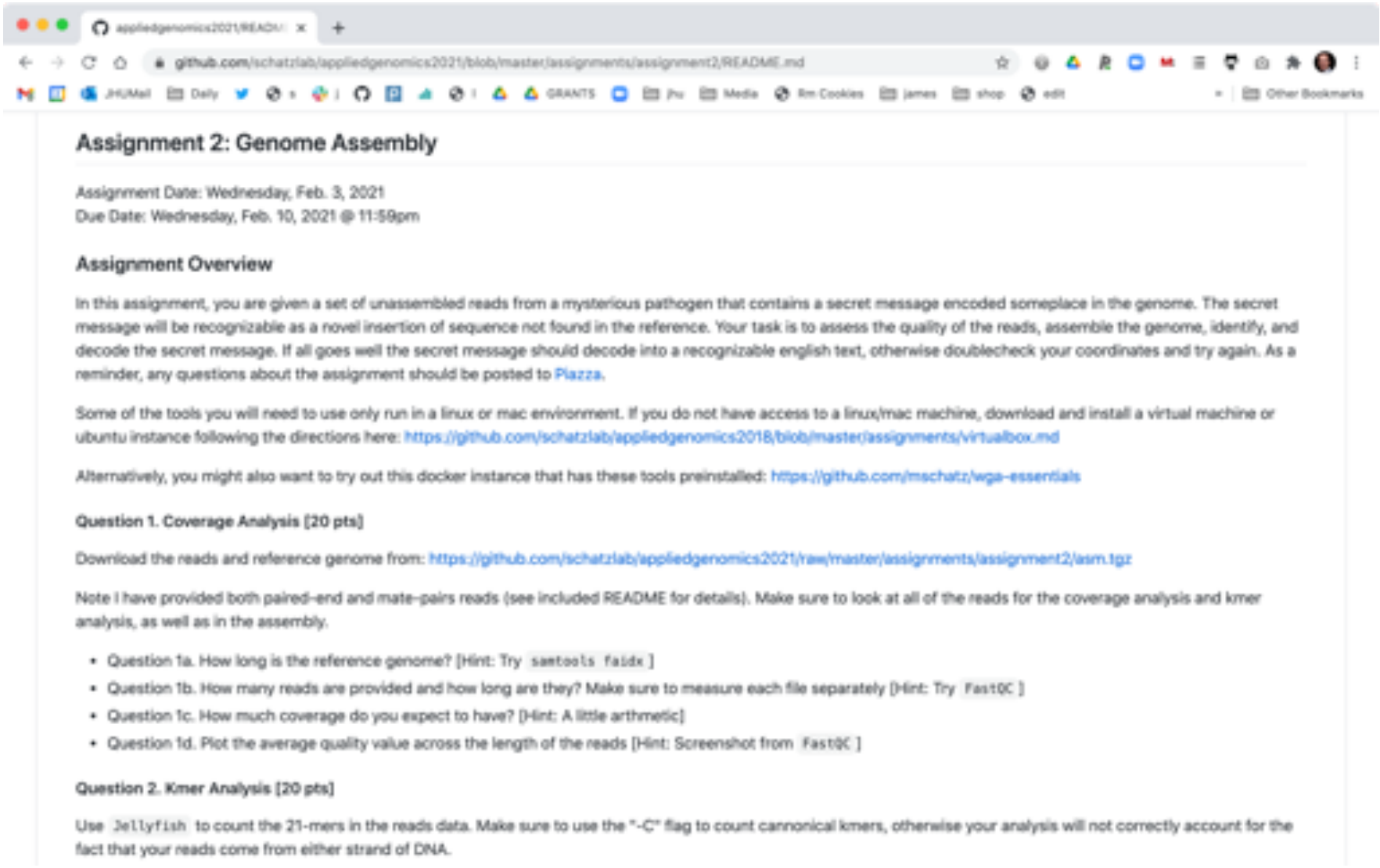
Long Read Assembly

Michael Schatz

Feb 15, 2021

Lecture 7: Computational Biomedical Research

Assignment 2: Genome Assembly
Due Feb 10 @ 11:59pm



Assignment 2: Genome Assembly

Assignment Date: Wednesday, Feb. 3, 2021
Due Date: Wednesday, Feb. 10, 2021 @ 11:59pm

Assignment Overview

In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded someplace in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text, otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza.

Some of the tools you will need to use only run in a linux or mac environment. If you do not have access to a linux/mac machine, download and install a virtual machine or ubuntu instance following the directions here: https://github.com/schatzlab/appliedgenomics2018/blob/master/assignments/virtualbox.md

Alternatively, you might also want to try out this docker instance that has these tools preinstalled: https://github.com/mschatz/wga-essentials

Question 1. Coverage Analysis [20 pts]

Download the reads and reference genome from: https://github.com/schatzlab/appliedgenomics2021/raw/master/assignments/assignment2/asm.tgz

Note I have provided both paired-end and mate-pairs reads (see included README for details). Make sure to look at all of the reads for the coverage analysis and kmer analysis, as well as in the assembly.

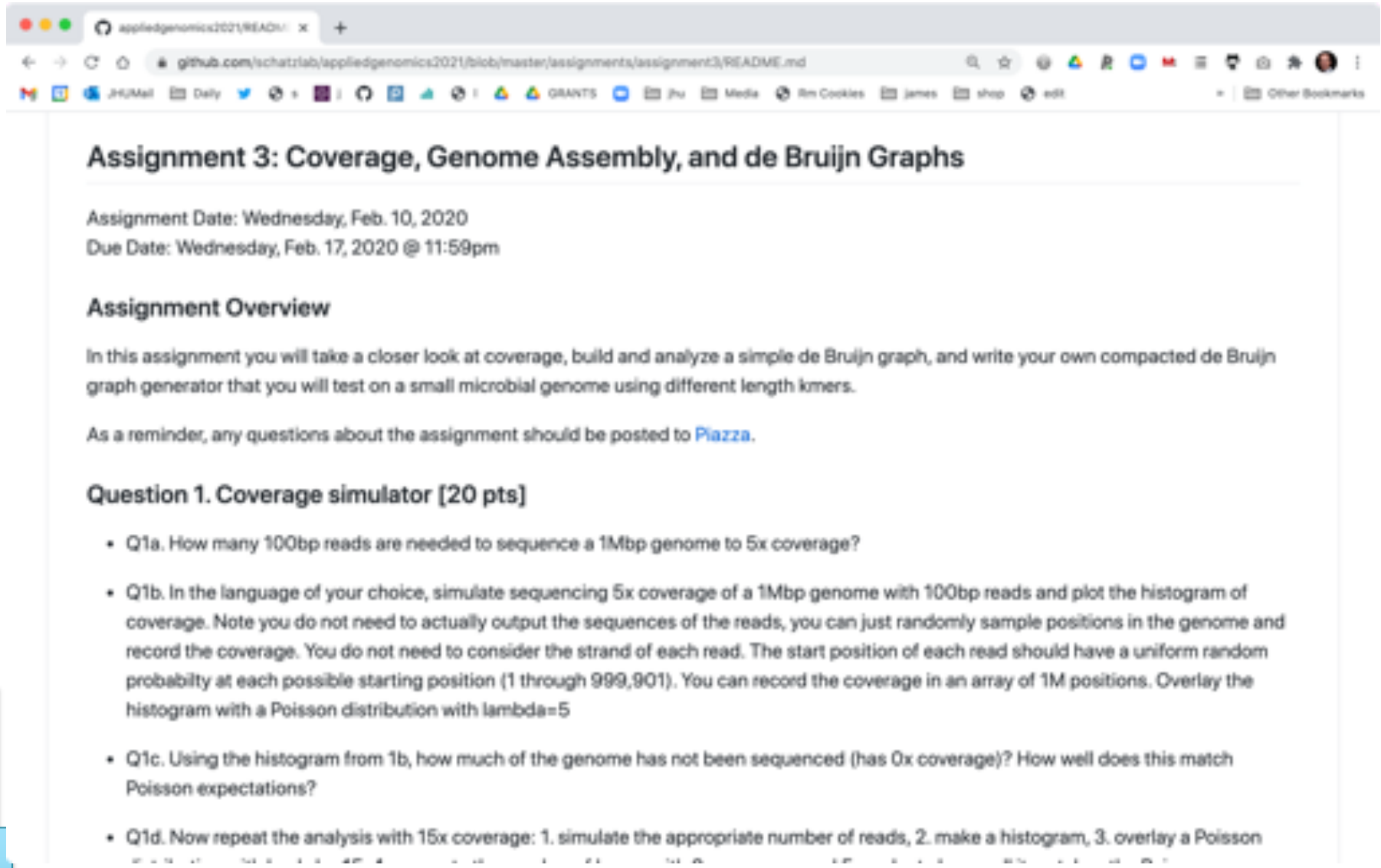
- Question 1a. How long is the reference genome? [Hint: Try `samtools faidx`]
- Question 1b. How many reads are provided and how long are they? Make sure to measure each file separately [Hint: Try `FastQC`]
- Question 1c. How much coverage do you expect to have? [Hint: A little arthmetic]
- Question 1d. Plot the average quality value across the length of the reads [Hint: Screenshot from `FastQC`]

Question 2. Kmer Analysis [20 pts]

Use `Jellyfish` to count the 21-mers in the reads data. Make sure to use the "-C" flag to count cannonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

https://github.com/schatzlab/appliedgenomics2021

Assignment 3: de Bruijn Graphs
Due Feb 17 @ 11:59pm



Assignment 3: Coverage, Genome Assembly, and de Bruijn Graphs

Assignment Date: Wednesday, Feb. 10, 2020
Due Date: Wednesday, Feb. 17, 2020 @ 11:59pm

Assignment Overview

In this assignment you will take a closer look at coverage, build and analyze a simple de Bruijn graph, and write your own compacted de Bruijn graph generator that you will test on a small microbial genome using different length kmers.

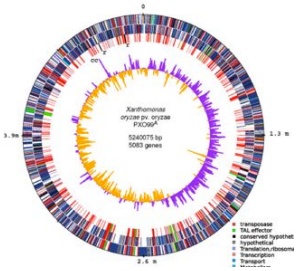
As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. Coverage simulator [20 pts]

- Q1a. How many 100bp reads are needed to sequence a 1Mbp genome to 5x coverage?
- Q1b. In the language of your choice, simulate sequencing 5x coverage of a 1Mbp genome with 100bp reads and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just randomly sample positions in the genome and record the coverage. You do not need to consider the strand of each read. The start position of each read should have a uniform random probabilty at each possible starting position (1 through 999,901). You can record the coverage in an array of 1M positions. Overlay the histogram with a Poisson distribution with lambda=5
- Q1c. Using the histogram from 1b, how much of the genome has not been sequenced (has 0x coverage)? How well does this match Poisson expectations?
- Q1d. Now repeat the analysis with 15x coverage: 1. simulate the appropriate number of reads, 2. make a histogram, 3. overlay a Poisson

