

# The human genome

Michael Schatz

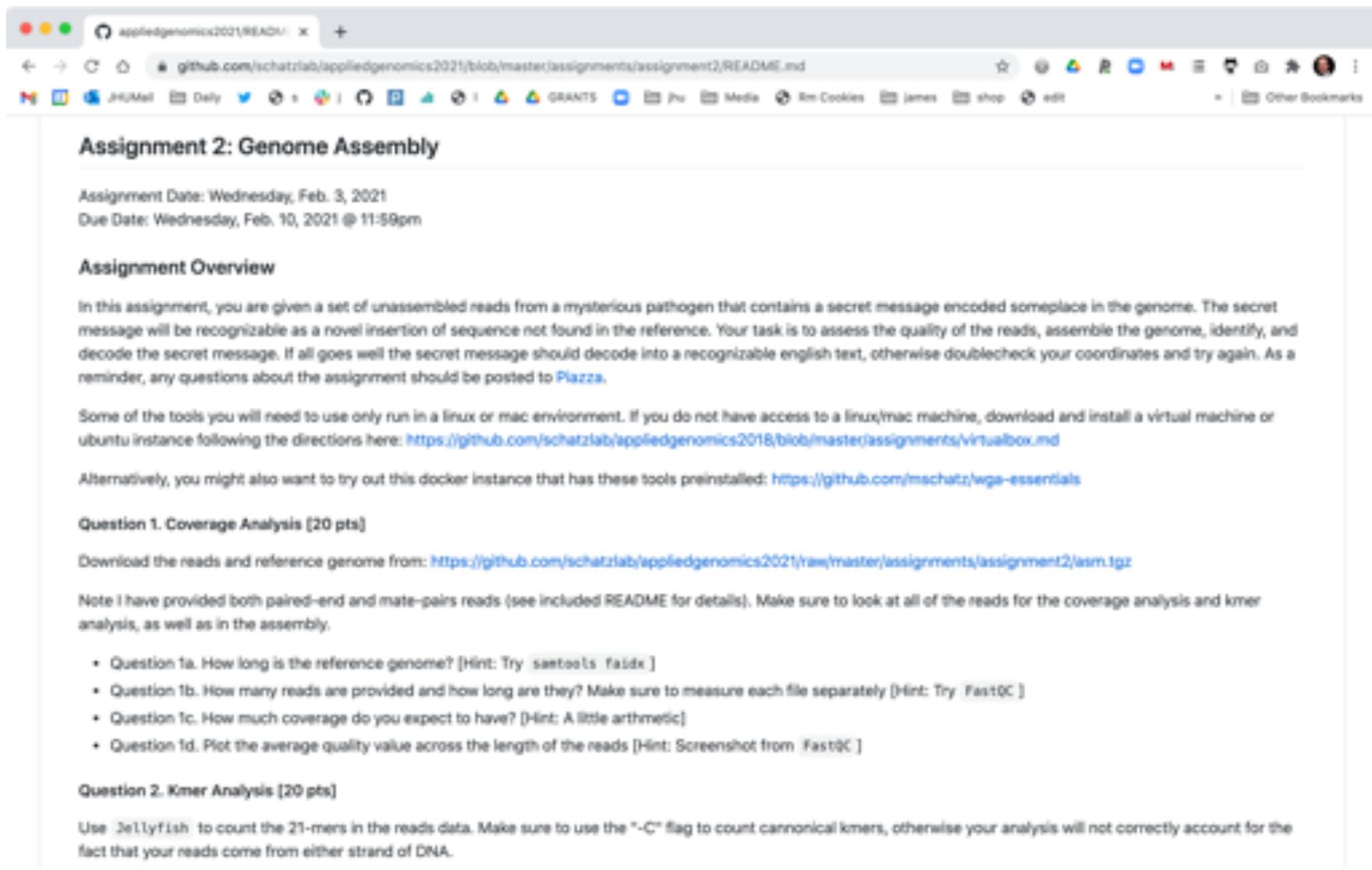
Feb 10, 2021

Lecture 6: Computational Biomedical Research



# Assignment 2: Genome Assembly

## Due Feb 10 @ 11:59pm



A screenshot of a web browser window displaying the assignment details. The title bar shows the URL: <https://github.com/schatzlab/appliedgenomics2021/blob/master/assignments/assignment2/README.md>. The page content includes:

### 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 arithmetic]
- 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 canonical kmers, otherwise your analysis will not correctly account for the fact that your reads come from either strand of DNA.

# Assignment 3: de Bruijn Graphs

## Due Feb 17 @ 11:59pm

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]

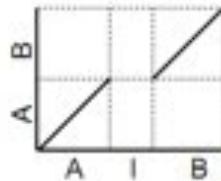
- Q1a. How many 100bp reads are needed to sequence a 1Mbp genome to 5x coverage?
- Q1b. In the language of your choice, simulate sequencing 5x coverage of a 1Mbp genome with 100bp reads and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just 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 probability 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

# Part I: Recap

# SV Types

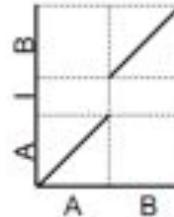
Insertion into Reference

R: AIB  
Q: AB



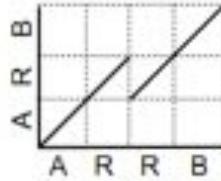
Insertion into Query

R: AB  
Q: AIB



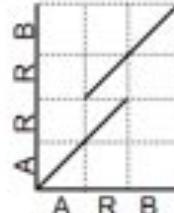
Collapse Query

R: ARRB  
Q: ARB



Collapse Reference

R: ARB  
Q: ARRB



Collapse Query w/ Insertion

R: ARIRB  
Q: ARB

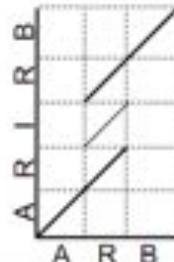
Exact tandem alignment if I=R



Collapse Reference w/ Insertion

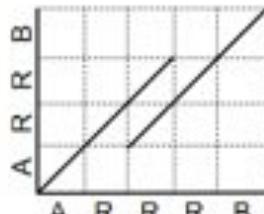
R: ARB  
Q: ARIRB

Exact tandem alignment if I=R



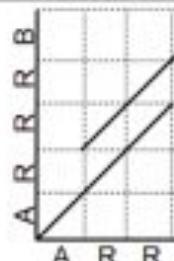
Collapse Query

R: ARRRB  
Q: ARRB



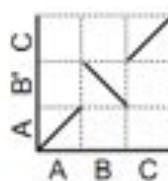
Collapse Reference

R: ARRB  
Q: ARRRB



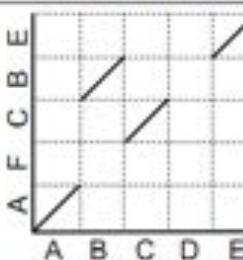
Inversion

R: ABC  
Q: AB'C



Rearrangement w/ Disagreement

R: ABCDE  
Q: AFCBE



- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints

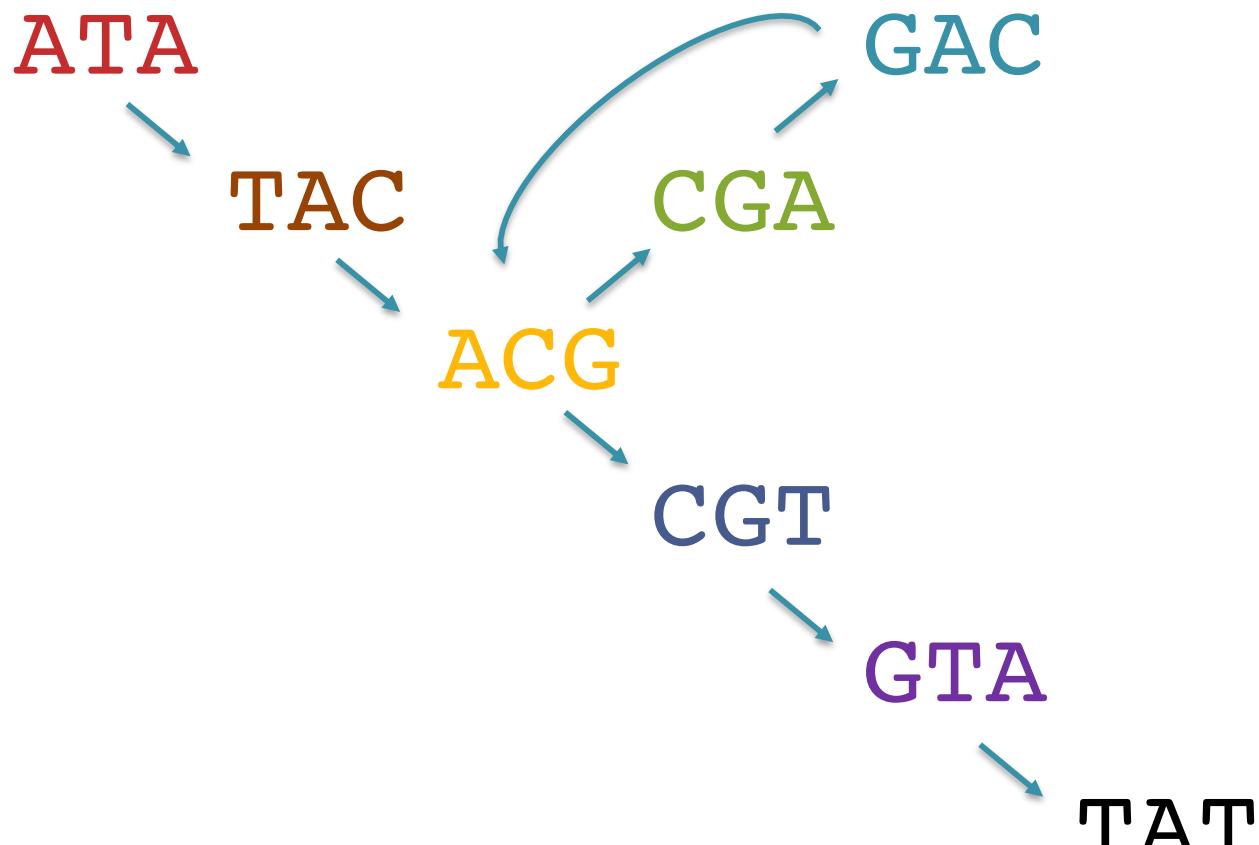
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

[http://mummer.sf.net/manual/  
AlignmentTypes.pdf](http://mummer.sf.net/manual/AlignmentTypes.pdf)

# Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~  
~~ACGT~~  
~~ATAC~~  
~~CGAC~~  
~~CGTA~~  
~~GACG~~  
~~GTAT~~  
~~TACG~~



Note: there is no edge from ATA to TAT

ATACGACGTAT

# K-mer counting

## Kmer-ize

Read 1: GATTACA => GAT, ATT, TTA, TAC, ACA  
Read 2: TACAGAG => TAC, ACA, CAG, AGA, GAG  
Read 3: TTACAGA => TTA, TAC, ACA, CAG, AGA



GAT	ACA	ACA: 3
ATT	ACA	
TTA	ACA	
TAC	AGA	AGA: 2
ACA	AGA	
TAC	ATT	ATT: 1
ACA	CAG	CAG: 2
CAG	CAG	
AGA	GAG	GAG: 1
GAG	GAT	GAT: 1
TTA	TAC	TAC: 3
TAC	TAC	
ACA	TAC	
CAG	TTA	TTA: 2
AGA	TTA	

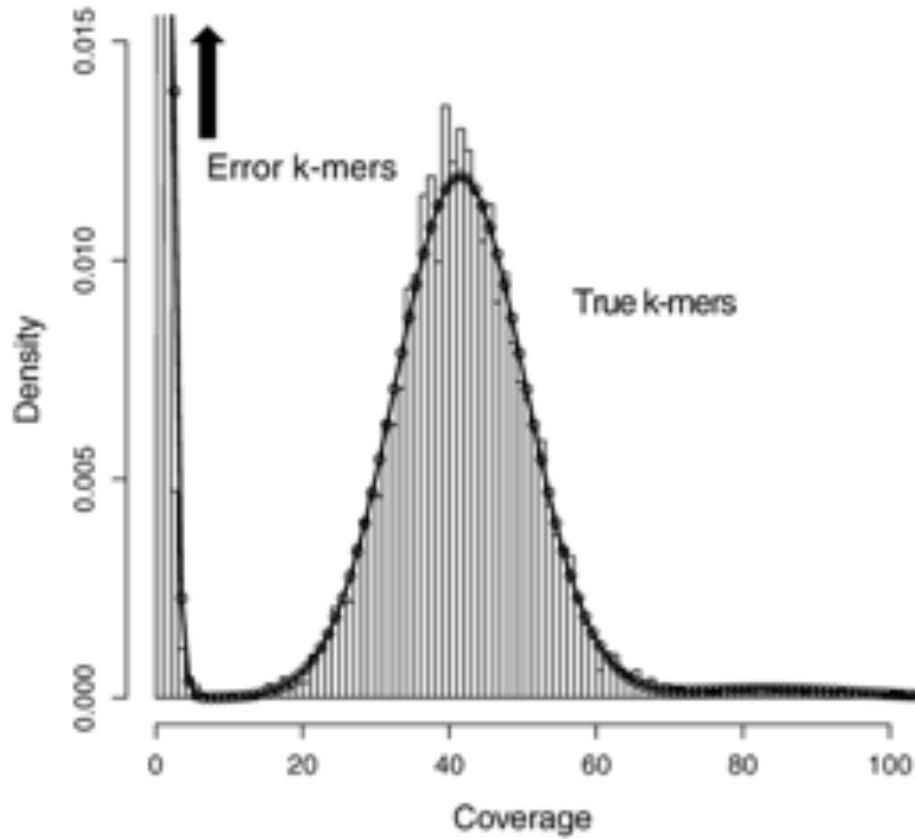
3 kmers occur 1x  
3 kmers occur 2x  
2 kmers occur 3x



From read k-mers alone, can learn something about how frequently different sequences occur (aka coverage)

Fast to compute even over huge datasets

# K-mer counting in real genomes

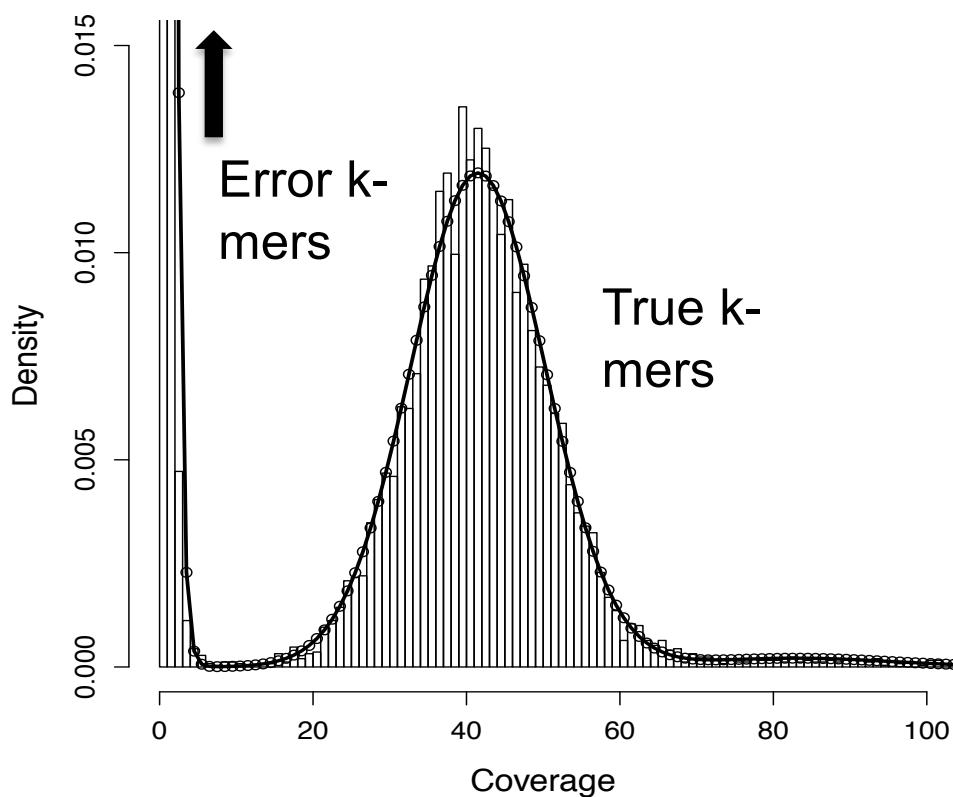


- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors in the reads
- There are also kmers that occur many times (>>70 times). These are repeats in the genome

# Error Correction with Quake

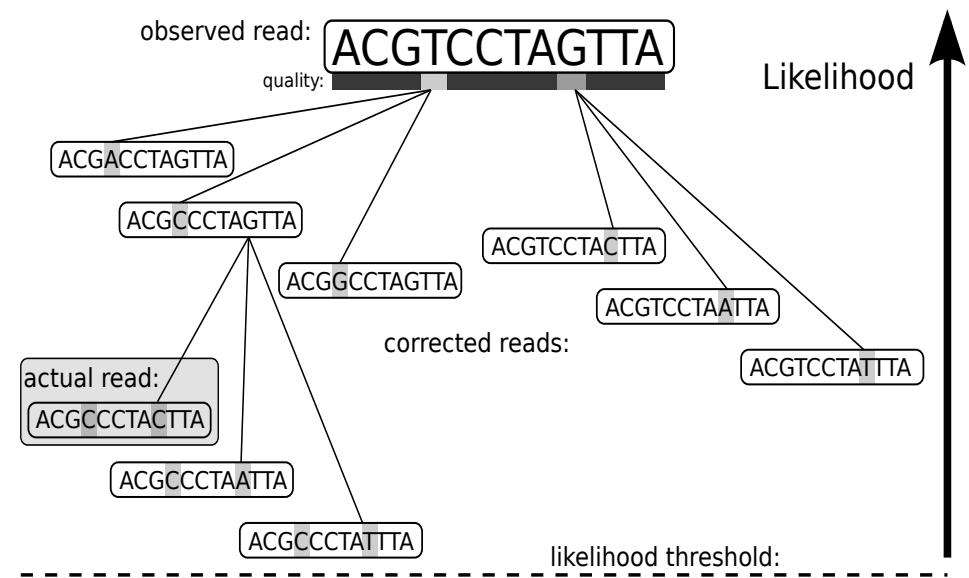
## 1. Count all “Q-mers” in reads

- Fit coverage distribution to mixture model of errors and regular coverage
- Automatically determines threshold for trusted k-mers



