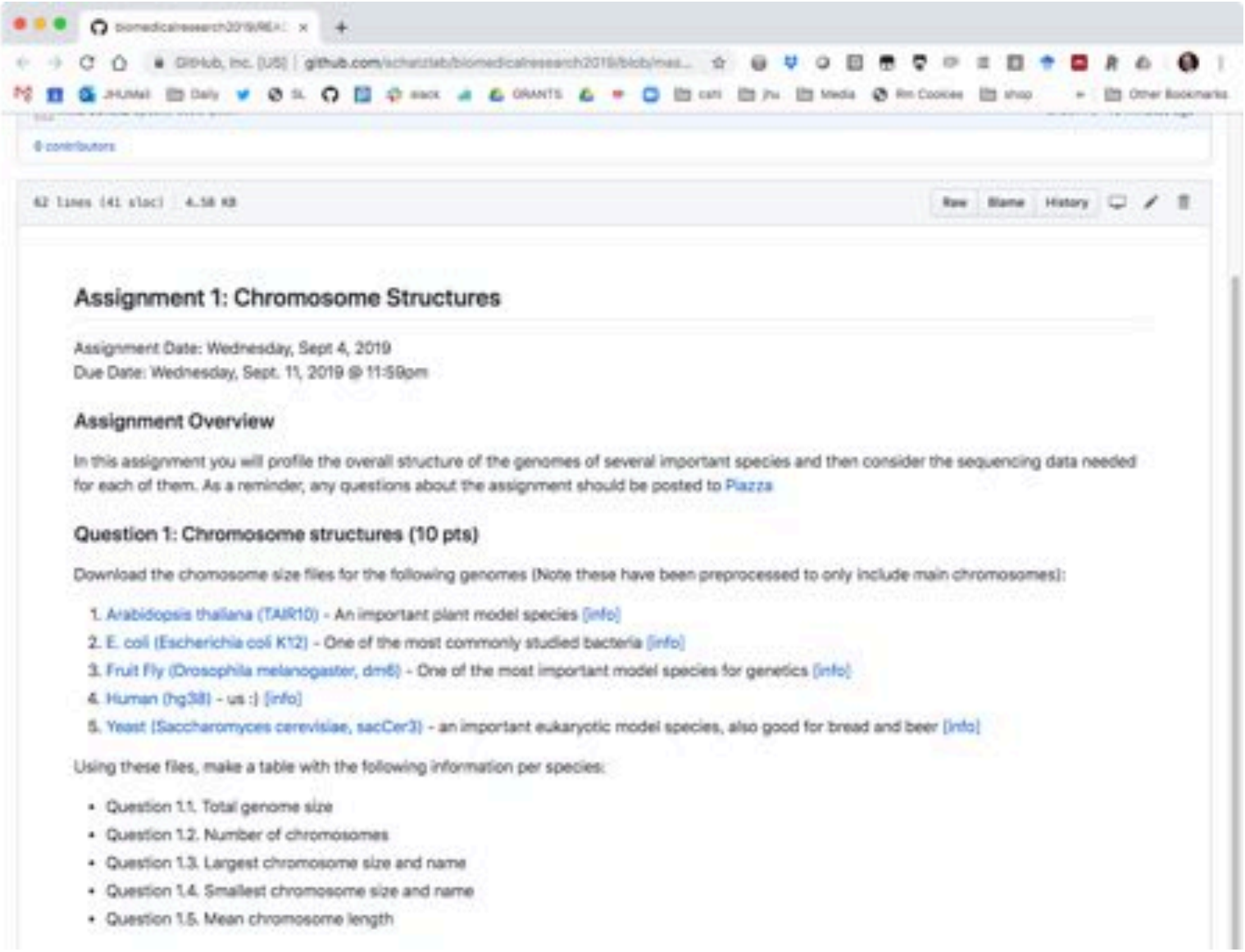
Whole Genome Alignment

Michael Schatz

Sept 16, 2019

Lecture 5: Computational Biomedical Research

Assignment 1: Chromosome Structures
Due Wed Sept 11 @ 11:59pm



Assignment 1: Chromosome Structures

Assignment Date: Wednesday, Sept 4, 2019
Due Date: Wednesday, Sept. 11, 2019 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then consider the sequencing data needed for each of them. As a reminder, any questions about the assignment should be posted to Piazza