https://github.com/schatzlab/appliedgenomics2021

Assembly Summary



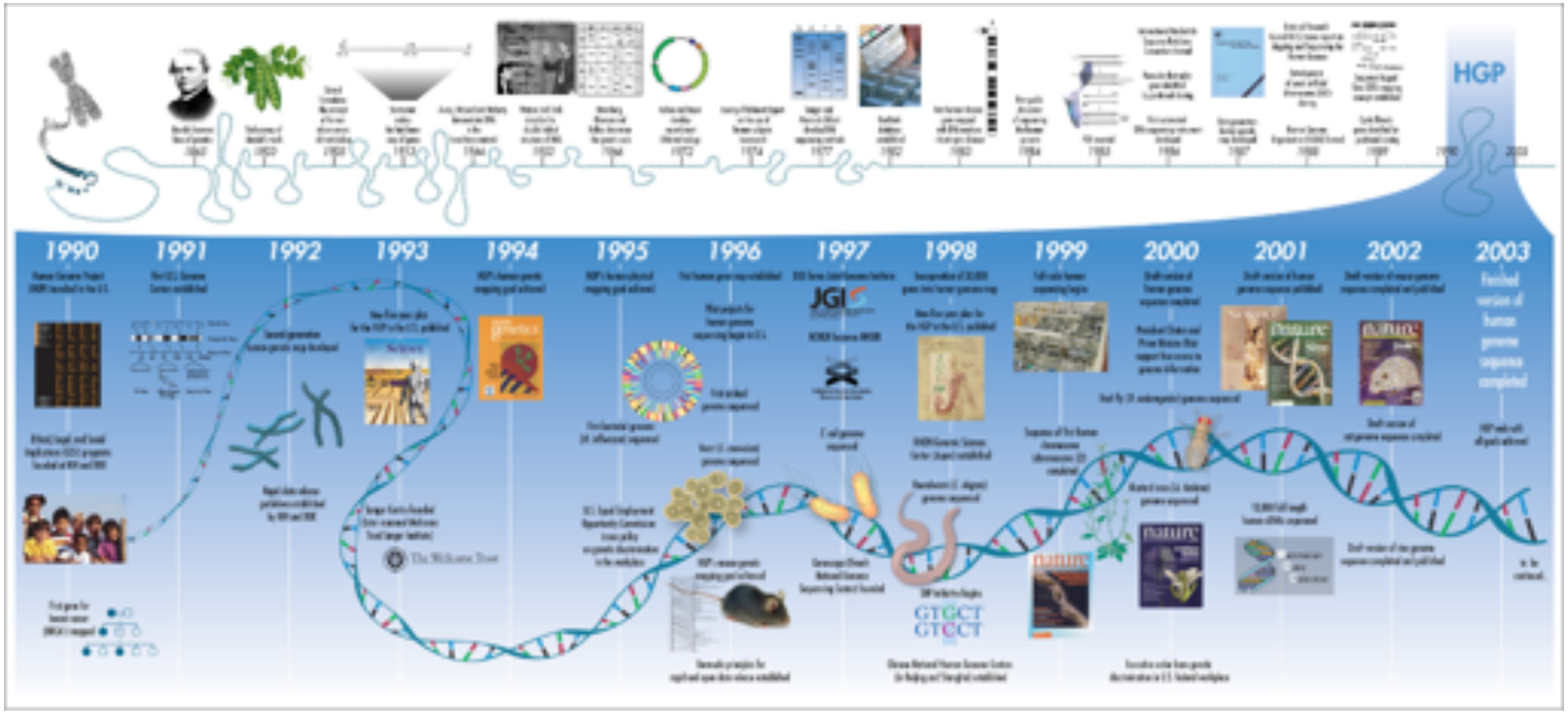
Assembly quality depends on

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

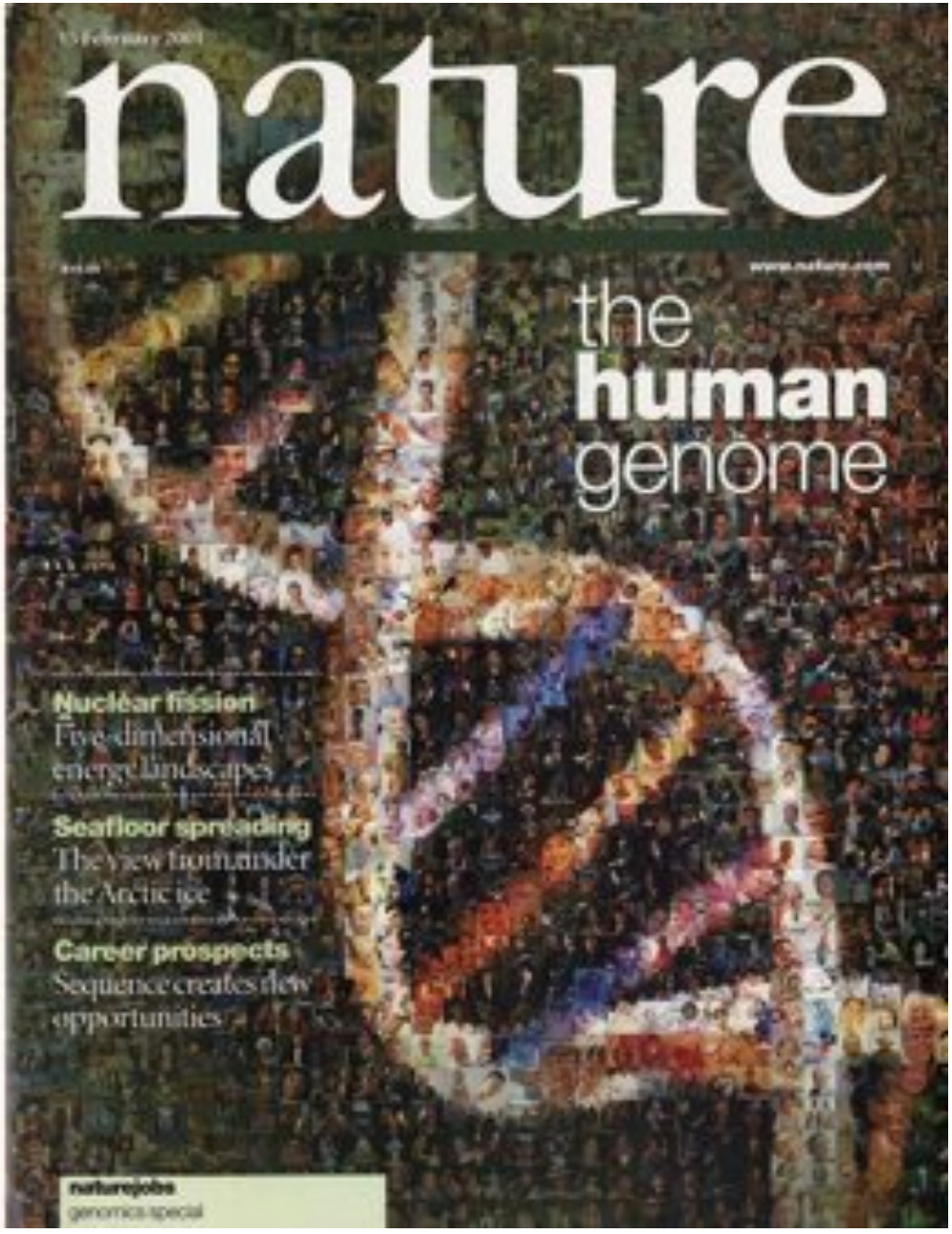
- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

History of the Human Genome Project



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 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

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.

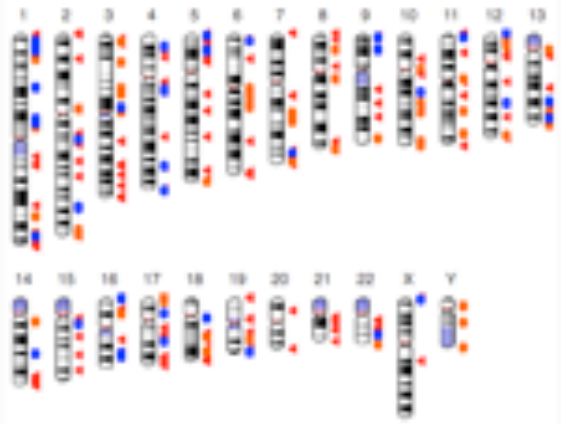
Genome Reference Consortium

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

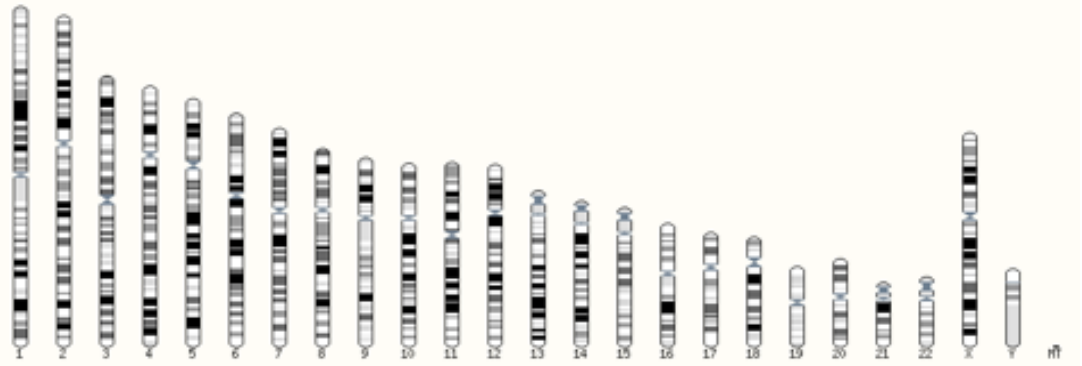
GRCh38.p11

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.

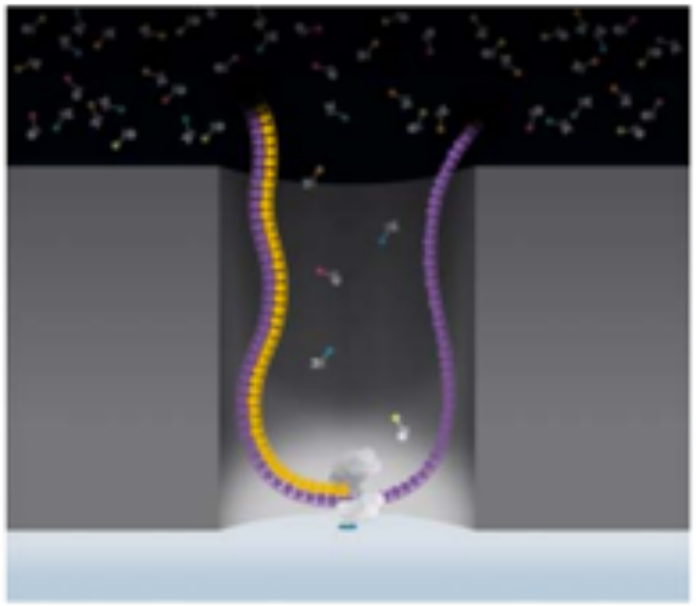
- Ongoing efforts to fill "gaps" and properly/thoroughly represent complex structures and loci in the genome (e.g., Major Histocompatibility Complex)
- Each improvement leads to a new genome "build". Currently on build 38.
- Experimental and computational methods provide new genome annotations
  - New 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?

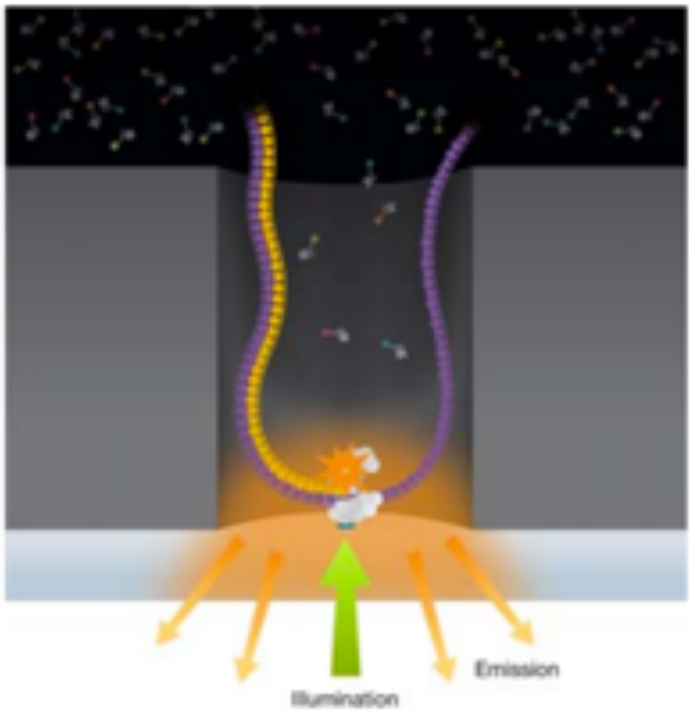
Genomics Arsenal in the Year 2021

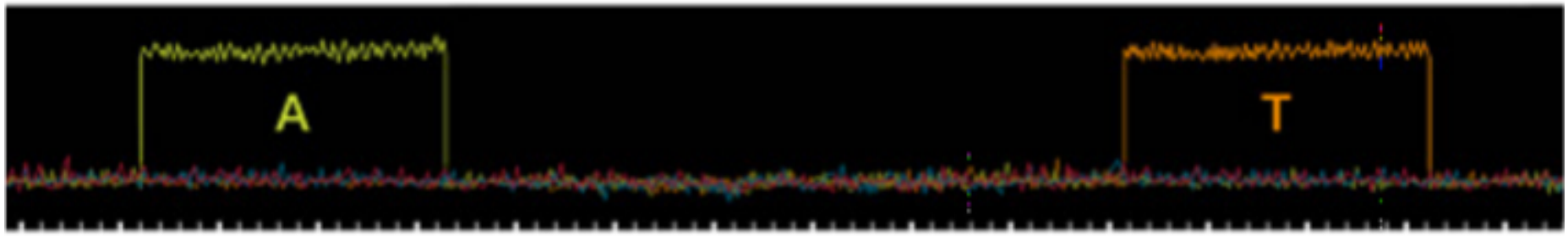
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