## 2. Correction Algorithm

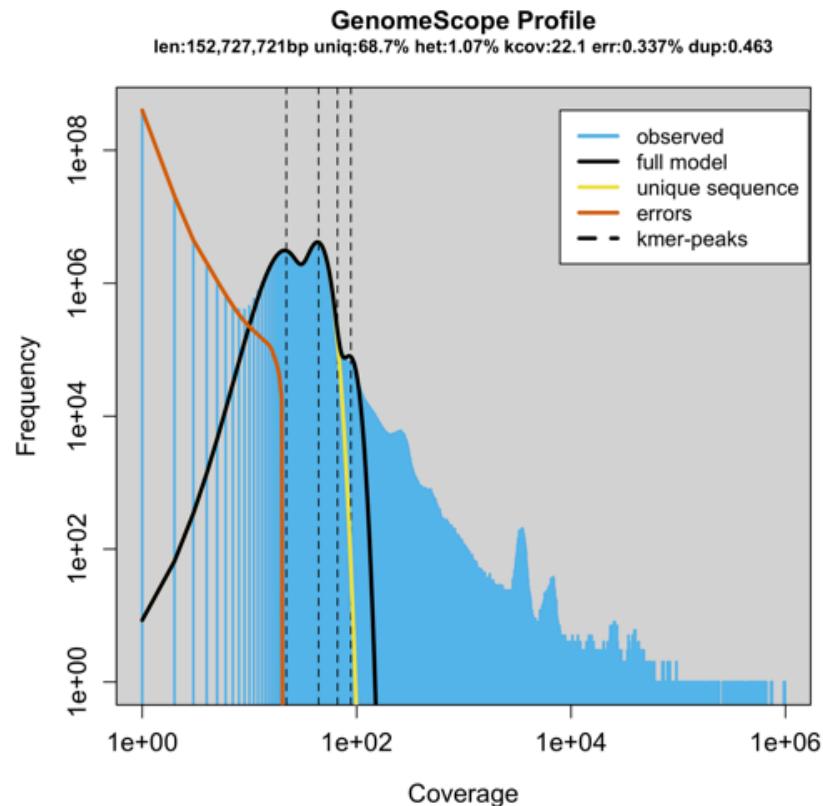
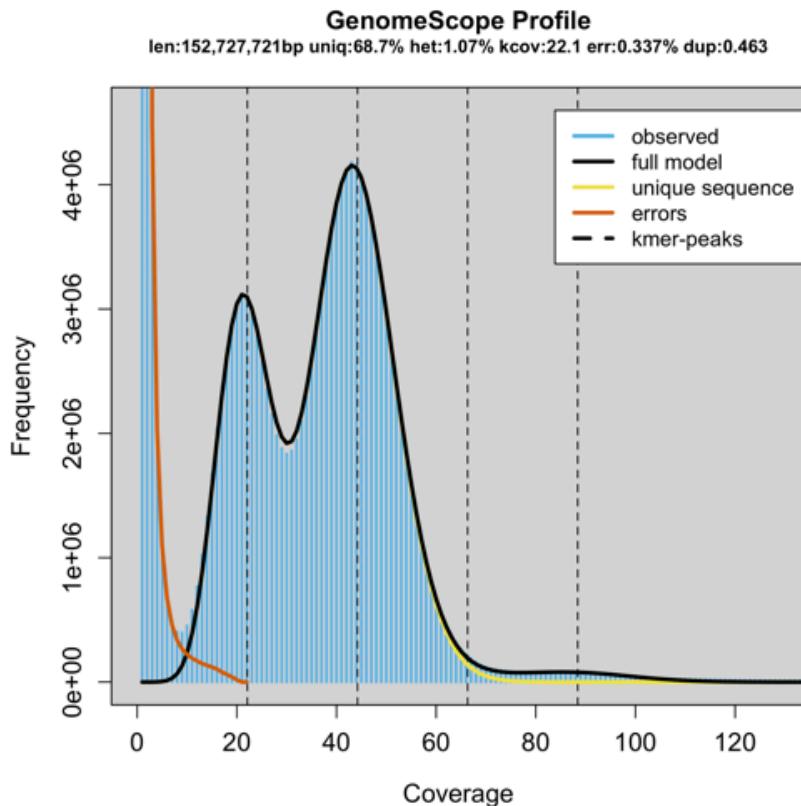
- Considers editing erroneous kmers into trusted kmers in decreasing likelihood
- Includes quality values, nucleotide/nucleotide substitution rate



**Quake: quality-aware detection and correction of sequencing reads.**  
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

# GenomeScope: Fast genome analysis from short reads

<http://genomescope.org>

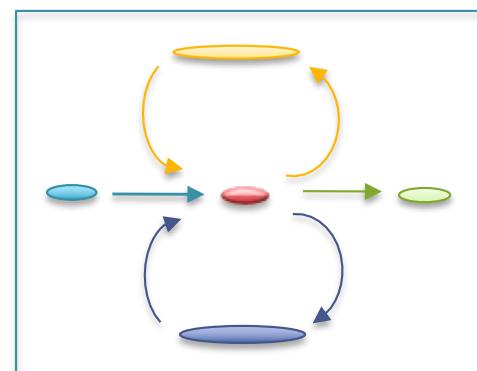
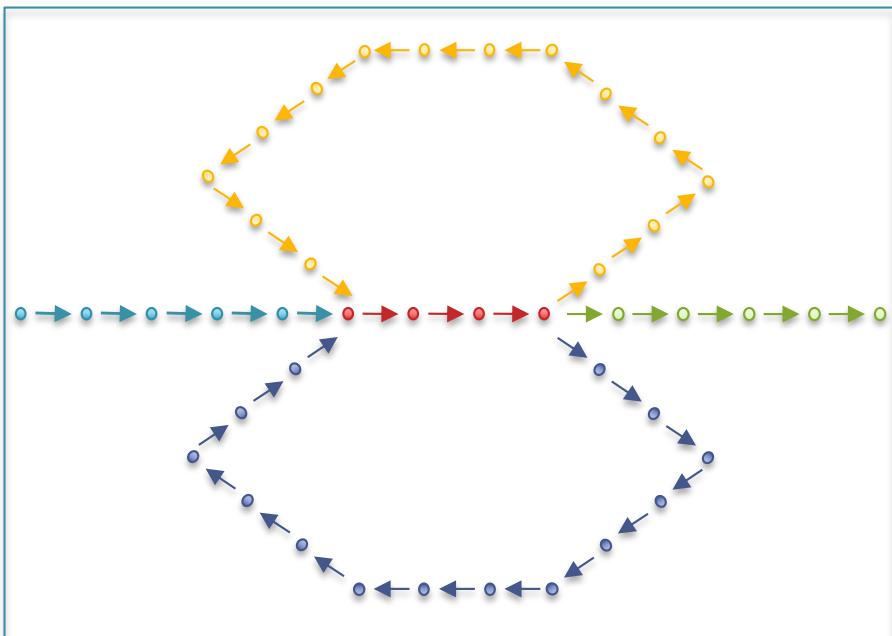


- Theoretical model agrees well with published results:
  - Rate of heterozygosity is higher than reported by other approaches but likely correct.
  - Genome size of plants inflated by organelle sequences (exclude very high freq. kmers)

Vulture, GW\*, Sedlazeck FJ\*, et al. (2017) *Bioinformatics*  
Ranallo-Benavidez, TR. et al. (2020) *Nature Communication*

# Unitigging / Unipathing

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



Why do contigs end?

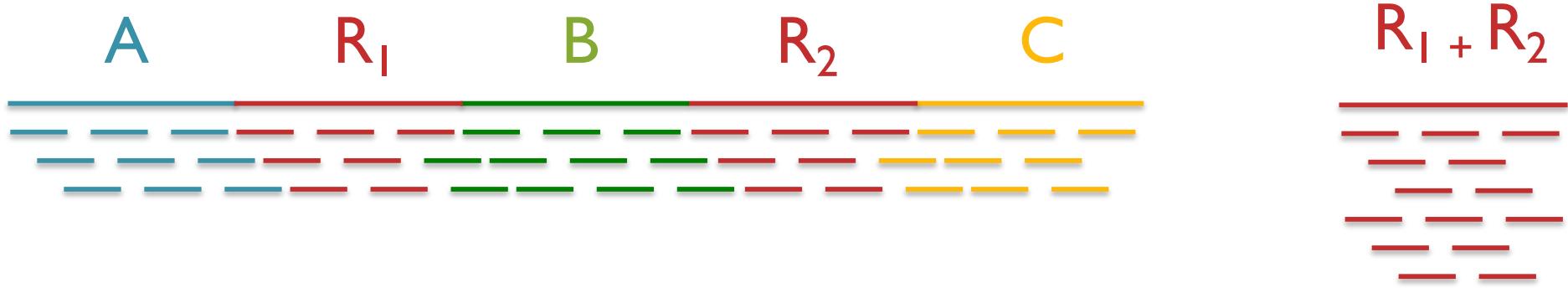
- (1) End of chromosome! ☺, (2) lack of coverage, (3) errors, (4) heterozygosity and (5) repeats

# Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1 b_2 \dots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	<i>Alu</i> sequence (~280 bp) <i>Mariner</i> elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp

# Repeats and Coverage Statistics



- If  $n$  reads are a uniform random sample of the genome of length  $G$ , we expect  $k = n \Delta/G$  reads to start in a region of length  $\Delta$ .
  - If we see many more reads than  $k$  (if the arrival rate is  $> A$ ) , it is likely to be a collapsed repeat

$$\Pr(X - \text{copy}) = \binom{n}{k} \left( \frac{X\Delta}{G} \right)^k \left( \frac{G - X\Delta}{G} \right)^{n-k}$$

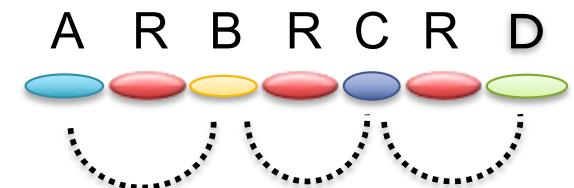
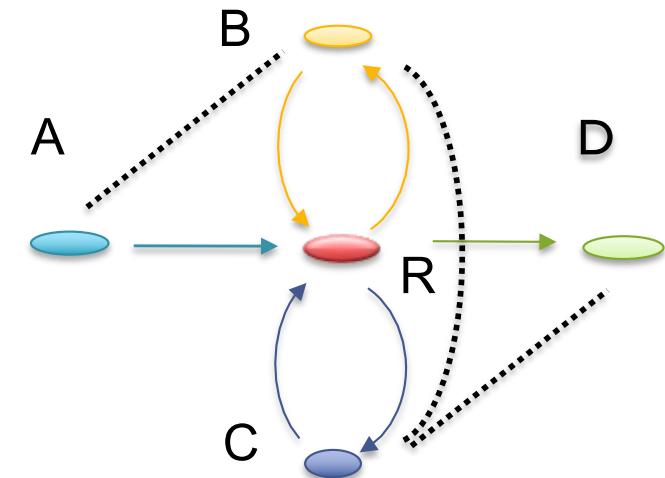
$$A(\Delta, k) = \ln \left( \frac{\Pr(1 - \text{copy})}{\Pr(2 - \text{copy})} \right) = \ln \left( \frac{\frac{(\Delta n/G)^k e^{-\Delta n}}{k!}}{\frac{(2\Delta n/G)^k e^{-2\Delta n}}{k!}} \right) = \frac{n\Delta}{G} - k \ln 2$$

**The fragment assembly string graph**

Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

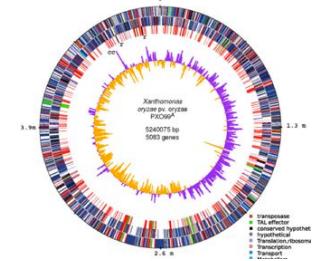
# Scaffolding

- 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



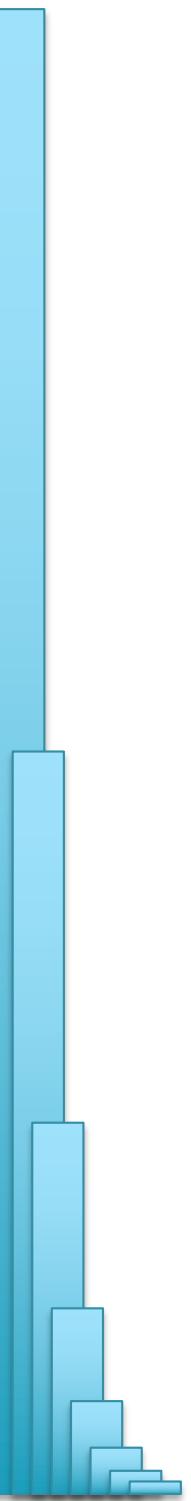
Why do scaffolds end?

# 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



## 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.
- $40 \text{ trillion} * 4 \text{ meters} = 160 \text{ trillion meters}$ .
- $160 \text{ trillion meters} / 1609.34 = 99,750,623,441 \text{ miles}$
- $99,750,623,441 / 92,960,000 = 1,073.05 \text{ trips to the sun.}$

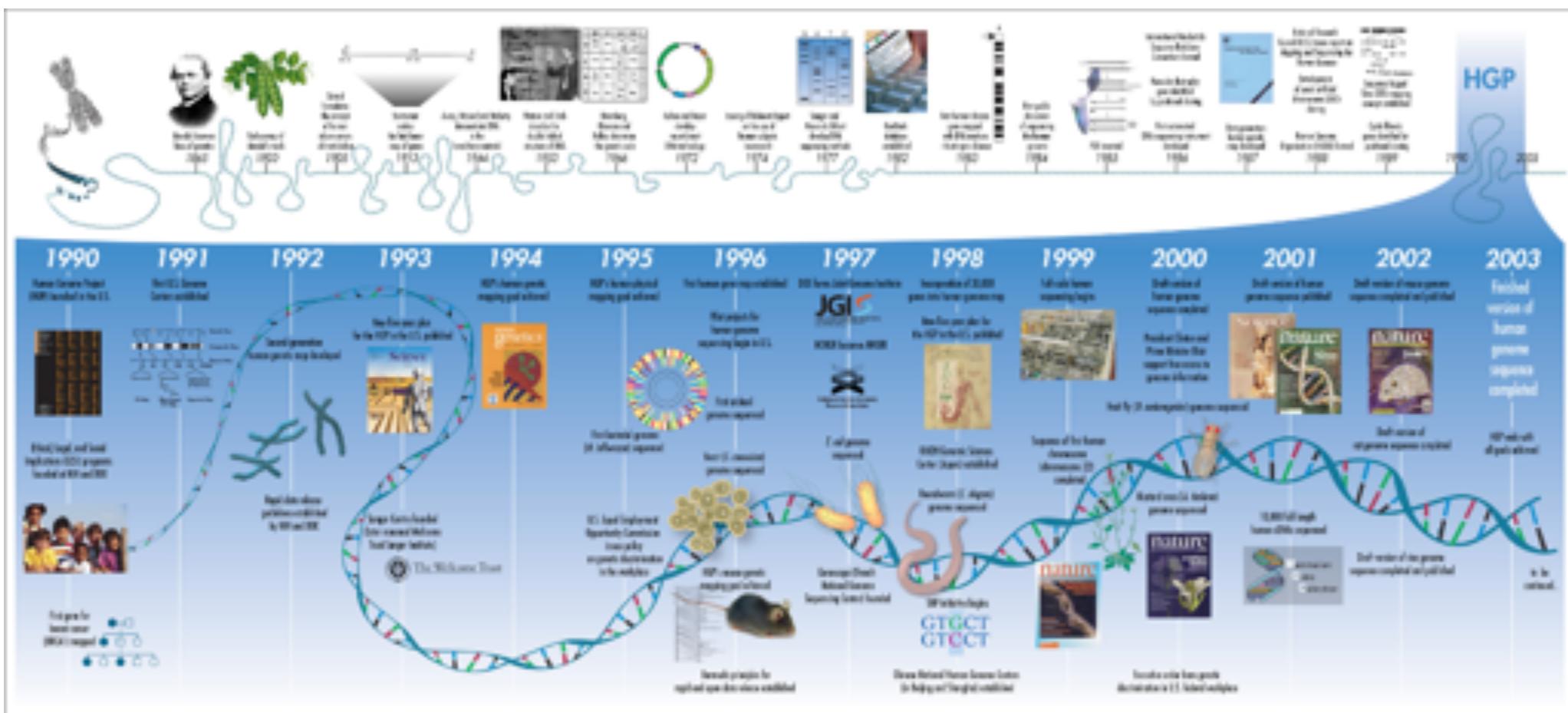
A typical cell replicates about 100 times

160 trillion meters x 100 =

**1.69123746 light years**

[More info](#)