Question 1: Chromosome structures (10 pts)

Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

1. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
2. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
3. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
4. Human (hg38) - us :) [info]
5. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes
- Question 1.3. Largest chromosome size and name
- Question 1.4. Smallest chromosome size and name
- Question 1.5. Mean chromosome length

https://github.com/schatzlab/biomedicalresearch2019

Assignment 2: De novo Assembly
Due Wed Sept 18 @ 11:59pm

Assignment 2: De novo assembly

Assignment Date: Wednesday, Sept 11, 2019
Due Date: Wednesday, Sept. 18, 2019 @ 11:59pm

Assignment Overview

In this assignment you will assemble a few simple datasets by hand and then by using an automated assembler called Spades. As a reminder, any questions about the assignment should be posted to Piazza

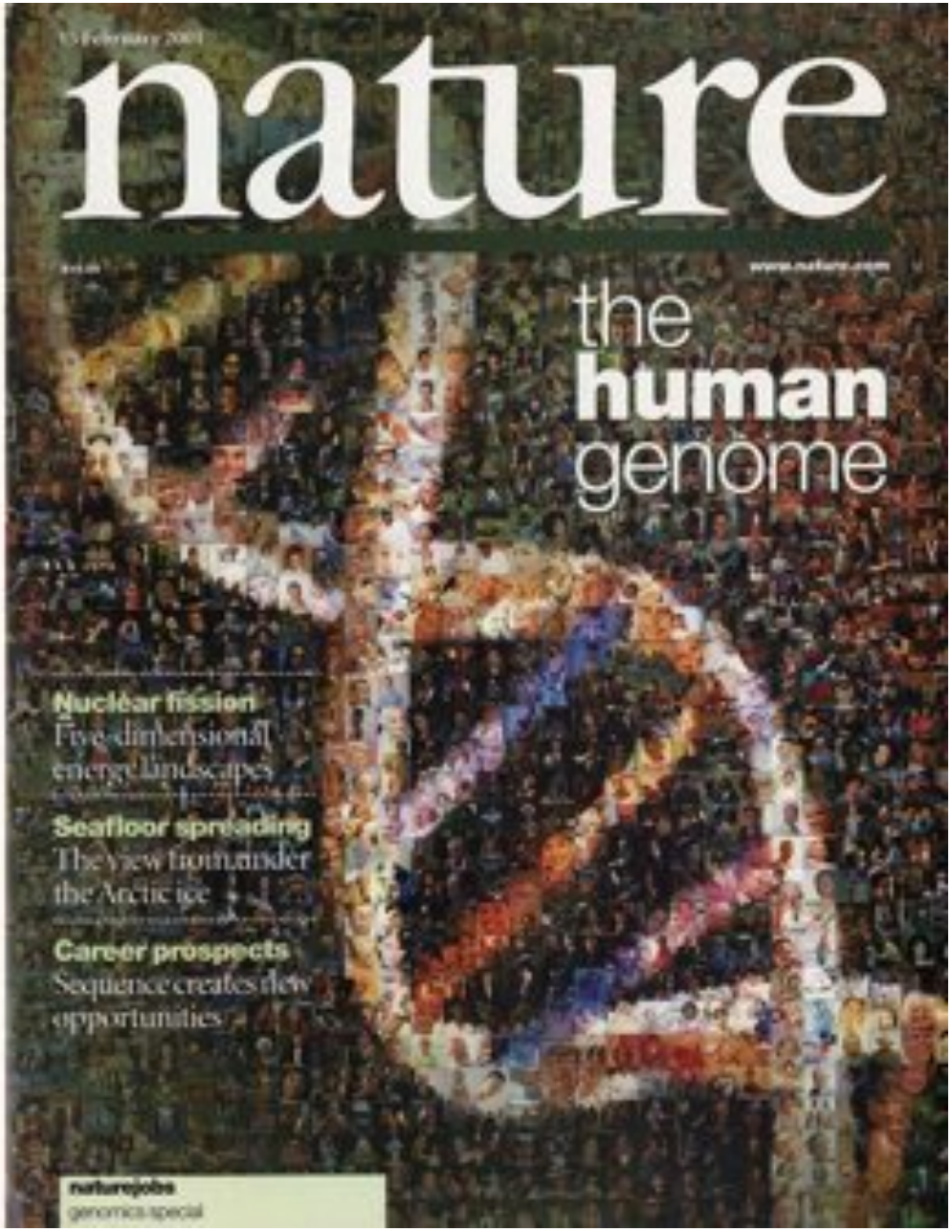
Question 1: de Bruijn graph construction (10 pts)

