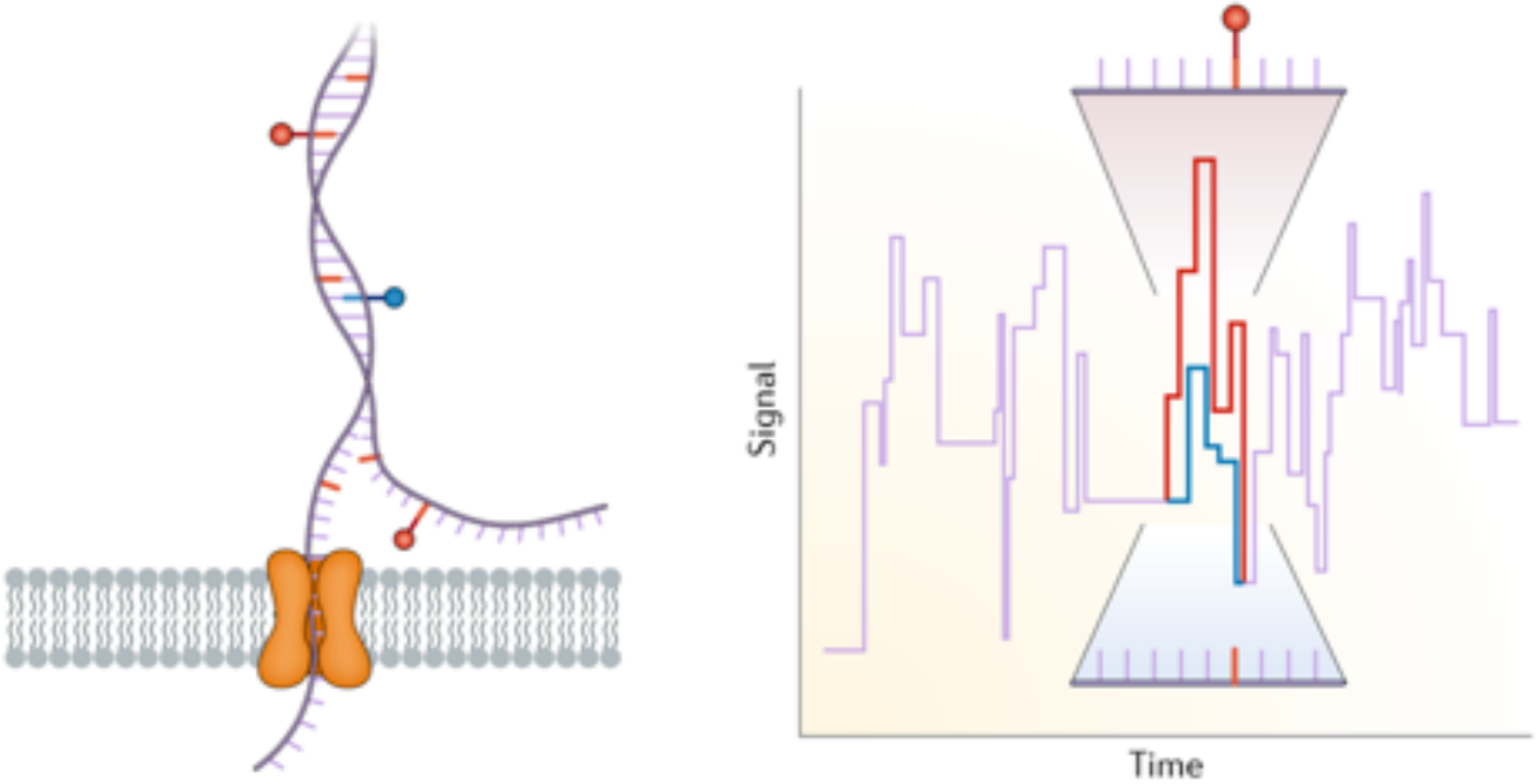


From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology.* https://doi.org/10.1186/s13059-018-1462-9

DNA Modification Detection

Like PacBio, ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



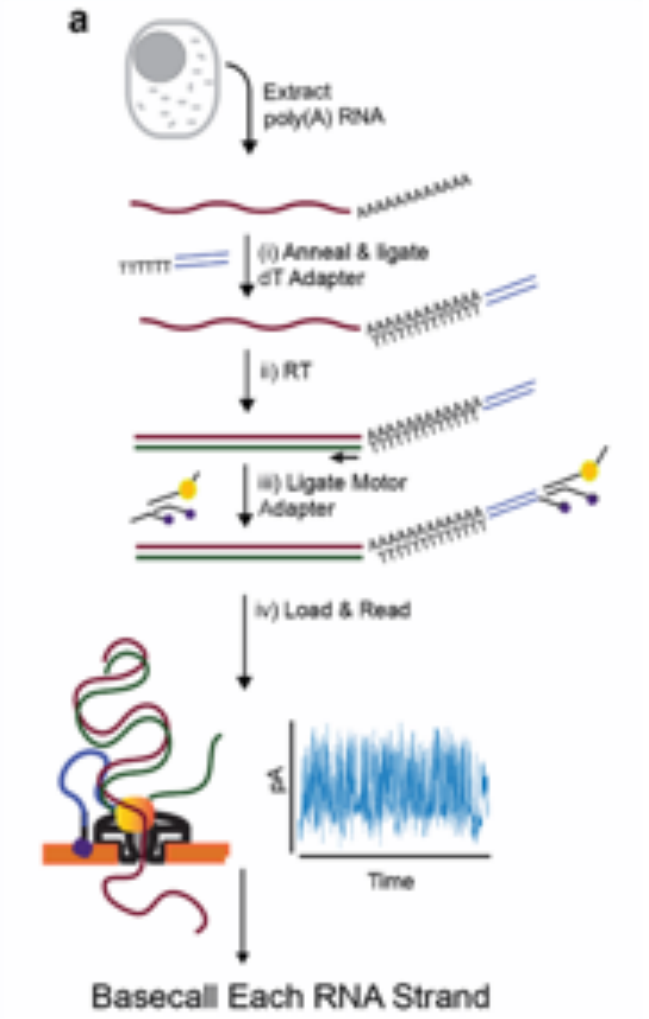
Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

Direct RNA-seq

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

ONT recently introduced direct RNA sequencing

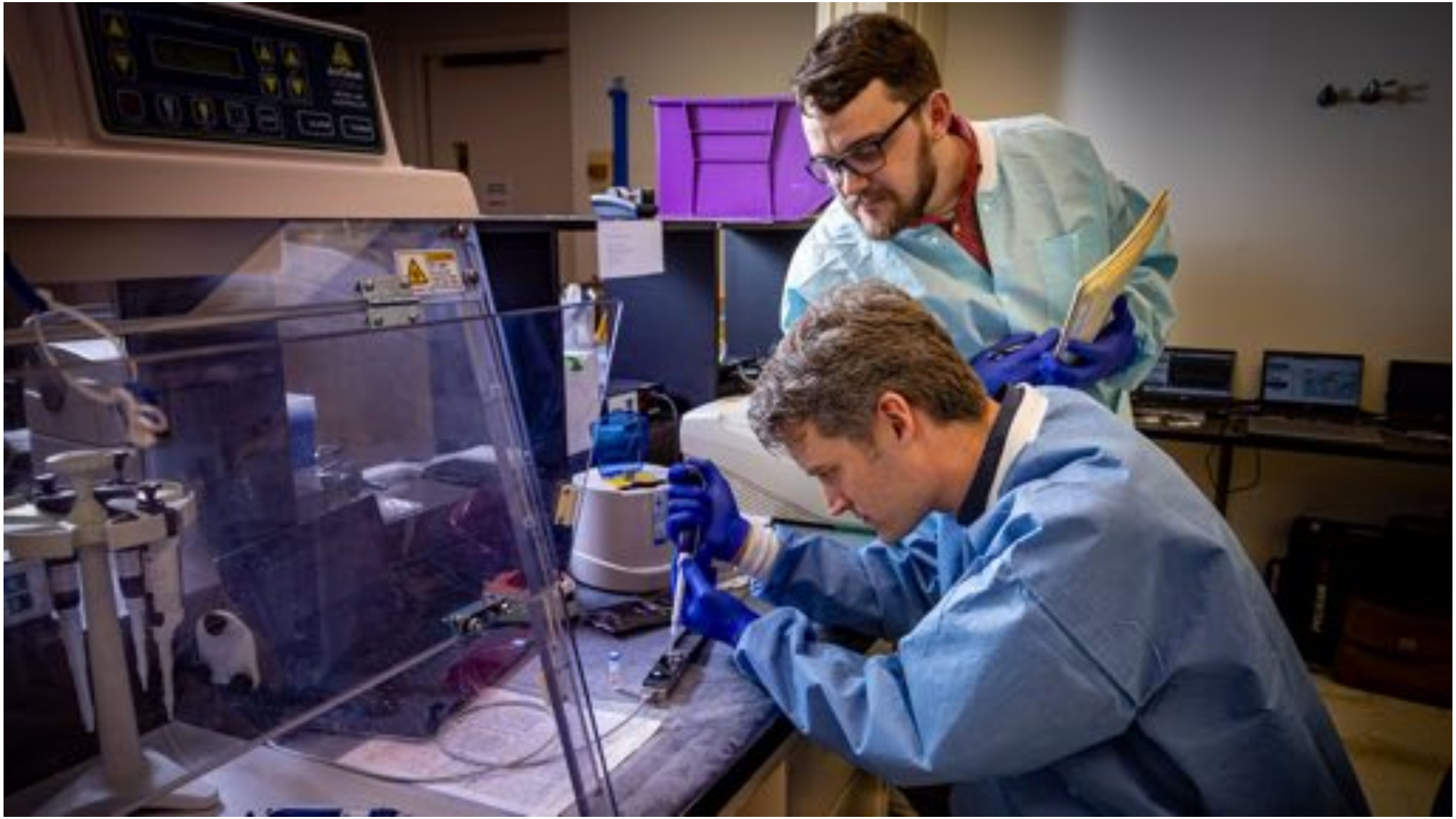
Allows detection of RNA modifications, and potentially secondary structure



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

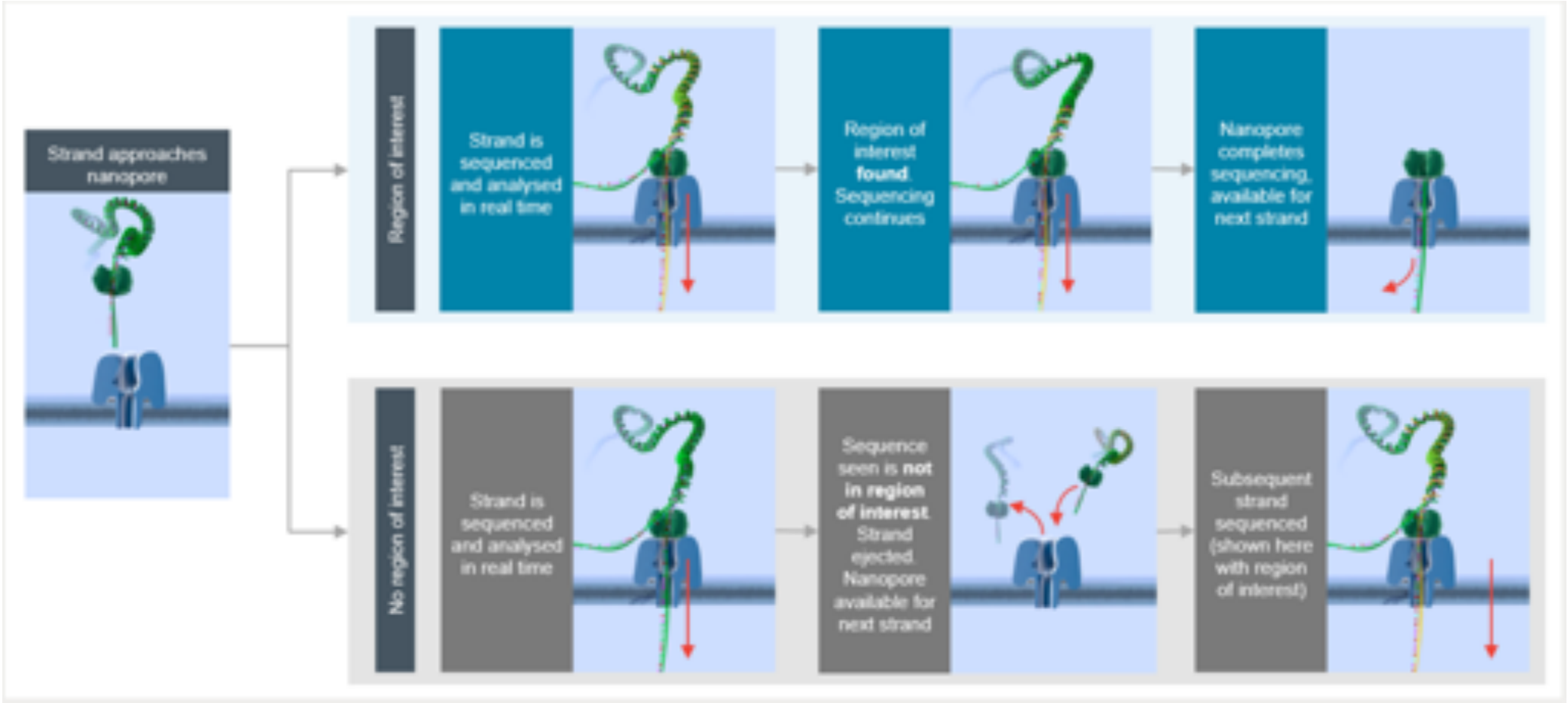
Workman et al. *Nature Methods.* 16:1297–1305

Realtime Surveillance of COVID-19



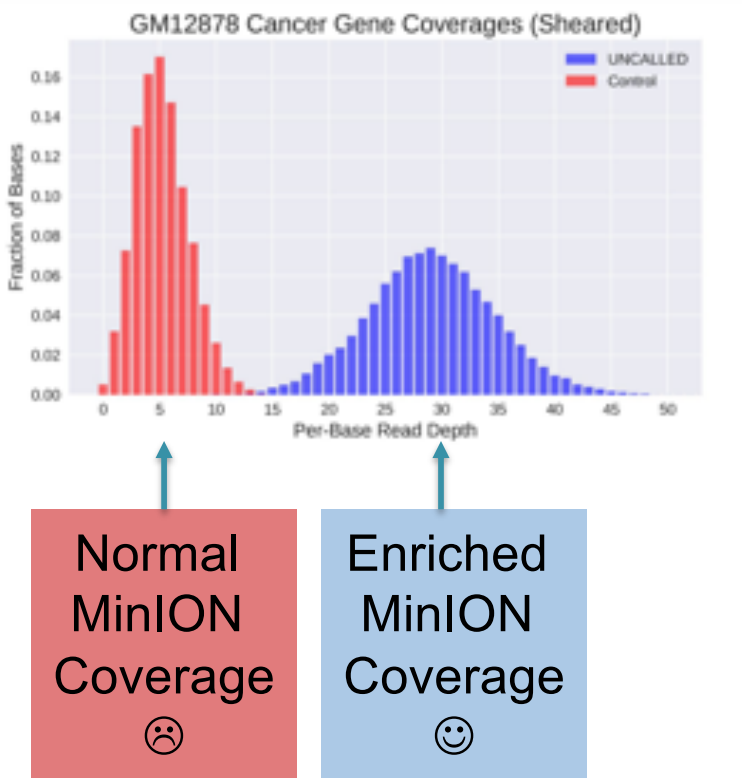
Genomic Diversity of SARS-CoV-2 During Early Introduction into the United States National Capital Region
Thielen, et al. (2020) medRxiv doi: https://doi.org/10.1101/2020.08.13.20174136

Adaptive Sequencing



Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED
Kovaka, S, Fan, Y, Ni, B, Timp, W, Schatz, MC (2020) bioRxiv doi: https://doi.org/10.1101/2020.02.03.931923

Adaptive Sequencing



Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED
Kovaka, S, Fan, Y, Ni, B, Timp, W, Schatz, MC (2020) bioRxiv doi: https://doi.org/10.1101/2020.02.03.931923

Oxford Nanopore Technologies Raises £109.5M

Jan 02, 2020 | staff reporter

Save for later

NEW YORK – Oxford Nanopore Technologies said Thursday it has raised a total of £109.5 million ($144.4 million) between new capital investments and the sale of secondary shares.

The privately held, UK-based firm said it raised £29.3 million in capital and sold £80.2 million in shares. The investors include both new and existing shareholders from the US, Europe, and Asia/Pacific regions, the firm said in a statement.

Other details, including how the firm plans to use the proceeds, were not disclosed.

Oxford Nanopore said it has raised approximately £480 million to date. In October 2018, the firm announced that Amgen would take a £50 million stake, or approximately 3 percent of the firm at the time, giving it a £1.5 billion valuation.

In July 2019, Oxford Nanopore announced 2018 revenues of $43.7 million, up from $17.8 million in 2017.

In a tweet following the announcement, Oxford Nanopore CTO Clive Brown suggested that the minimum amount needed to invest in the firm is $20 million.

Breaking News

- PerkinElmer Q4 Revenues Rise 7 Percent
- CMS to Cover FDA-Approved, -Cleared NGS Germline Tests for Breast, Ovarian Cancer Patients
- Meridian Stock Rises 21 Percent After Reagent Mix Used in Coronavirus Testing
- UK Newborn Trial to Assess PCR Test for Antibiotic-Induced Hearing Loss
- Whole-Genome Sequencing, Deep Phenotyping Shows Many Adults Harbor Pathogenic Genetic Variants
- QuantuMDx Raises $14M to Support Development of POC Genotyping Assay

What's Popular?
In Sequencing

1 Diagnostics Developers Leap Into Action on Novel Coronavirus Tests

2 Whole-Genome Sequencing, Deep Phenotyping Shows Many Adults Harbor Pathogenic Genetic Variants

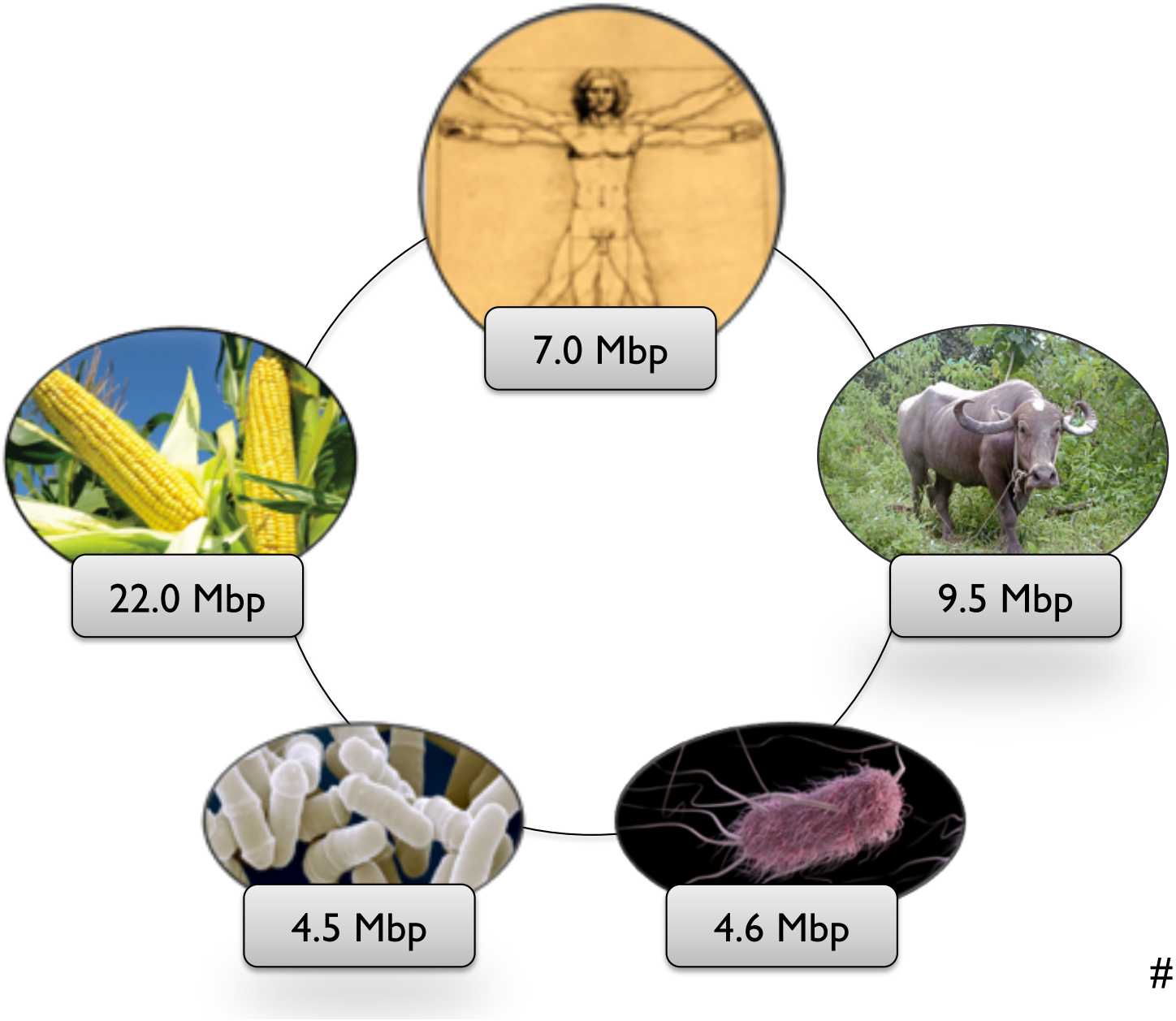
Oxford Nanopore sets sights on IPO

4th April 2019 Callum Cyrus

The Oxford University genetic sequencing spinout is reportedly mulling an IPO that would provide exits to investors including commercialisation firm IP Group.