# History of the Human Genome Project



# The reference human genome



*“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”*

*Bill Clinton  
June 26, 2000*

# The reference human genome



*“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”*

*Bill Clinton  
June 26, 2000*

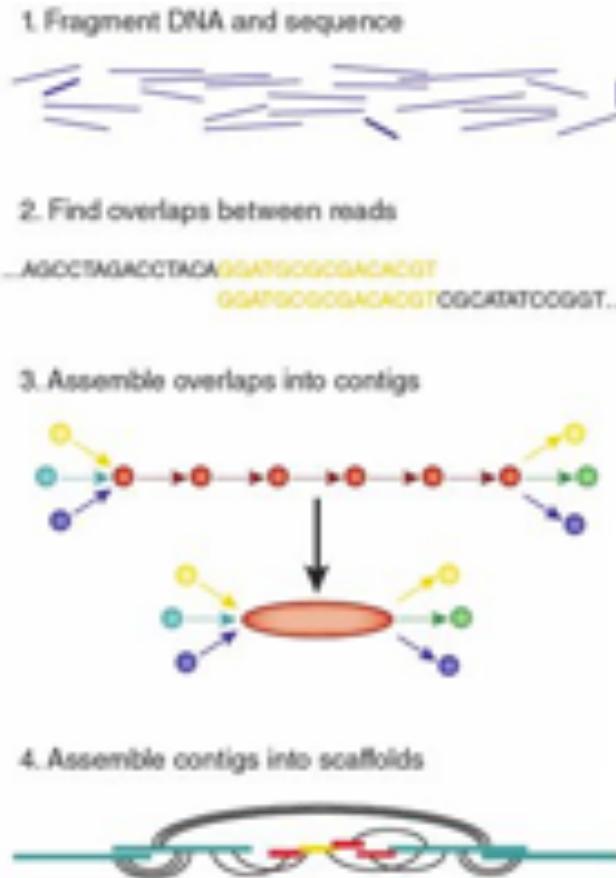


**The Sequence of the Human Genome**  
Venter et al.  
Science 291, pp 1304-1351 (2001)



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

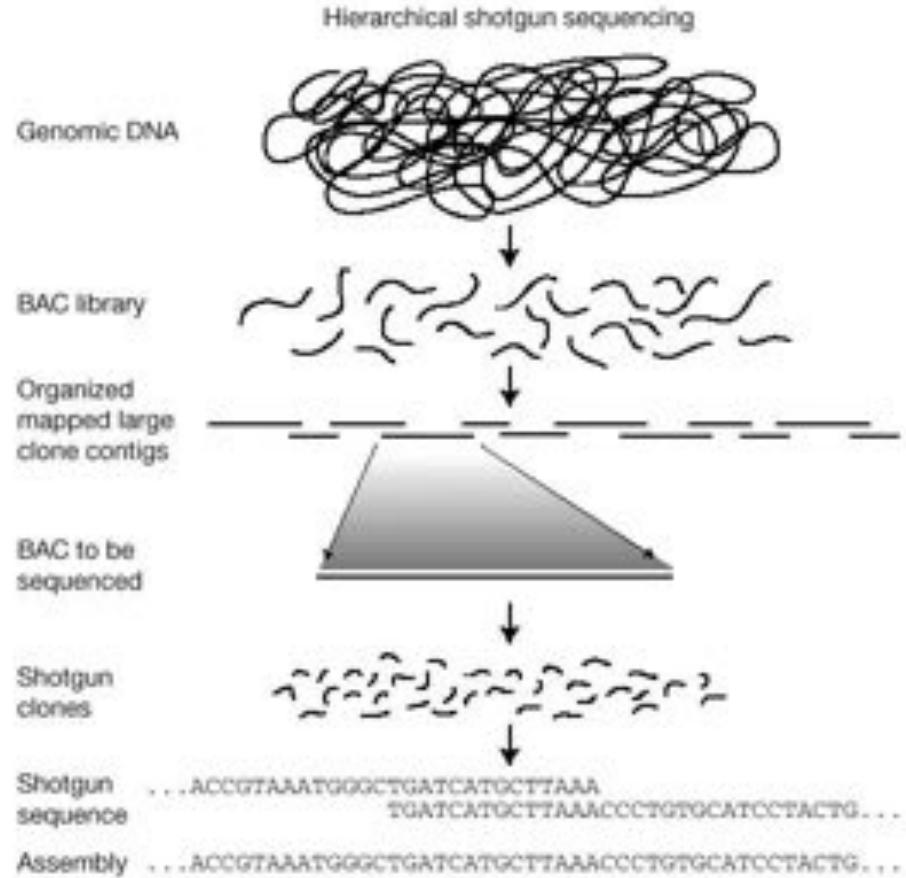
# Two Human Genomes?



# The Sequence of the Human Genome

Venter et al.  
Science 291, pp 1304-1351 (2001)

(Figure from Baker (2012) Nature Methods)



# **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

ople. But the very nature of government creates a mind set that insinuates increase their authority, always at the expense of the people," Parlato said. "The government has forgotten that it's not of the people," Parlato added, acting more like it's the master." So and the Lapps share an abiding non-violent civil disobedience.

"We insist on being respectful in our resistance," Barbara Lyn Lapp said. "If we claim to care about our rights, we must protest government instead of violence," she said.

Violence has to be the watchword, said, calling civil disobedience the heart of the violent militia movement. Non-violence can serve as an anti-government oppression, he added. "If law is unjust or you're given an order without 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 CANCER INSTITUTE**

To receive information please contact the  
Clinical Genetics Service  
643-5730 (7:00 am - 10:00 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 CANCER INSTITUTE**

To receive information please contact the  
Clinical Genetics Service  
643-5730 (7:00 am - 10:00 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

ople. But the very nature of government creates a mind set that inspires increase their authority, always at the expense of the people."

"The government has forgotten that it's the servant of the people," Parlato added. "It's acting more like it's the master." Parlato and the Lapps share an abiding non-violent civil disobedience.

"We must insist on being respectful in our acts of resistance," Barbara Lyn Lapp said. "But if we claim to care about our rights, we must protest government instead of violence."

Violence has to be the watchword, said, calling civil disobedience the act of the violent militia movement. Non-violence can serve as an anti-government oppression, he added.

"If the law is unjust or you're given an order without 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 genome) 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.*

For more information please contact the  
Clinical Genetics Service  
843-7520 (7:00 am - 10:00 pm)  
March 24 - 26, 1997

**ROSWELL PARK CANCER INSTITUTE**



*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 government creates a mind set that inspires increase their authority, always at the expense of the people," Parlato added. "The government has forgotten that it's not of the people," Parlato added, acting more like it's the master." So and the Lapps share an abiding non-violent civil disobedience.

"We insist on being respectful in our resistance," Barbara Lyn Lapp said. "If we claim to care about our rights, we must protest government instead of violence," she said.

"Non-violence has to be the watchword, said, calling civil disobedience the heart of the violent militia movement. Non-violence can serve as an anti-government oppression, he added. "If law is unjust or you're given an order without 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 genome) 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.

To receive information please contact the Clinical Genetics Service 843-5731 (7:00 am - 10:00 pm)  
March 24 - 26, 1997

**ROSWELL PARK CANCER INSTITUTE**

### 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.

Pieter de Jong, RPCI

### 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?

The screenshot shows the homepage of the **nature methods** journal. At the top right, a welcome message says "Welcome back: Michael Schatz" with a "Logout" link and a shopping cart icon. Below the header, there's a search bar with "Search" and "Advanced search" buttons.

The main content area features a sidebar on the left with links to "Journal content" (Journal home, Advance online publication, Current issue, Archive, Focuses and Supplements, Methagora blog, Method of the Year 2016, Multimedia, Press releases), "Journal information" (Guide to authors, Reporting checklist, Online submission, Subscribe, Permissions, For referees, Contact the journal, About this site), and "Nature Research services" (Authors & Referees, Advertising).

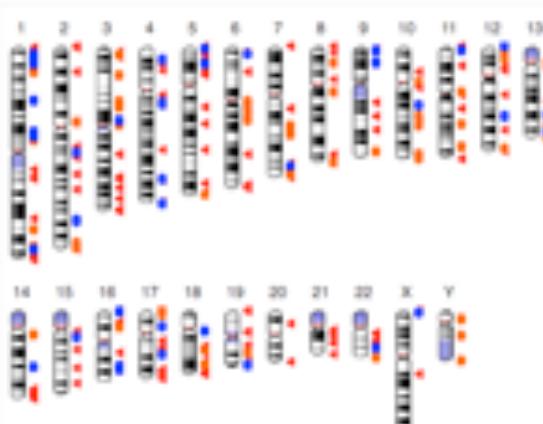
The main article, titled "EDITORIAL", is from **Nature Methods** 7, 331 (2010) doi:10.1038/nmeth0510-331. The title of the article is ***E pluribus unum***. The text discusses the need for the human reference genome to reflect more of the actual genomic diversity in humans, emphasizing community participation. It notes that the genome is ten years old and has been improved through resequencing projects, but its improvement depends on the research community providing data for less accessible regions. The article also mentions the GRCh37 reference genome and the HuRef genome.

On the right side, there are two columns: "Subscribe to Nature Methods" with a "Subscribe" button, and "This Issue" with links to the Table of contents and the next article. Below that is a "Article tools" section with links to Download PDF, Send to a friend, CrossRef lists (11 articles citing this article), Scopus lists (9 articles citing this article), Export citation, and Rights and permissions.

At the bottom right is a "naturejobs" section listing recruitment opportunities at Sun Yat-sen University and Institut franco-chinois de l'énergie nucléaire, along with links to More science jobs and Post a job.

## Human Genome Overview

Information about the continuing improvement of the human genome



- ◀ Region containing alternate loci
- Region containing fix patches
- Region containing novel patches

Karyogram of the latest human assembly, GRCh38.p11

The GRC is working hard to provide the best, possibly by both generating multiple representations (alternatives) for each locus, and by allowing users to represent by a single path. Additionally, we are now allowing users who are interested in a specific locus to affect users who need chromosome coordinate sets.

### 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 Remap](#) 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

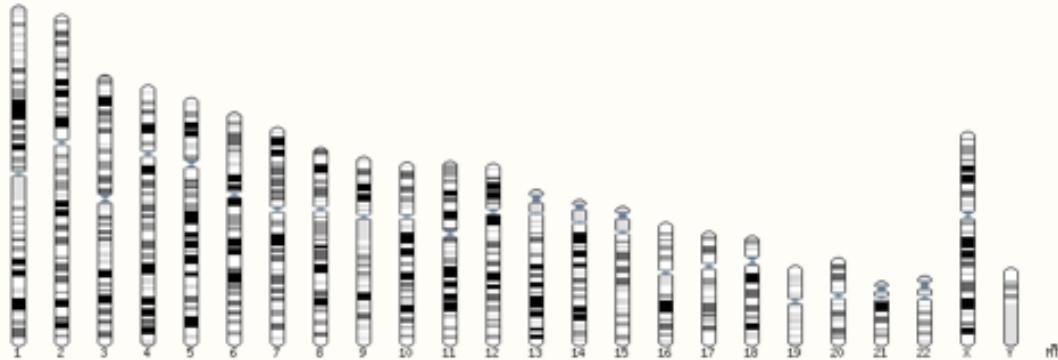
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordinates of patch scaffolds is now: 64 FIX and 59 NOVEL.

Assembly accessions: GenBank: [GCA\\_000001405.26](#), RefSeq: [GCF\\_000001405.37](#)

### Pseudoautosomal regions

Name	Chr	Start	Stop
PAR81	X	10,001	2,781,479
PAR82	X	158,701,363	158,030,895
PAR81	Y	10,001	2,781,479
PAR82	Y	56,887,903	57,217,415

# The human genome - basic stats



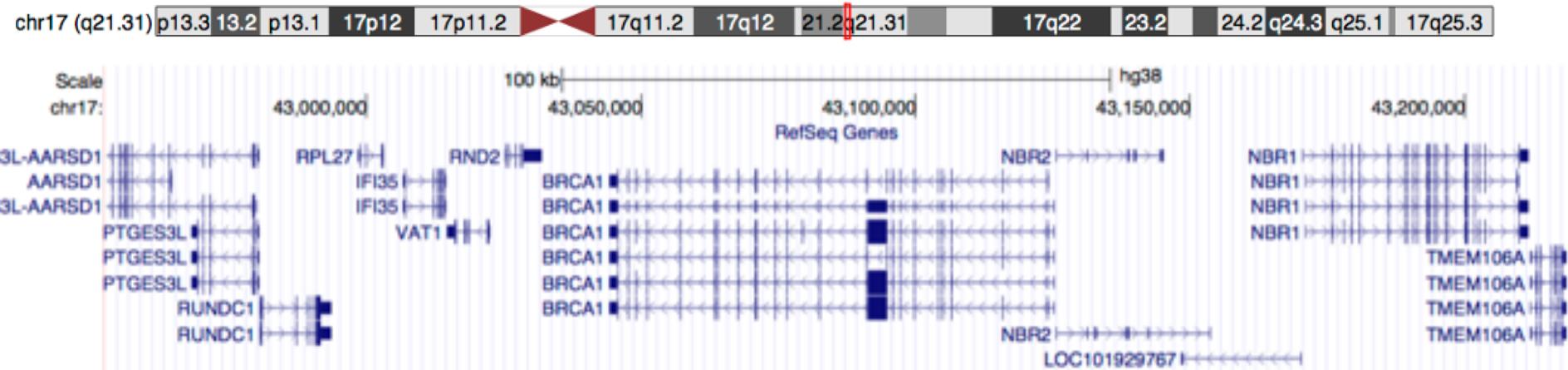
- 3.096 billion base pairs (haploid)
- 20,454 protein coding genes
- 226,950 coding transcripts  
(isoforms of a gene that each encode a distinct protein product)

Assembly	GRCh38.p12 (Genome Reference Consortium Human Build 38), INSDC Assembly <a href="#">GCA_000001405.27</a> , 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

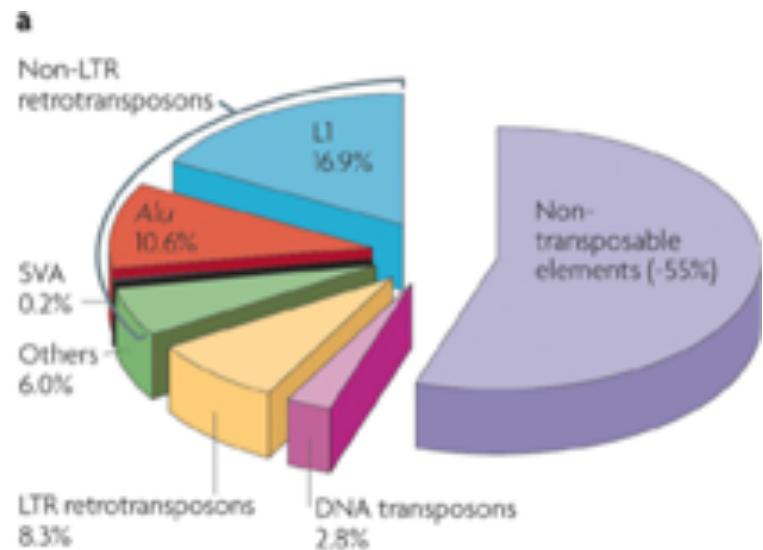
## Gene counts (Primary assembly)

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

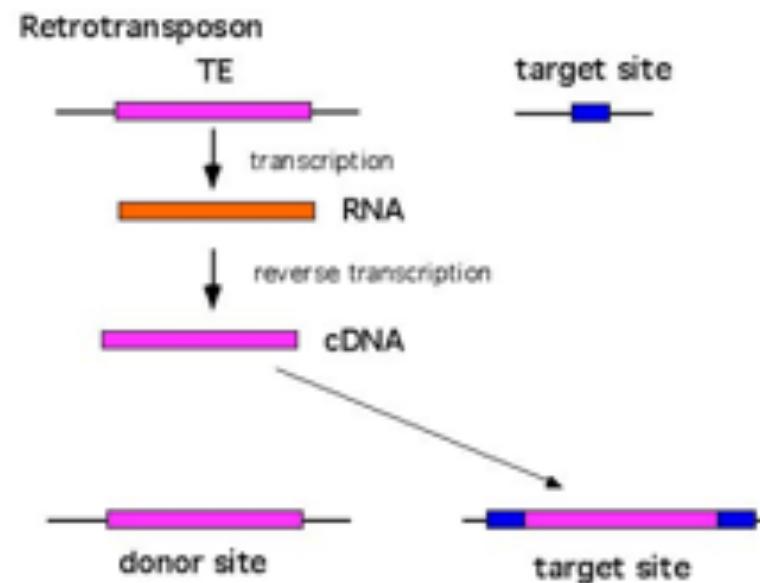
# Solely 2% of the human genome encodes proteins.