PacBio: SMRT Sequencing

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

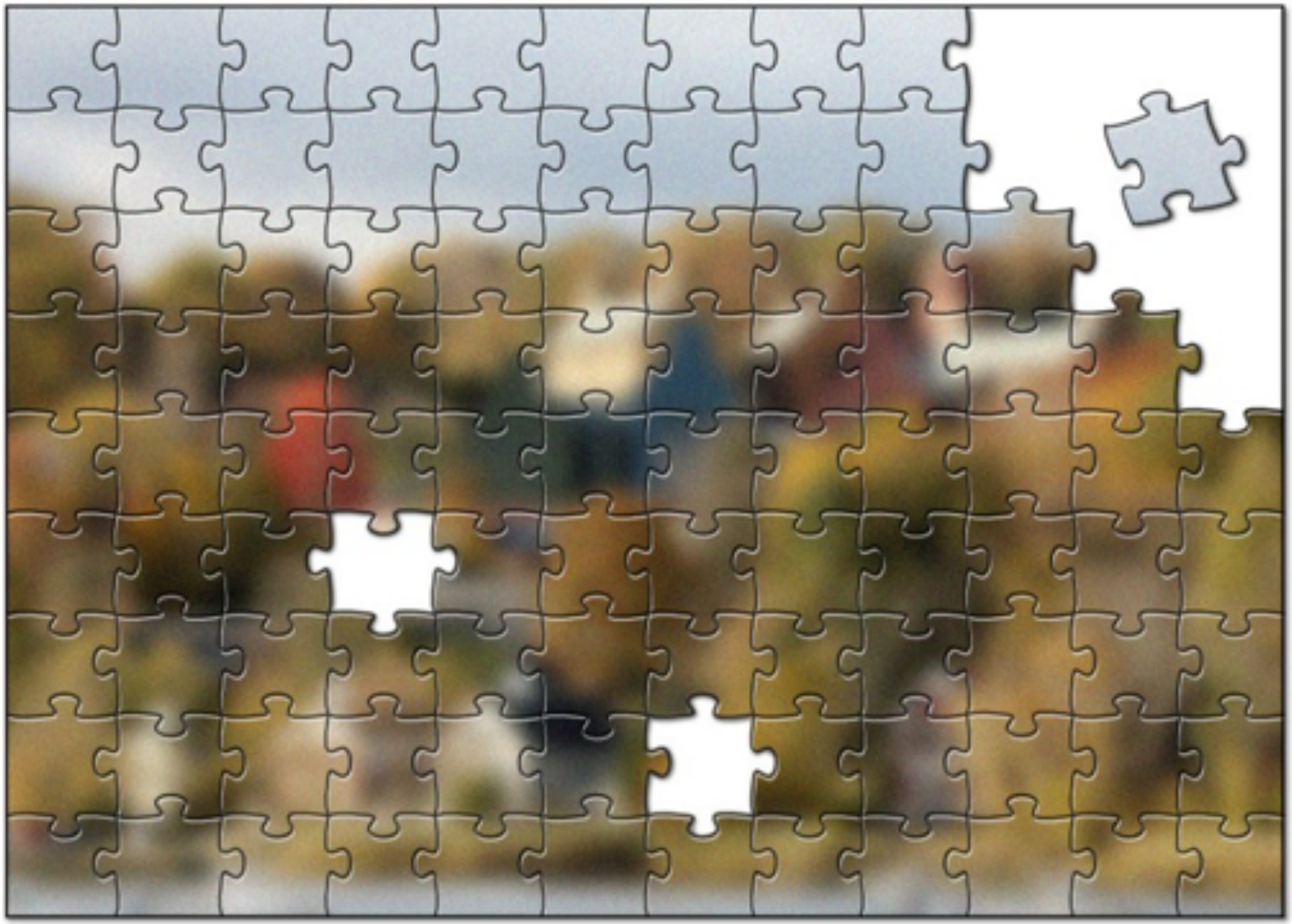




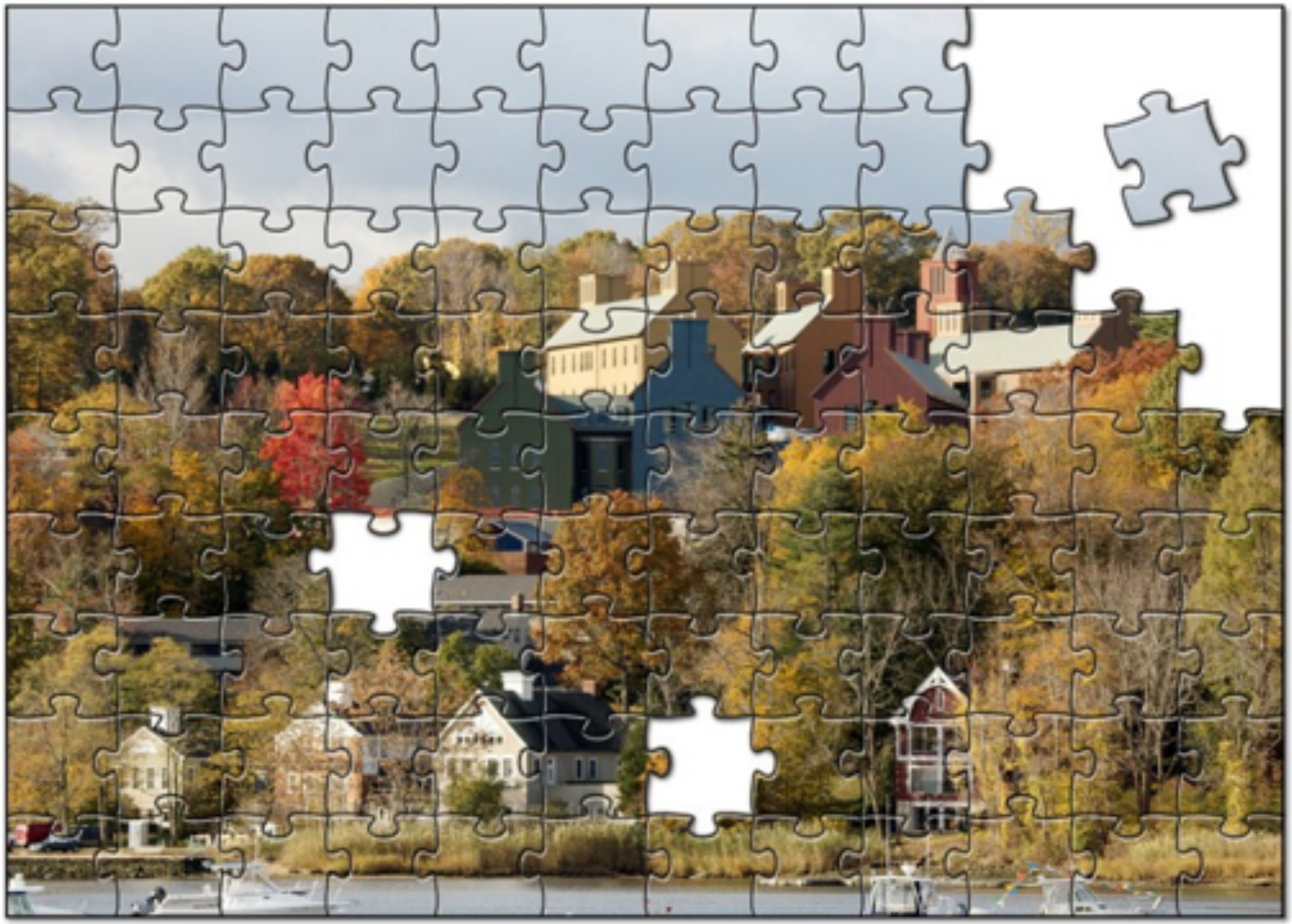


http://www.youtube.com/watch?v=v8p4ph2MAvI

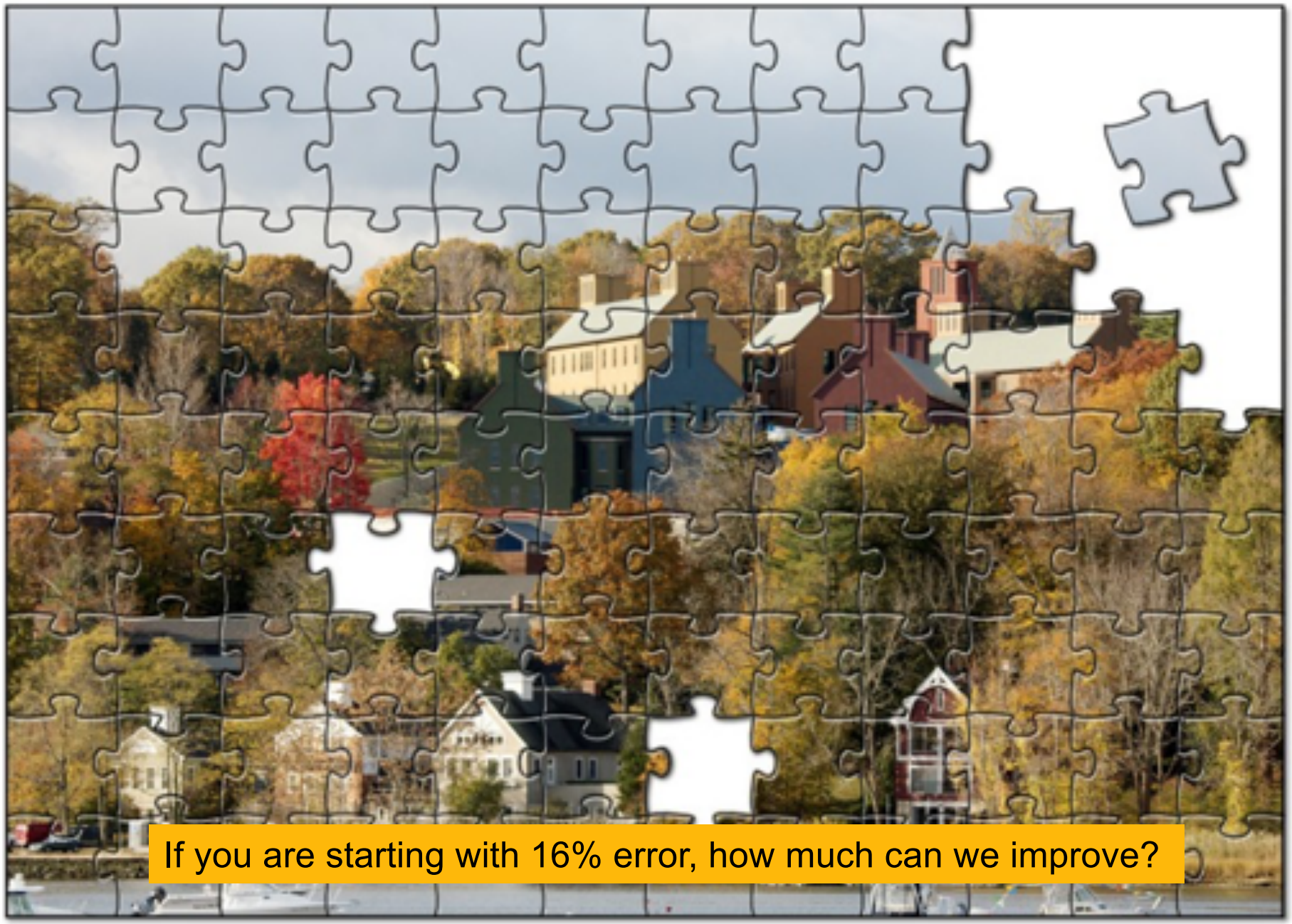
Single Molecule Sequences



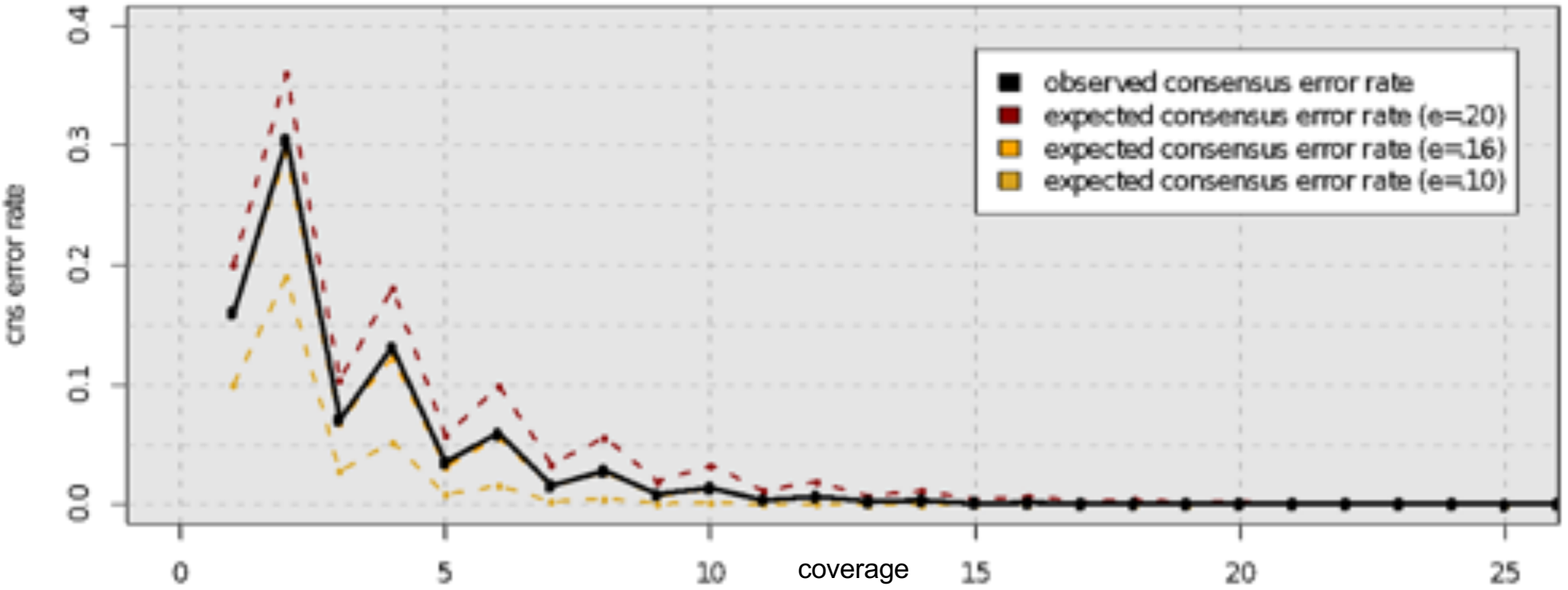
“Corrective Lens” for Sequencing



“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

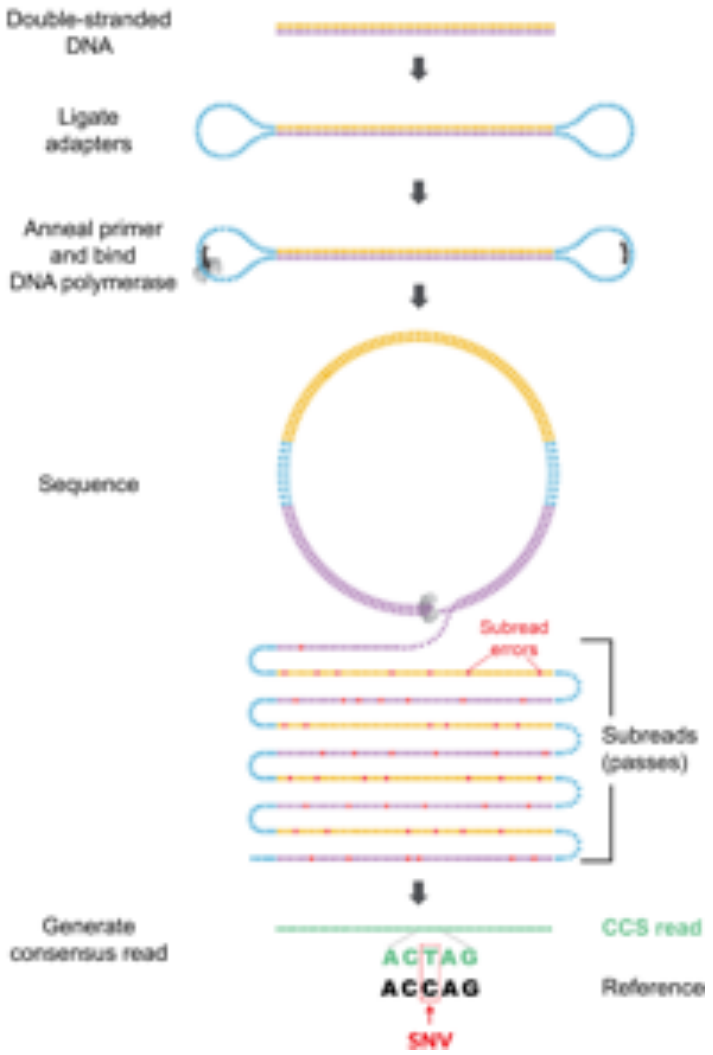
$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