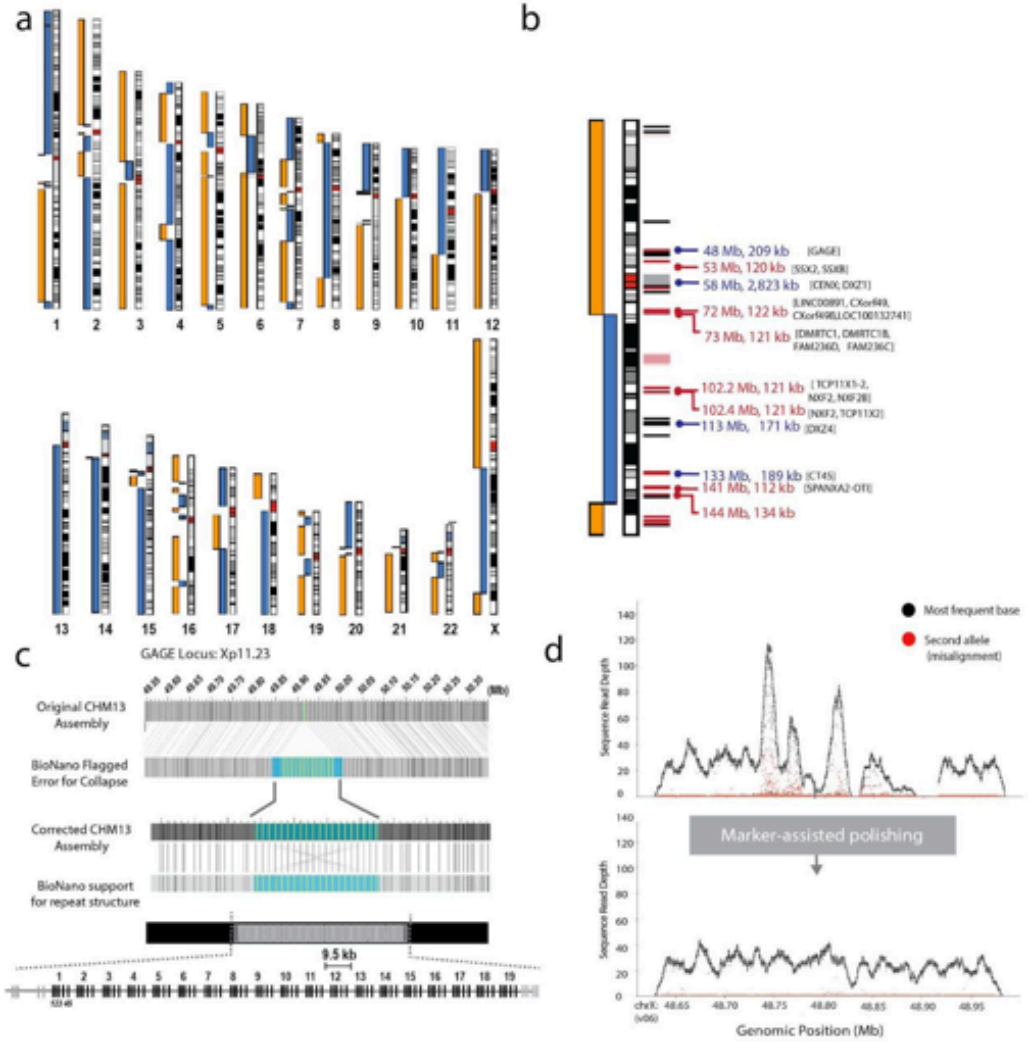
Oxford Nanopore Technologies, a UK-based genetic sequencing technology developer spun out from University of Oxford, is considering floating its shares in an initial public offering (IPO), The Telegraph has reported. Founded in 2005, Oxford Nanopore has developed real-time DNA and RNA sequencing technology that offers biological analyses at a relatively low cost. It has applications...

(A few) Recent Long Read Assemblies



#1mbctgclub

First Telomere-to-Telomere Human Chromosome



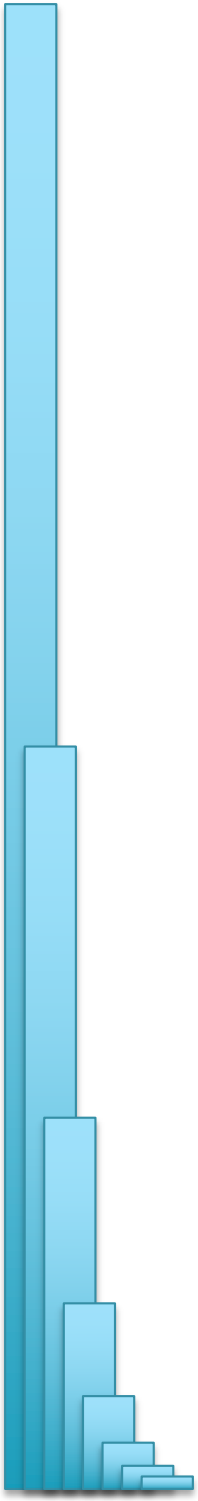
Telomere-to-telomere assembly of a complete human X chromosome
Miga et al. (2020) Nature. doi https://doi.org/10.1038/s41586-020-2547-7

t2t-hprc-2020conference.com

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

SEPTEMBER 21 - 23RD, 2020

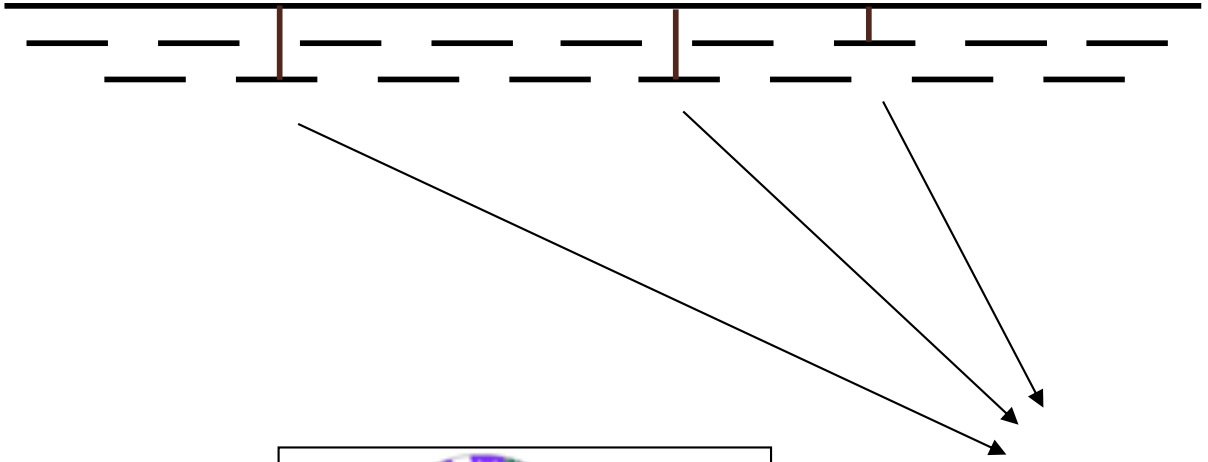
T2T / HPRC TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY IS OPEN FOR REGISTRATION.

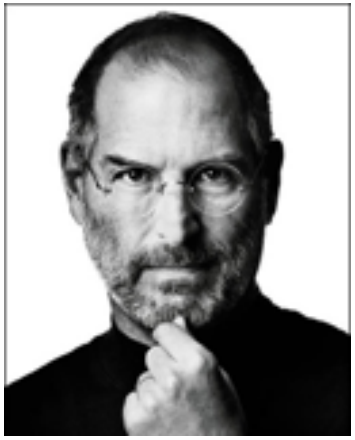


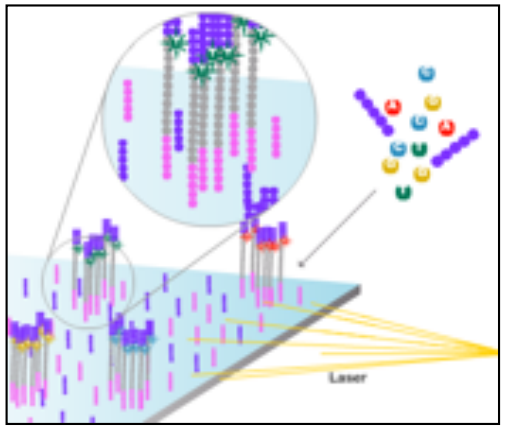
Part 4. Read Alignment

Personal Genomics

How does your genome compare to the reference?







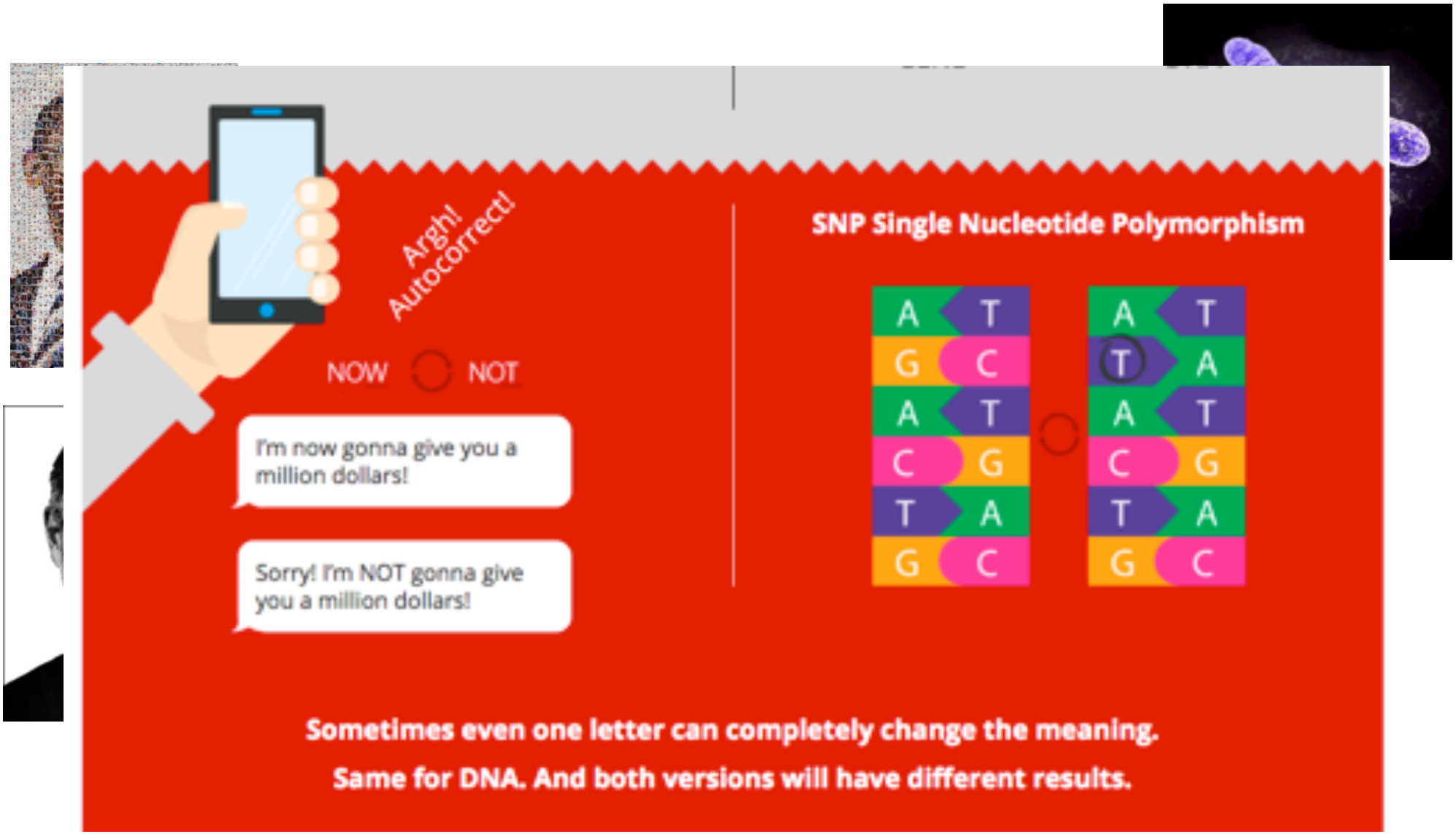
Heart Disease

Cancer

Creates magical technology

Personal Genomics

How does your genome compare to the reference?



Searching for GATTACA

- Where is GATTACA in the human genome?
- Strategy 1: Brute Force

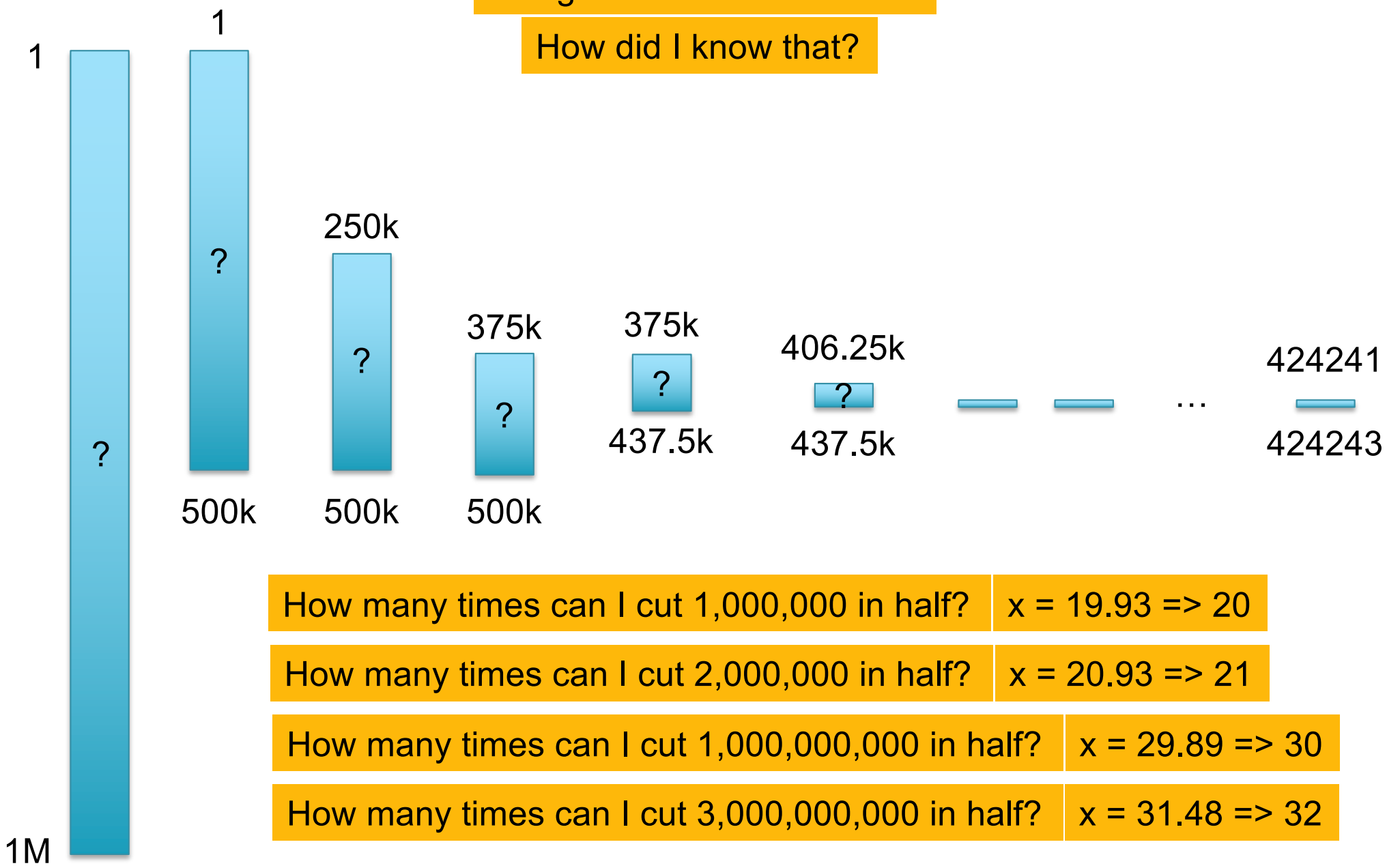
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
	G	A	T	T	A	C	A								

Match at offset 2

Hi/Lo Game

<20 guesses to find 424242

How did I know that?