# Half of the human genome is comprised of repeats

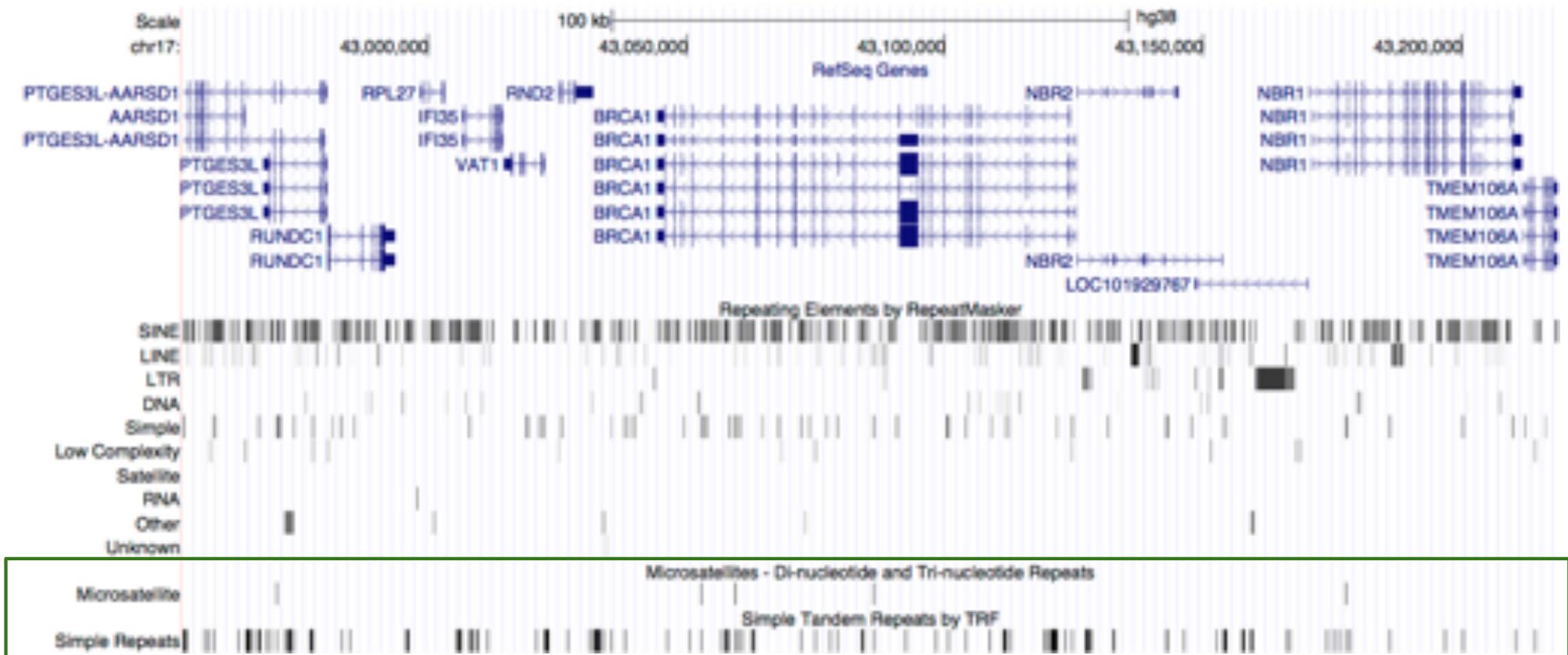


McClintock's  
"jumping  
genes" in  
maize



Retrotransposons use a "copy/paste" mechanism  
DNA transposons use a "cut/paste" mechanism

# Half of the human genome is comprised of repeats



Repetitive DNA not driven by retrotransposition (e.g., ATATATATATATATAT...)

# GC content varies dramatically in the genome

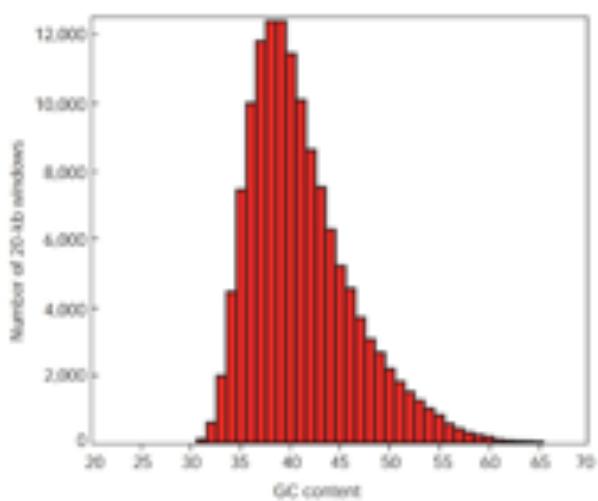
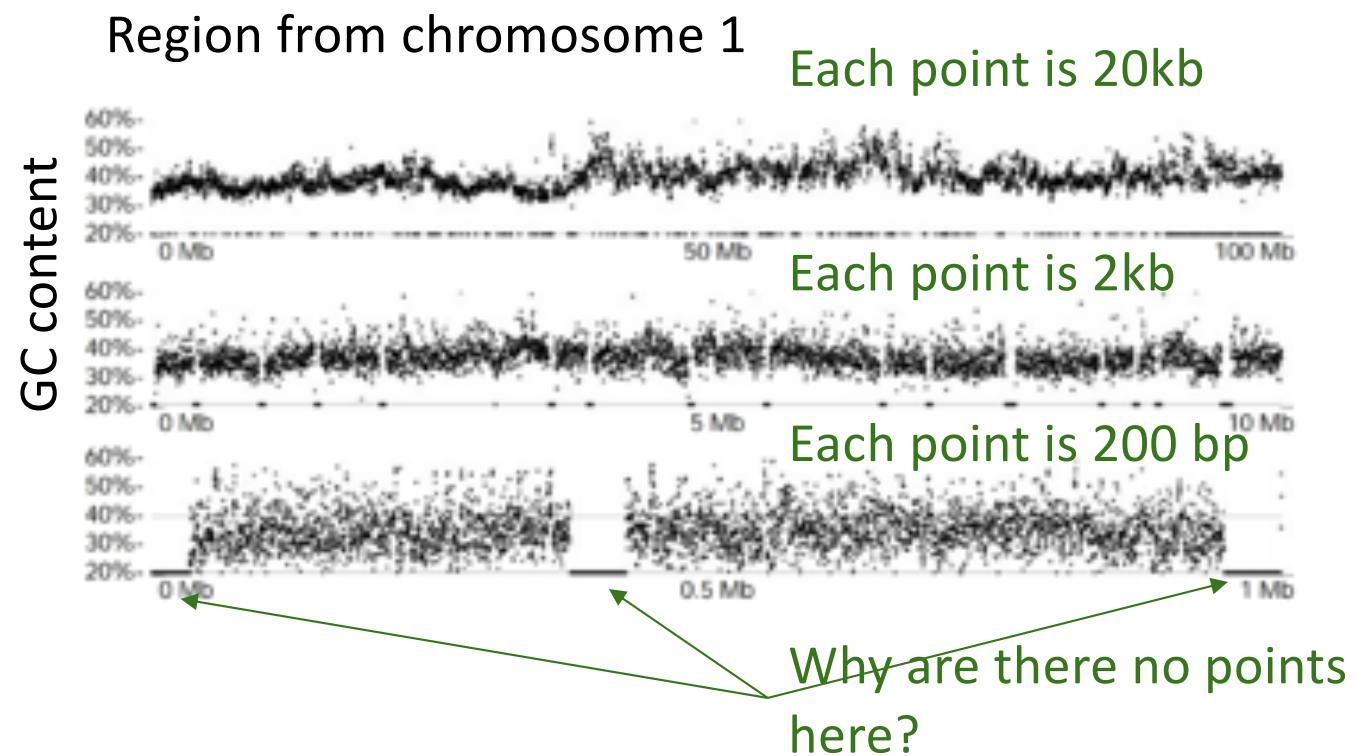


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.



# 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

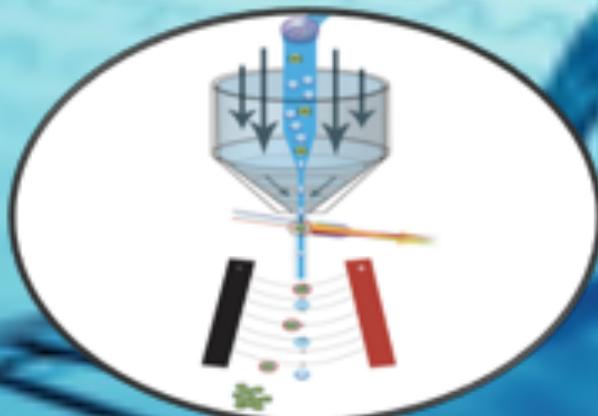
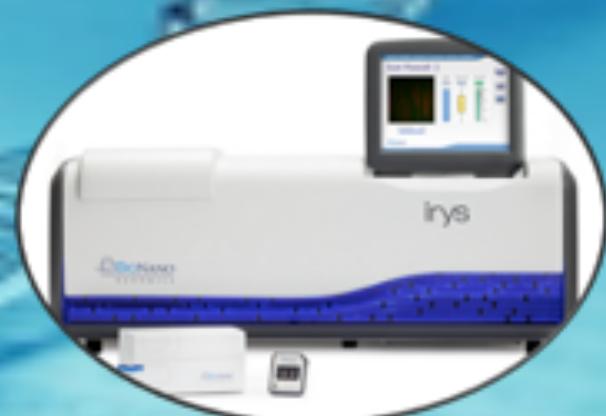
Sample Preparation



Sequencing



Chromosome Mapping

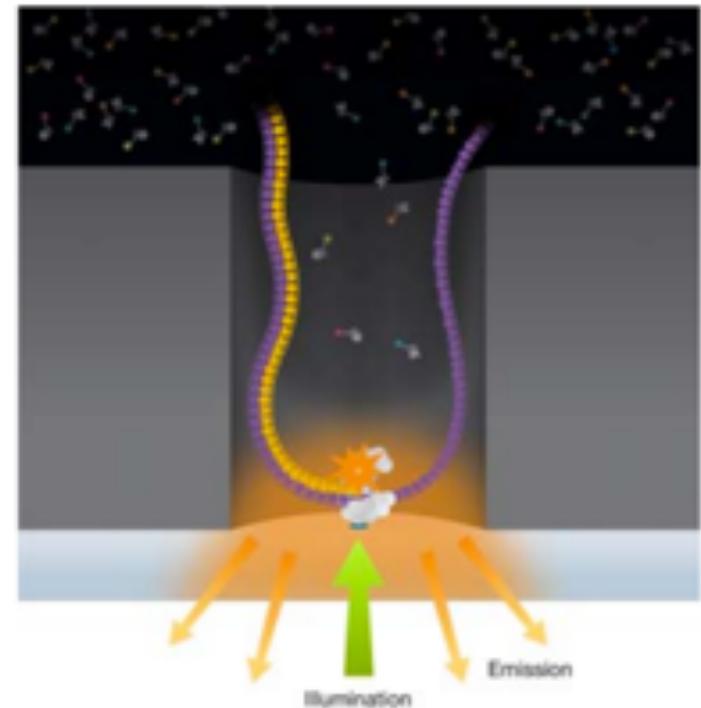
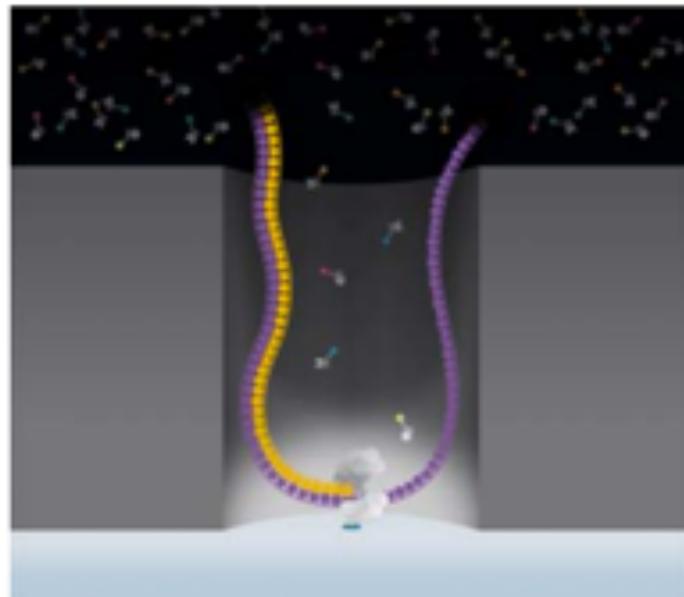


# PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

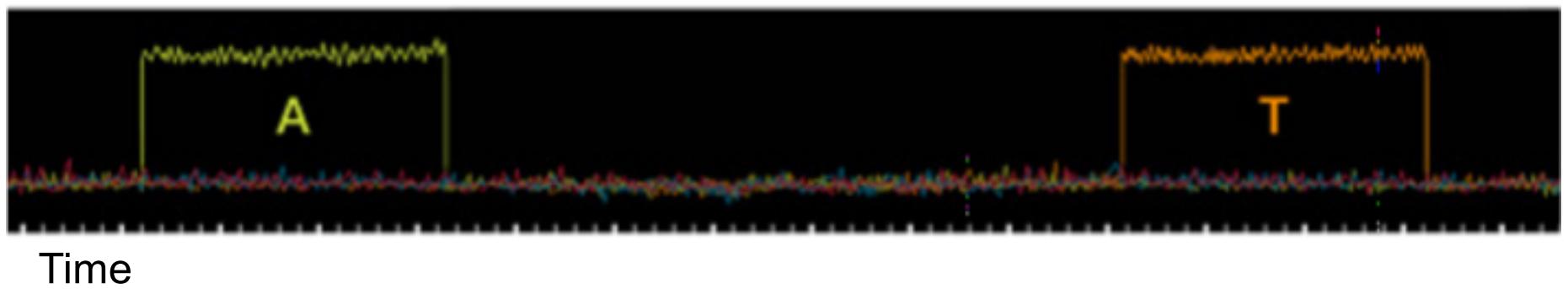


# PacBio: SMRT Sequencing

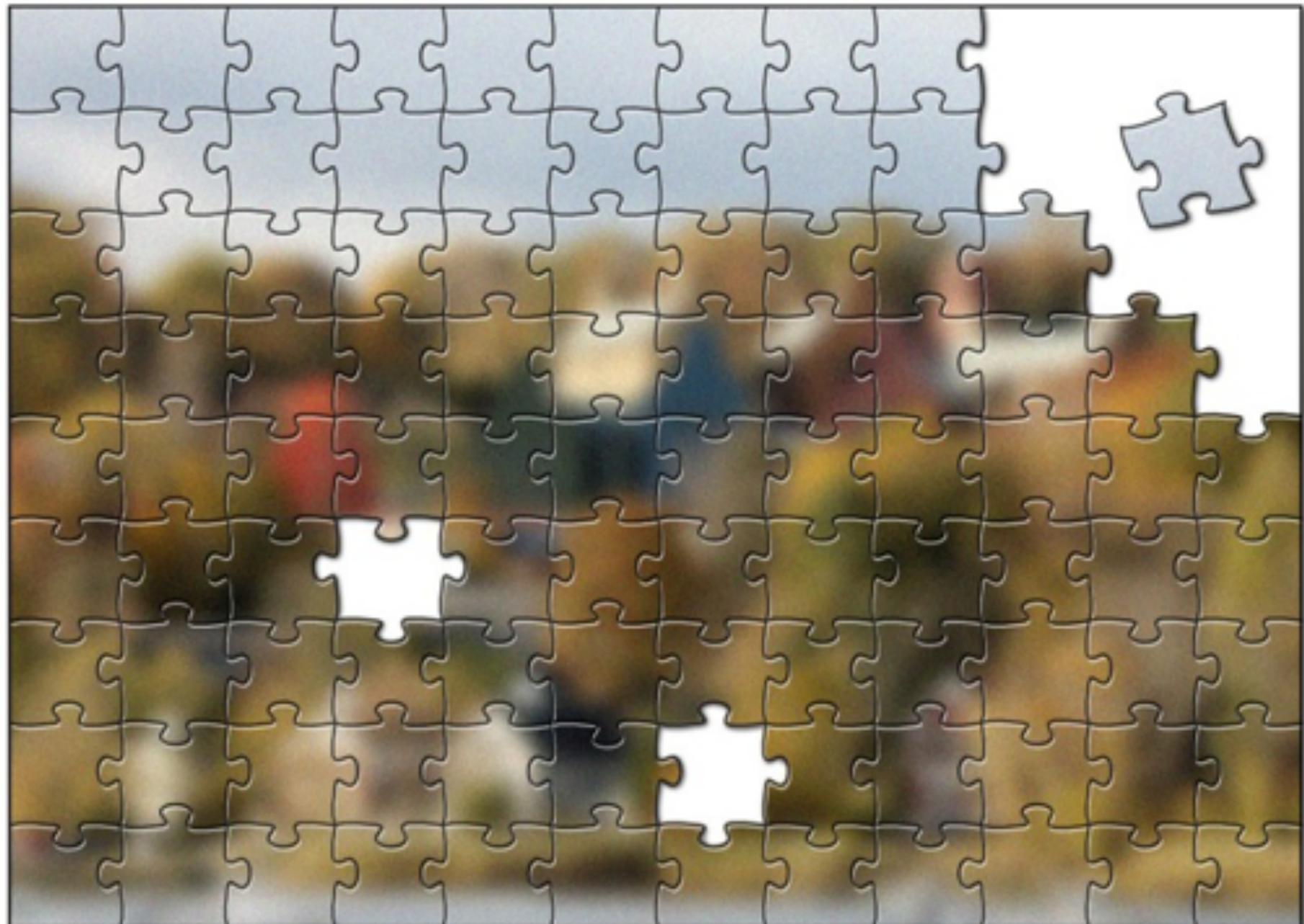
Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