“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, better & faster assembly

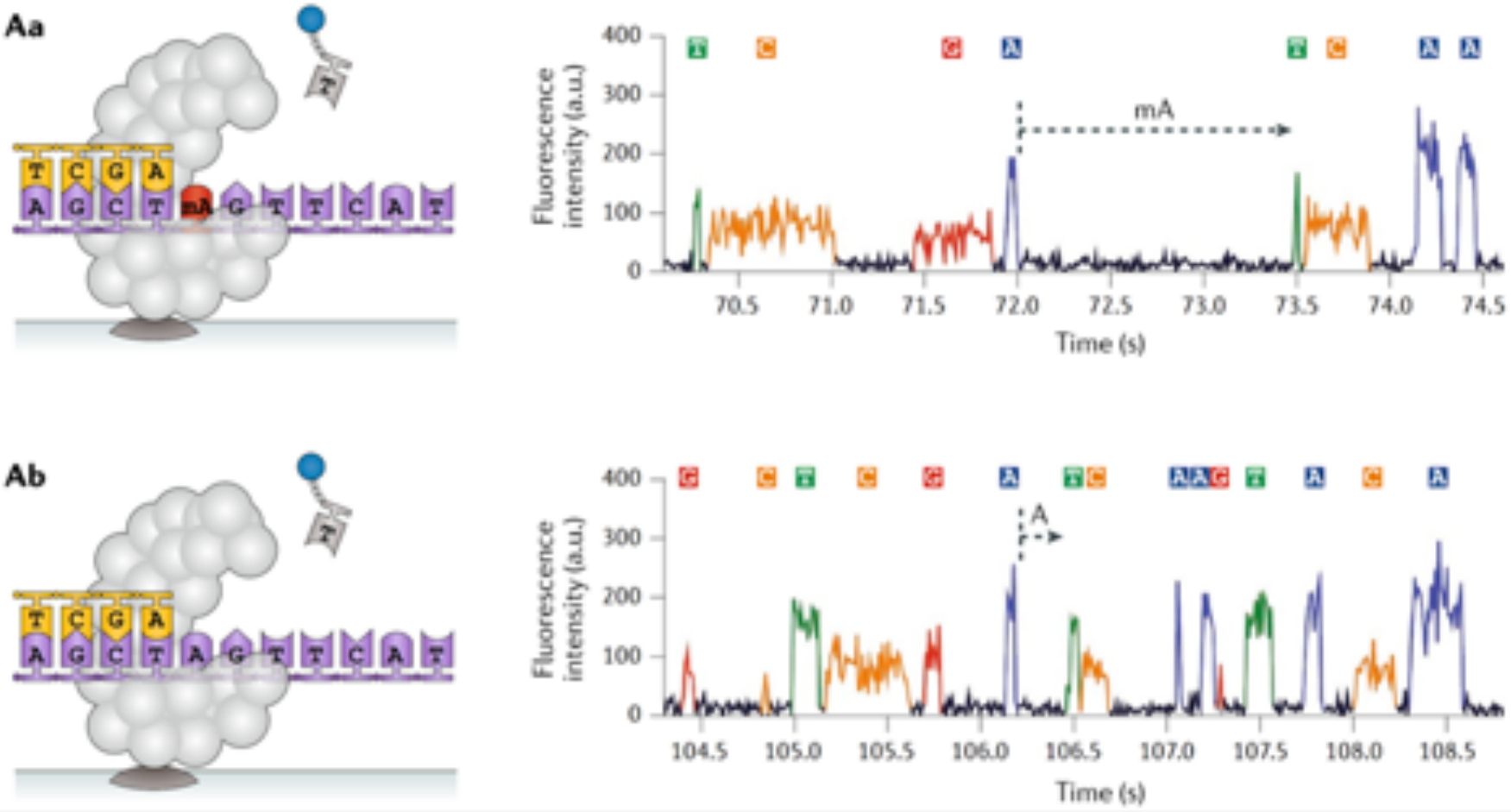
Limits read length, used to be very expensive but more manageable now



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

Methylation Detection

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



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

Following

5.66 USD +0.11 (1.89%) ↑
Sep 3, 2:45 PM EDT · Disclaimer

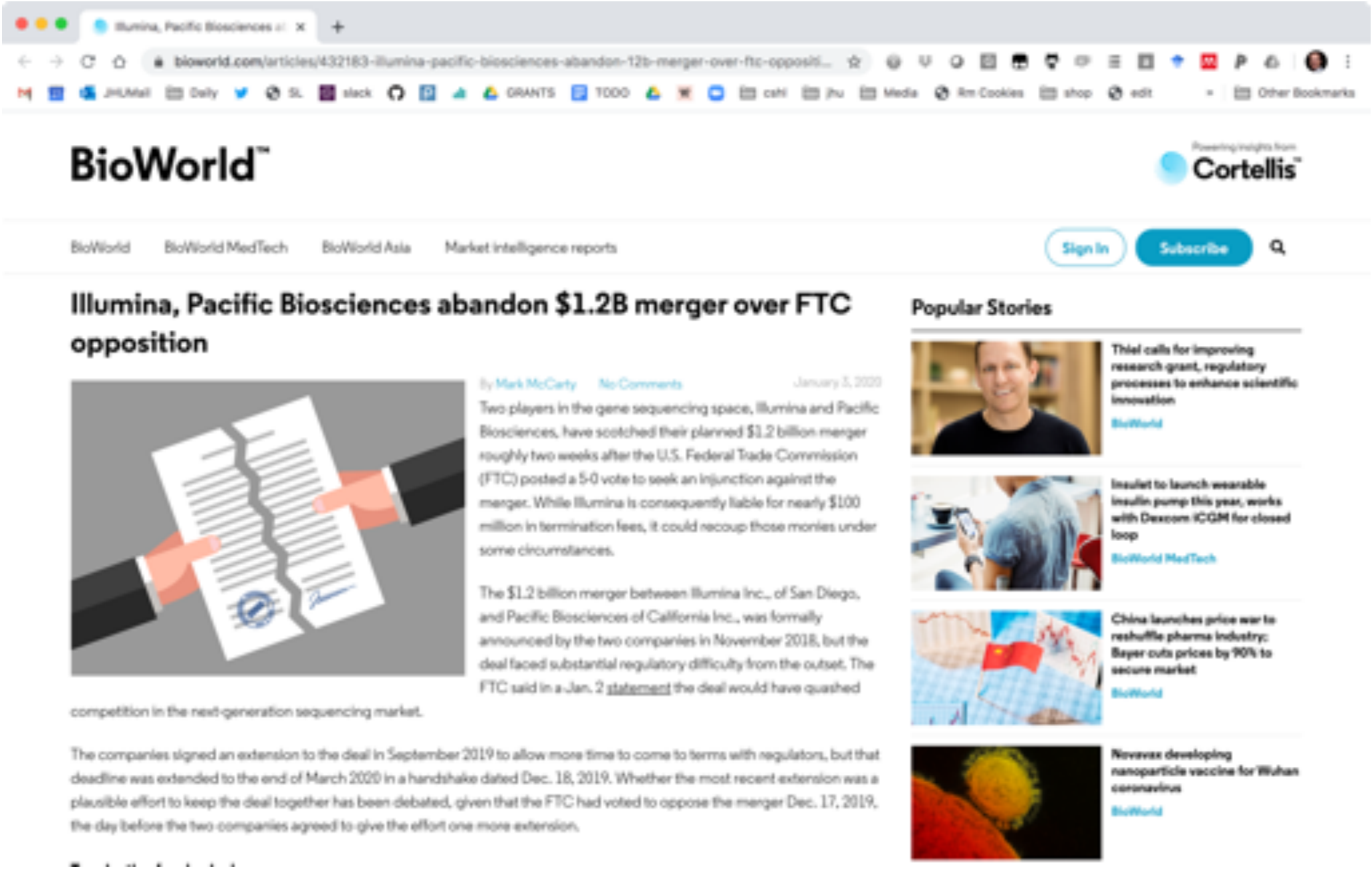
1 day | 5 days | 1 month | 6 months | YTD | 1 year | 5 years | **Max**

Open	5.53	Div yield	-
High	5.68	Prev close	5.55
Low	5.53	52-wk high	7.84
Mkt cap	865.02M	52-wk low	3.90
P/E ratio	-		

→ Financial news, comparisons and more

Feb 10, 2020

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

PACB

wsj.com/articles/softbank-to-make-900-million-investment-in-pacific-biosciences-11612921031

THE WALL STREET JOURNAL.

Subscribe | Sign In

Home World U.S. Politics Economy Business Tech **Markets** Opinion Life & Arts Real Estate WSJ Magazine

Search

SHARE

WSJ NEWS EXCLUSIVE | MARKETS

SoftBank to Make $900 Million Investment in Pacific Biosciences

PacBio investment will be made through asset-management arm SB Northstar

Market Summary > Pacific Biosciences of California Inc

NASDAQ: PACB

Following

50.32 USD −0.83 (1.62%)

Closed: Feb 12, 7:46 PM EST · Disclaimer
After hours 50.25 −0.070 (0.14%)

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

Open	49.03	Div yield	-
High	51.27	Prev close	51.15
Low	45.75	52-wk high	53.69
Mkt cap	9.72B	52-wk low	2.20
P/E ratio	299.43		

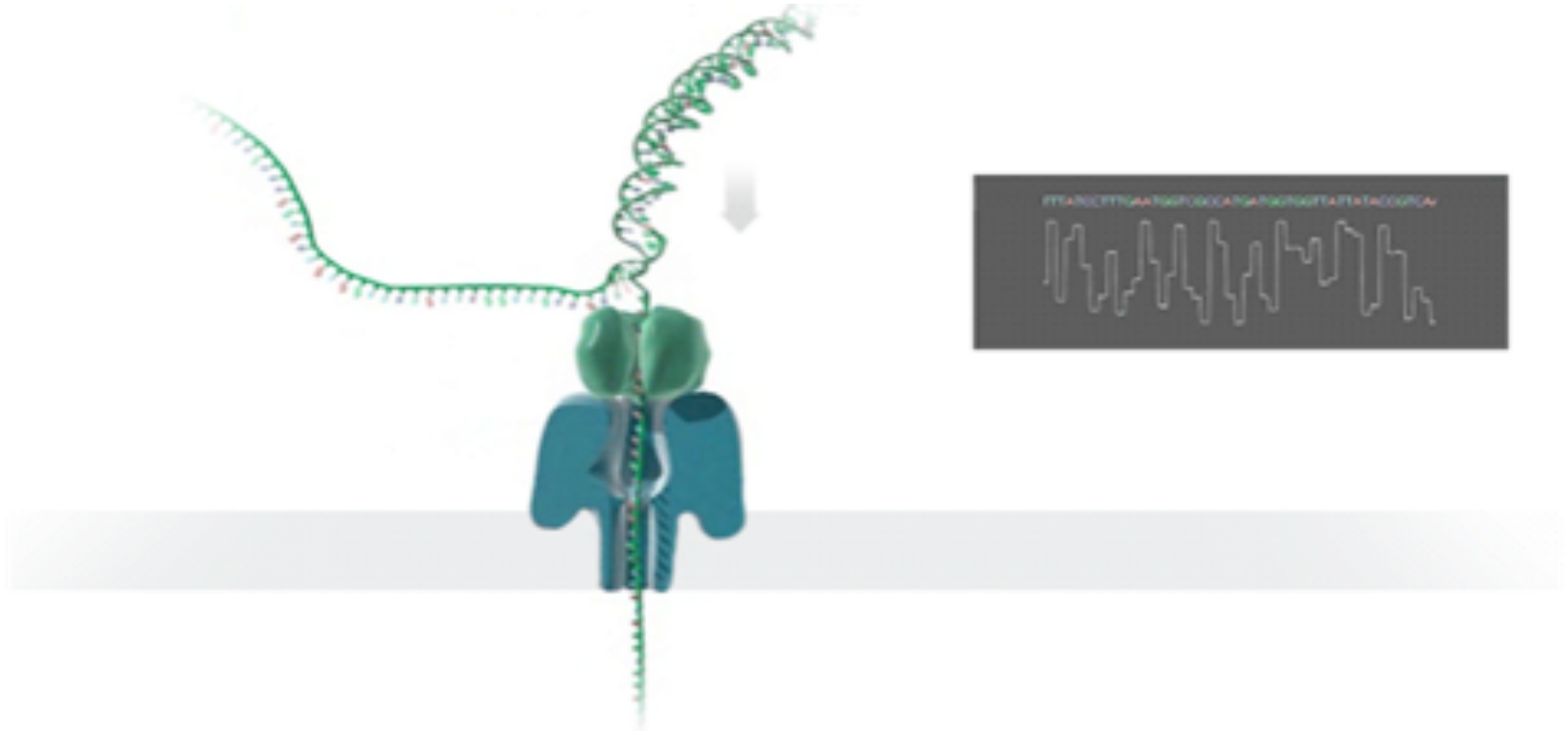
→ More about Pacific Biosciences o...