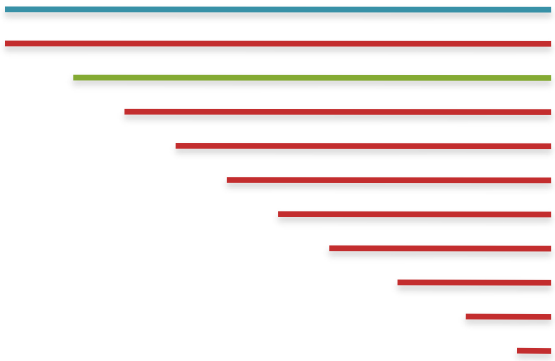


How many times can I cut 1,000,000 in half?	x = 19.93 => 20
How many times can I cut 2,000,000 in half?	x = 20.93 => 21
How many times can I cut 1,000,000,000 in half?	x = 29.89 => 30
How many times can I cut 3,000,000,000 in half?	x = 31.48 => 32

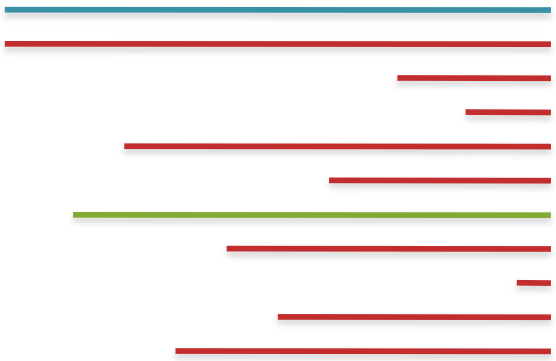
Searching the Phone Book

- What if we need to check many queries?
  - We don't need to check every page of the phone book to find 'Schatz'
  - Sorting alphabetically lets us immediately skip 96% (25/26) of the book *without any loss in accuracy*

- Sorting the genome: Suffix Array (Manber & Myers, 1991)
  - Sort every suffix of the genome



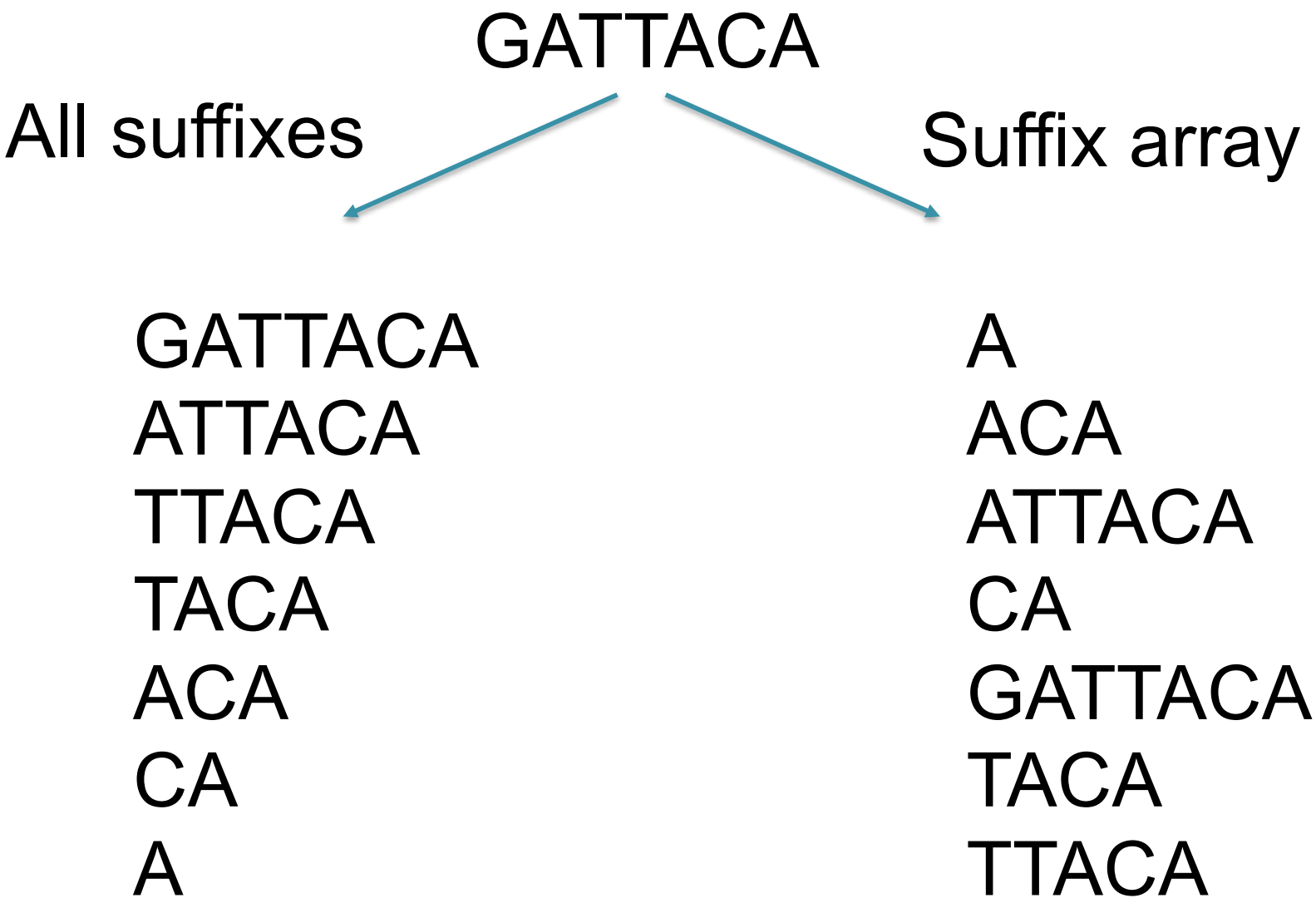
Split into n suffixes



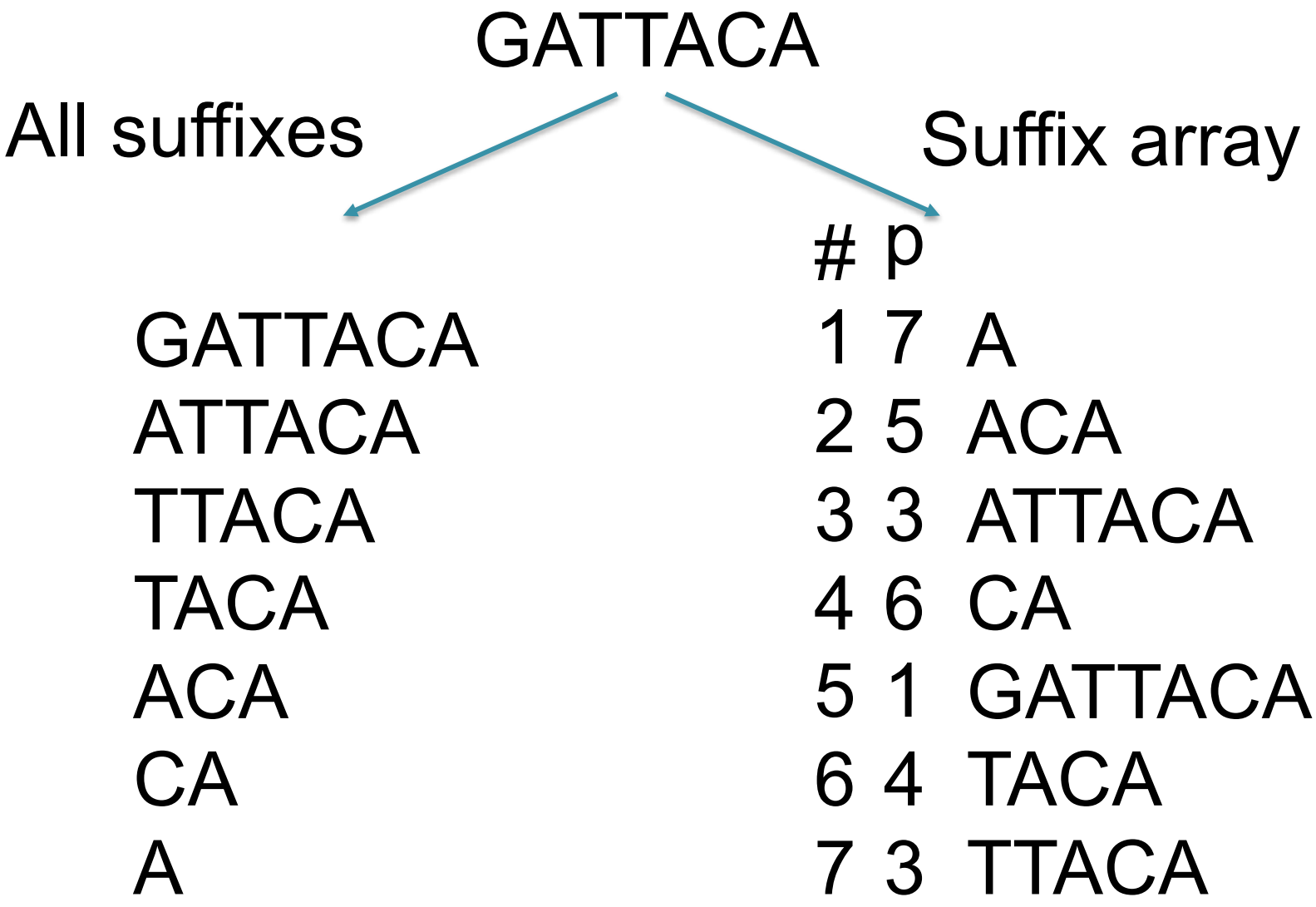
Sort suffixes alphabetically

[Challenge Question: How else could we split the genome?]

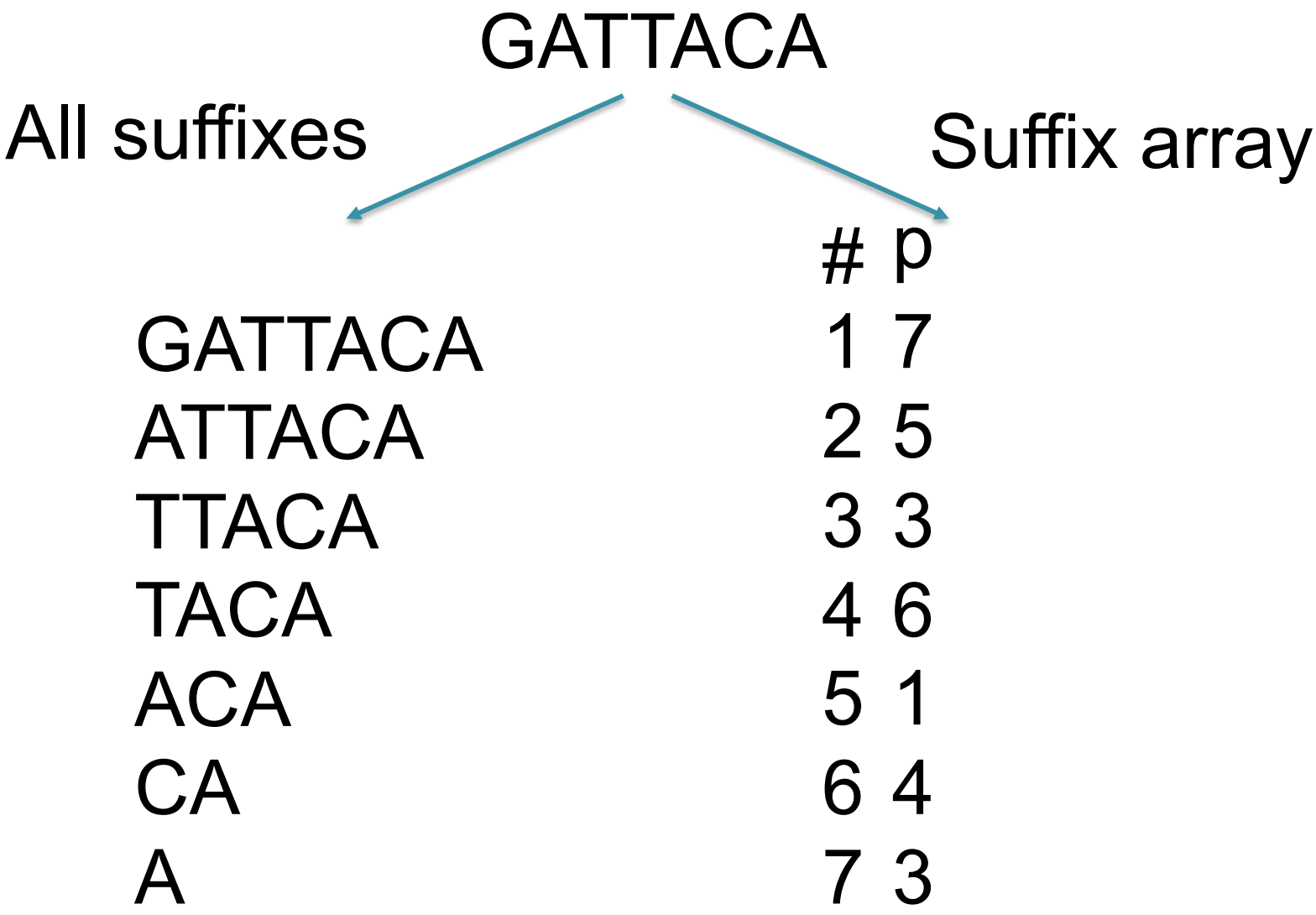
Suffix Arrays: Searching the Phone Book



Suffix Arrays: Searching the Phone Book



Suffix Arrays: Searching the Phone Book



Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15;

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

Lo →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15;

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Lo →

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Lo →

Hi →

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11;

Lo →

Hi →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC

Lo →

Hi →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 9;

Lo →
Hi →

#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Searching the Index

- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower

- Searching for GATTACA
  - Lo = 1; Hi = 15; Mid = (1+15)/2 = 8
  - Middle = Suffix[8] = CC
    => Higher: Lo = Mid + 1

  - Lo = 9; Hi = 15; Mid = (9+15)/2 = 12
  - Middle = Suffix[12] = TACC
    => Lower: Hi = Mid - 1

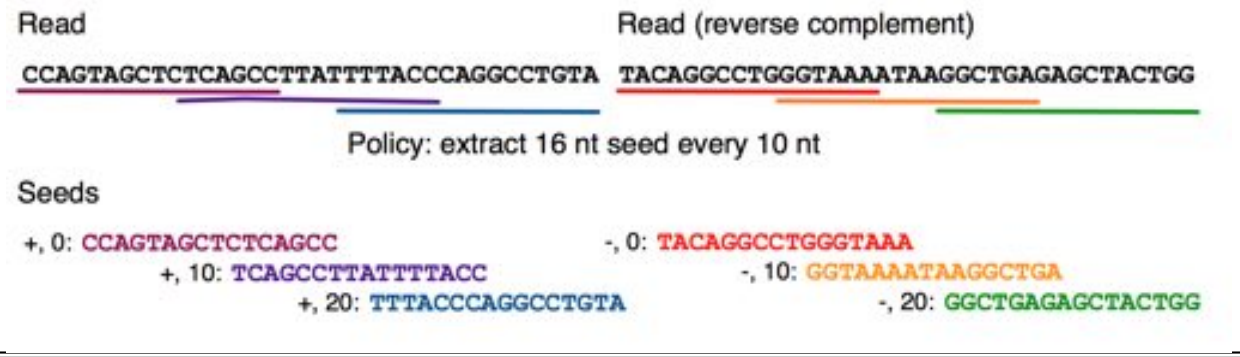
  - Lo = 9; Hi = 11; Mid = (9+11)/2 = 10
  - Middle = Suffix[10] = GATTACC
    => Lower: Hi = Mid - 1

  - Lo = 9; Hi = 9; Mid = (9+9)/2 = 9
  - Middle = Suffix[9] = GATTACA…
    => Match at position 2!

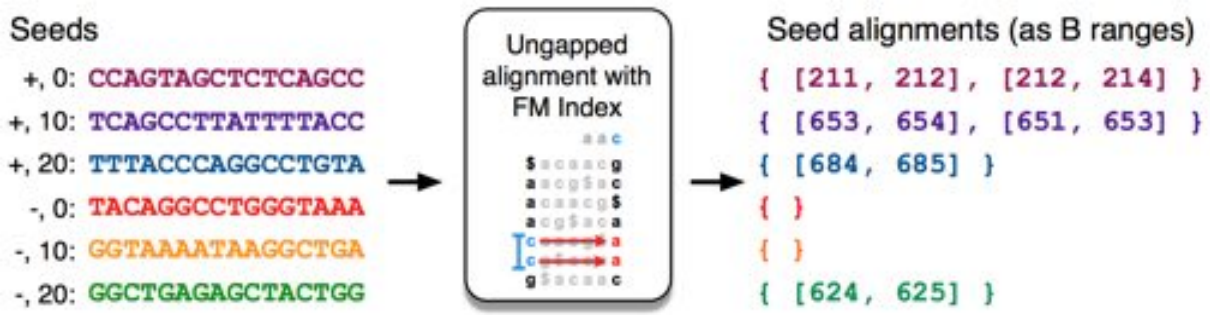
#	Sequence	Pos
1	ACAGATTACC…	6
2	ACC…	13
3	AGATTACC…	8
4	ATTACAGATTACC…	3
5	ATTACC…	10
6	C…	15
7	CAGATTACC…	7
8	CC…	14
9	GATTACAGATTACC…	2
10	GATTACC…	9
11	TACAGATTACC…	5
12	TACC…	12
13	TGATTACAGATTACC…	1
14	TTACAGATTACC…	4
15	TTACC…	11

Algorithm Overview

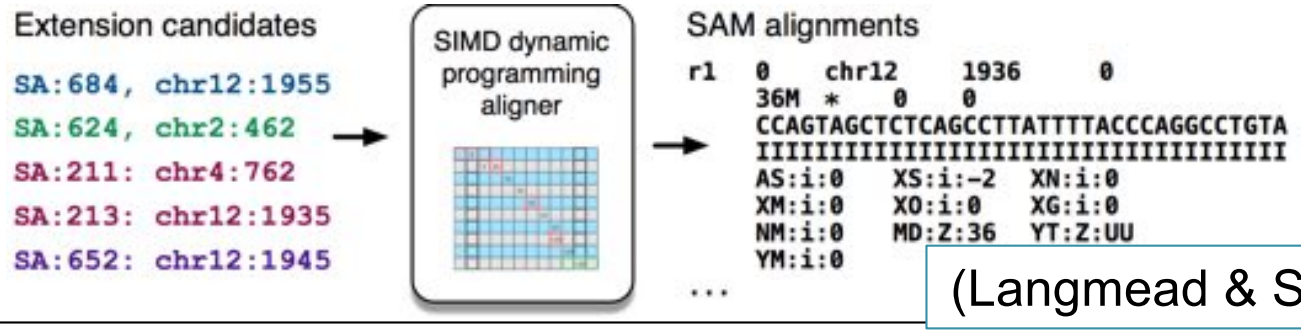
1. Split read into segments



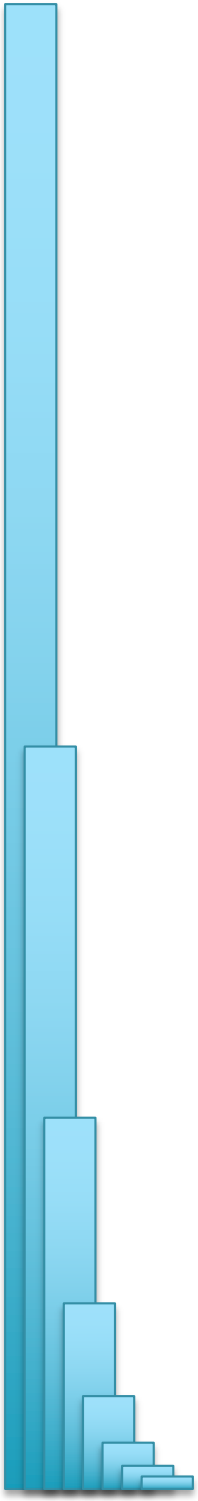
2. Lookup each segment and prioritize



3. Evaluate end-to-end match

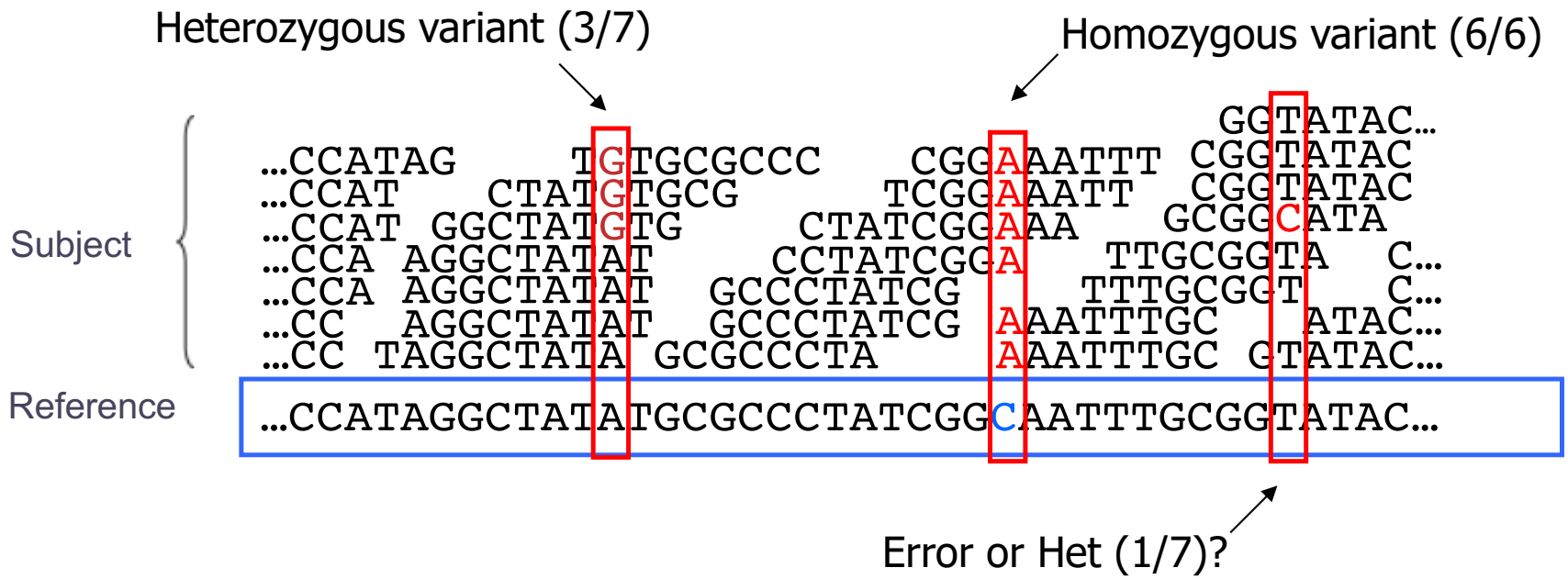


(Langmead & Salzberg, 2012)