1a. Draw (by hand or in code) the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ACTA
ATCA
ATCG
ATCT
CATC
CTAC
CTAT
GATC
TACT
TATC
TCAT
TCTA
```

https://github.com/schatzlab/biomedicalresearch2019

Part 1: Recap



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)

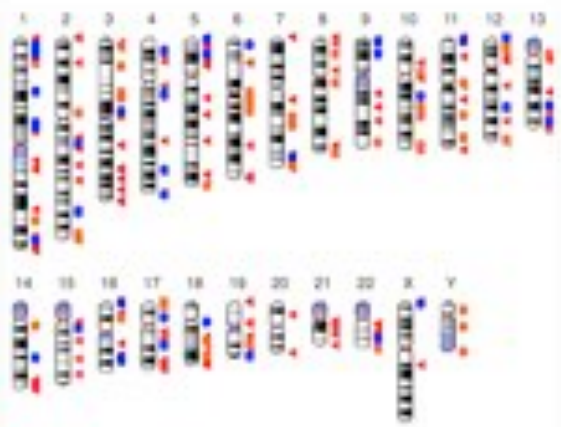
Genome Reference Consortium

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 to
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

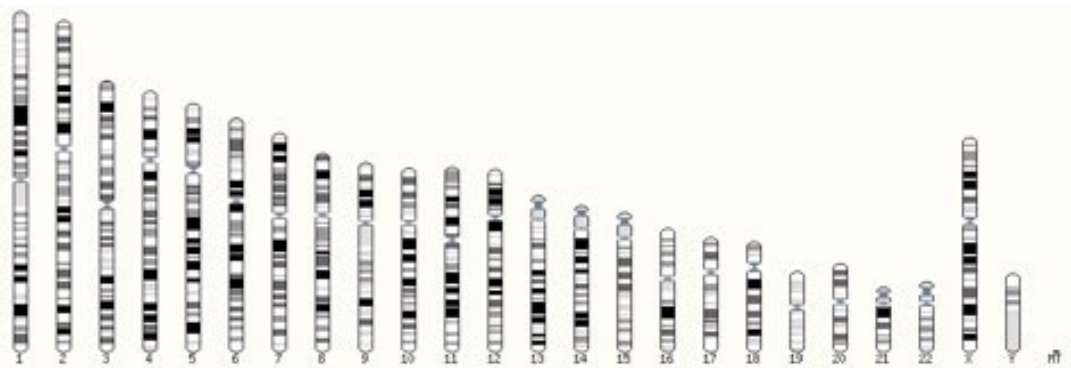
GRCh38.p11

Release date: June 14, 2017
Release type: minor
Release notes: GRCh38.p11 is the eleventh patch release for the GRCh38 reference assembly. No chromosome coordin