Feb 13, 2021

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

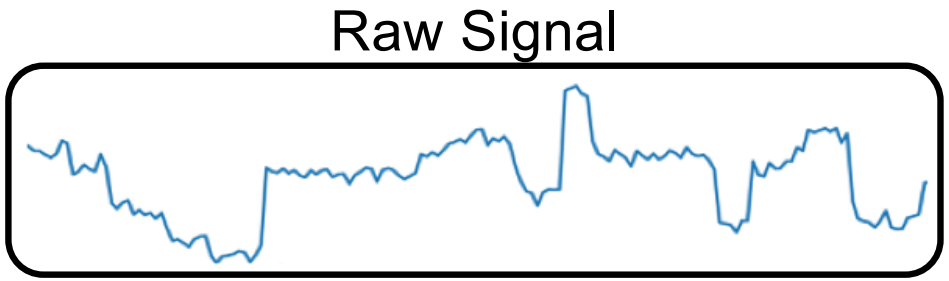
Sequences DNA/RNA by measuring changes in ionic current as nucleotide strand passes through a pore



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

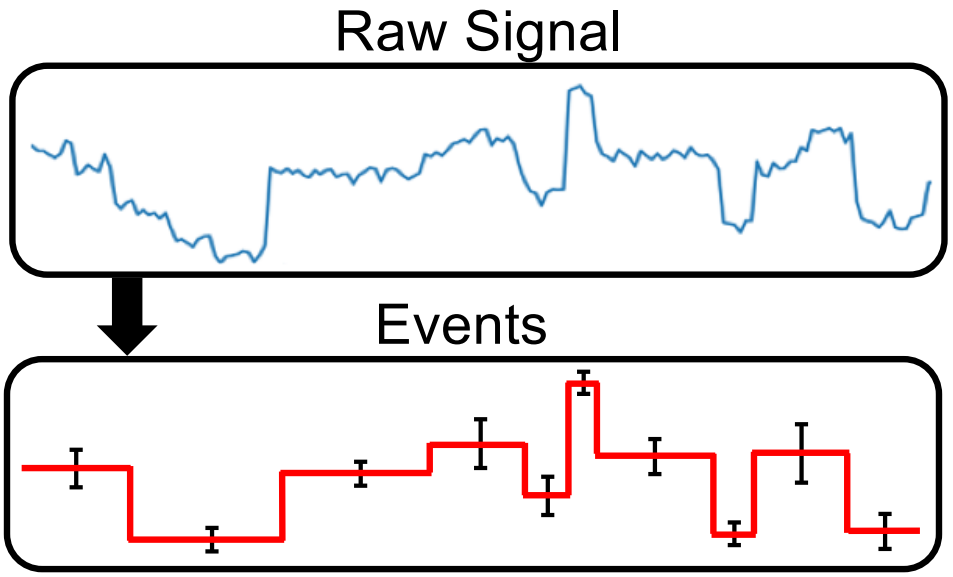
nanoporetech.com/applications/dna-nanopore-sequencing

Nanopore Basecalling



Translation of raw signal into basepairs

Nanopore Basecalling

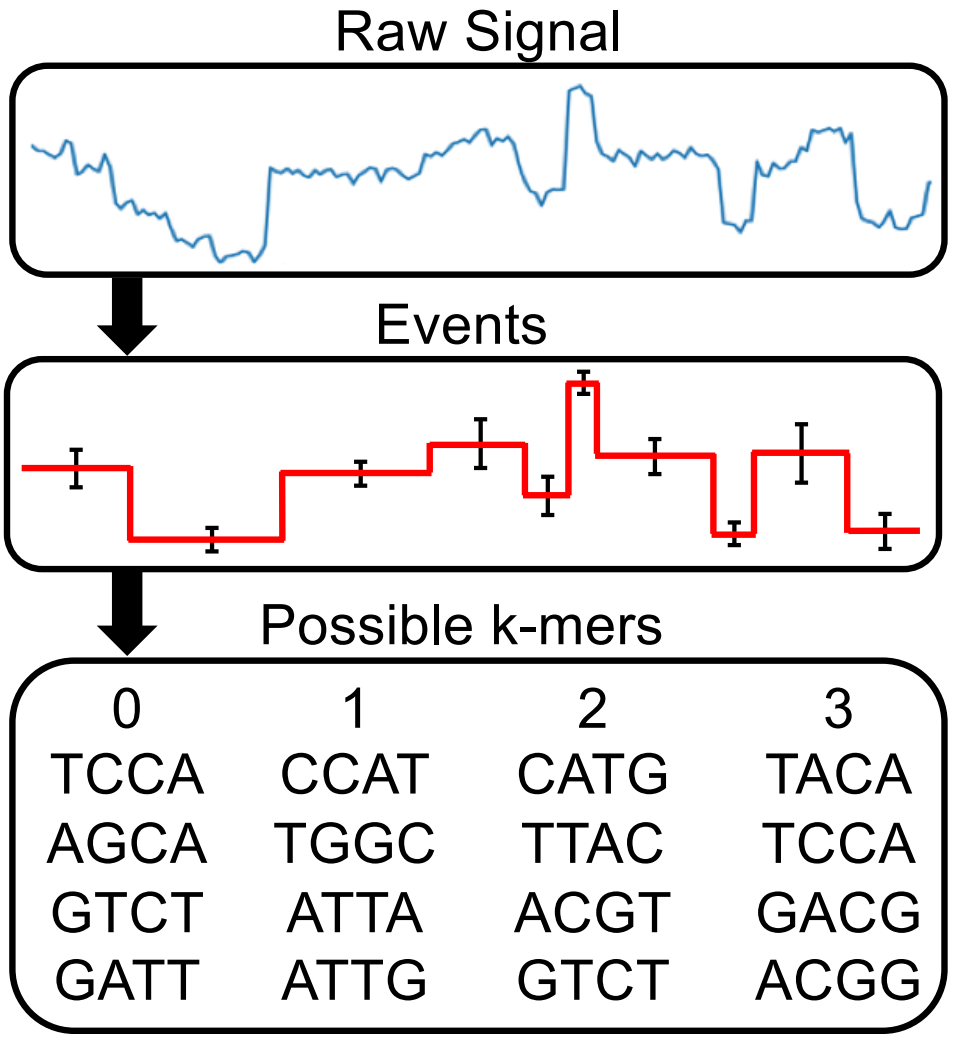


Translation of raw signal into basepairs

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

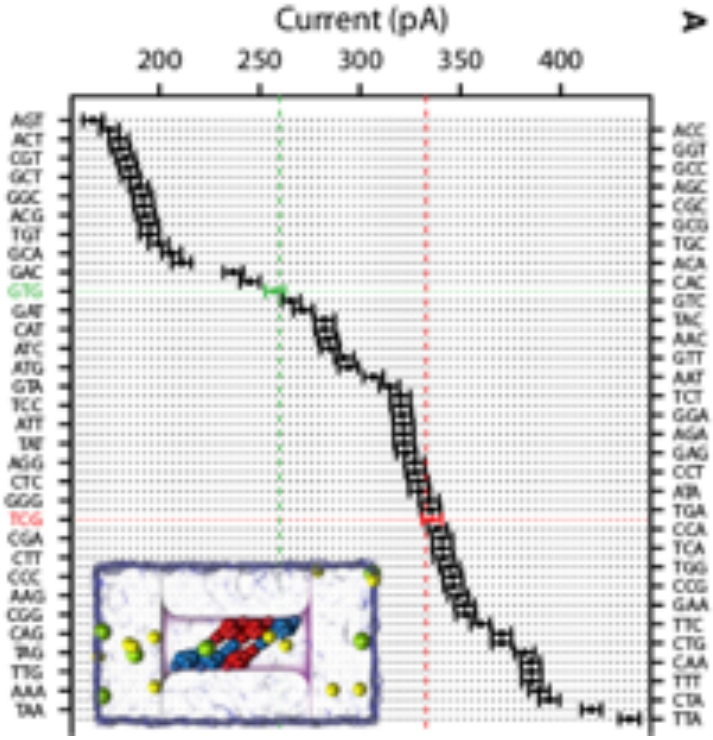
Modern basecalers use neural networks directly on raw signal

Nanopore Basecalling



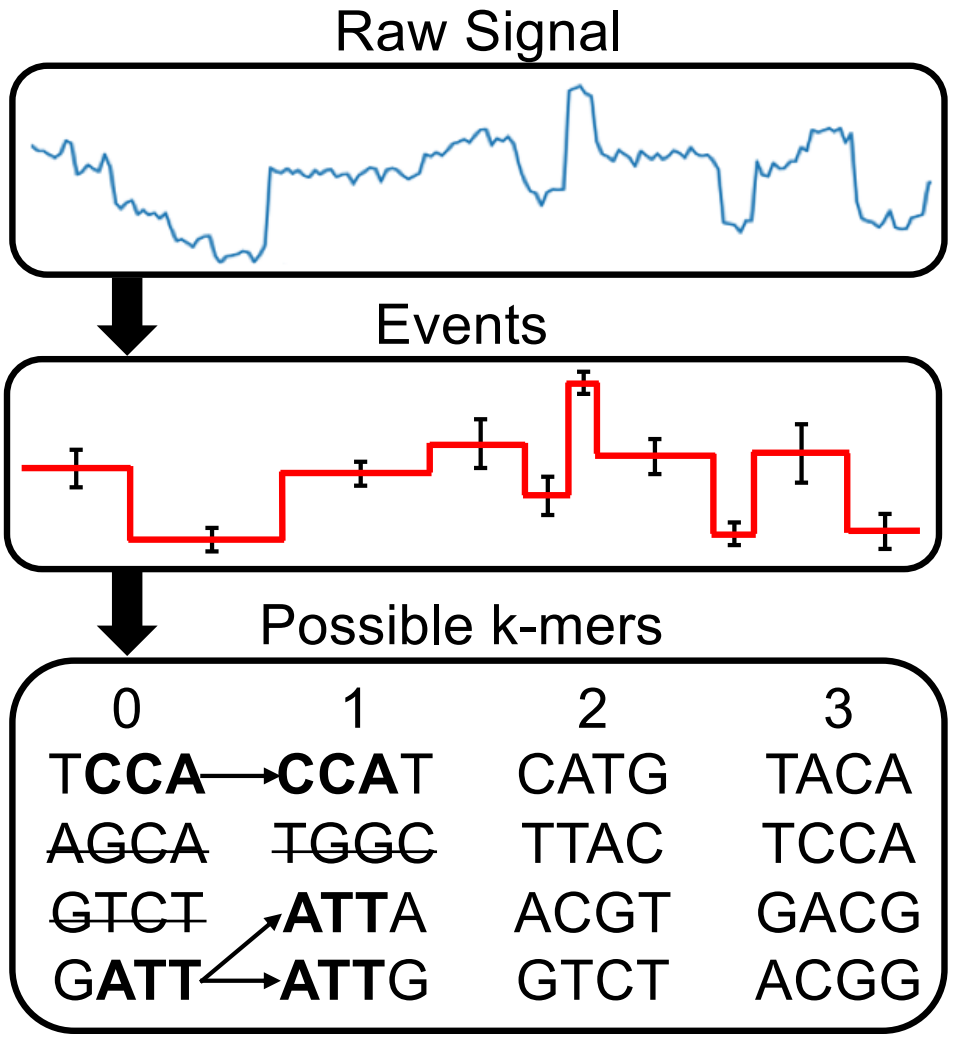
(Based on probability of event matches)