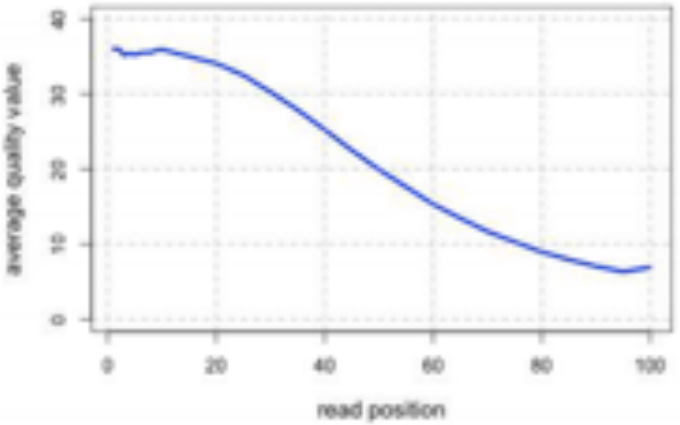
Part 5. Variant Calling

Genotyping Theory



- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!

- Sequencing instruments make mistakes
  - Quality of read decreases over the read length



- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

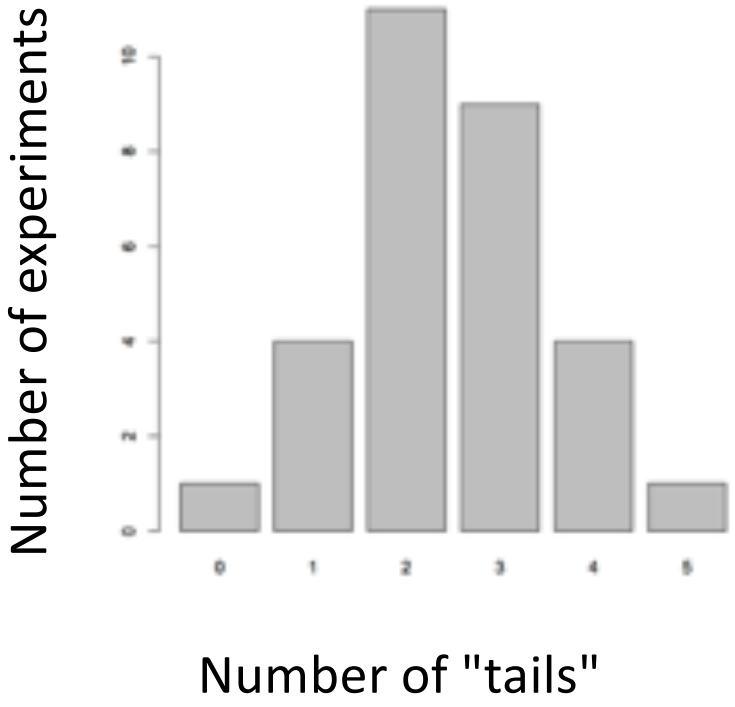
The Binomial Distribution: Adventures in Coin Flipping

P(heads) = 0.5

P(tails) = 0.5

Aaron Quinlan

What is the distribution of tails (alternate alleles) do we expect to see after 5 tosses (sequence reads)?



R code:

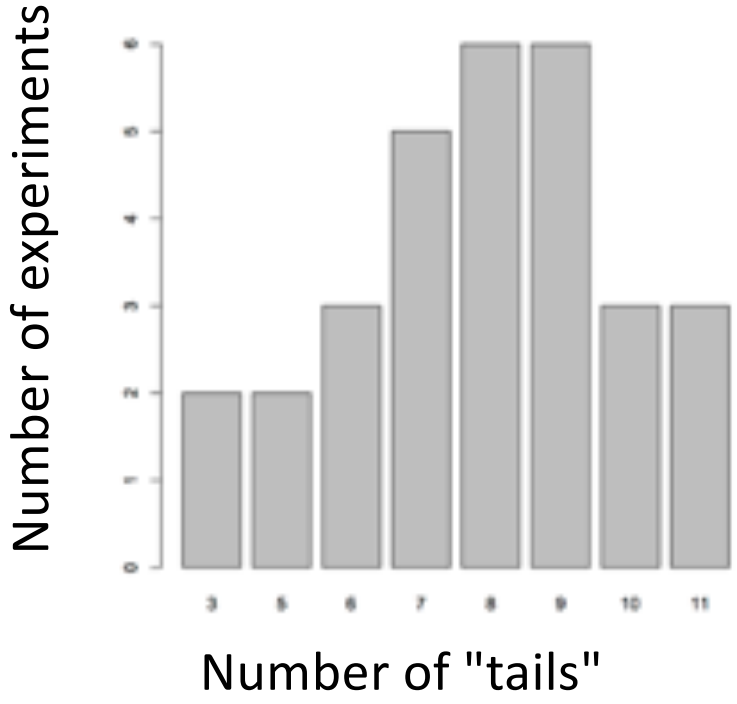
```
barplot(table(rbinom(30, 5, 0.5)))
```

30 experiments (students tossing coins)

5 tosses each

Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 15 tosses (sequence reads)?



R code:

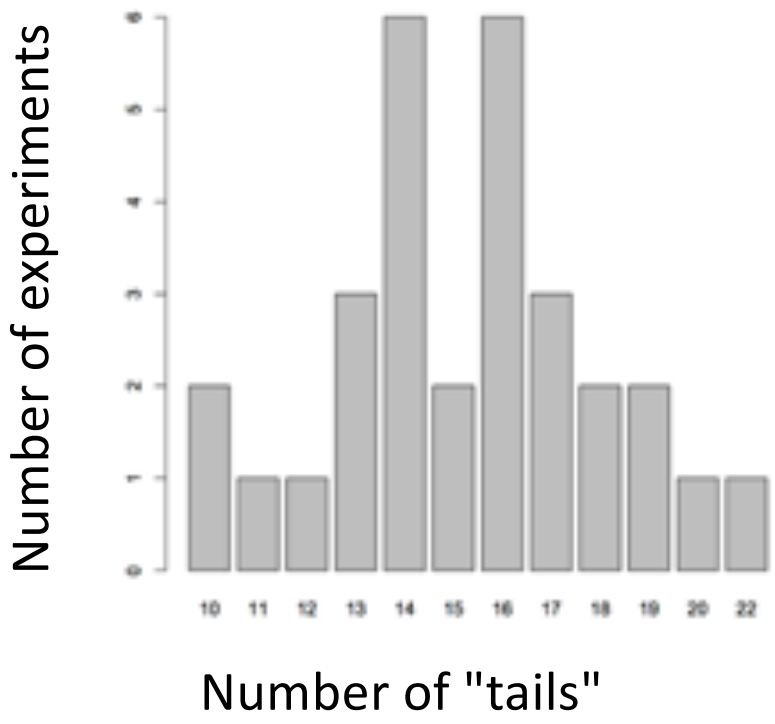
```
barplot(table(rbinom(30, 15, 0.5)))
```

30 experiments (students tossing coins)

15 tosses each

Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?



R code:

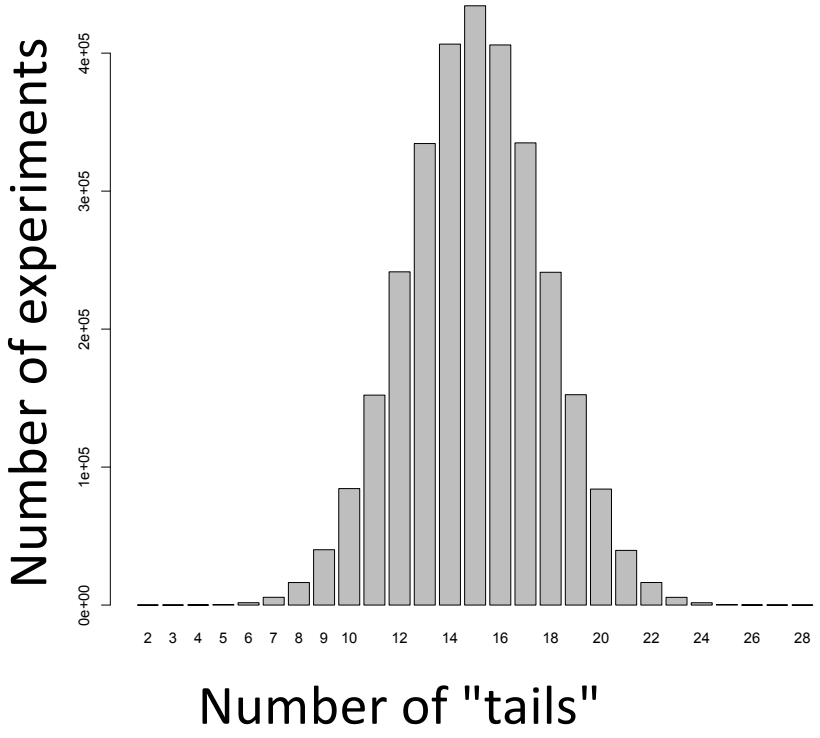
```
barplot(table(rbinom(30, 30, 0.5)))
```

30 experiments (students tossing coins)

30 tosses each

Probability of Tails

What is the distribution of tails (alternate alleles) do we expect to see after 30 tosses (sequence reads)?



R code:

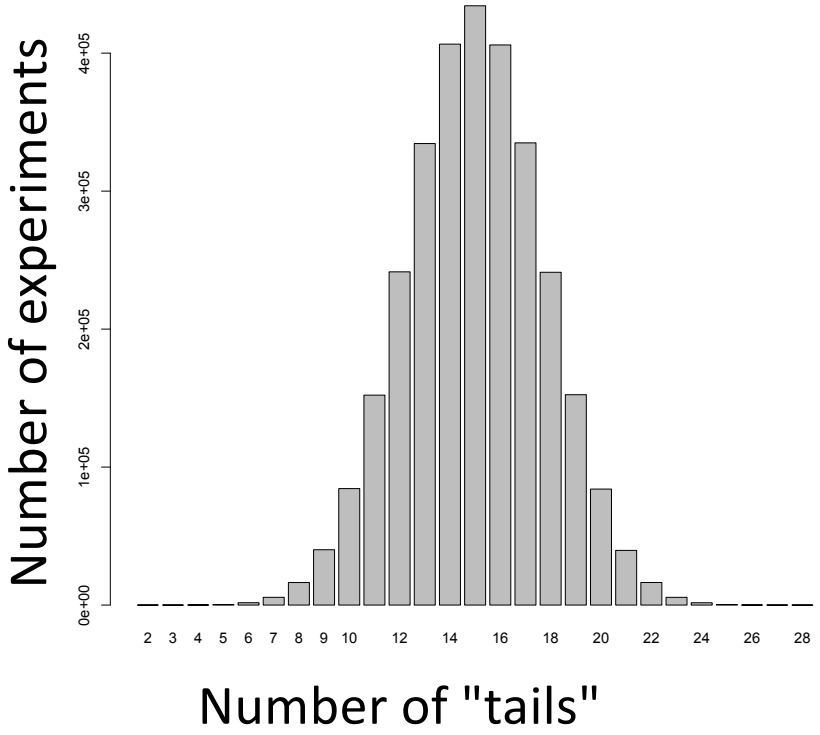
```
barplot(table(rbinom(3e6, 30, 0.5)))
```

3M experiments (students tossing coins)

30 tosses each

Probability of Tails

So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome

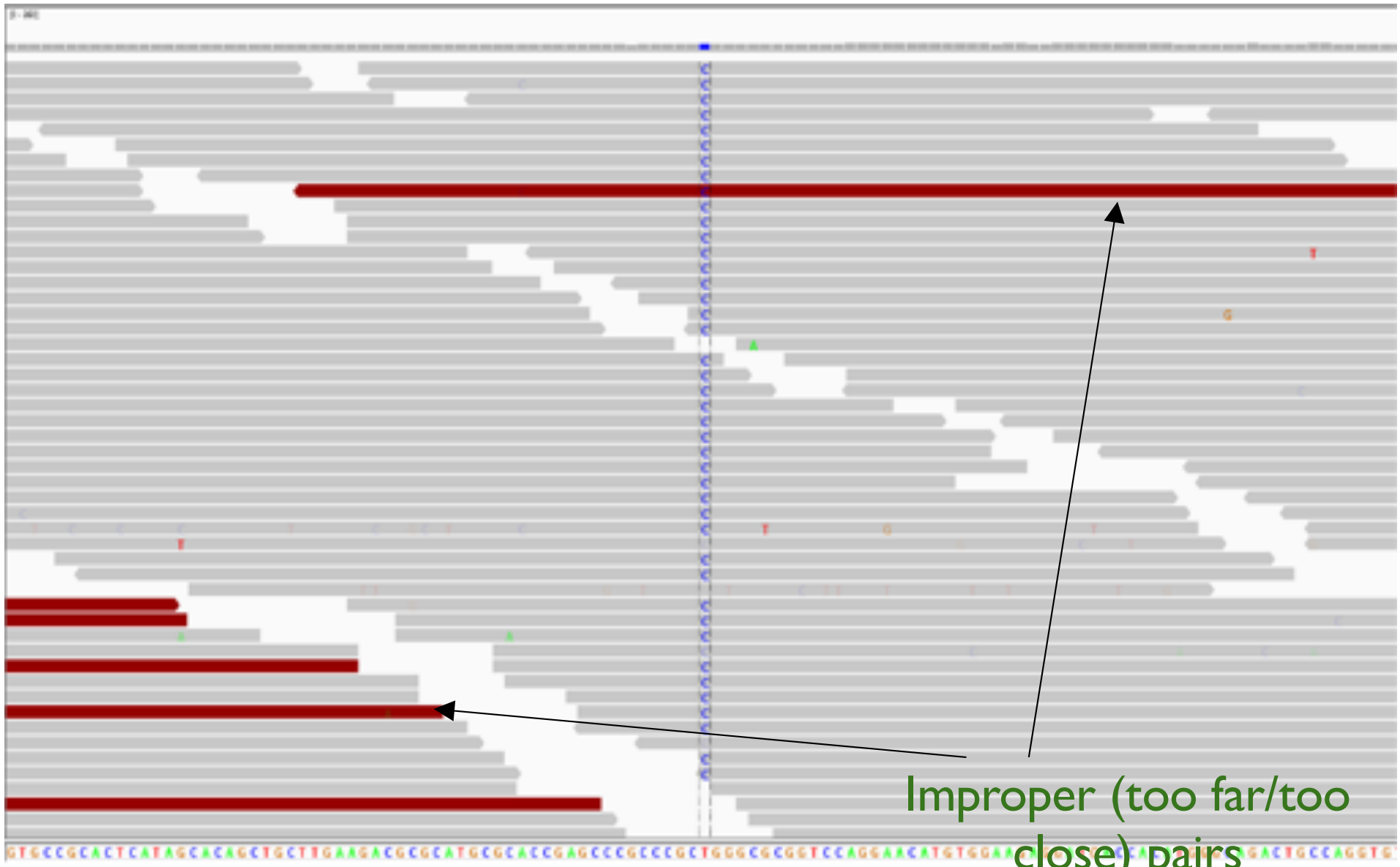


This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

P(3/30 het) <?> P(3/30 err)