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

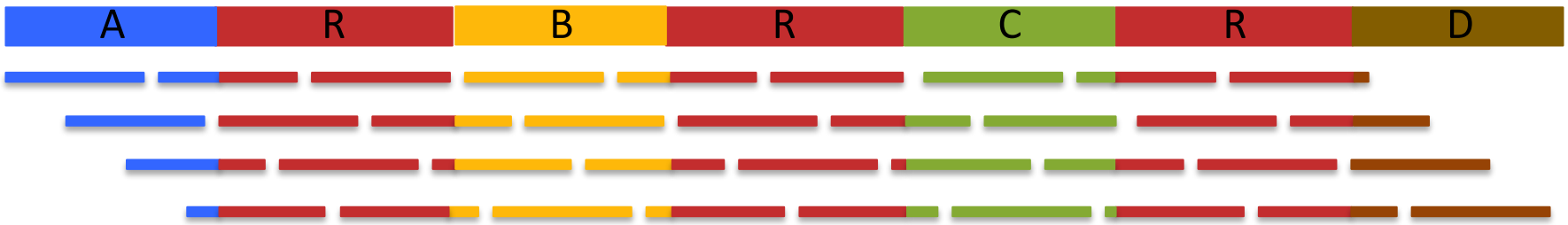
Gene counts (Primary assembly)

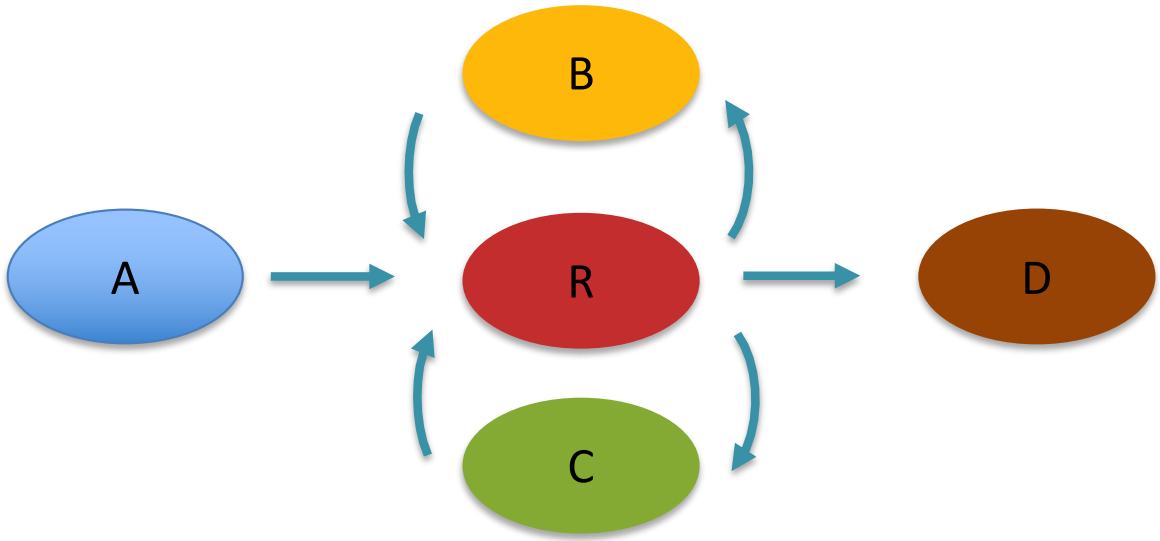
Coding genes	20,454 (incl 660 readthrough)
Non coding genes	23,940
Small non coding genes	4,871
Long non coding genes	16,848 (incl 302 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,204 (incl 8 readthrough)