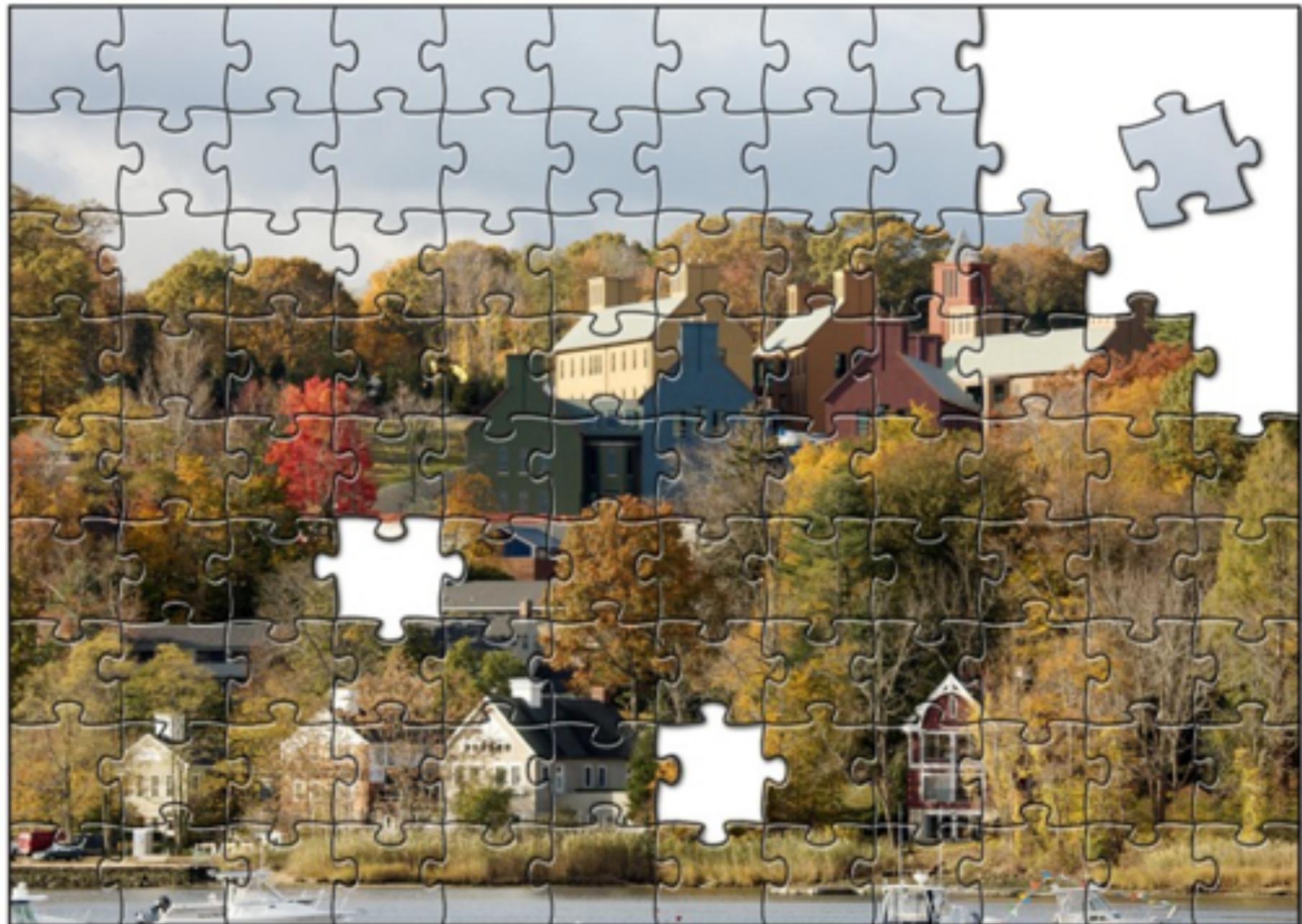
Intensity



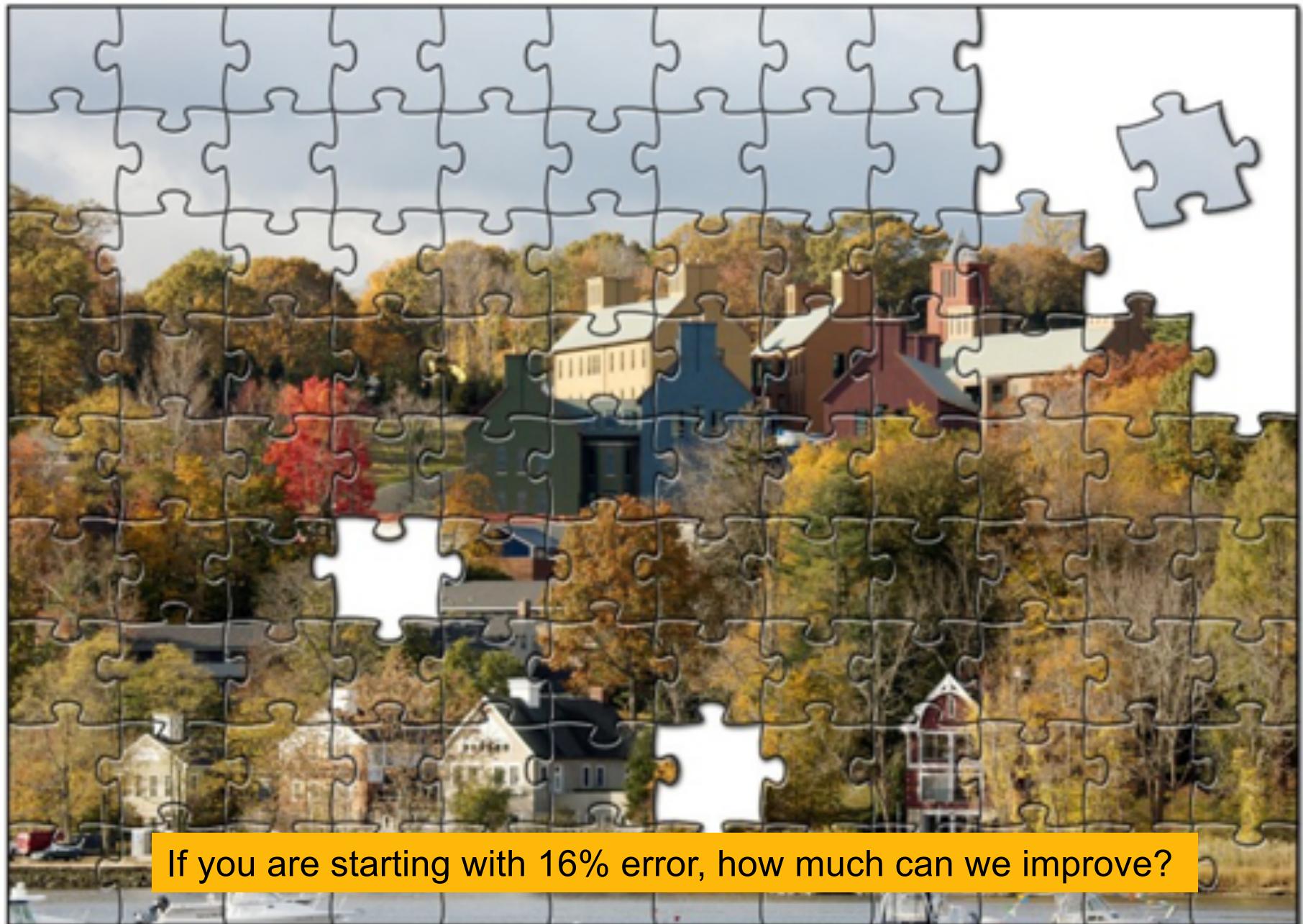
# Single Molecule Sequences



# “Corrective Lens” for Sequencing

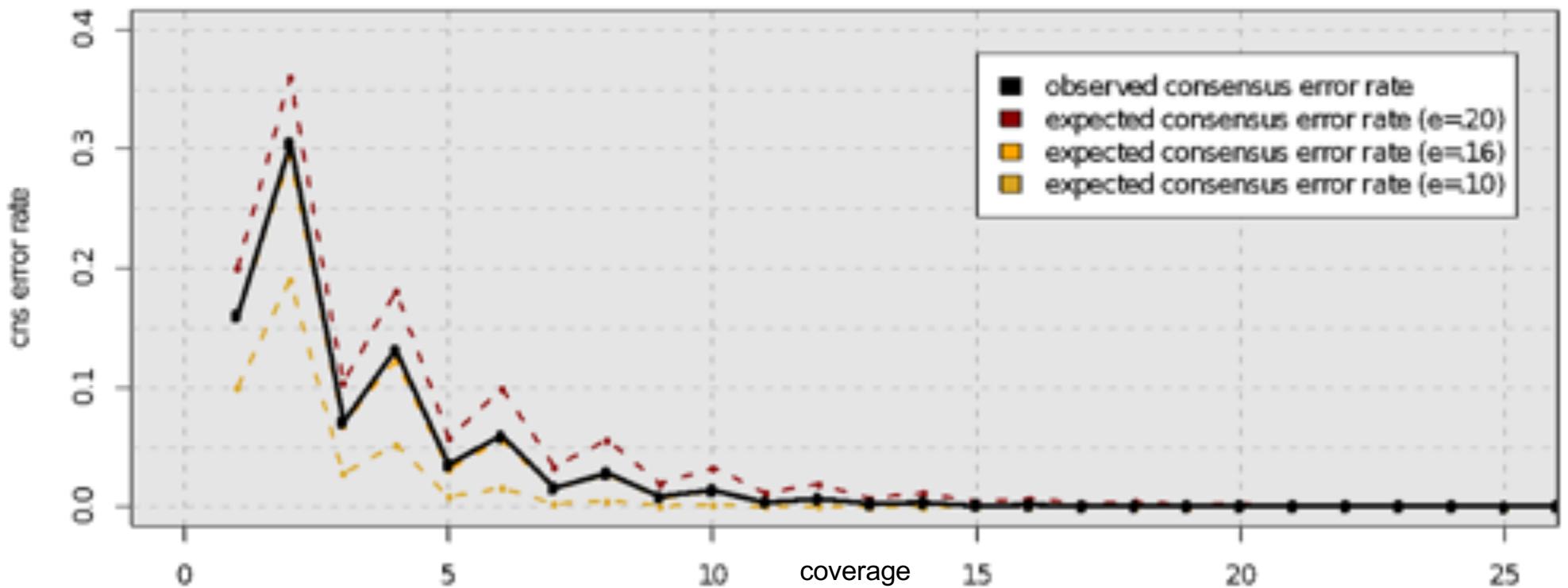


# “Corrective Lens” for Sequencing



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

# Consensus Accuracy and Coverage



Coverage can overcome random errors

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

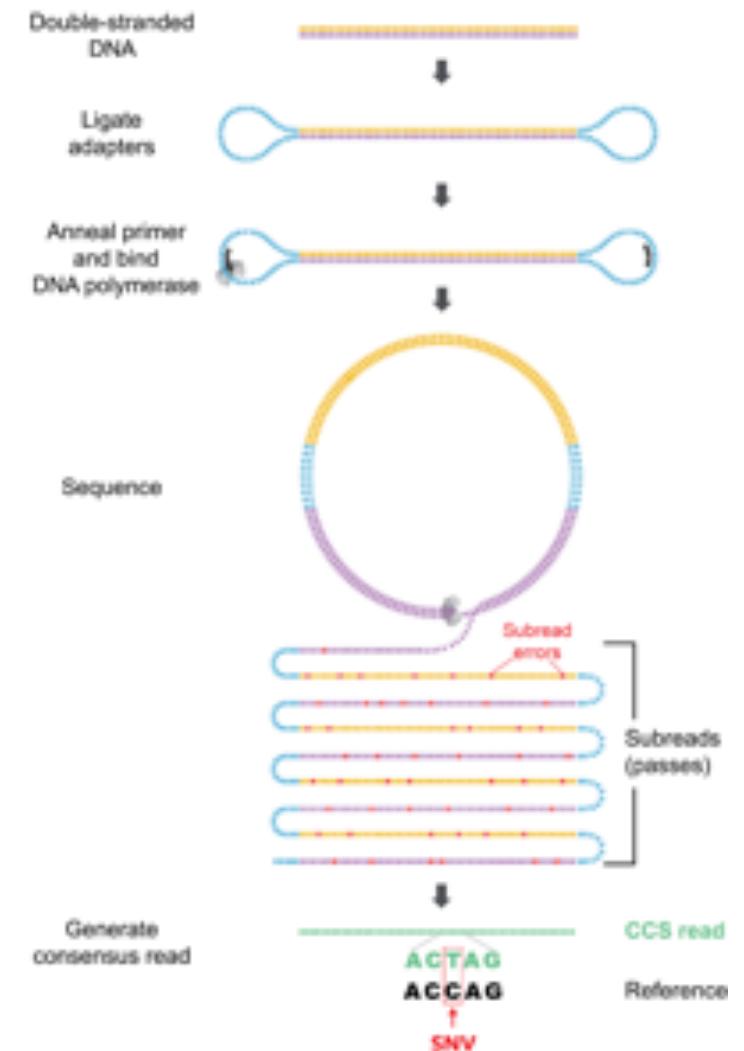
$$CNS Error = \sum_{i=\lceil c/2 \rceil}^c \binom{c}{i} (e)^i (1-e)^{n-i}$$

# “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

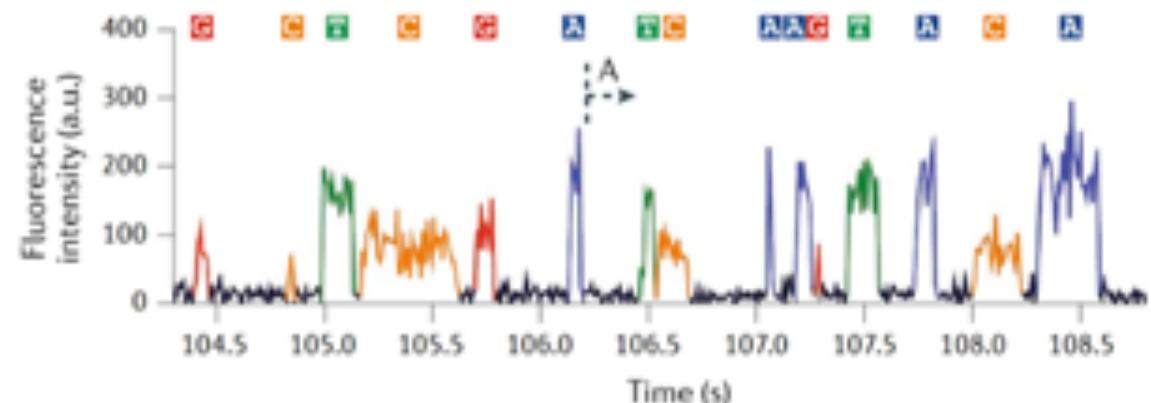
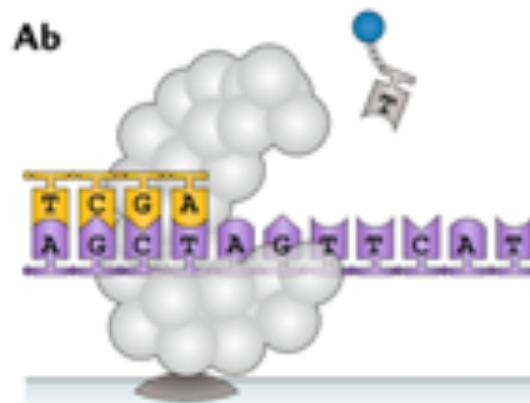
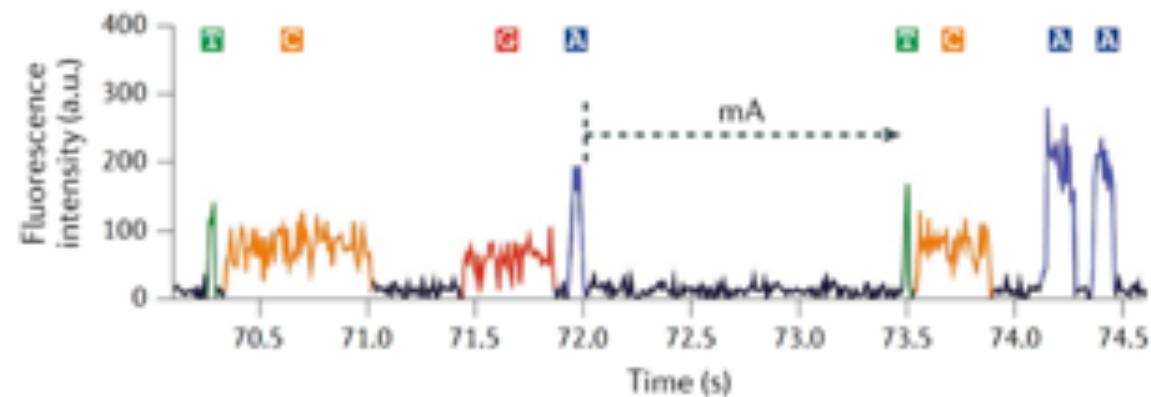
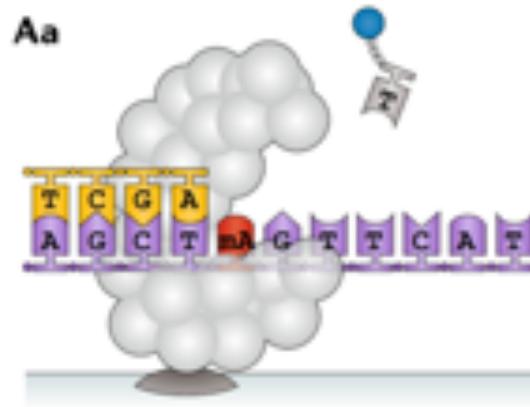


Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome

Wenger et al (2019) Nature Biotechnology doi:10.1038/s41587-019-0217-9

# 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 Inc

NASDAQ: PACB

Following

**39.54** USD **+0.89 (2.30%) ↑**

Closed: Feb 9, 7:47 PM EST · Disclaimer

After hours 38.90 **-0.64 (1.62%)**

1 day

5 days

1 month

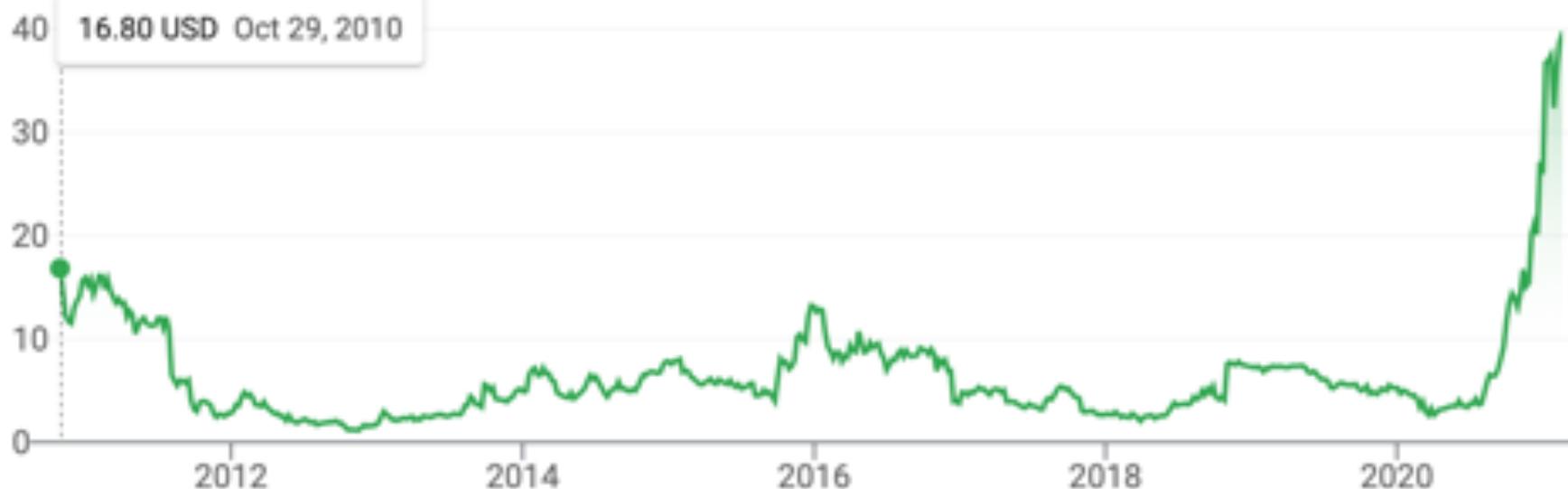
6 months

YTD

1 year

5 years

Max



Open

38.54

Div yield

-

High

40.53

Prev close

38.65

Low

37.67

52-wk high

41.65

Mkt cap

7.40B

52-wk low

2.20

P/E ratio

-

# PACB

A screenshot of a web browser window. The address bar shows the URL: [bioworld.com/articles/432183-illumina-pacific-biosciences-abandon-12b-merger-over-ftc-opposi...](https://bioworld.com/articles/432183-illumina-pacific-biosciences-abandon-12b-merger-over-ftc-opposi...). The page content is about Illumina and Pacific Biosciences abandoning a \$1.2 billion merger due to FTC opposition. The BioWorld logo is at the top left, and a Cortellis logo with the text "Powering insights from Cortellis" is at the top right. The browser interface includes various tabs and icons.