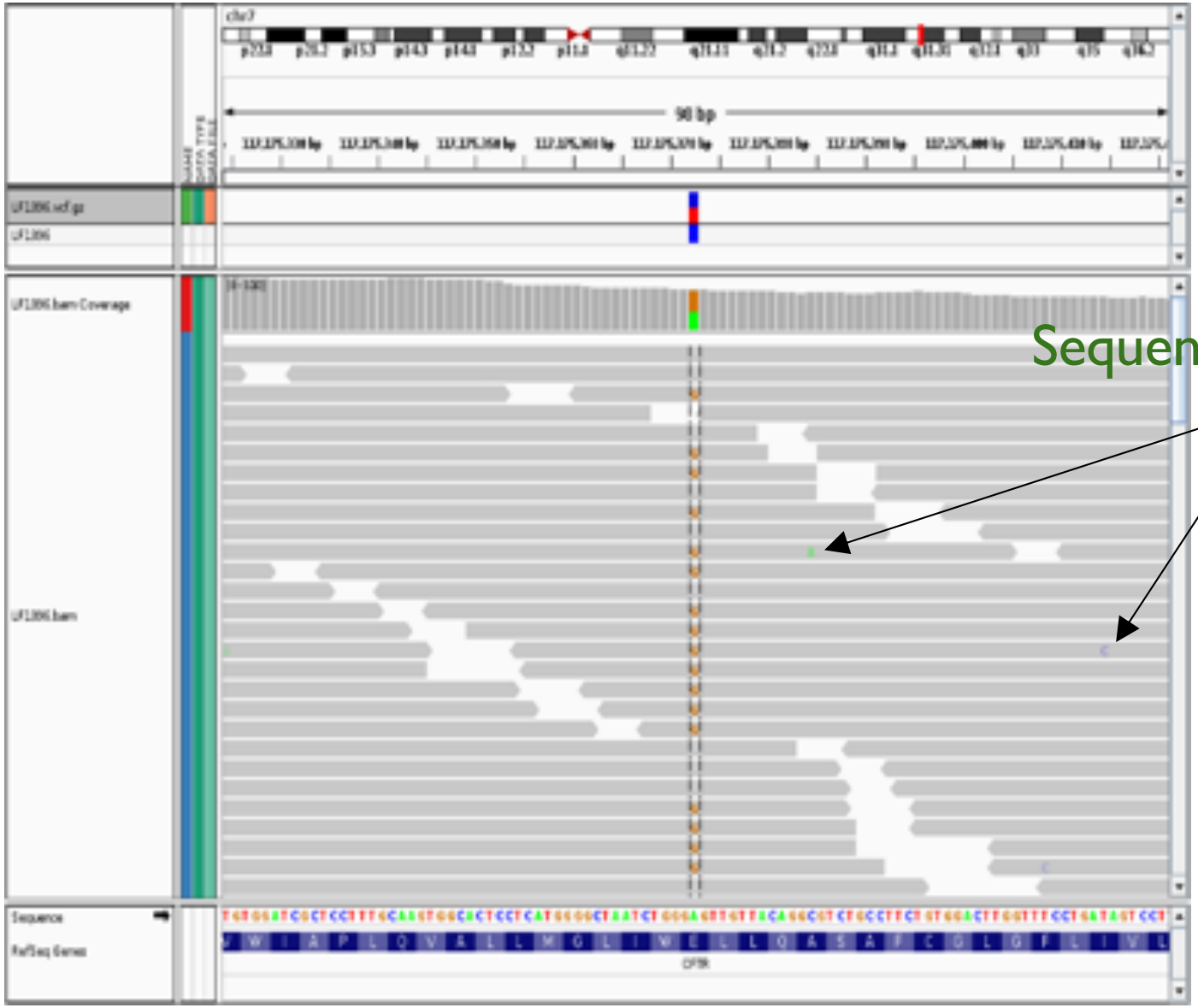
Some real examples of SNPs in IGV

Homozygous for the "C" allele



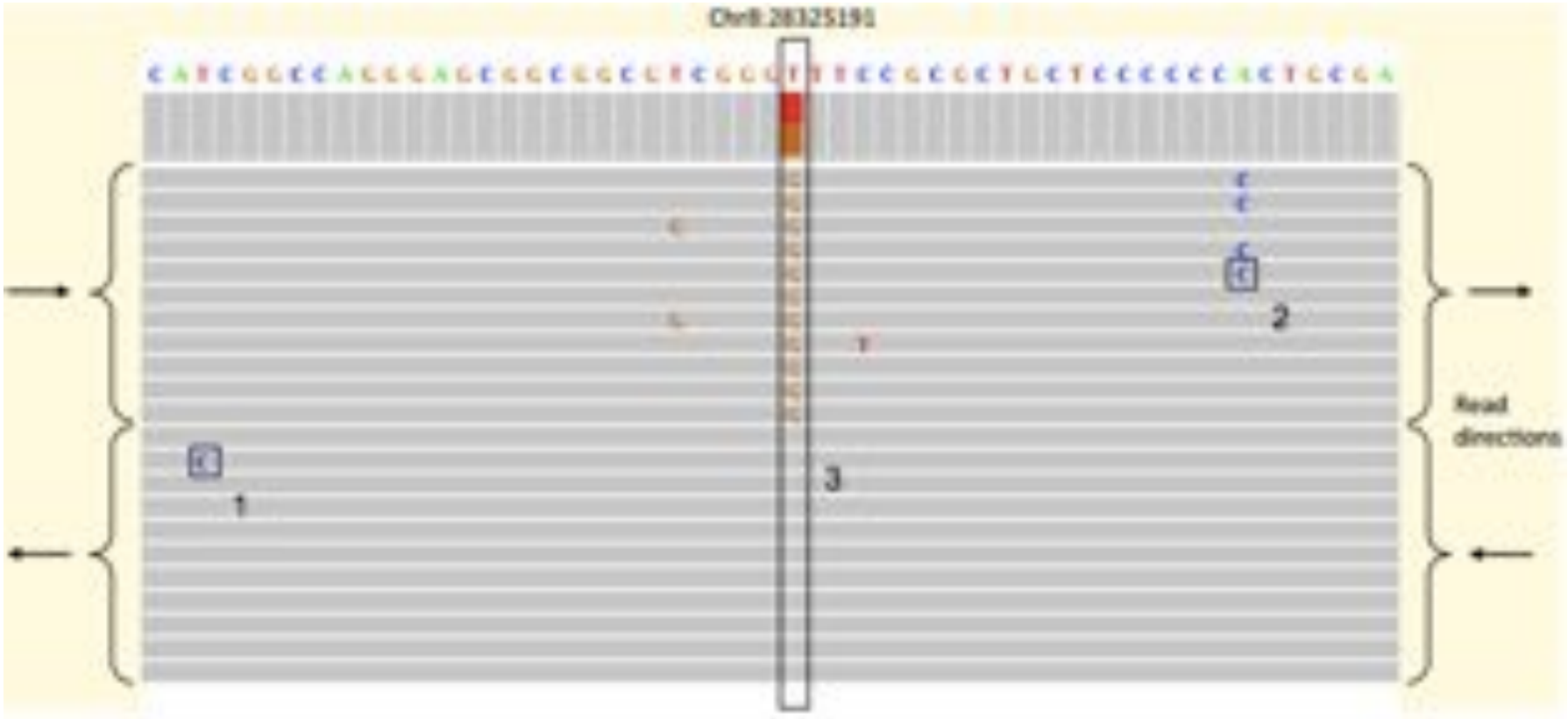
What else do you notice?

Sequencing errors fall out as noise (most of the time)



https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

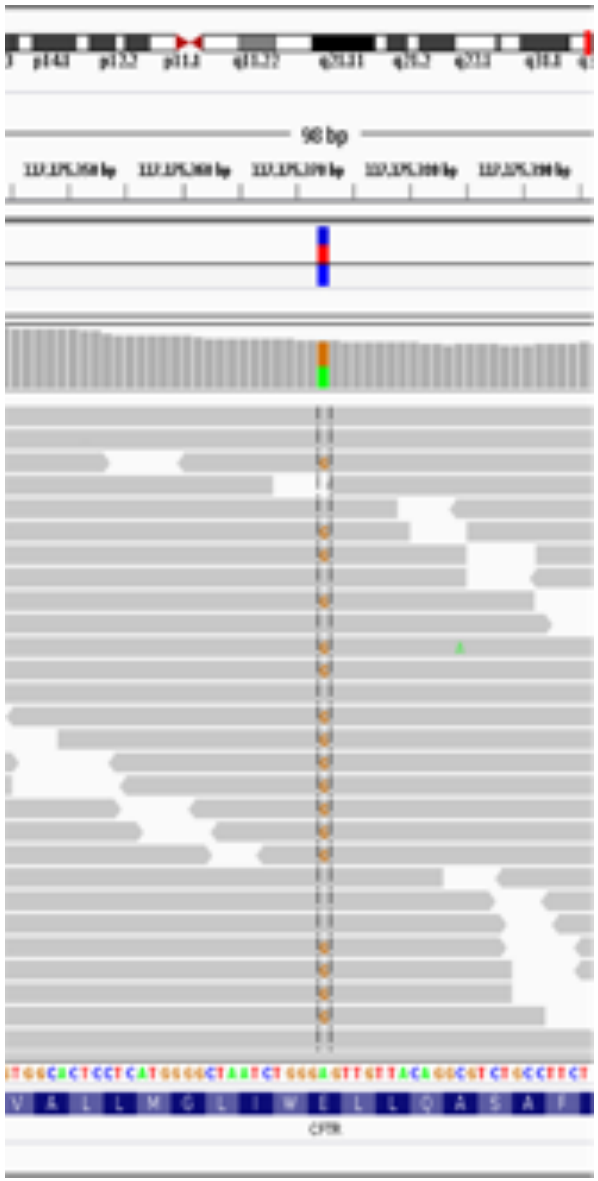
Beware of Systematic Errors



Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics.* 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

What information is needed to decide if a variant exists?



- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

PolyBayes: The first statistically rigorous variant detection tool.

letter

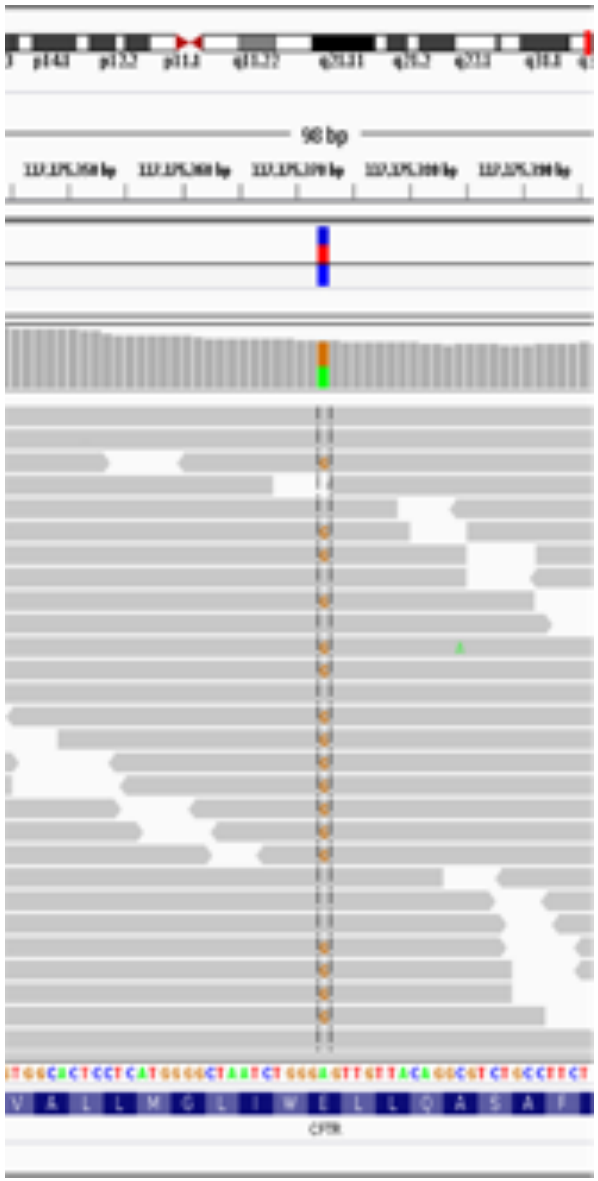
A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Its main innovation was the use of Bayes's theorem

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

Bayesian SNP calling



$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- Transition or Transversion? Which type?
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

PolyBayes: The first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

Bayesian posterior probability

Base call + Base quality

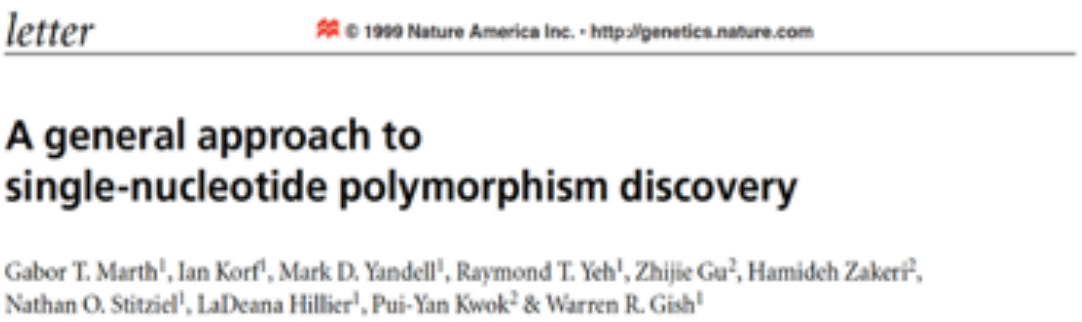
Expected (prior) polymorphism rate

$$P(SNP) = \sum_{all\ variable\ S} \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1,\ldots,S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \ldots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} \cdot P_{Prior}(S_{i_1},\ldots,S_{i_N})}$$

Probability of observed base composition (should model sequencing error rate)

http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

PolyBayes: The first statistically rigorous variant detection tool.



letter © 1999 Nature America Inc. • http://genetics.nature.com

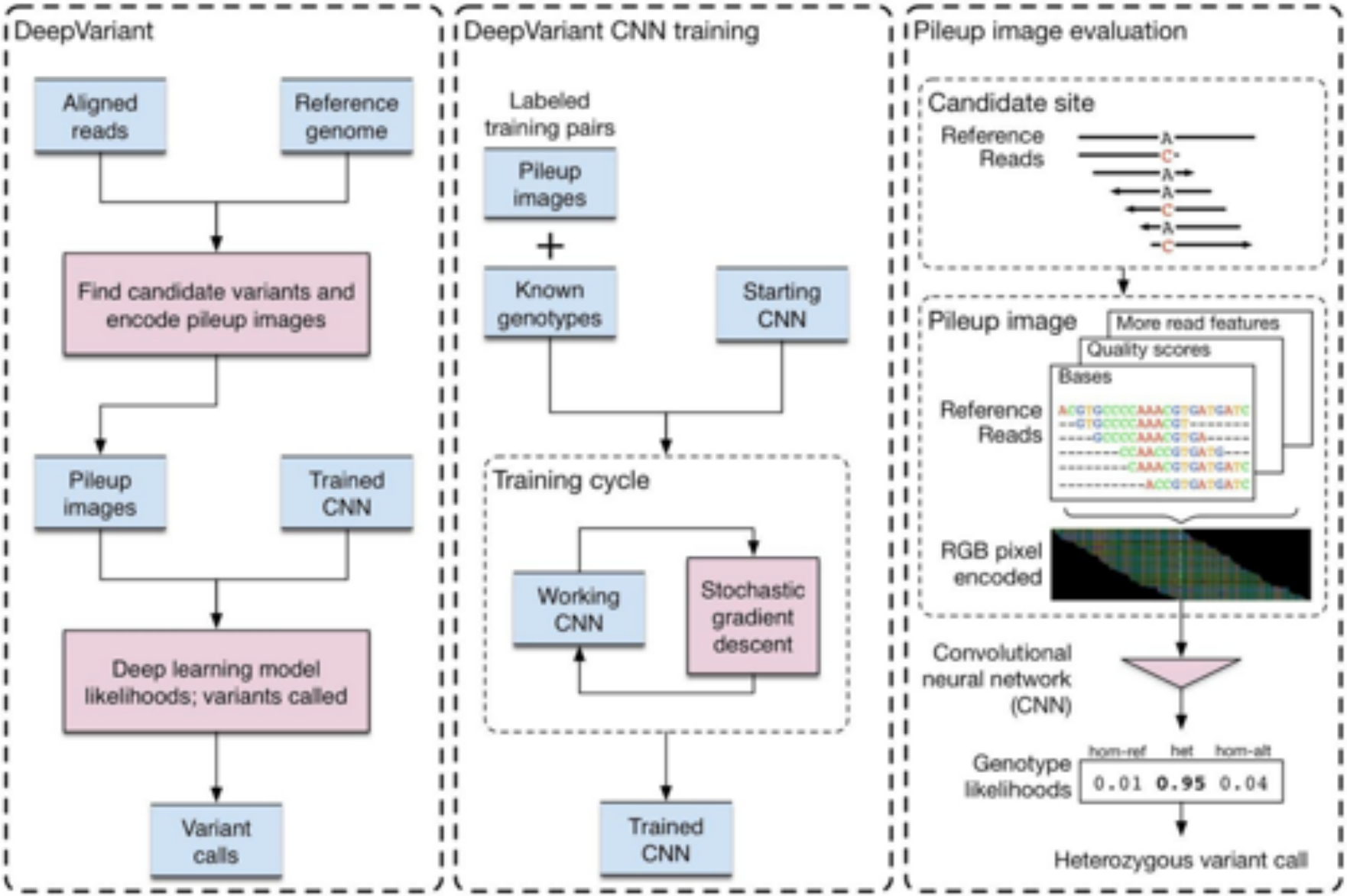
A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

This Bayesian statistical framework has been adopted by other modern SNP/INDEL callers such as FreeBayes, GATK, and samtools

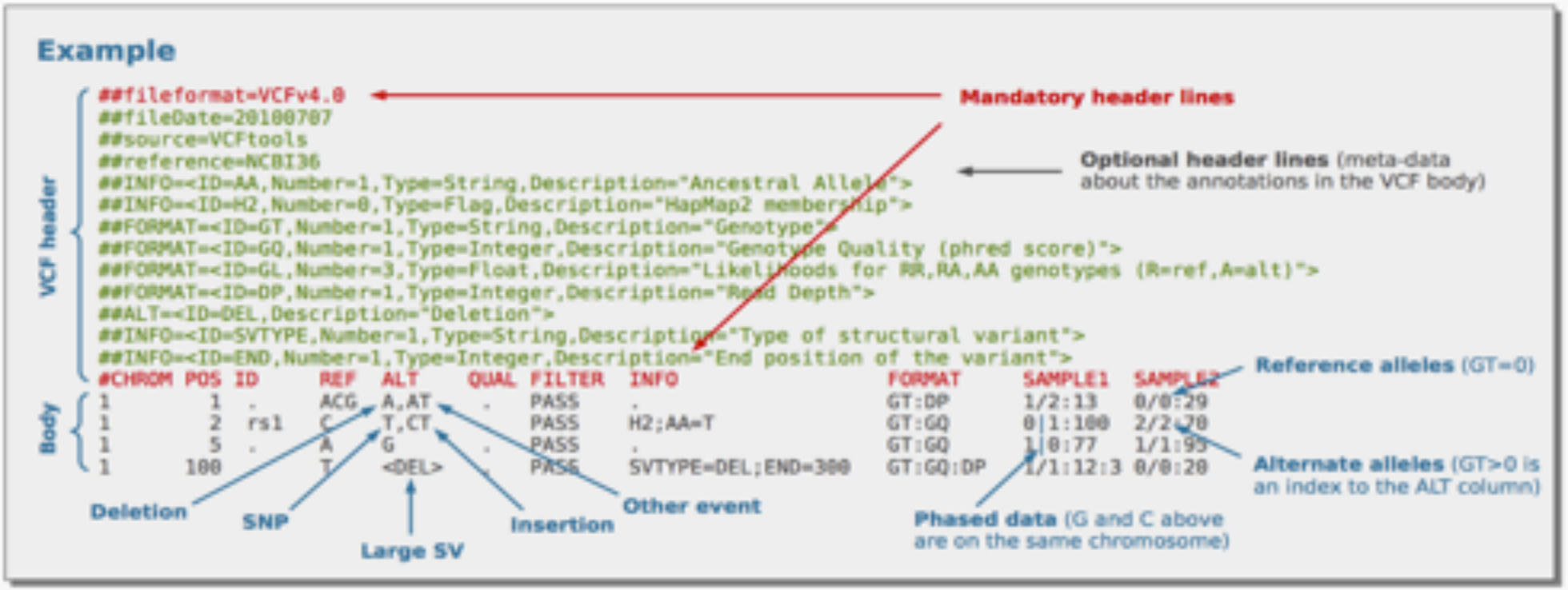
http://www.nature.com/ng/journal/v23/n4/full/ng1299_452.html