Gene transcripts	226,950

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

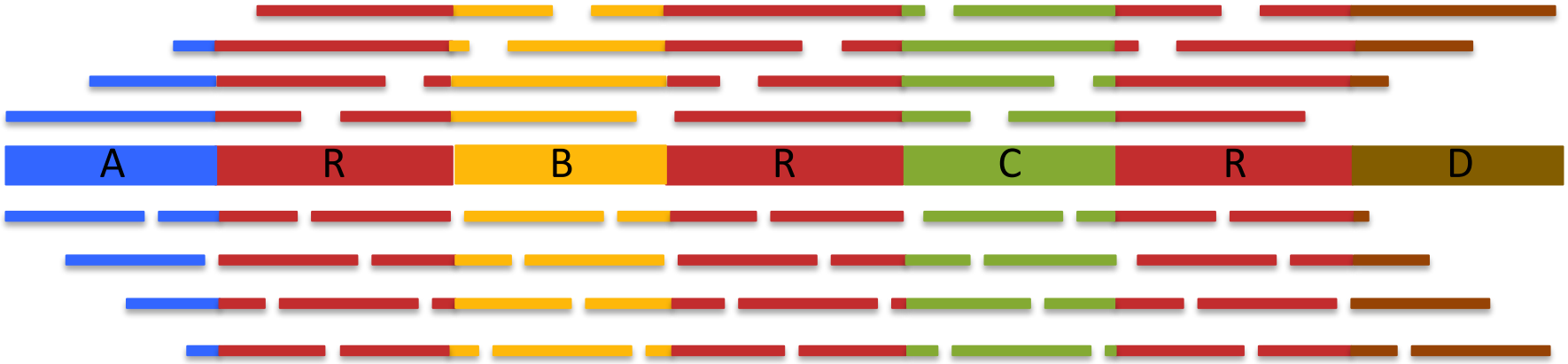
Genomics Arsenal in the Year 2019

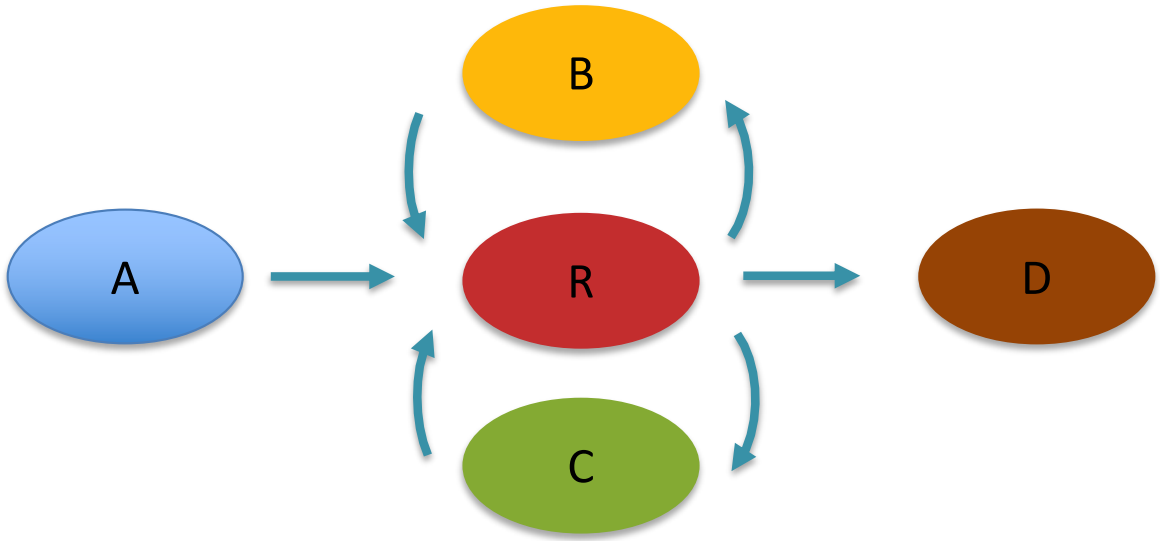
Assembly Complexity



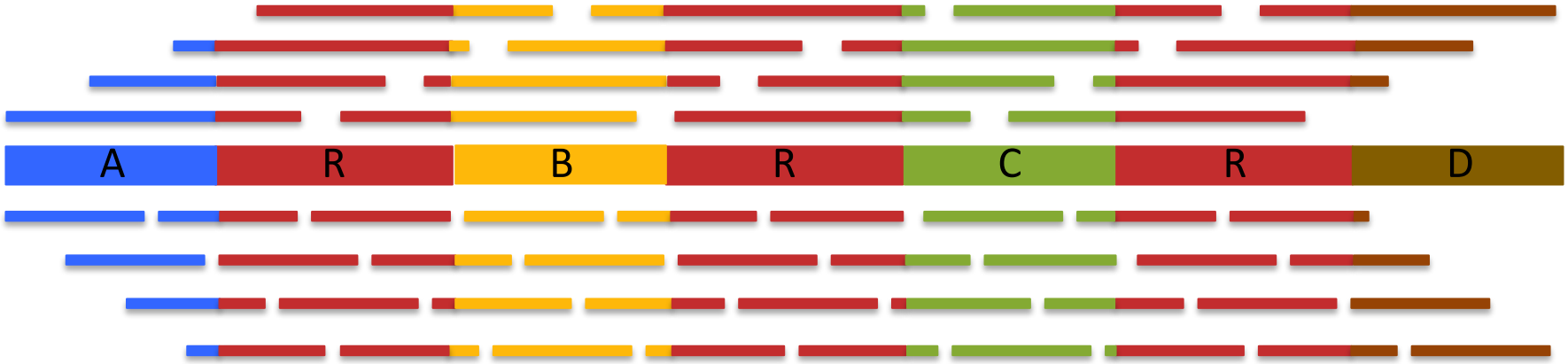


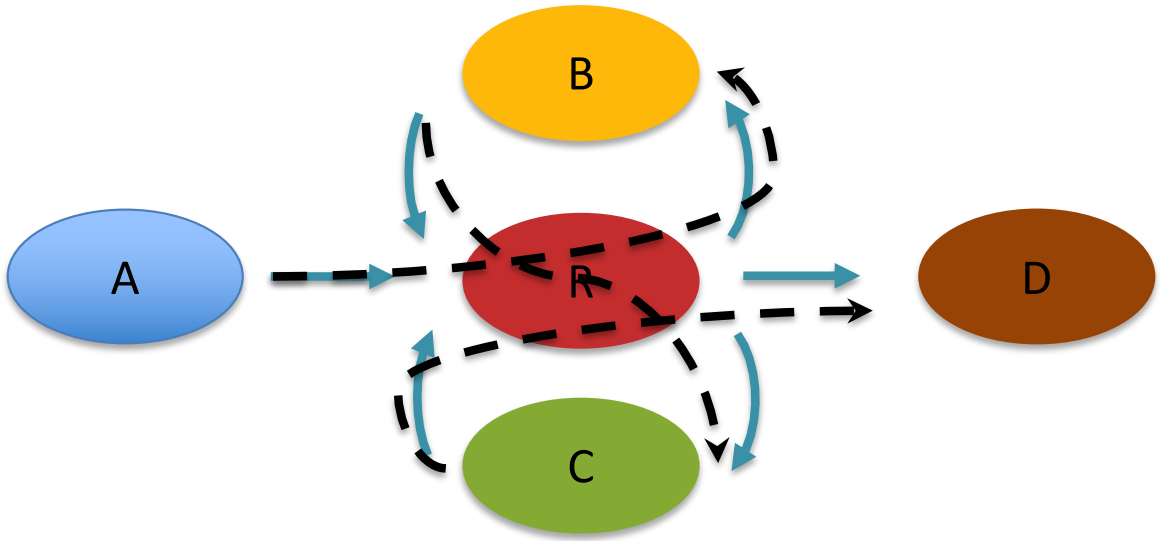