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

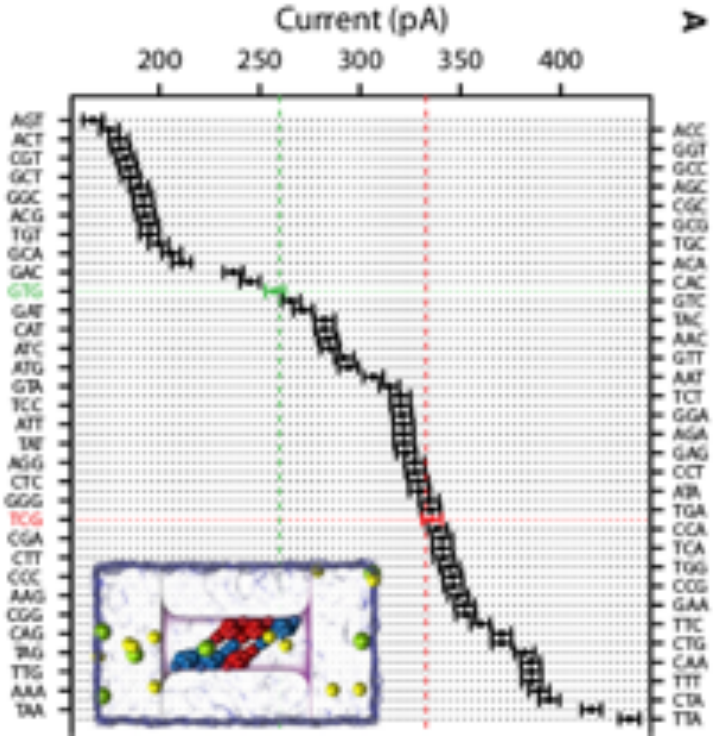


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling

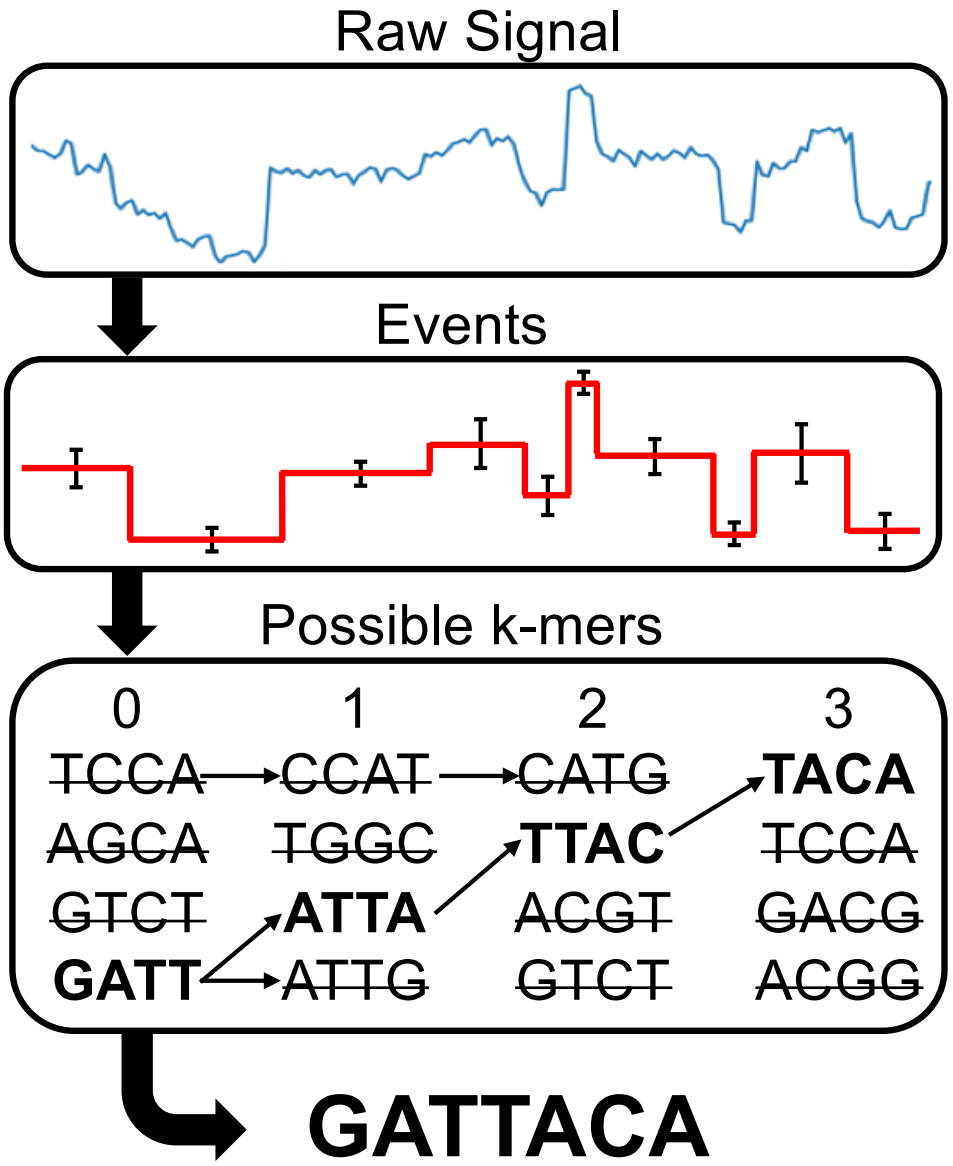


Certain k-mers can be eliminated based on possible transitions

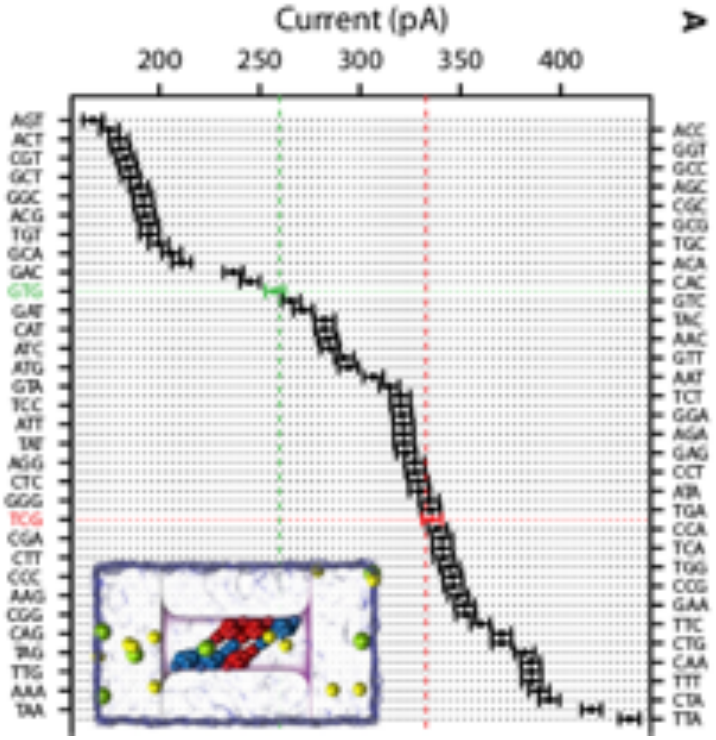


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm
Timp et al. (2012) *Biophysical Journal*

Nanopore Basecalling



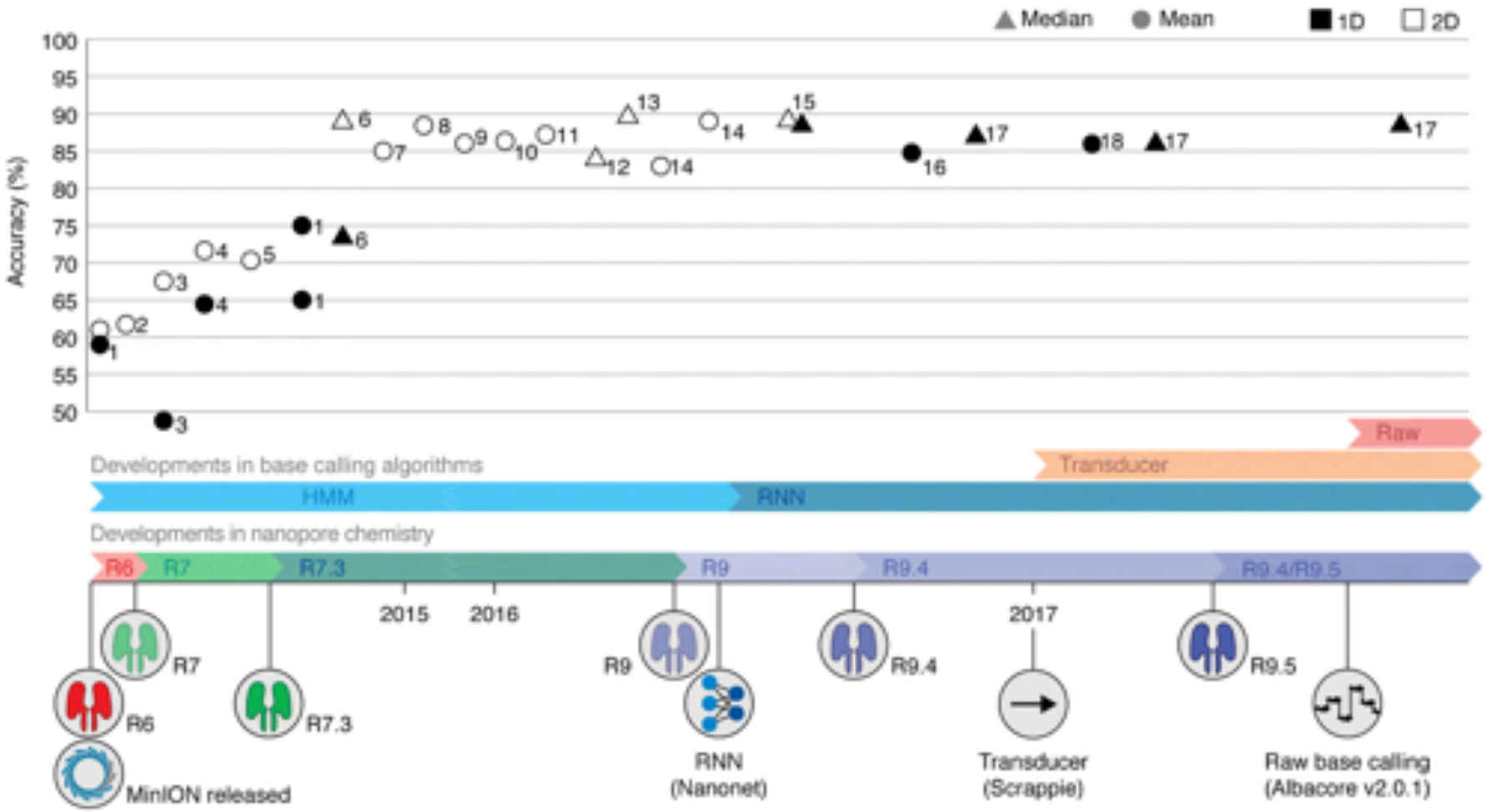
Final sequence determined by most probable k-mers



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

Basecaller/Pore Timeline

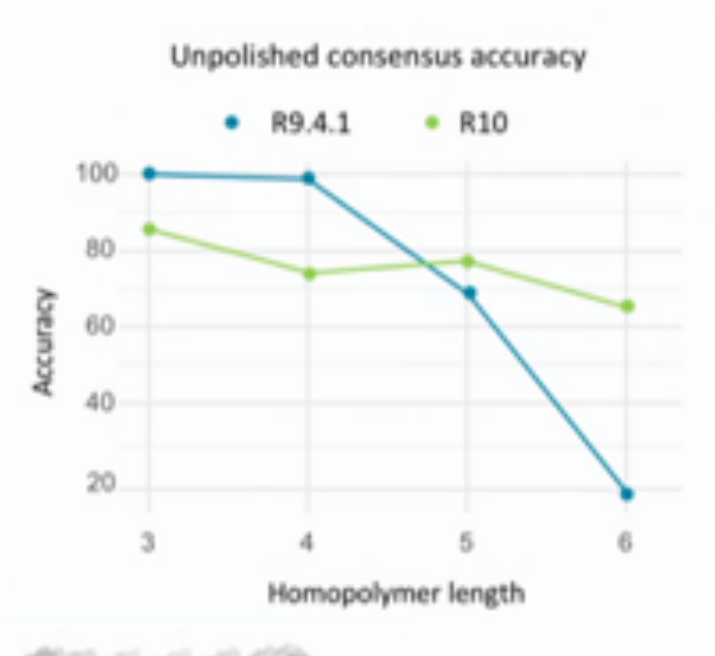
Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy

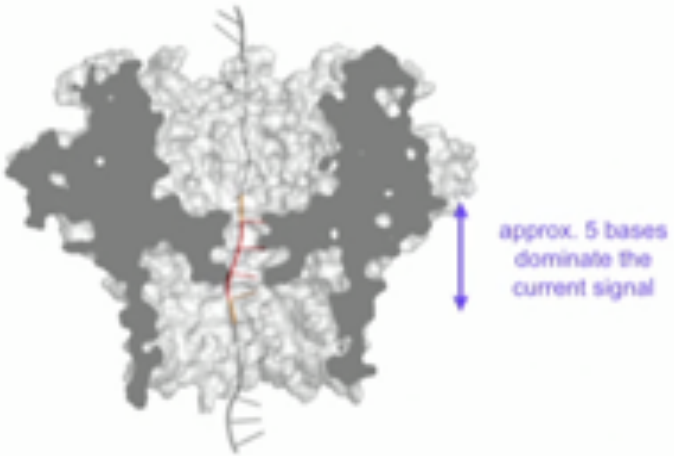


From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy
Rang *et al* (2018) *Genome Biology.* https://doi.org/10.1186/s13059-018-1462-9

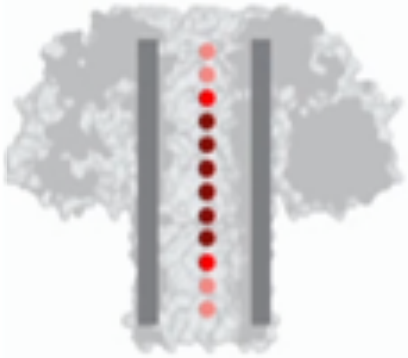
New Pore Chemistries

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

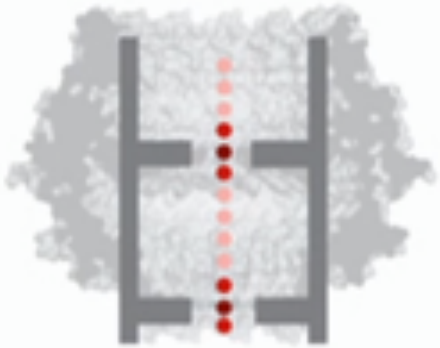




Standard pore chemistry "R9"