Deep Variant

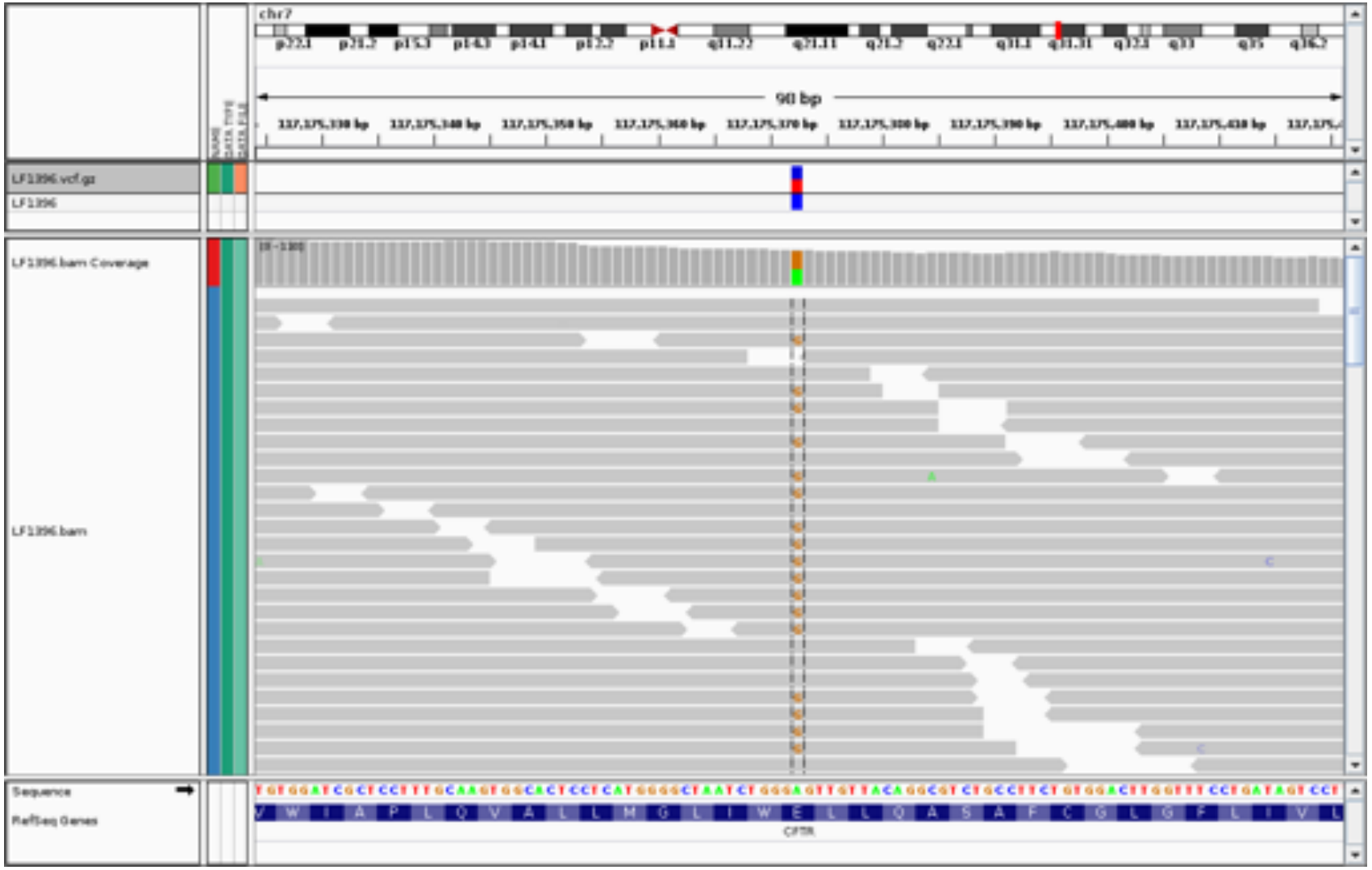


Creating a universal SNP and small indel variant caller with deep neural networks
Poplin et al. (2018) Nature Biotechnology. https://www.nature.com/articles/nbt.4235

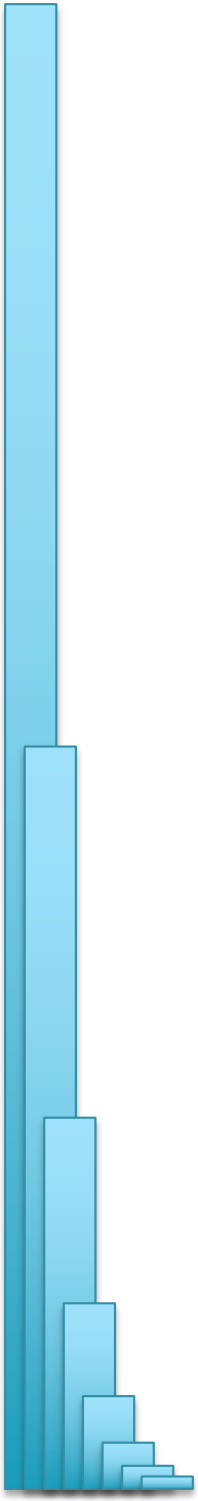
VCF Format


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

VCF Format




```
#CHROM POS        ID  REF ALT QUAL   FILTER INFO    FORMAT  LF1396
chr7   117175373  .   A   G   90     PASS   AF=0.5  GT      0/1
```



Part 6. Structural Variant Calling

Breast Cancer Risk

- ***Breast cancer affects 1 in 8 women and is one of the major causes of death in women***

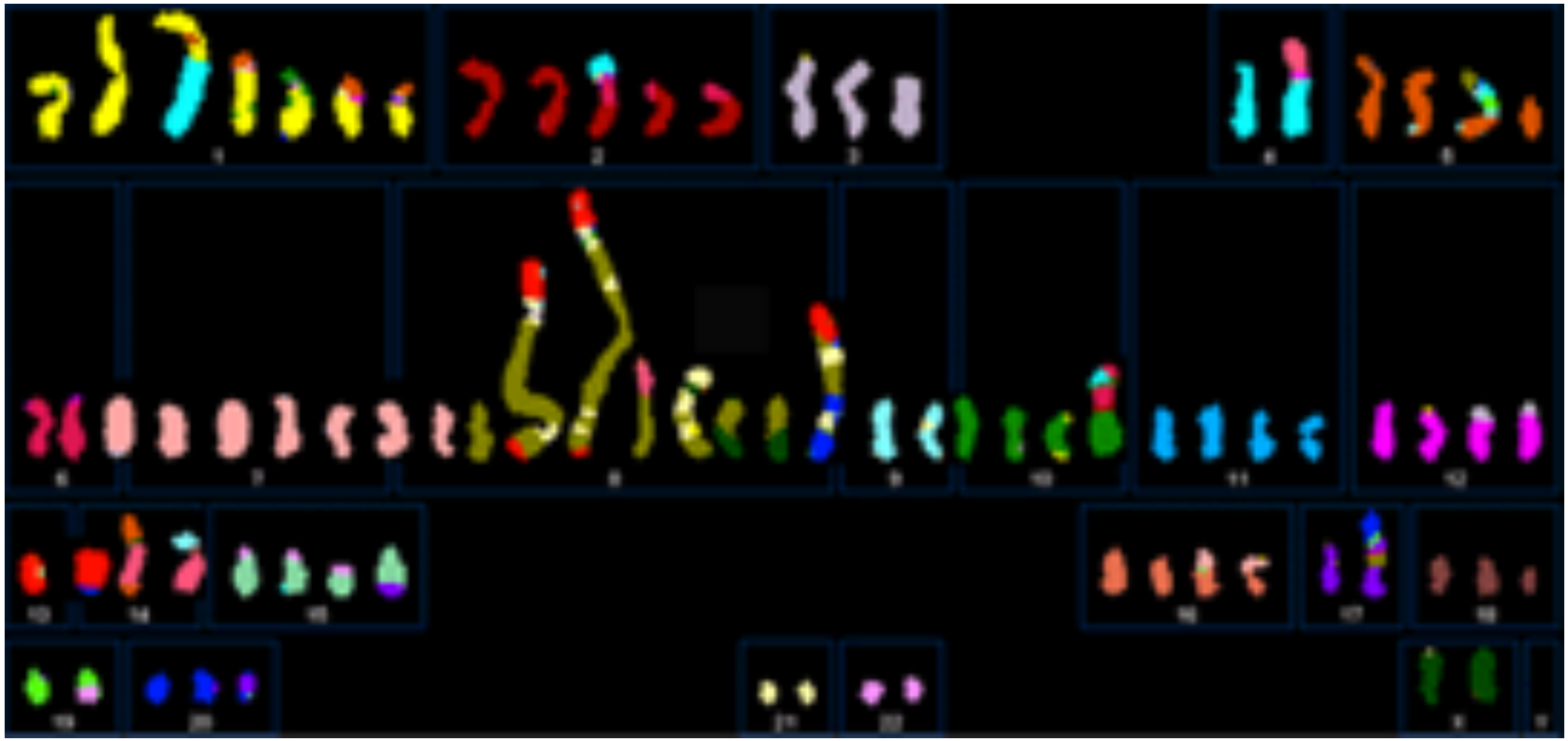
- ***Family history is one of the most important risk factors***
  - Reported in about 20% of cases
  - The genetic risk is probably much higher because we don't observe cases along the paternal lineage

- ***Within high risk families, it is now standard practice to screen for mutations in BRCA1, BRCA2, and other important genes***
  - Better cancer management: more regular screening, and in some cases prophylactic surgery to minimize the possibility of developing the cancer.

Only a small fraction of familial cases have a recognizable mutation that explains the predisposition

Long Read Sequencing of the SK-BR-3 Breast Cancer Cell Line

Most commonly used Her2-amplified breast cancer cell line



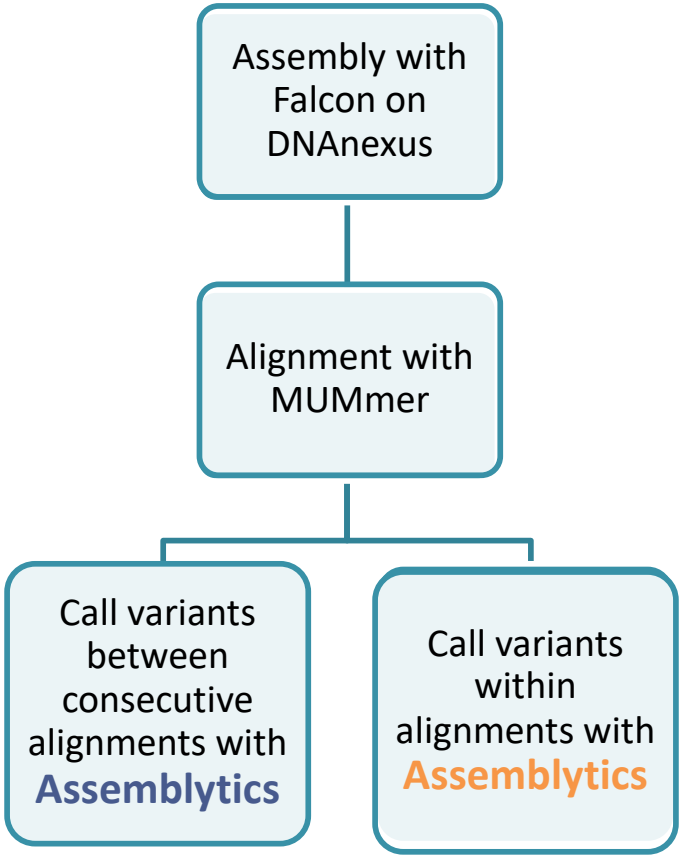
(Davidson et al, 2000)

Can we resolve the complex structural variations, especially around Her2?

Recent collaboration between JHU, CSHL and OICR to *de novo* assemble and analyze the complete cell line genome with PacBio long reads

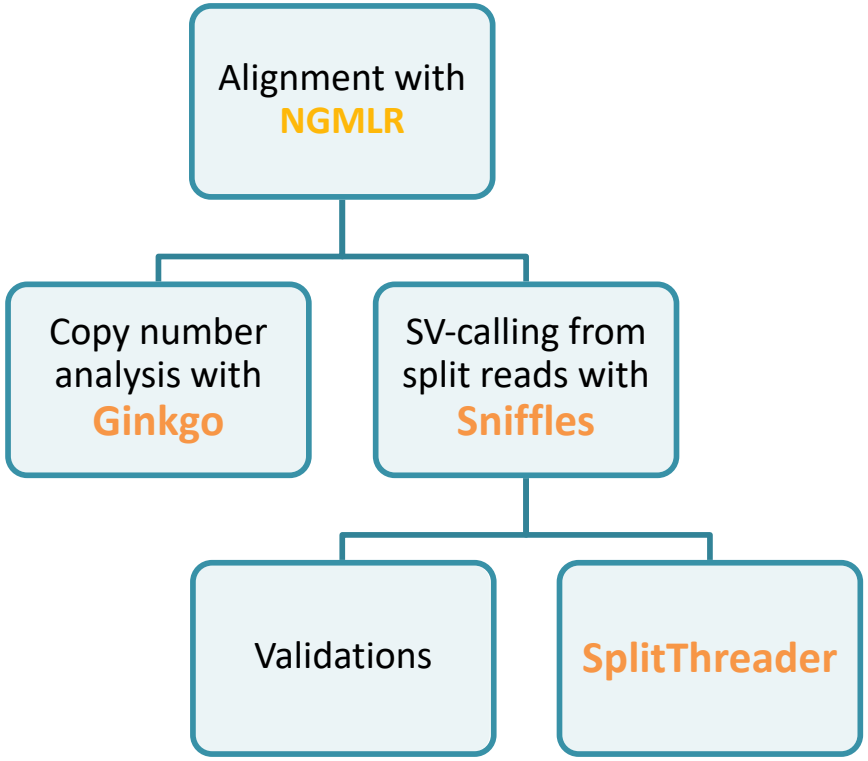
Structural Variant Analysis

Assembly-based



~ 11,000 structural variants
50 bp to 10 kbp

Split-Read based

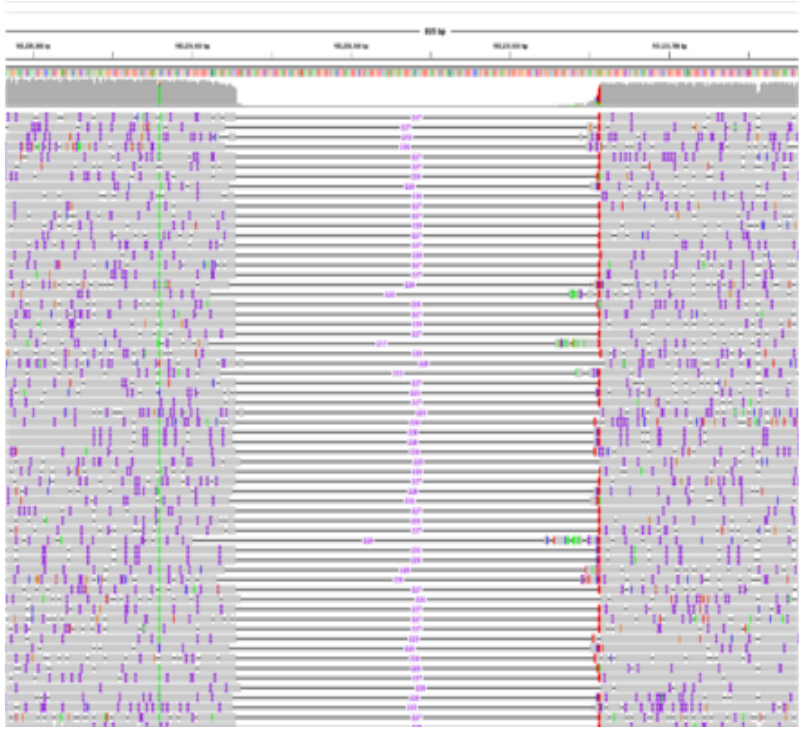


~ 20,000 structural variants
Including many inter-chromosomal rearrangements

NGMLR + Sniffles

BWA-MEM:

NGMLR:



NGMLR: Convex scoring model to accommodate many small gaps from sequencing errors along with less frequent but larger SVs

Accurate detection of complex structural variations using single molecule sequencing
Sedlazeck, Rescheneder et al (2018) *Nature Methods. doi:10.1038/s41592-018-0001-7*

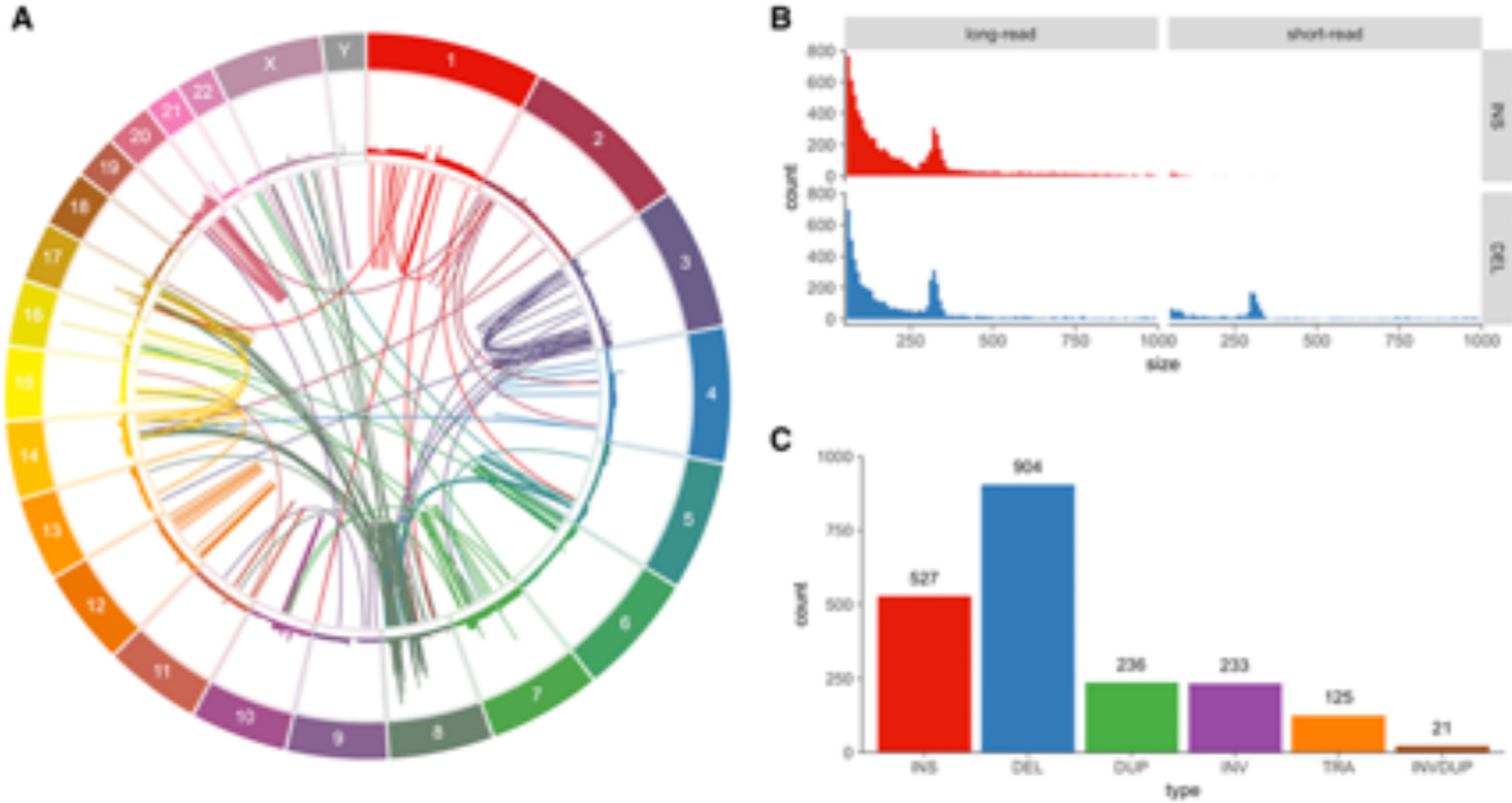


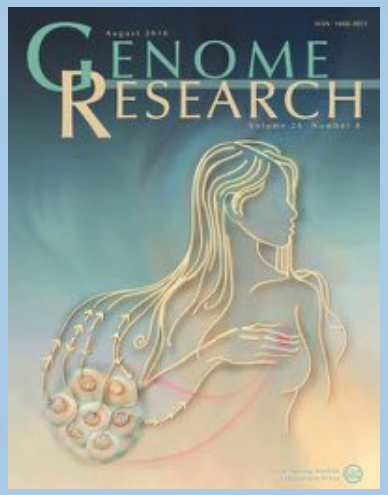
Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (*C*) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Nattestad et al. (2018) *Genome Research.* doi: 10.1101/gr.231100.117