Assembly Complexity



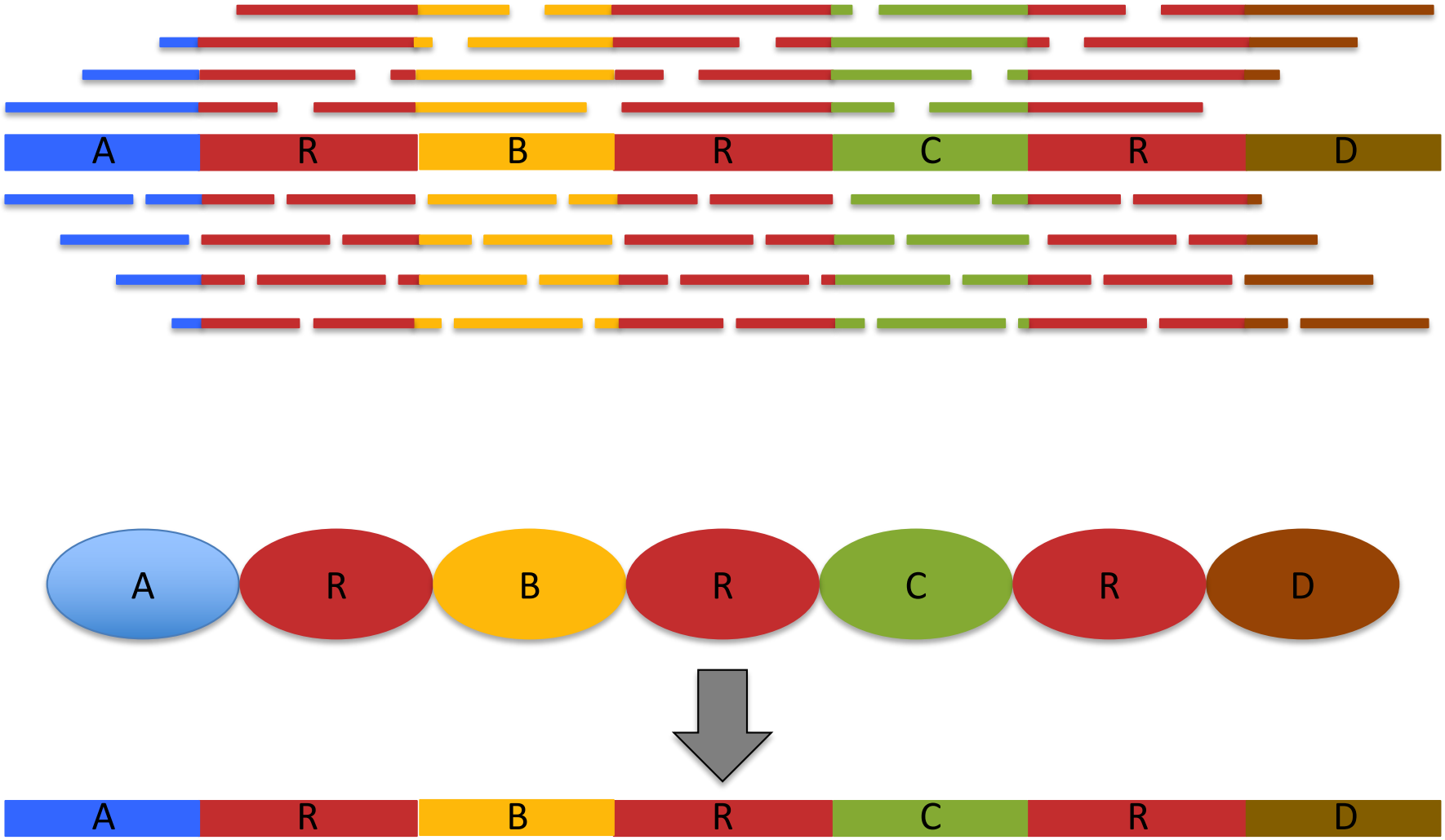


Assembly Complexity





Assembly Complexity

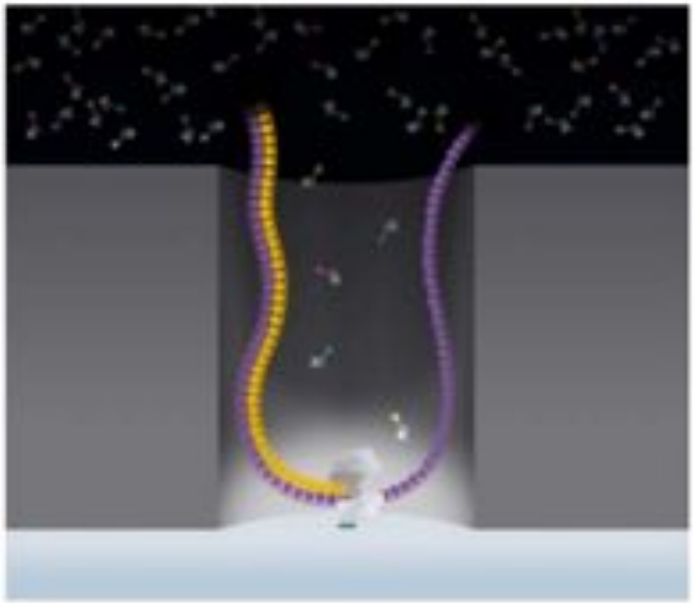


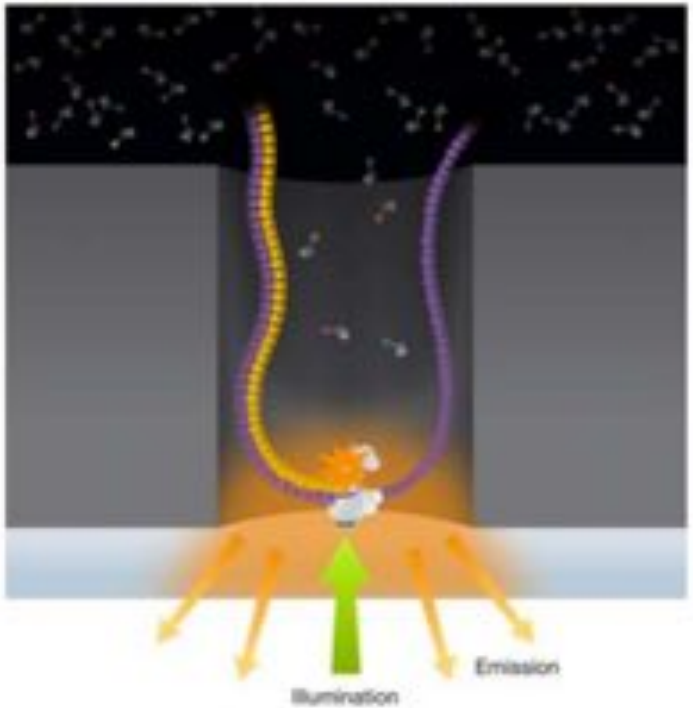
The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

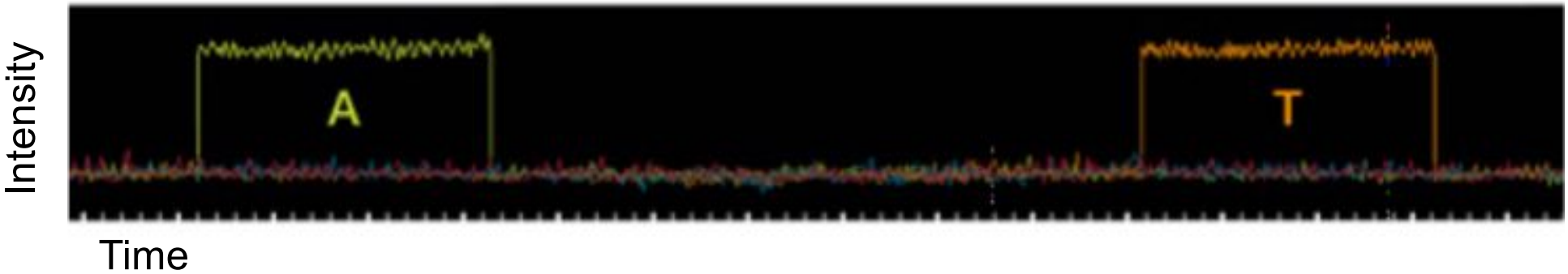
PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