Pore with long reader head Lysenin – "R8"



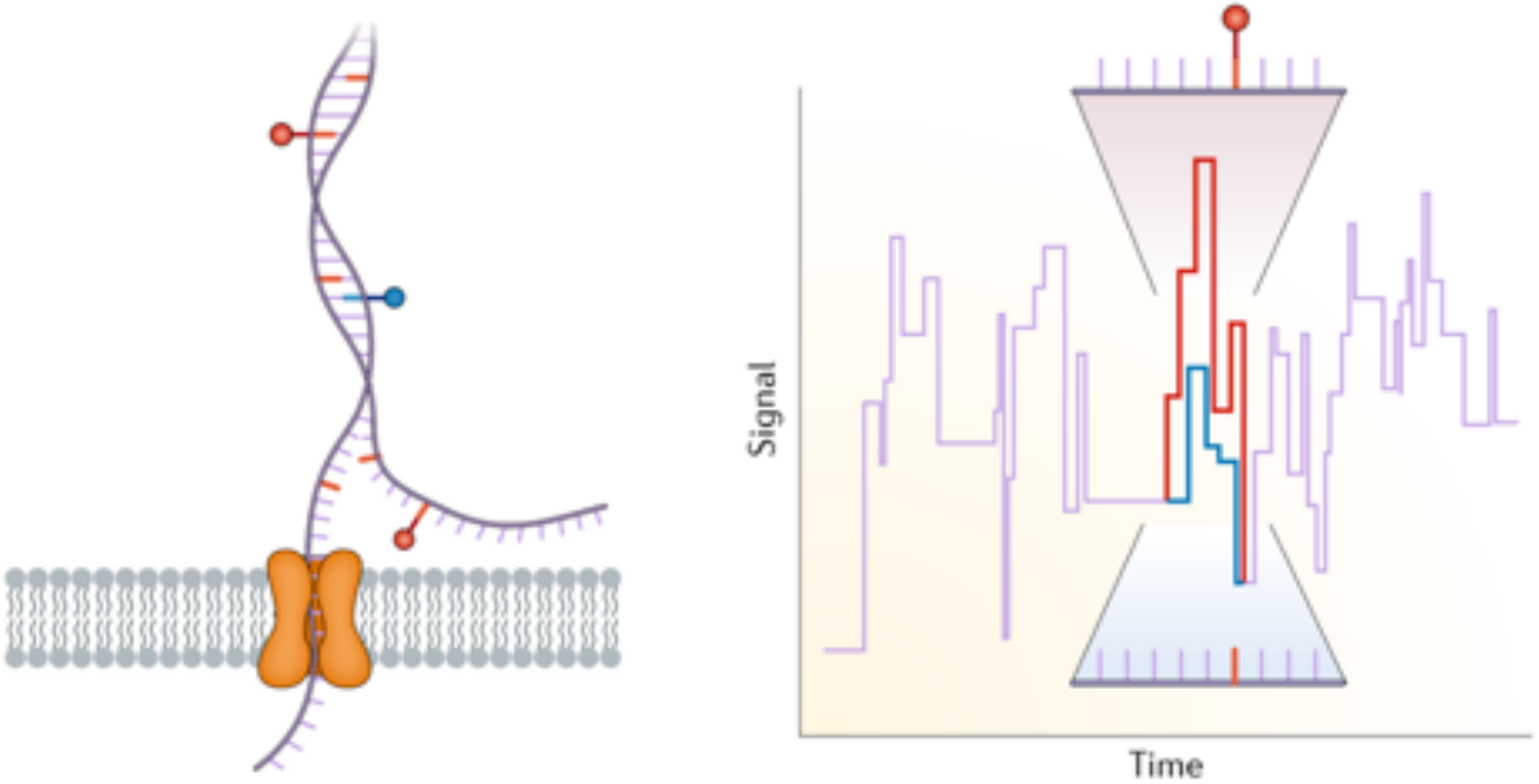
Multiple points of contribution "R10"

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

DNA Modification Detection

Like PacBio, ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



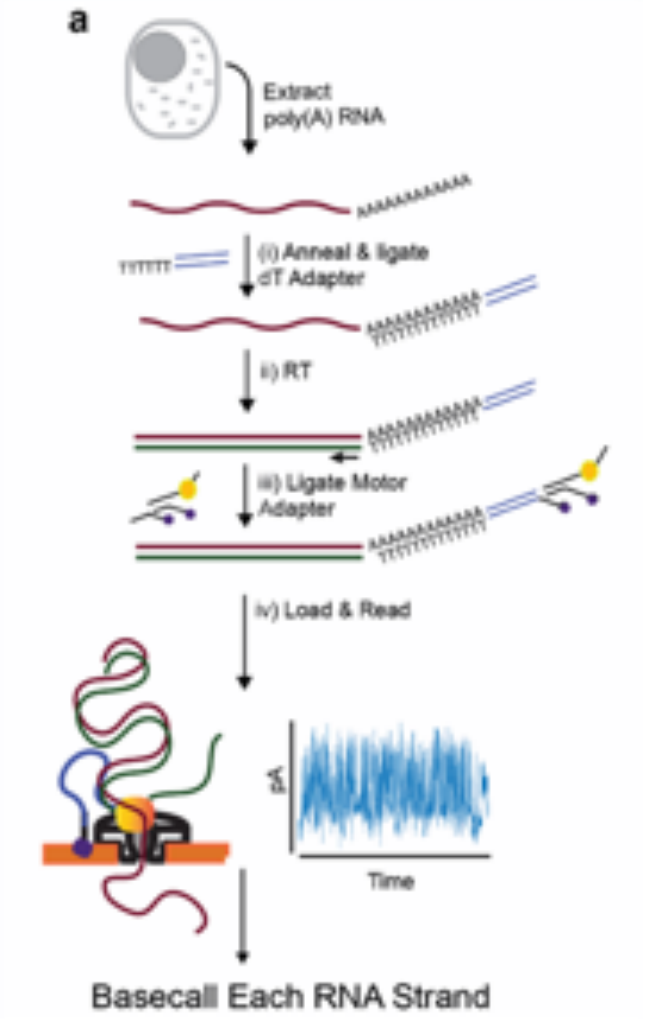
Piercing the dark matter: bioinformatics of long-range sequencing and mapping
Sedlazeck et al. (2018) *Nature Reviews Genetics.* 19:329

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

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

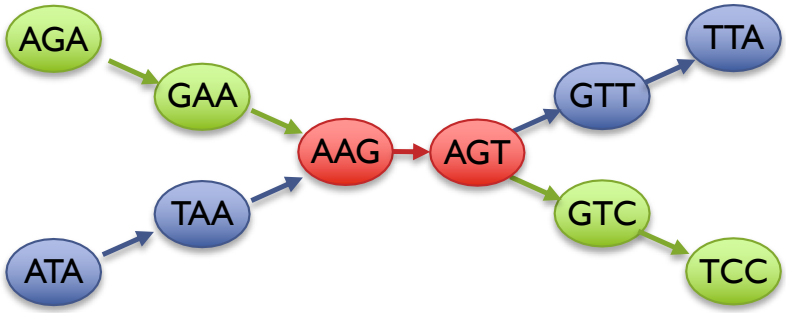
4th April 2019 Callum Cyrus

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...

Two Paradigms for Assembly

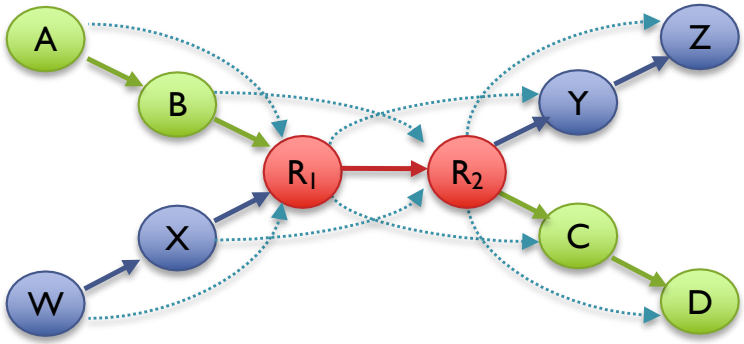
de Bruijn Graph



Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Overlap Graph

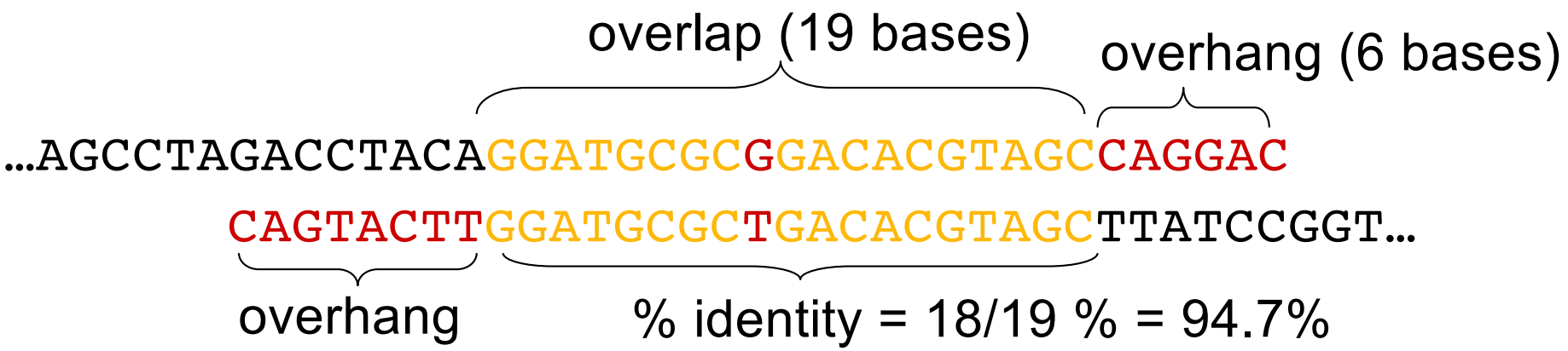


Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

Overlap between two sequences



overlap - region of similarity between regions
overhang - un-aligned ends of the sequences

The assembler screens merges based on:

- length of overlap
- % identity in overlap region
- maximum overhang size.

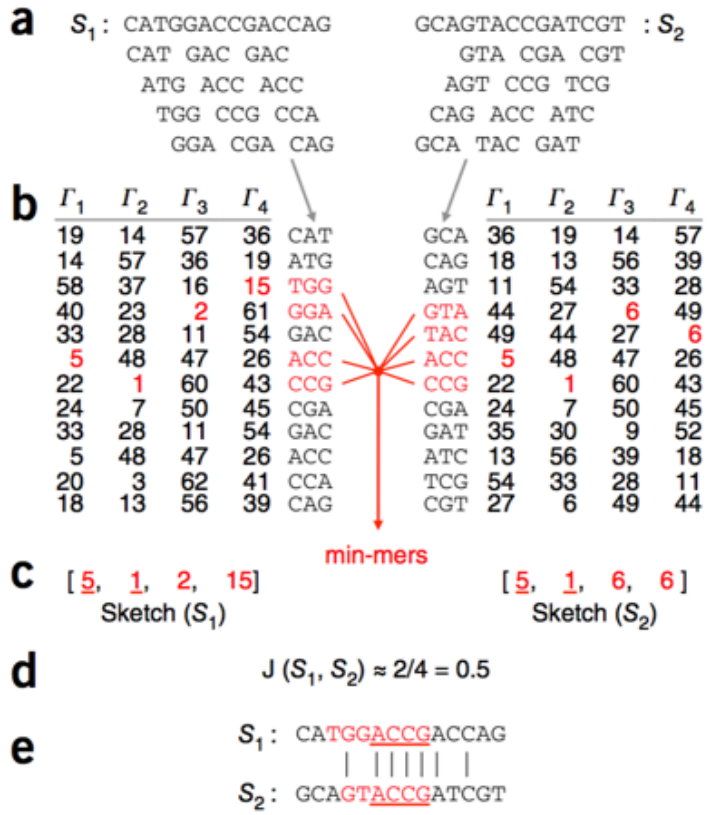
[How do we compute the overlap?]

[Do we really want to do all-vs-all?]