Highlights

- Finding 10s of thousands of additional variants
- PCR validation confirms high accuracy of long reads
- Detect many novel gene fusions
- Identify early vs late mutations in the cancer



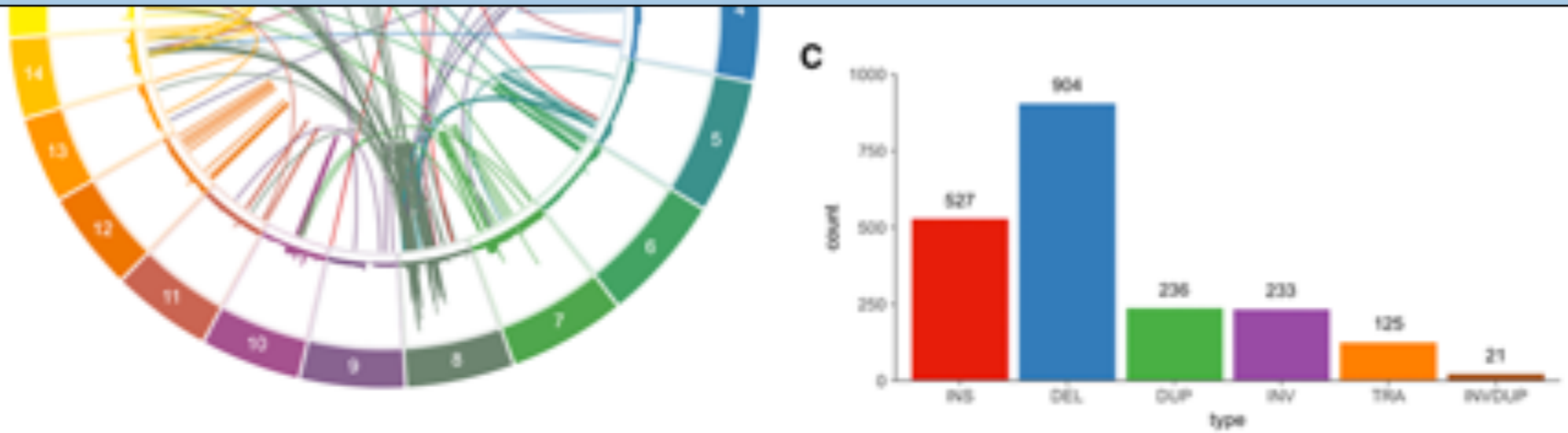
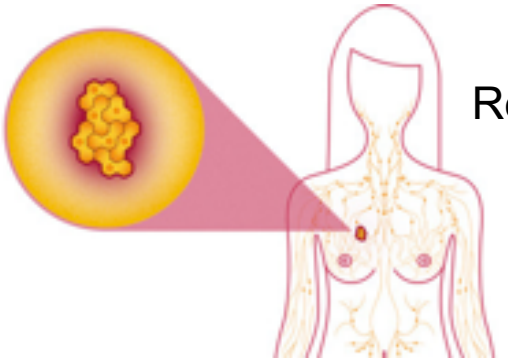


Figure 1. Variants found in SK-BR-3 with PacBio long-read sequencing. (*A*) Circos (Krzywinski et al. 2009) plot showing long-range (larger than 10 kbp or inter-chromosomal) variants found by Sniffles from split-read alignments, with read coverage shown in the *outer* track. (*B*) Variant size histogram of deletions and insertions from size 50 bp up to 1 kbp found by long-read (Sniffles) and short-read (SURVIVOR 2-caller consensus) variant calling, showing similar size distributions for insertions and deletions from long reads but not for short reads, where insertions are greatly underrepresented. (C) Sniffles variant counts by type for variants above 1 kbp in size, including translocations and inverted duplications.

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

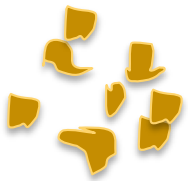
Nattestad et al. (2018) *Genome Research*. doi: 10.1101/gr.231100.117

Taking Long Read Sequencing into the Clinic

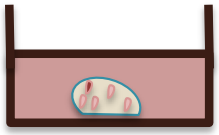


Resected primary tumor

Mechanical/ enzymatic



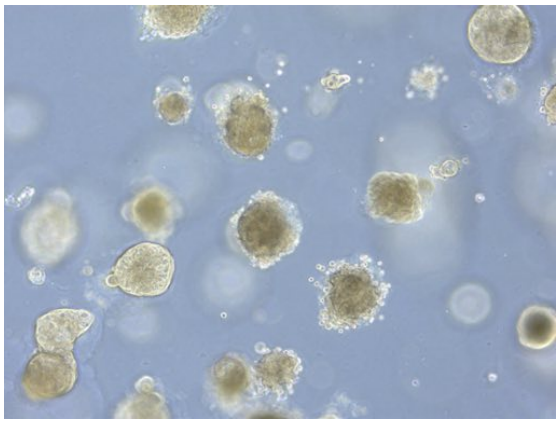
Purification



Plating on Matrigel

Add growth factors

Tumor organoids in culture



- ✓ Stable Growth in 3D
- ✓ Recapitulate tumor pathology & treatment response
- ✓ Maintenance of tissue/tumor heterogeneity
- ✓ "2017 Method of the Year" - Nature Methods

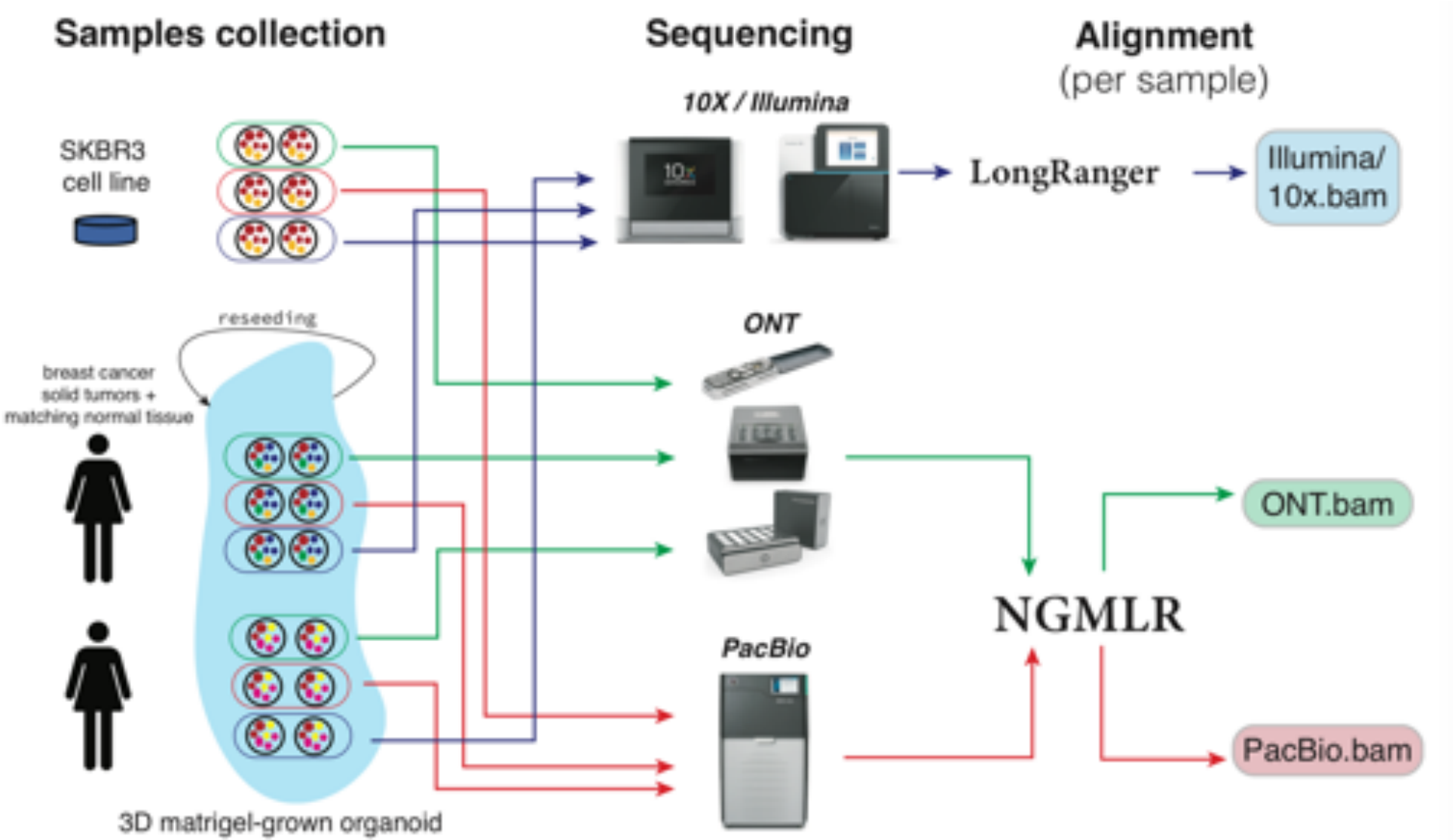
CSH

David Spector

Northwell Health

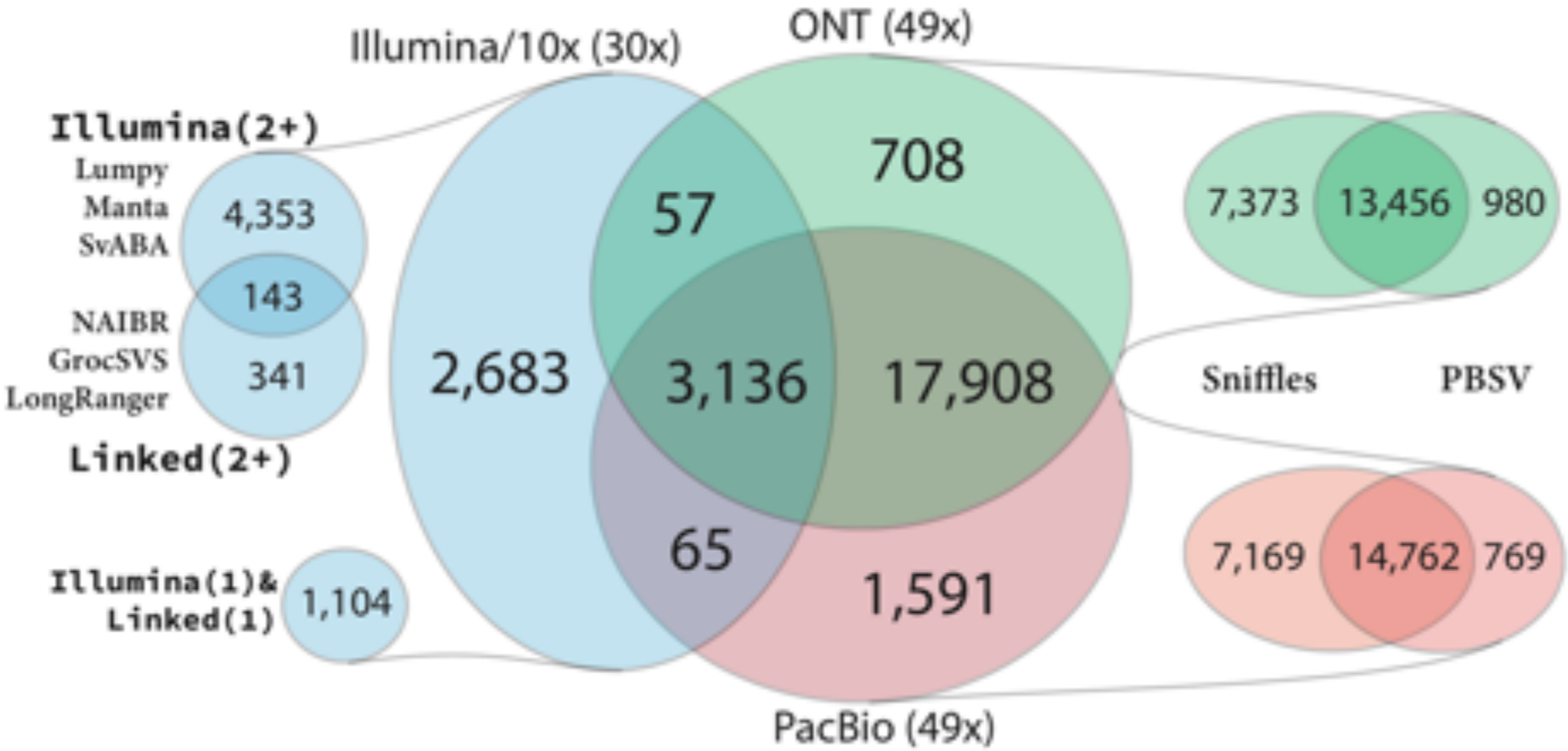
Karen Kostroff

Breast Cancer Patient Sequencing

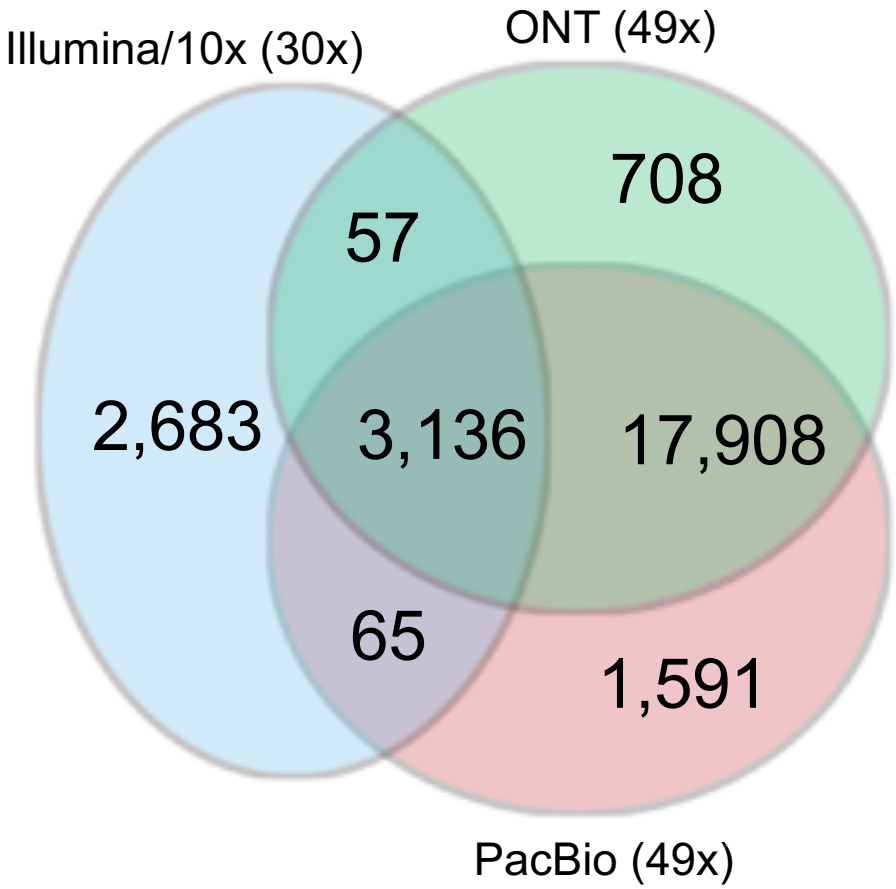


Comprehensive analysis of structural variants in breast cancer genomes using single molecule sequencing
Aganezov, S et al. (2020) *Genome Research. doi: 10.1101/gr.260497.119*