## BioWorld™

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 5, 2020

Two players in the gene sequencing space, Illumina and Pacific Biosciences, have scuttled 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](#)



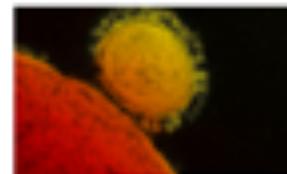
Insulet to launch wearable insulin pump this year, works with Dexcom CGM for closed loop

[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

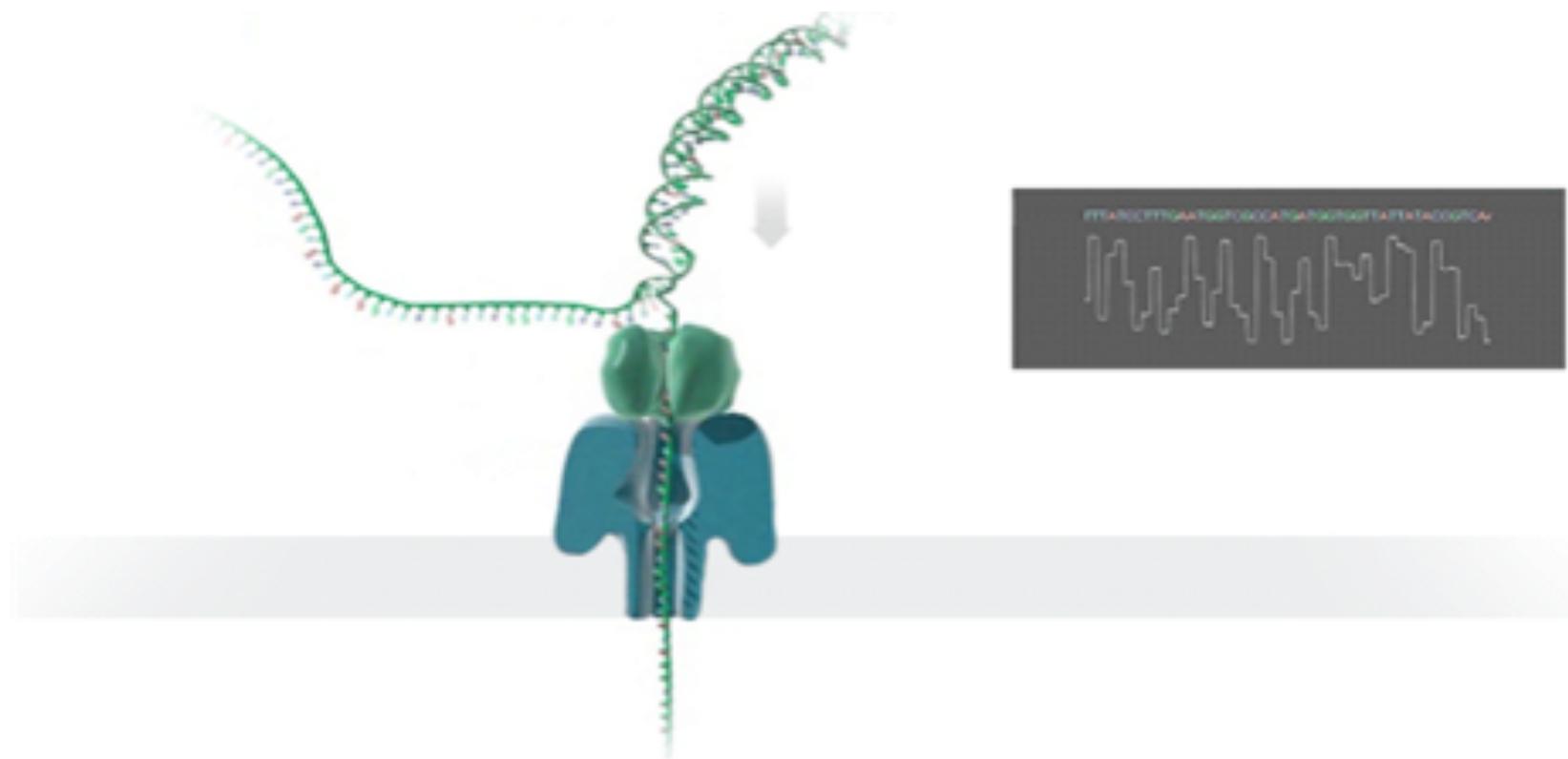
[BioWorld](#)

# Oxford Nanopore Technologies (ONT)



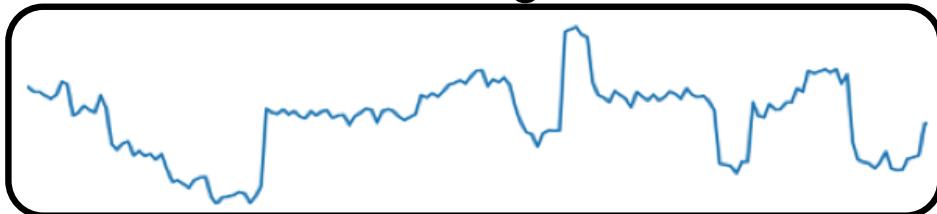
# Nanopore Sequencing

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



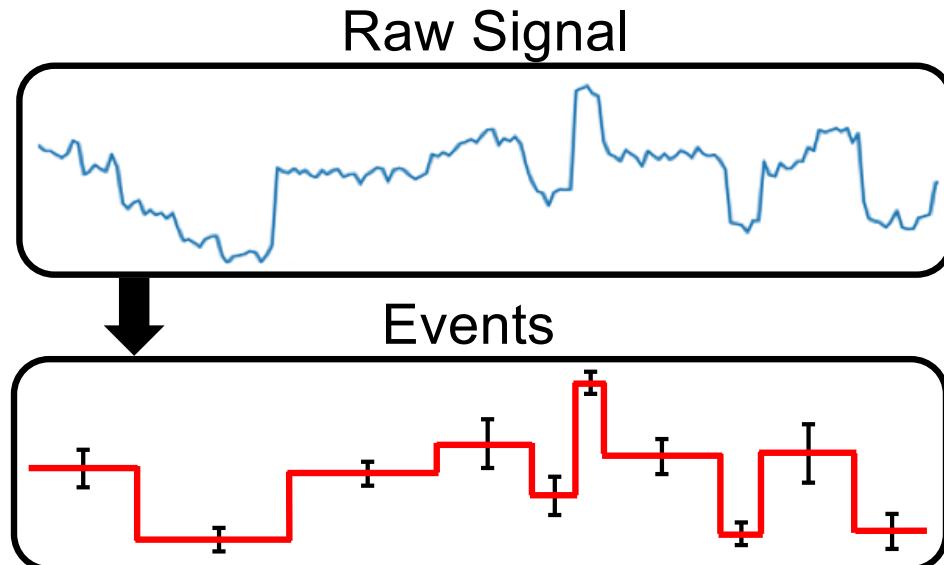
# Nanopore Basecalling

Raw Signal



Translation of raw signal  
into basepairs

# Nanopore Basecalling

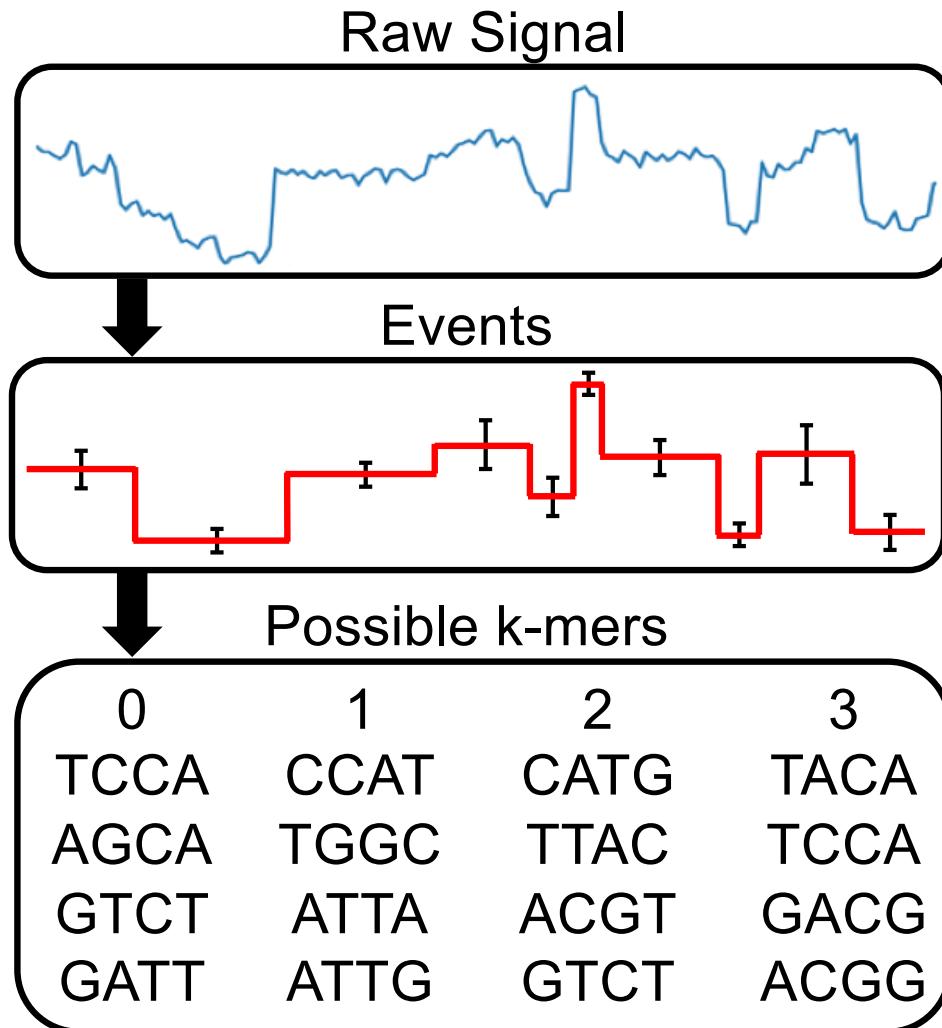


Translation of raw signal  
into basepairs

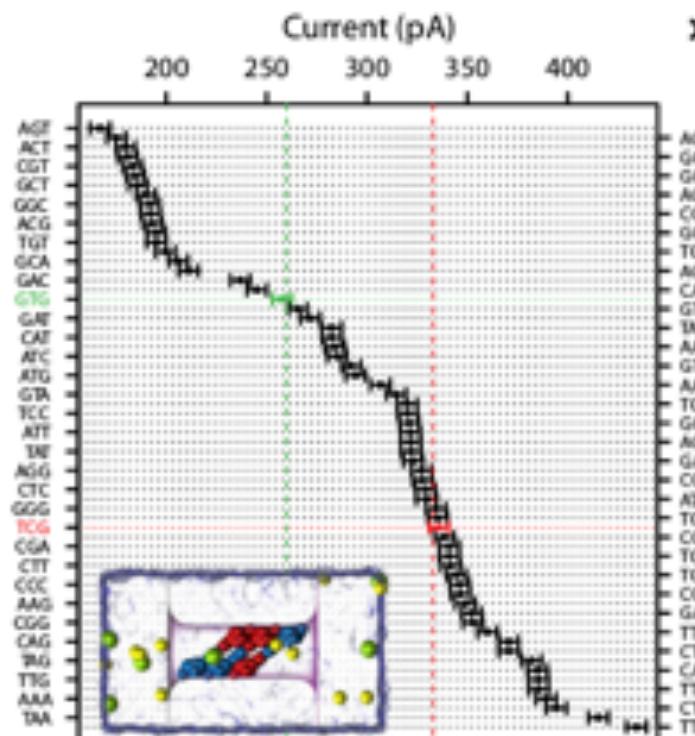
Early basecallers began by  
estimating k-mer boundaries  
using “events”, which were  
then input to an HMM

Modern basecallers use  
neural networks directly  
on raw signal

# Nanopore Basecalling

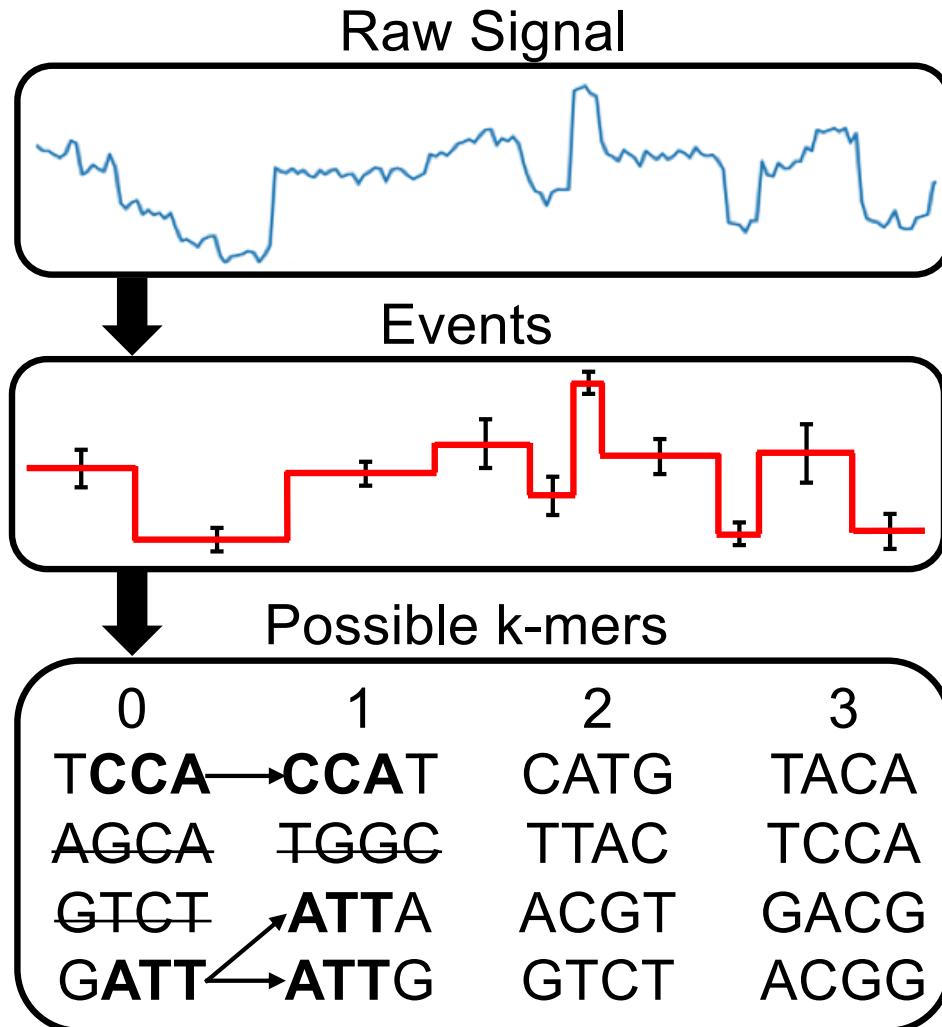


ONT releases k-mer models with expected current distribution of every k-mer

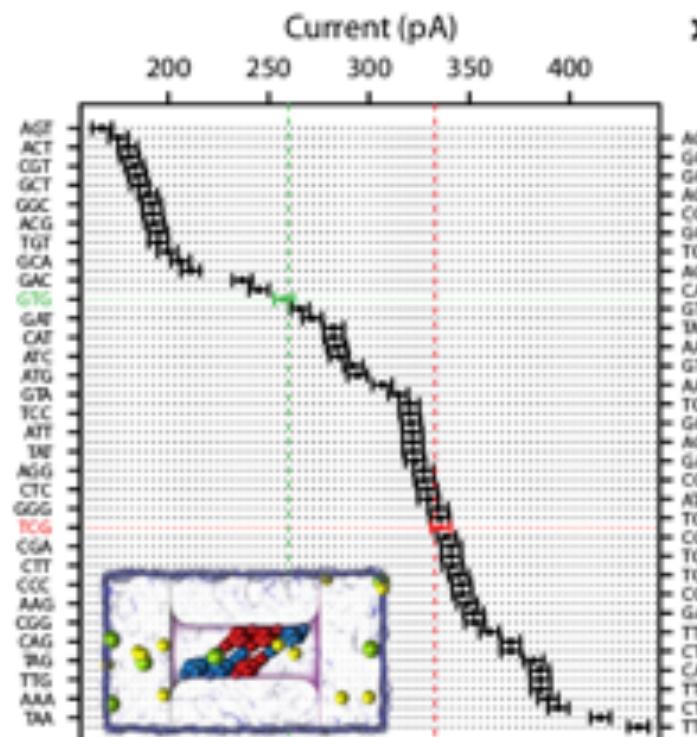


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm  
Timp et al. (2012) *Biophysical Journal*