Very fast approximate overlapping

Maybe we don't need to compute the exact identity of the overlap region, just approximate it

- If two reads overlap, they should share many of the same kmers: Their Jaccard coefficient should be high: |intersection| / |union|
- But tracking all of the kmers for a read is a lot of overhead
- Instead, compare the "sketch" of the reads: a small fraction of kmers carefully chosen
- LSH: Find the sketch by applying N hash functions to the kmers, and keeping the minimum hash values reported from each (N=4 in example)
- This forms a nice "random" sample of the reads, and the Jaccard coefficient is a good approximation of the sequence similarity



Assembling large genomes with single-molecule sequencing and locality-sensitive hashing
Berlin et al (2015) *Nature Biotechnology*

Unitigging: Pruning the Overlap Graph

The overlap graph has many redundant edges:

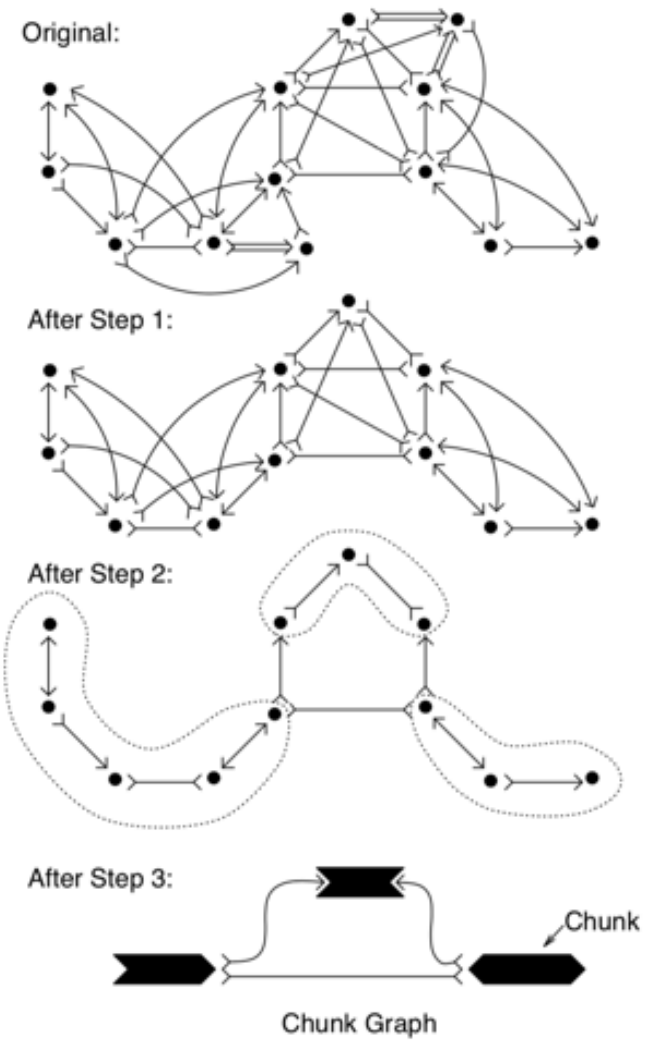
- If the average coverage is D, we should expect D overlaps at the beginning of the read, and D at the end

Transform the graph to simplify the assembly problem (without changing the valid solutions):

1. ***Contained reads removal:*** Short reads that are substrings of longer reads don't advance the assembly, remove those nodes and all of the edges

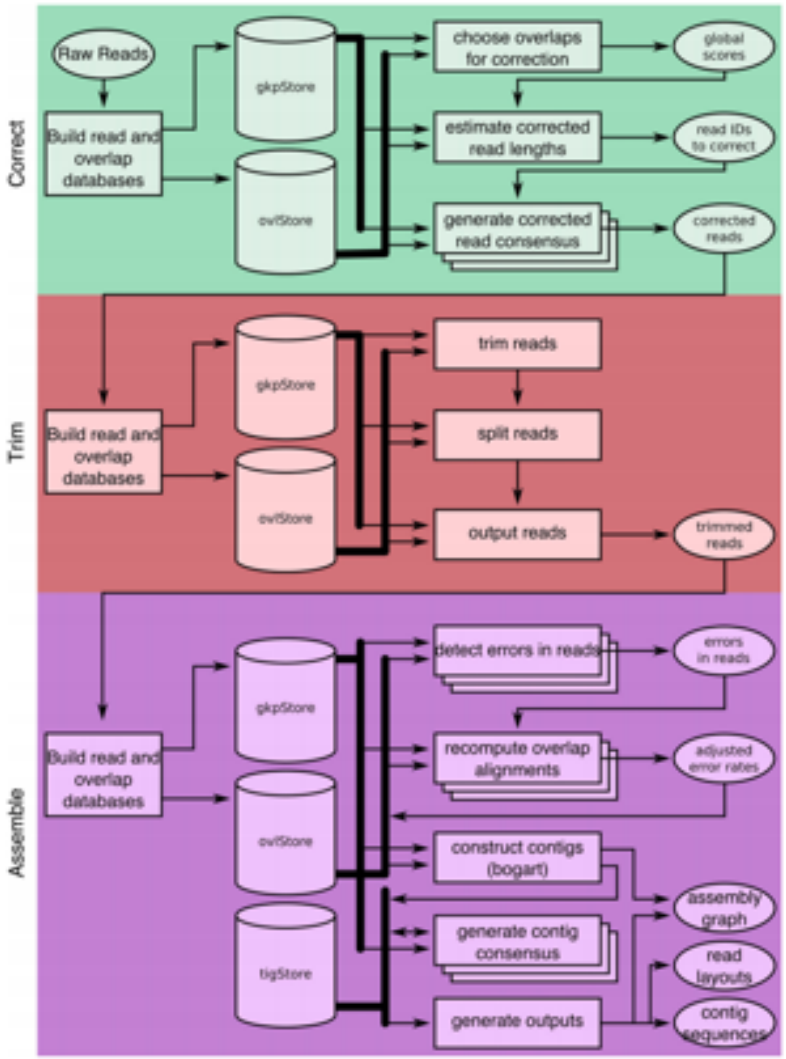
2. ***Transitive edge removal:*** If A -> B, and B->C, remove the transitive edge A->C

3. ***"Chunkification":*** Linear subgraphs define uniquely assemblible segments: "unitigs"



Towards Simplifying and Accurately Formulating Fragment Assembly
Myers (1995) *J Comput Biol.* Summer;2(2):275-90.

Canu Workflow



Three rounds of analysis:

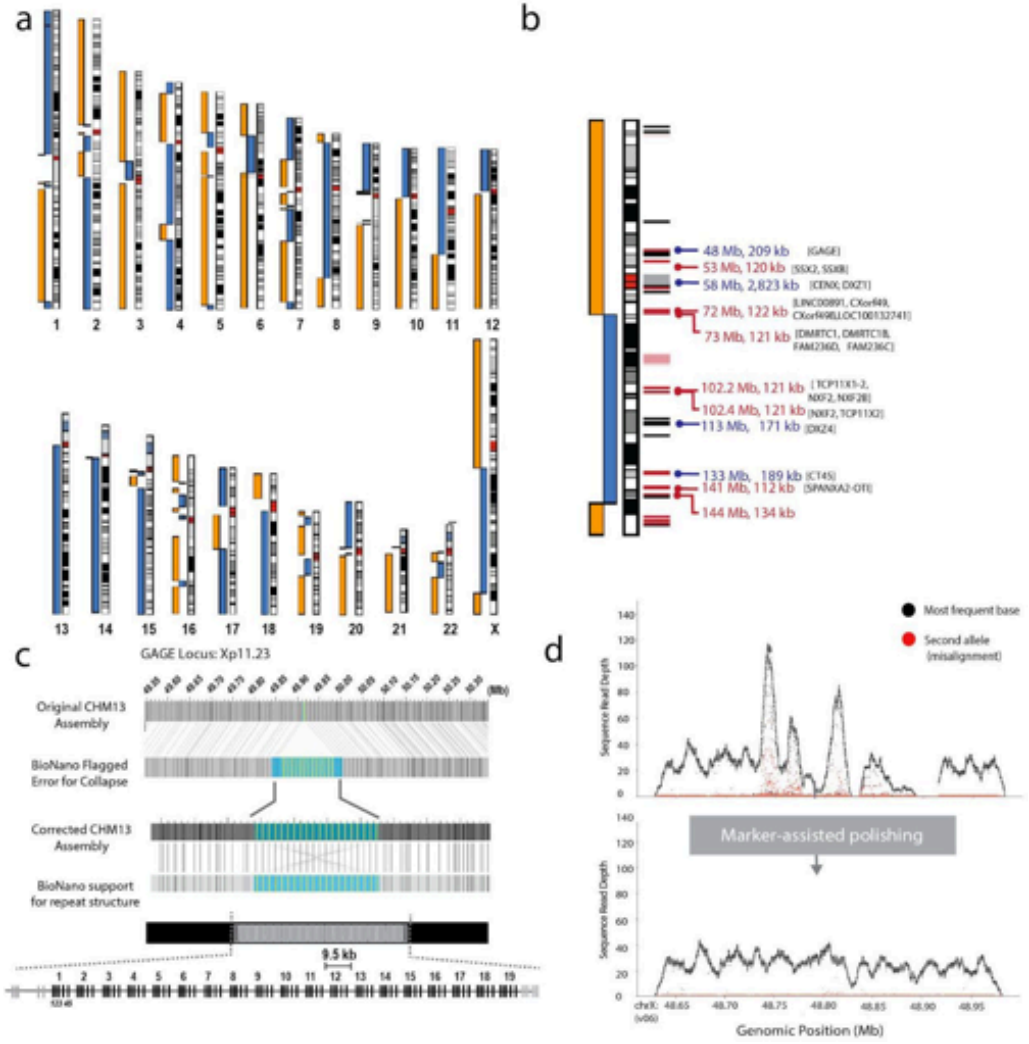
1. ***Error Correction:*** Use MHAP to overlap the reads, then compute a mini assembly centered around each read of good overlaps to error correct

2. ***Trim:*** Use MHAP to recompute overlaps to find regions that are not well supported and discard

3. ***Unitigging:*** Use Dynamic Programming to carefully overlap the error corrected reads, construct overlap graph, and then "unitig" those overlaps to build the contigs

Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation
Koren et al (2017) *Genome Research*

First Telomere-to-Telomere Human Chromosome



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

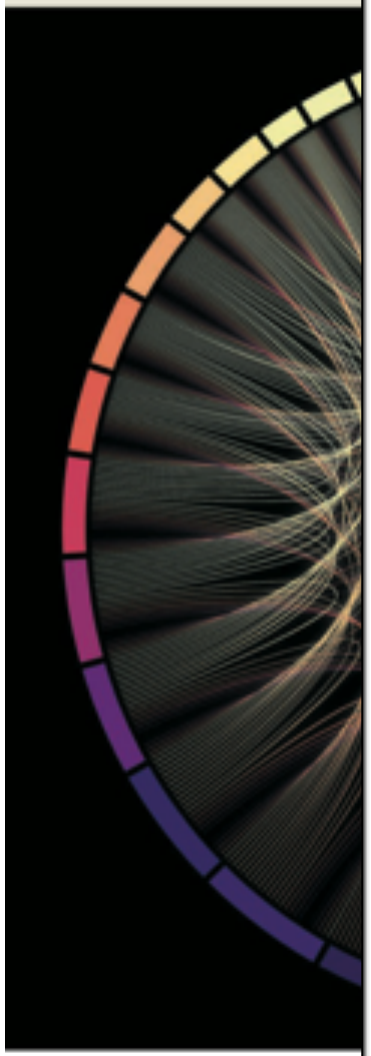
First Telomere-to-Telomere Human Genome

"We estimate that the quality of our polished assembly approaches Q70 (1 error per 10 million bases), with no known structural errors."

The (near) complete sequence of a human genome
https://genomeinformatics.github.io/CHM13v1/

www.nature.com/collections/g

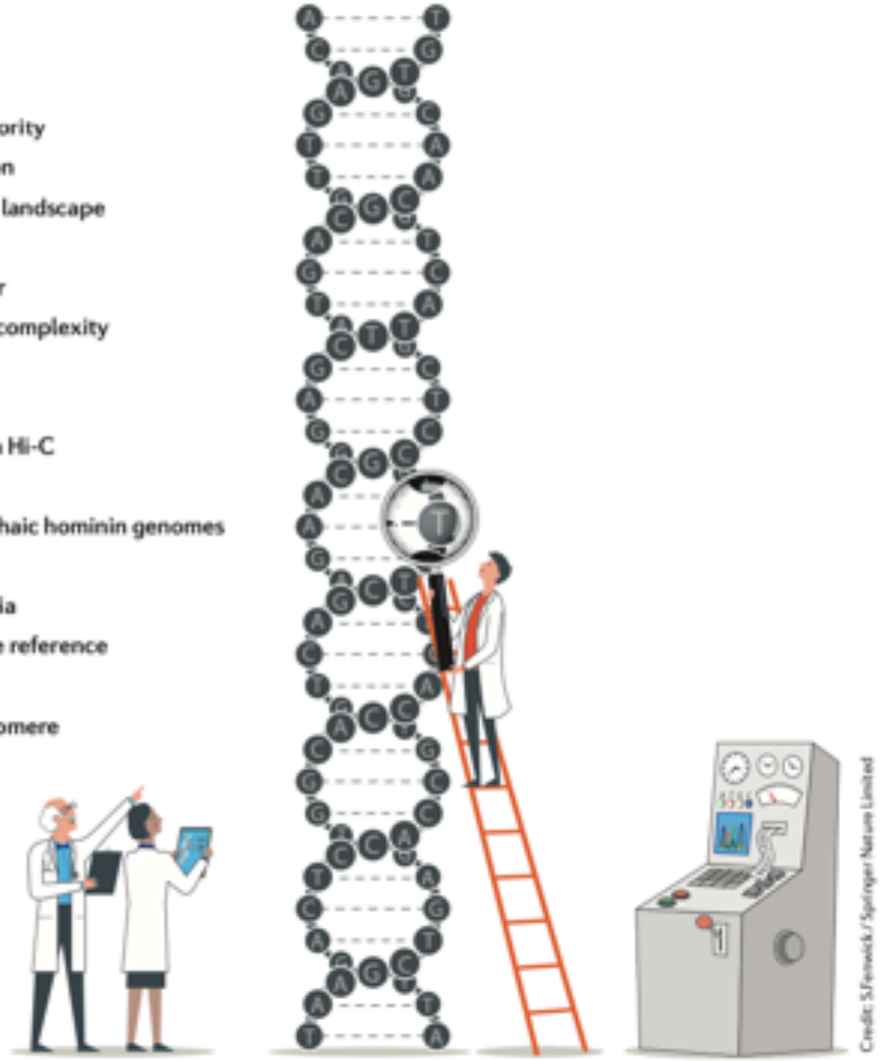
nature



Produced by:
Nature, *Nature Genetics* and
Nature Revie

Genomic sequencing

MILESTONES



https://www.nature.com/immersive/d42859-020-00099-0/index.html