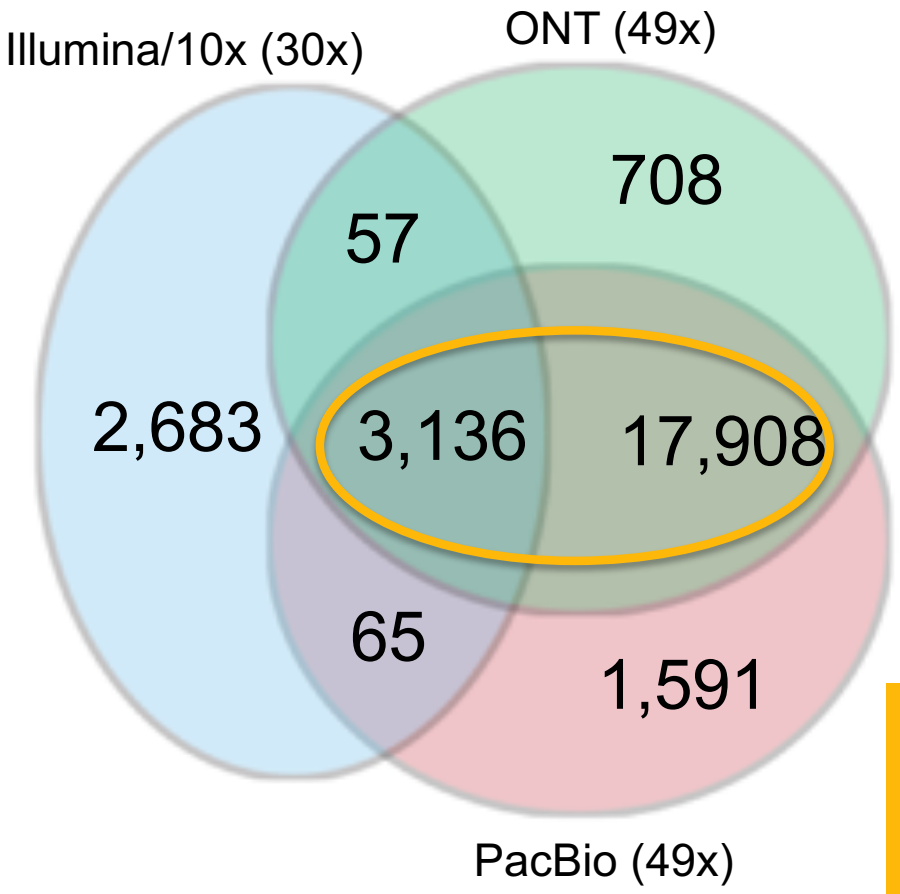
Structural Variation Consistency



Structural Variation Consistency

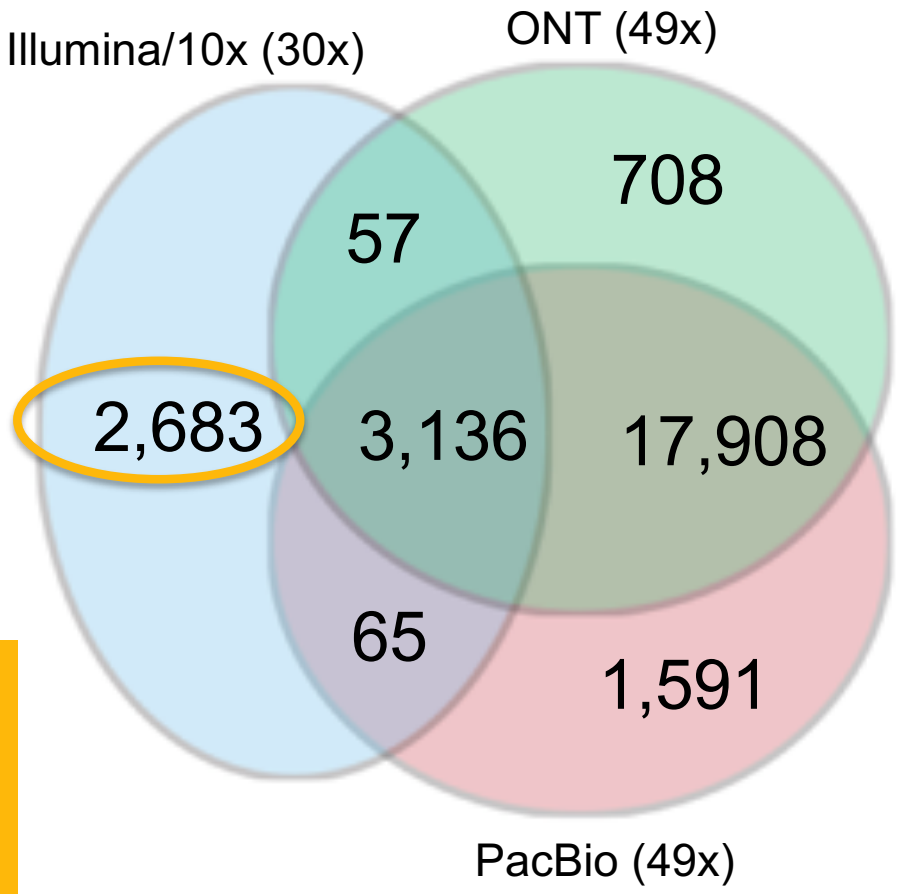


Structural Variation Consistency



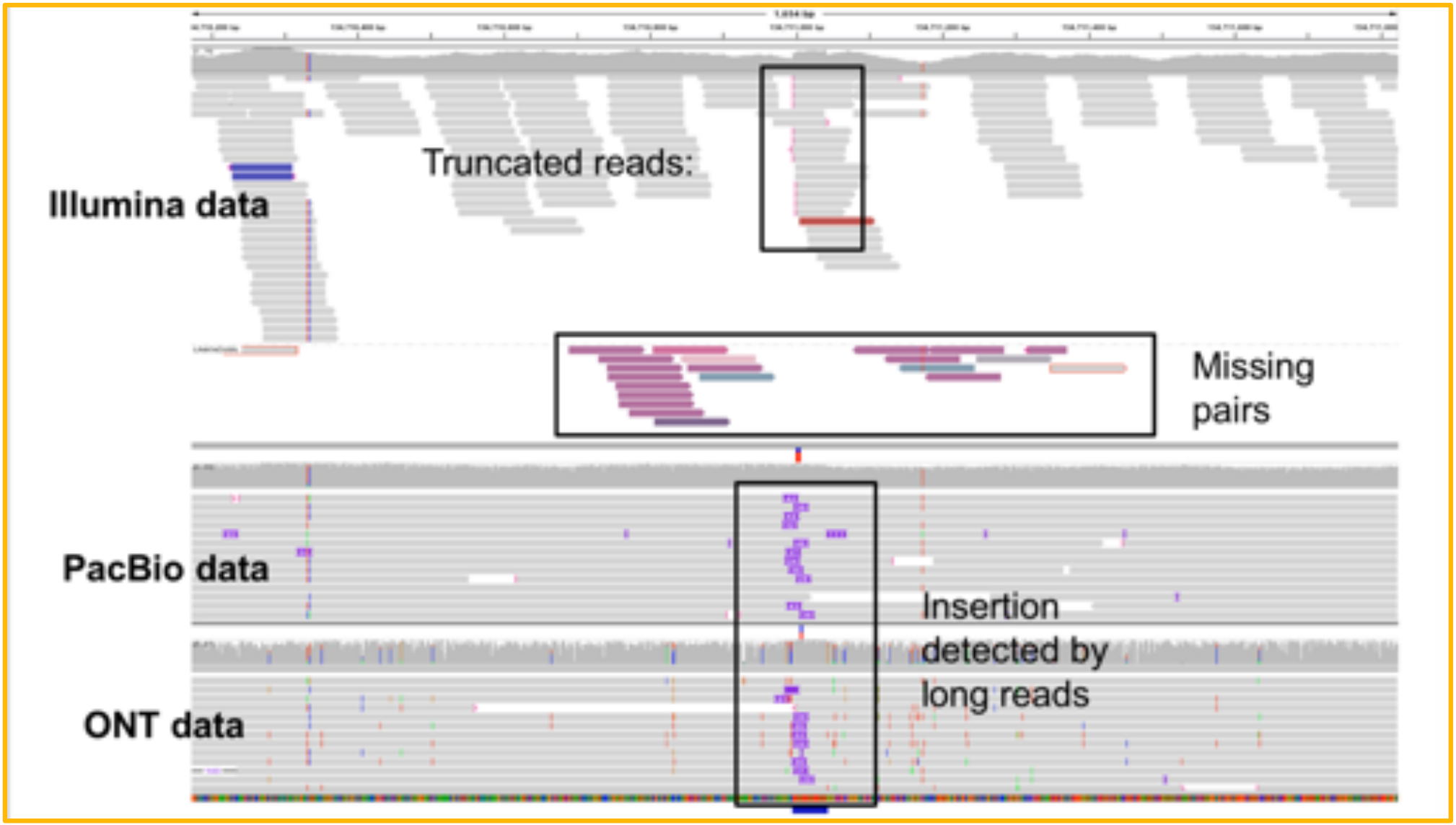
- Very strong concordance between long read platforms
- Substantially more variants than detected by short reads

Structural Variation Consistency

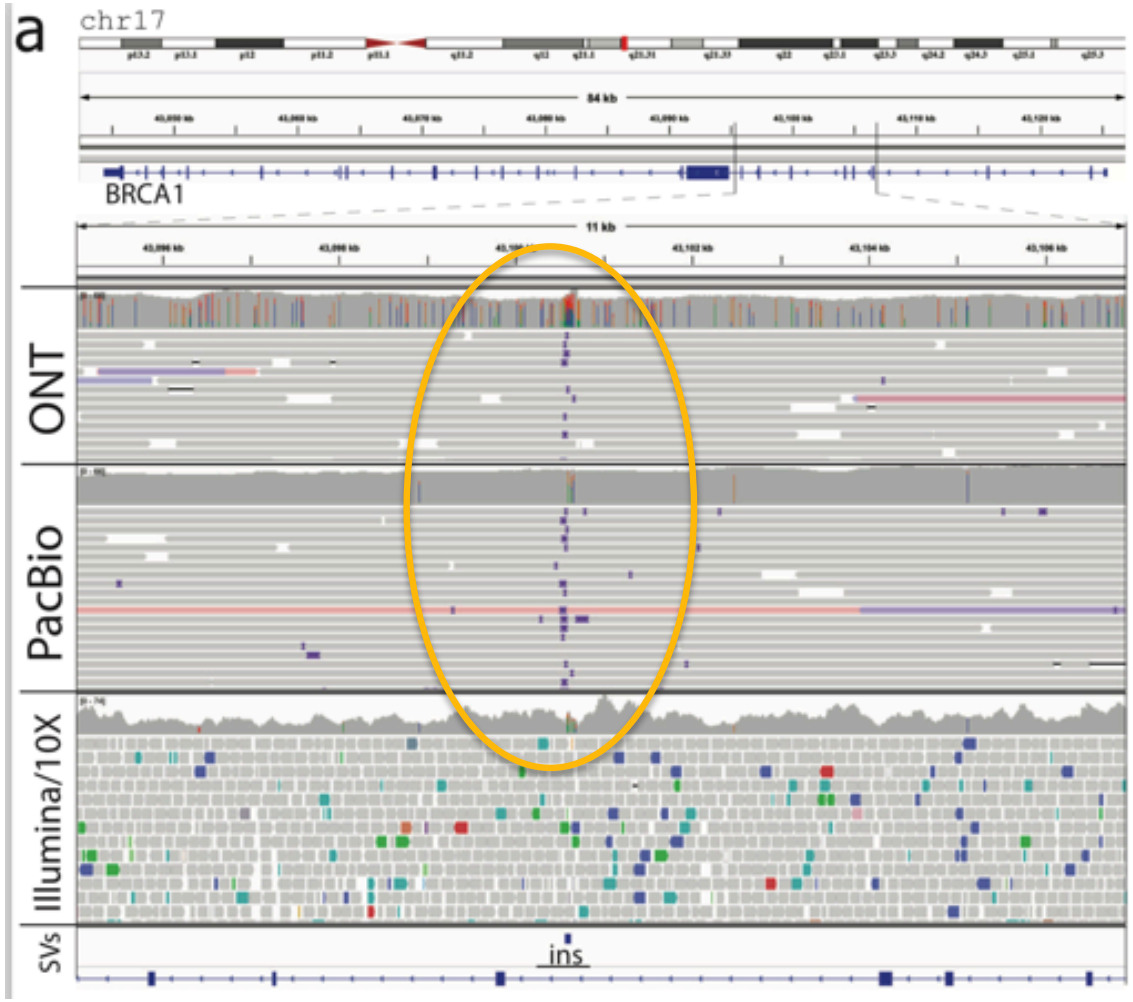


- PCR validation shows most Illumina-only calls are false positives

Structural Variation Consistency

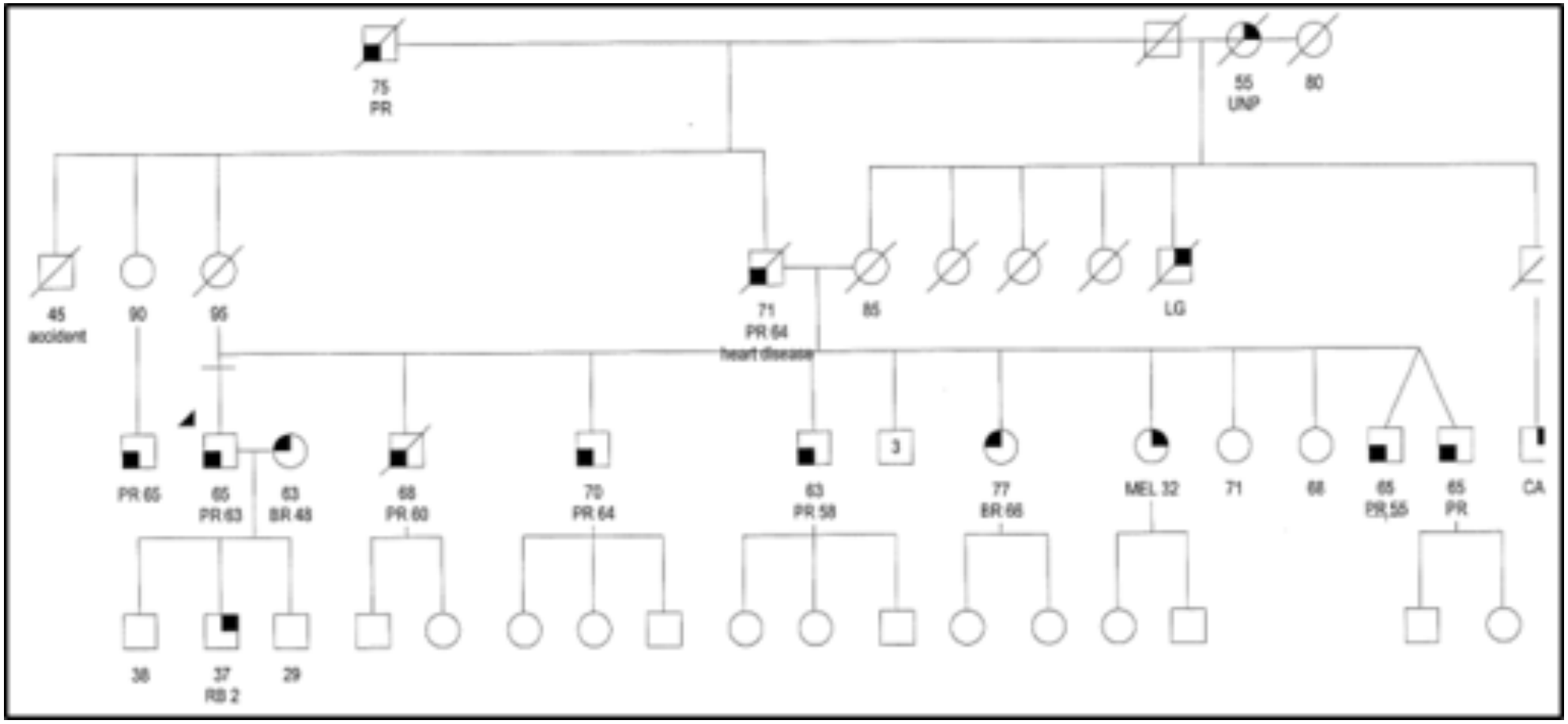


Hidden Variants in Breast Cancer Genes



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

What causes "outlier" families?



The Mark Foundation® for Cancer Research

Eli Van Allen,
Dana-Farber Cancer Institute

100 Genomes in 100 Days

The Structural Variant Landscape of Tomato Genomes

Photo by Uli Westphal

Major impacts of widespread structural variation on gene expression and crop improvement in tomato

Alonge, Wang, et al. (2020) *Cell.* https://doi.org/10.1016/j.cell.2020.05.021

COMPUTATIONAL TOOLS

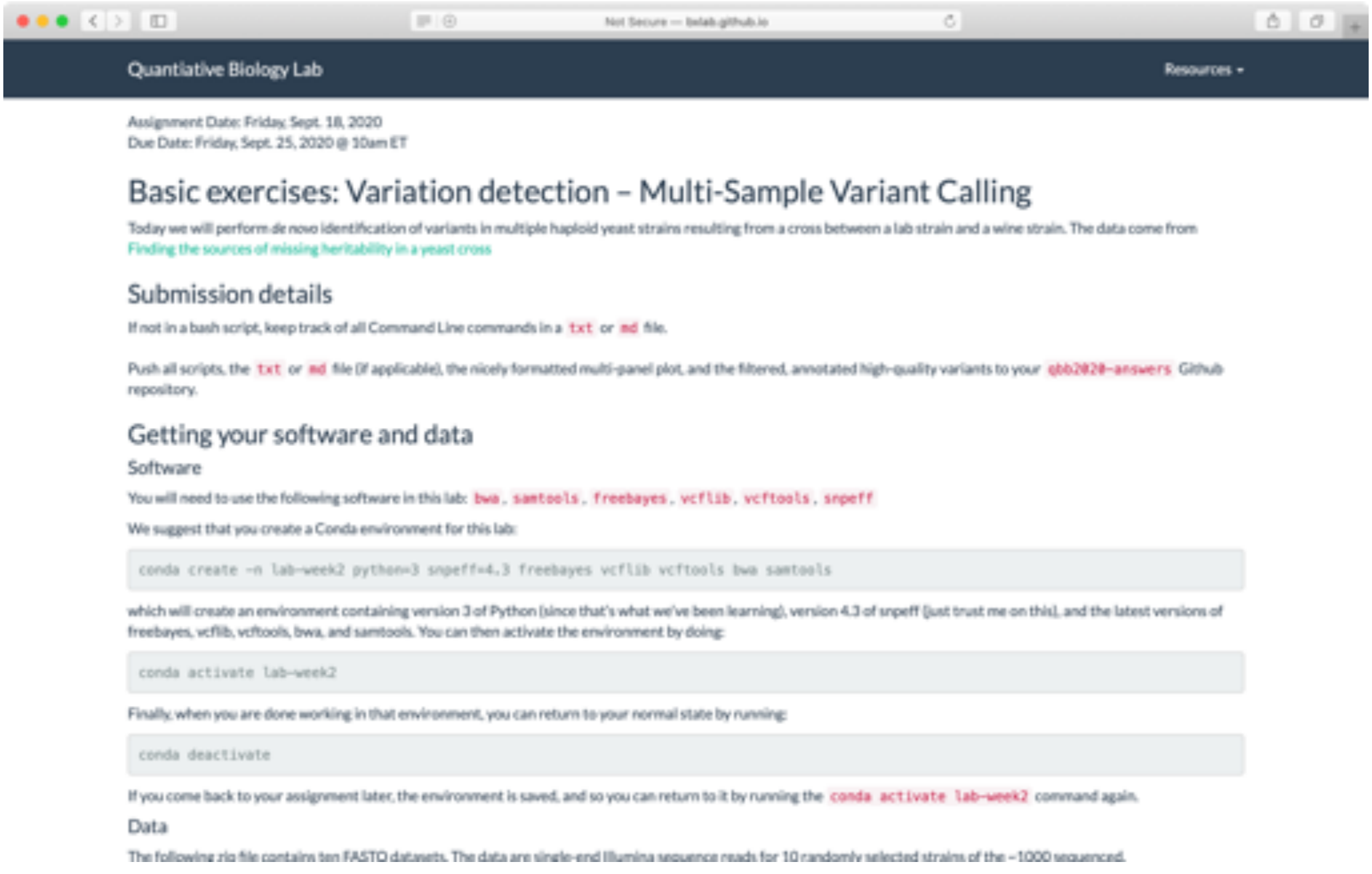
Piercing the dark matter: bioinformatics of long-range sequencing and mapping

*Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]**

Abstract | Several new genomics technologies have become available that offer long-read sequencing or long-range mapping with higher throughput and higher resolution analysis than ever before. These long-range technologies are rapidly advancing the field with improved reference genomes, more comprehensive variant identification and more complete views of transcriptomes and epigenomes. However, they also require new bioinformatics approaches to take full advantage of their unique characteristics while overcoming their complex errors and modalities. Here, we discuss several of the most important applications of the new technologies, focusing on both the currently available bioinformatics tools and opportunities for future research.

Piercing the dark matter: bioinformatics of long- range sequencing and mapping
Sedlazeck et al (2018) *Nature Reviews Genetics. doi:10.1038/s41576-018-0003-4*

Assignment 2



Quantiative Biology Lab

Resources

Assignment Date: Friday, Sept. 18, 2020
Due Date: Friday, Sept. 25, 2020 @ 10am ET

Basic exercises: Variation detection – Multi-Sample Variant Calling

Today we will perform de novo identification of variants in multiple haploid yeast strains resulting from a cross between a lab strain and a wine strain. The data come from Finding the sources of missing heritability in a yeast cross

Submission details

If not in a bash script, keep track of all Command Line commands in a `txt` or `md` file.

Push all scripts, the `txt` or `md` file (if applicable), the nicely formatted multi-panel plot, and the filtered, annotated high-quality variants to your `qbb2020-answers` Github repository.

Getting your software and data

Software

You will need to use the following software in this lab: `bwa`, `samtools`, `freebayes`, `vcflib`, `vcftools`, `snpeff`

We suggest that you create a Conda environment for this lab:

```
conda create -n lab-week2 python=3 snpeff=4.3 freebayes vcflib vcftools bwa samtools
```

which will create an environment containing version 3 of Python (since that's what we've been learning), version 4.3 of snpeff (just trust me on this), and the latest versions of freebayes, vcflib, vcftools, bwa, and samtools. You can then activate the environment by doing:

```
conda activate lab-week2
```

Finally, when you are done working in that environment, you can return to your normal state by running:

```
conda deactivate
```

If you come back to your assignment later, the environment is saved, and so you can return to it by running the `conda activate lab-week2` command again.

Data

The following zip file contains ten FASTQ datasets. The data are single-end Illumina sequence reads for 10 randomly selected strains of the ~1000 sequenced.