# Nanopore Basecalling

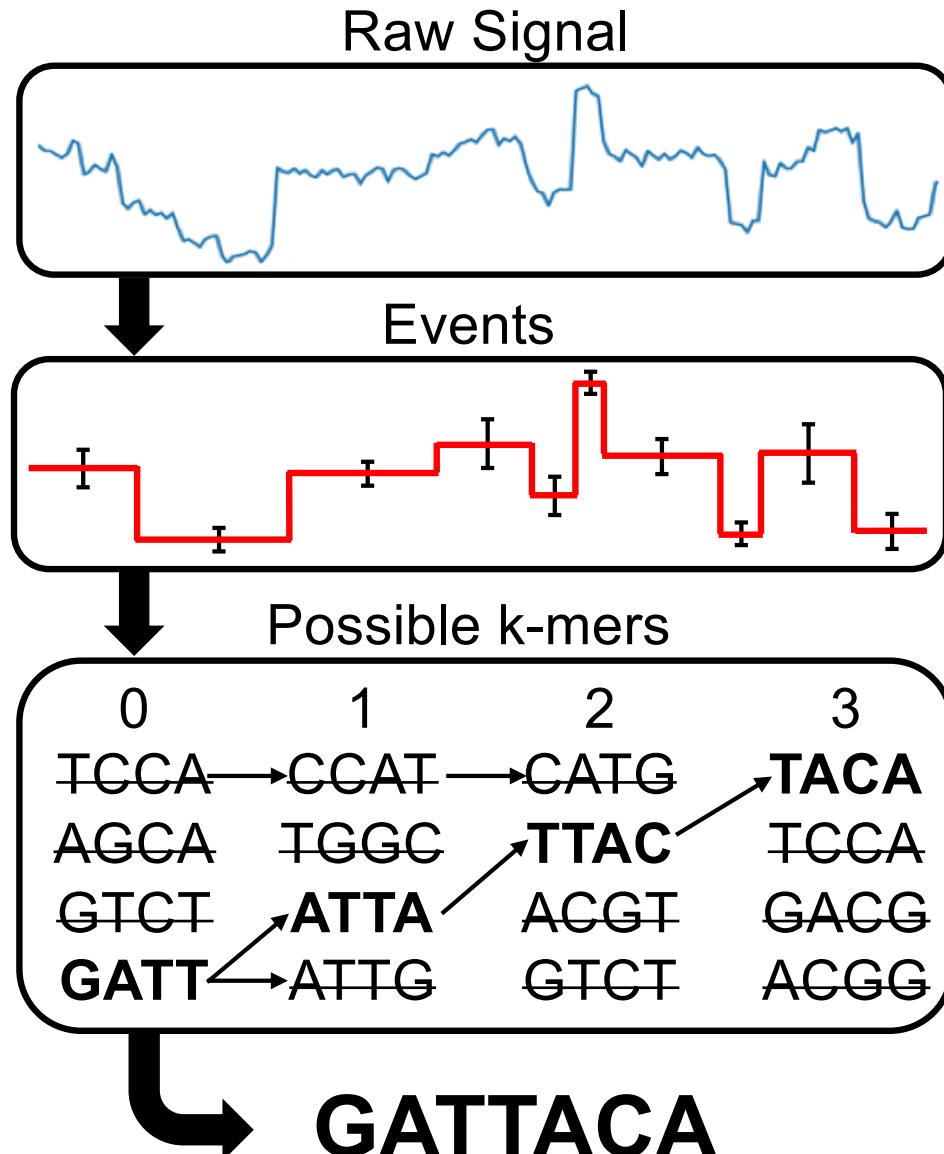


Certain k-mers can be eliminated based on possible transitions

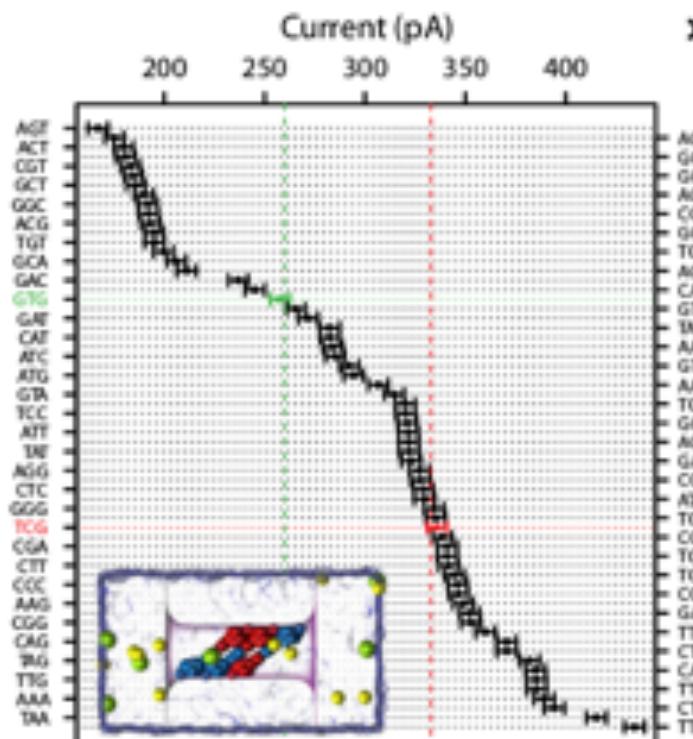


DNA Base-Calling from a Nanopore Using a Viterbi Algorithm  
Timp et al. (2012) *Biophysical Journal*

# Nanopore Basecalling



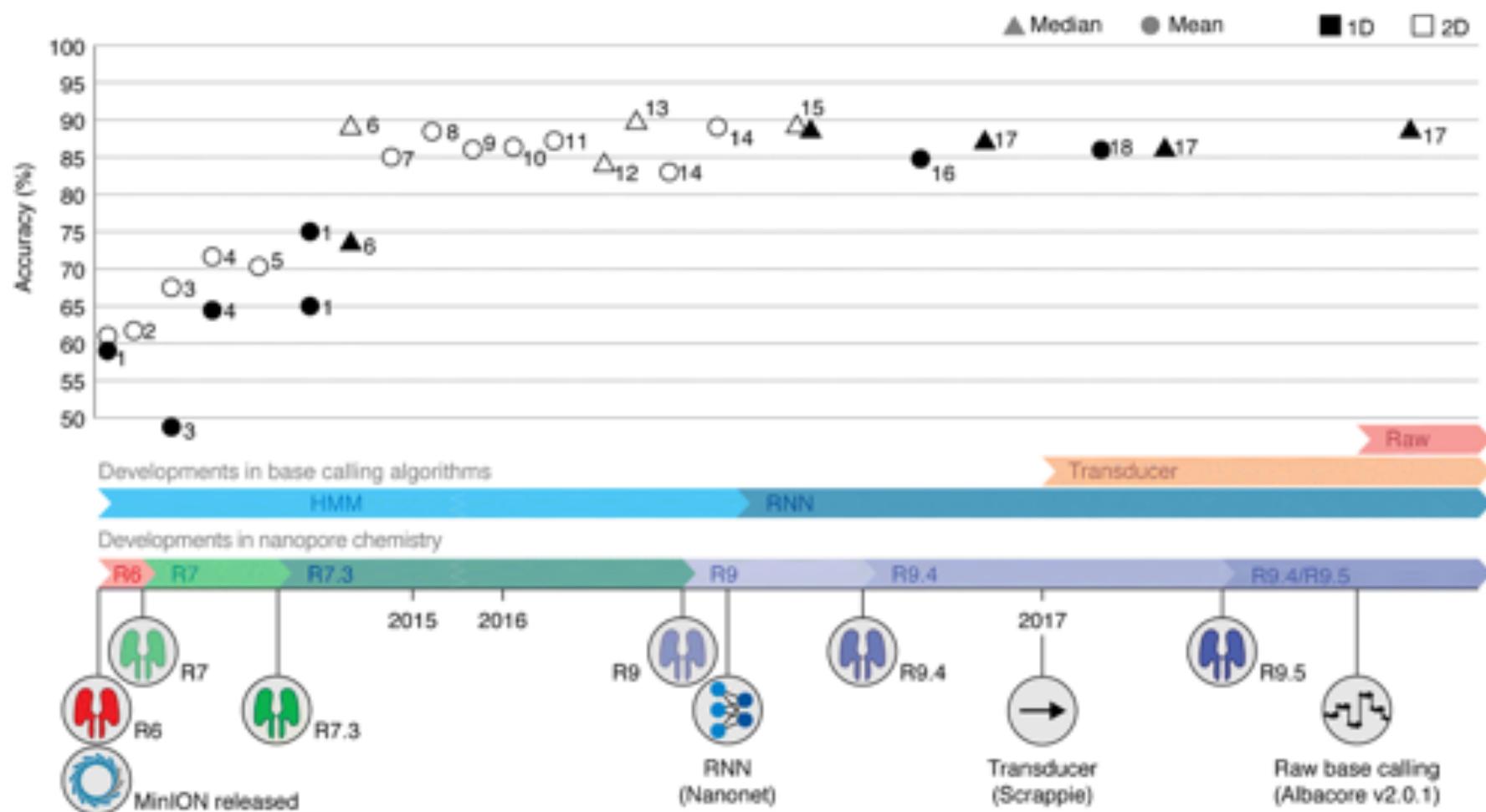
Final sequence determined by most probable k-mers



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

# Basecaller/Pore Timeline

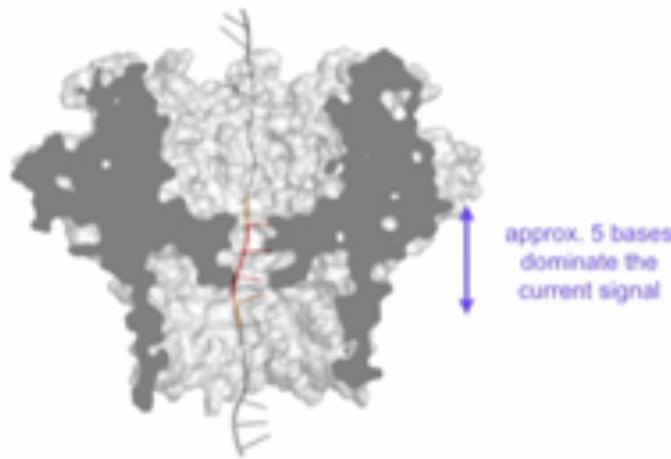
Development of both pore chemistry and basecalling algorithms is responsible for improvement in accuracy



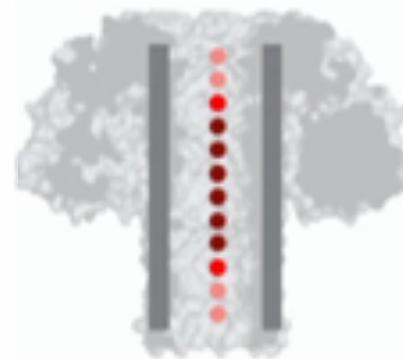
*From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy*  
Rang et al (2018) Genome Biology. <https://doi.org/10.1186/s13059-018-1462-9>

# New Pore Chemistries

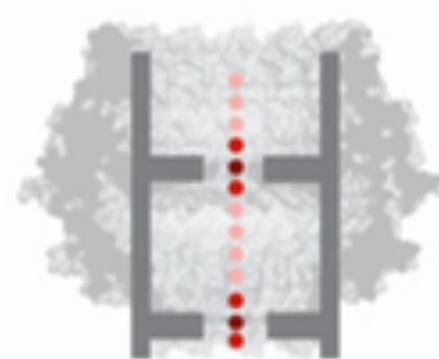
ONT is developing alternate pore chemistries to improve accuracy, particularly for homopolymers



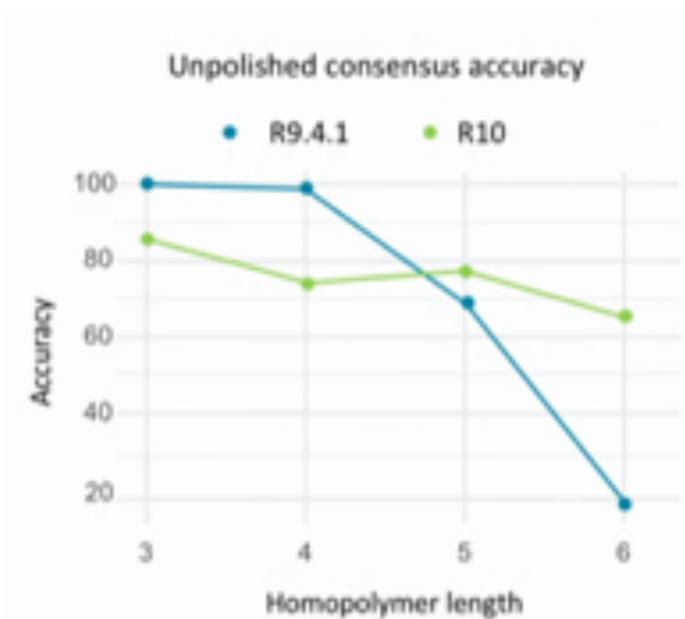
Standard pore  
chemistry  
“R9”



Pore with long  
reader head  
Lysenin –  
“R8”



Multiple points of  
contribution  
“R10”

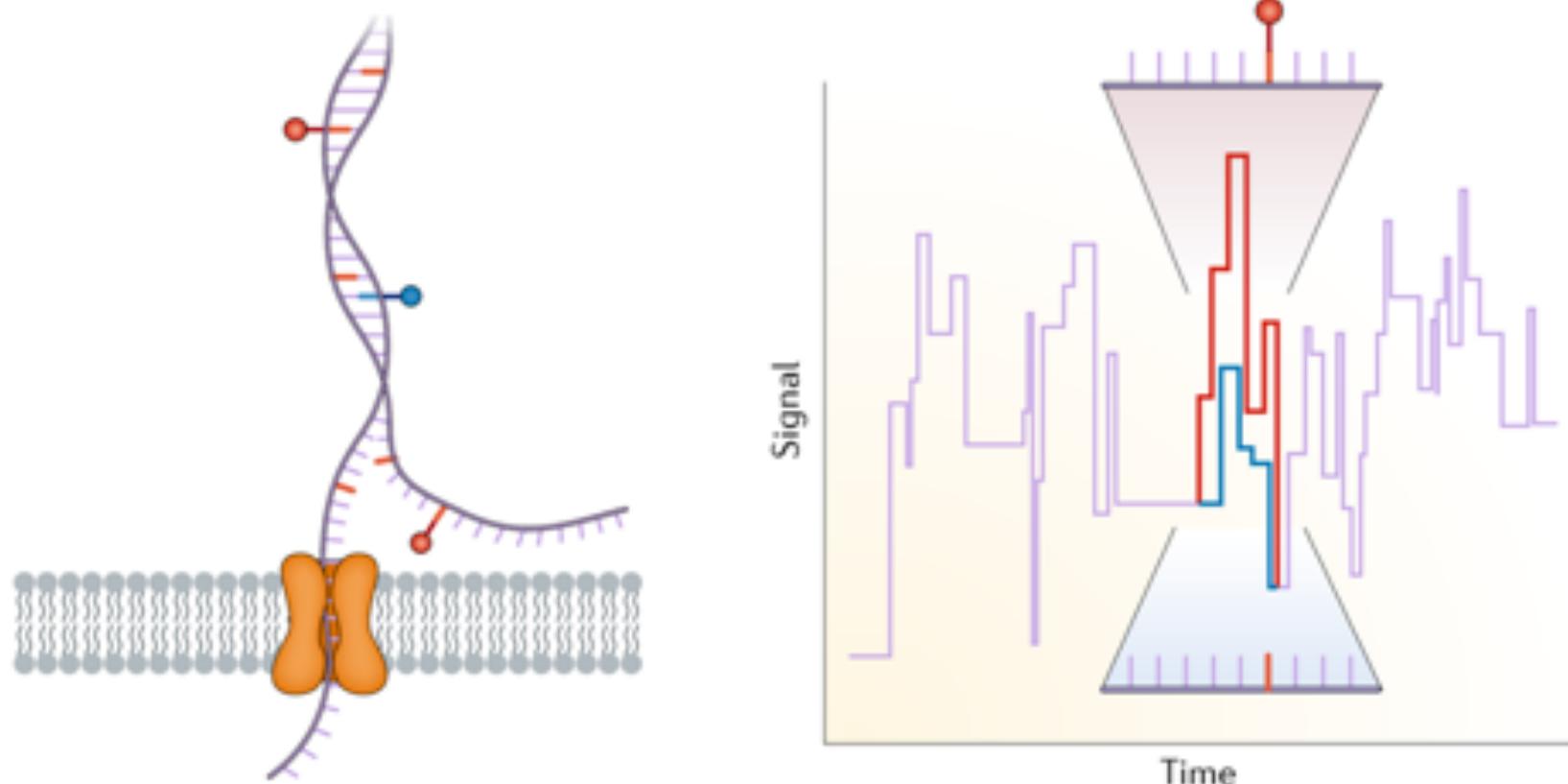


From 2018 London Calling Keynote  
<https://vimeo.com/272526835>

# DNA Modification Detection

Like PacBio, ONT can detect methylation from raw signal

- Or any other modification that changes ionic current



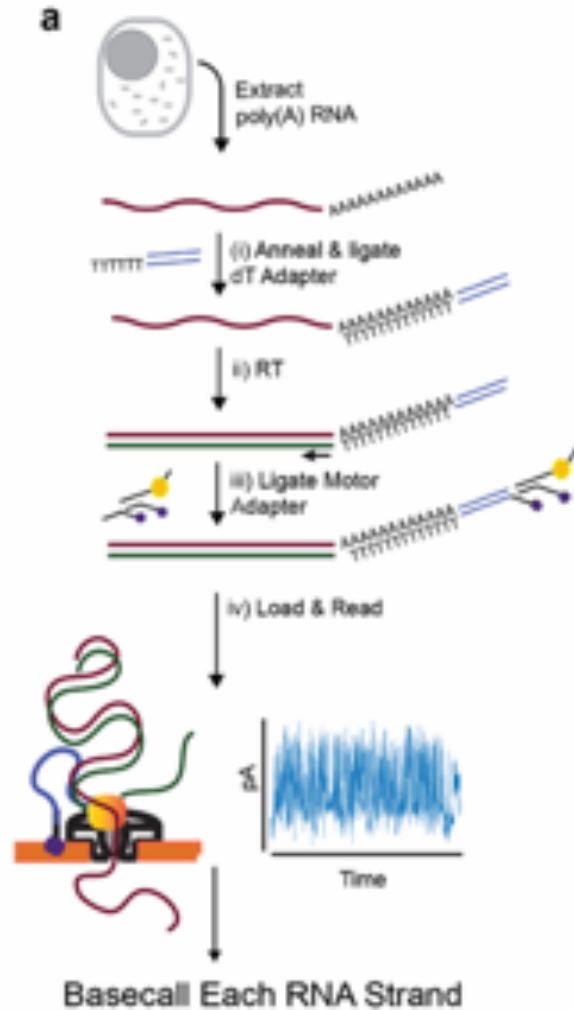
Piercing the dark matter: bioinformatics of long-range sequencing and mapping  
Sedlazeck et al. (2018) *Nature Reviews Genetics*. 19:329

# 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



genomeweb



[Business & Policy](#)   [Technology](#)   [Research](#)   [Diagnostics](#)   [Disease Areas](#)   [Applied Markets](#)   [Resources](#)

Enter your location

Home > Business, Policy & Funding > Business Plans - Oxford Nanopore Technologies Raises £100m

Oxford Nanopore Technologies Raises £109.5M

Jan 02, 2020 | staff reporter



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.

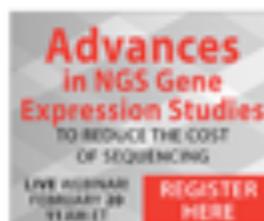
Oxford Nanopore said it has raised approximately £480 million to date. In October 2018, the firm announced that [Amgen would take a £50 million stake](#), or approximately 3 percent of the firm at the time, giving it a £1.5 billion valuation.

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

In a [brief following the announcement](#), Oxford Nanopore CTO Clive Brown suggested that the minimum amount needed to invest in the firm is \$20 million.

Breaking News 8

- 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



LIVE INDIANA  
FEBRUARY 20  
11:00 a.m.

[REGISTER  
HERE](#)



[Sign Up for  
Topical  
Newsletters](#)

## 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

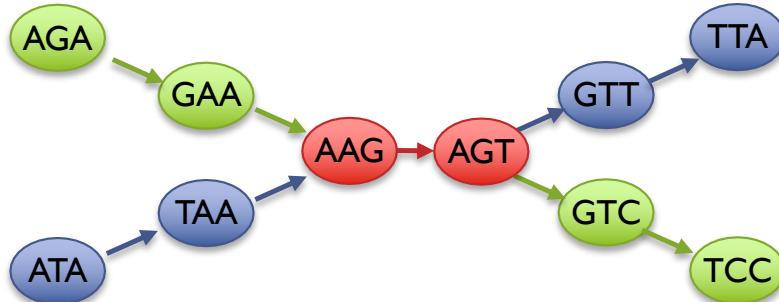
4th April 2019 ▲ Callum Cyrus

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

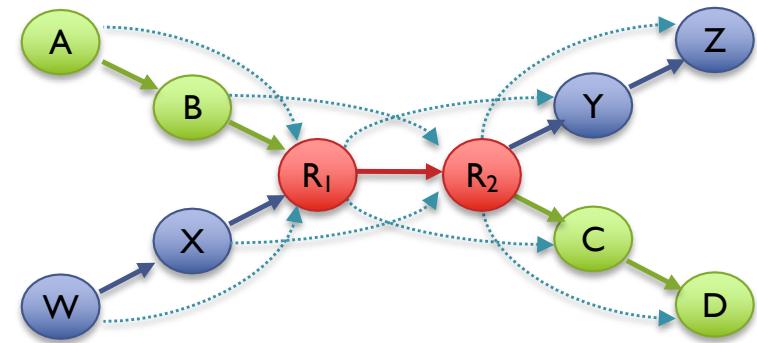
## de Bruijn Graph



Short read assemblers

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

## Overlap Graph

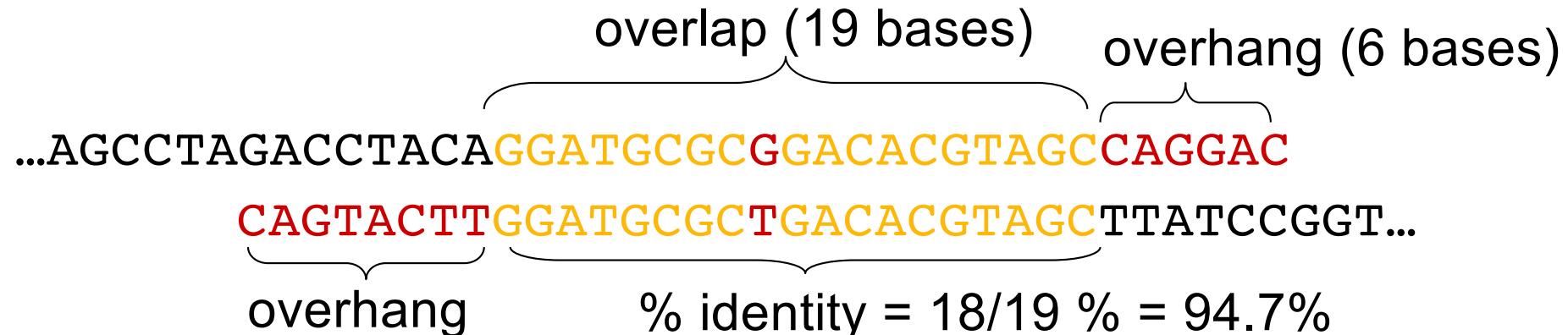


Long read assemblers

- 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.

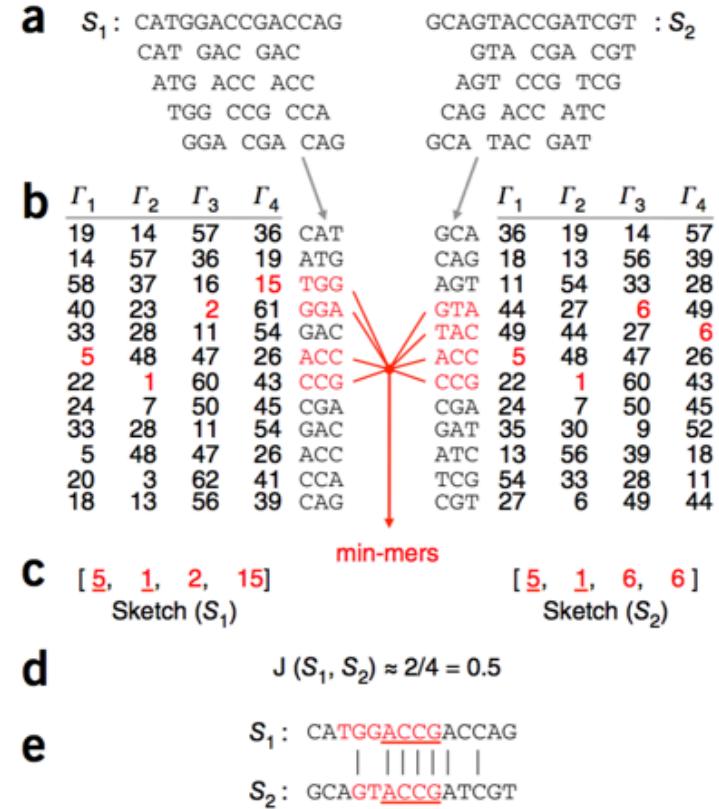
[How do we compute the overlap?]

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

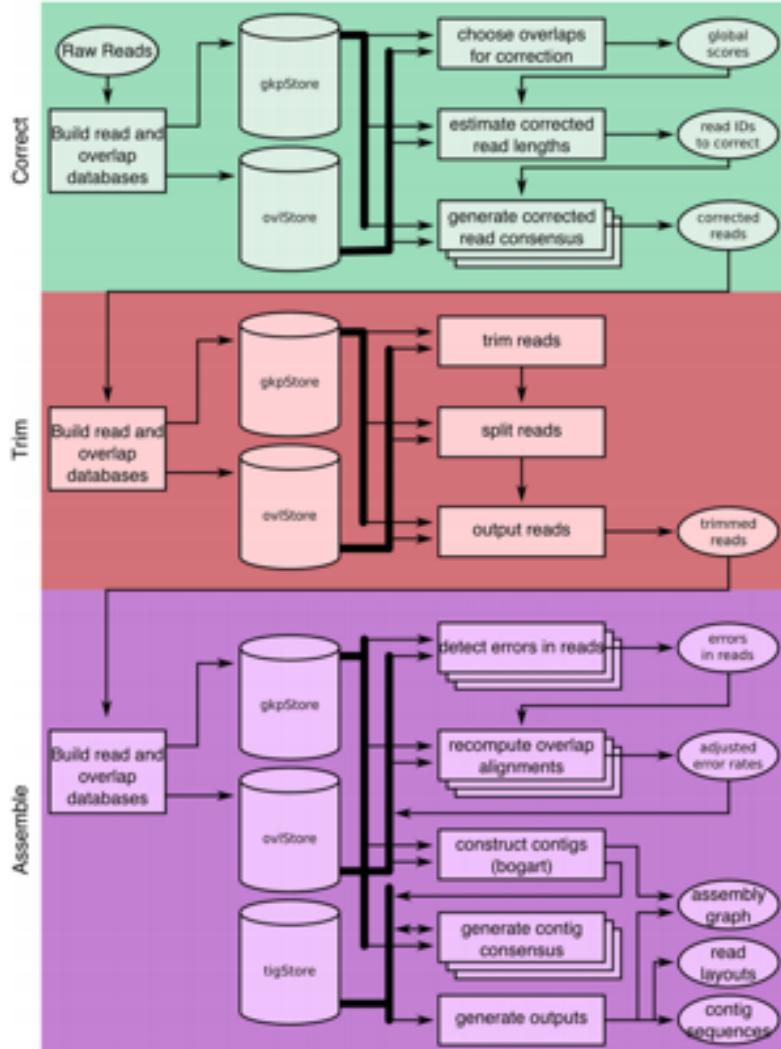
# Very fast approximate overlapping

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:  $|\text{intersection}| / |\text{union}|$
- 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



# 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

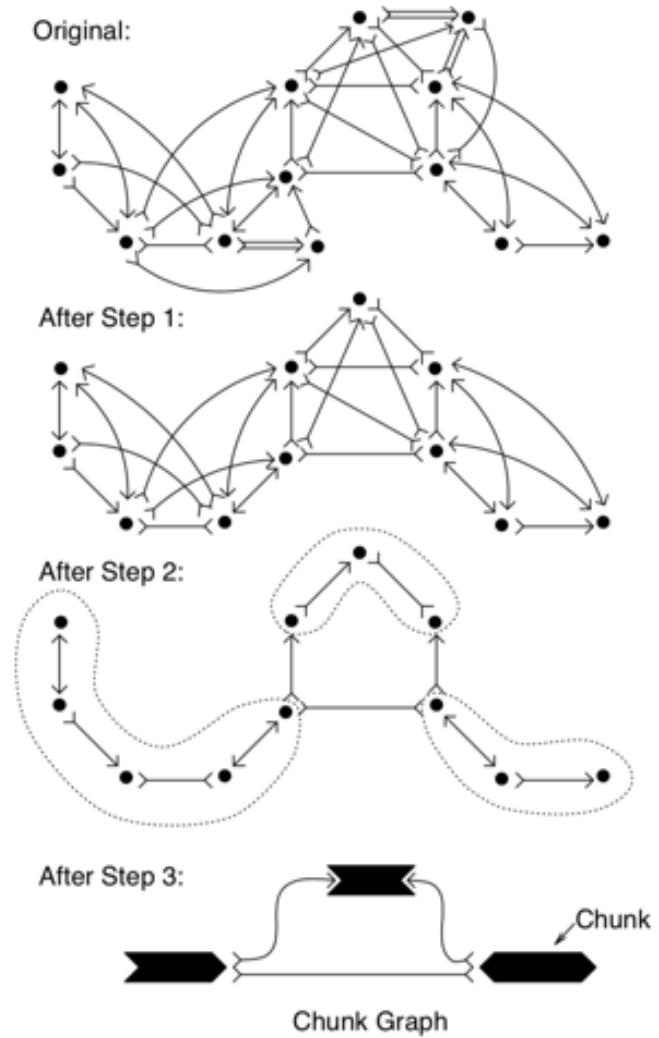
# Unitigging: Pruning the Overlap Graph

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

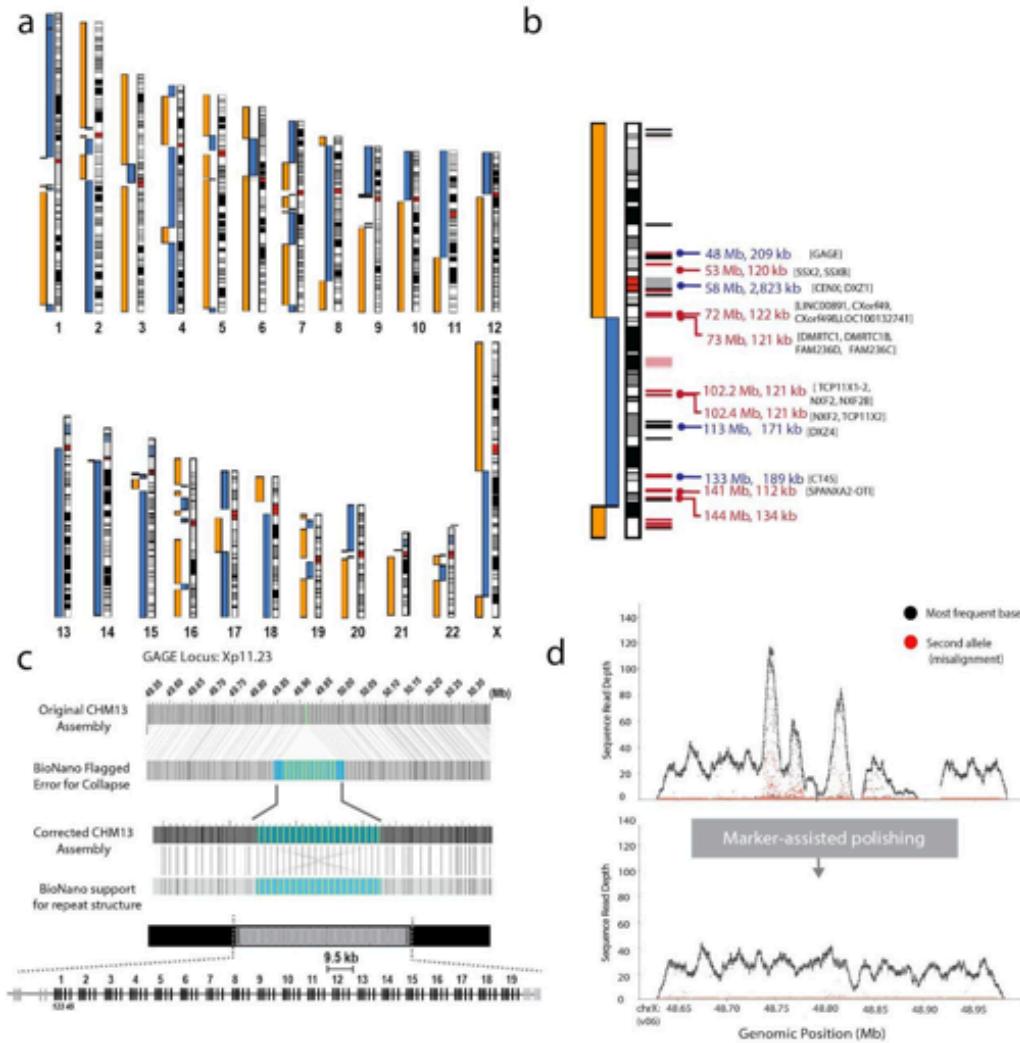
Transform the graph to simplify the assembly problem (without changing the valid solutions):

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  $\rightarrow$  B, and B  $\rightarrow$  C, remove the transitive edge A  $\rightarrow$  C
3. **“Chunkification”:** Linear subgraphs define uniquely assemblable segments: “unitigs”



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

# 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



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