PacBio: SMRT Sequencing

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







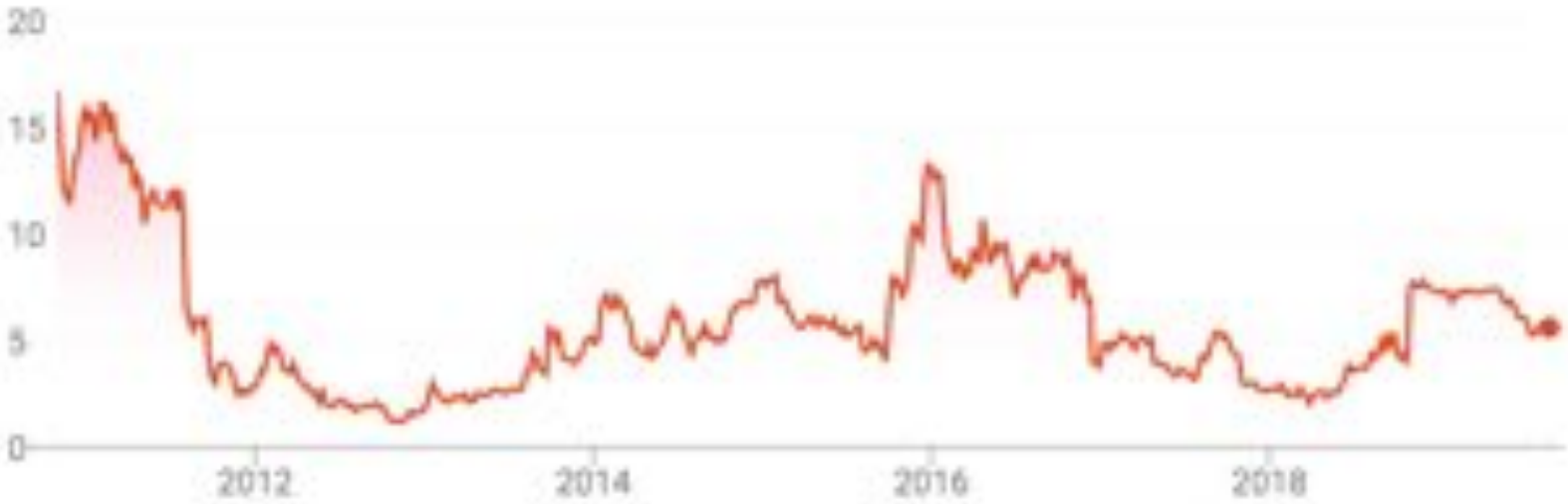
http://www.youtube.com/watch?v=v8p4ph2MAvI

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

Market Summary > Pacific Biosciences of California
NASDAQ: PACB

Following

5.66 USD +0.11 (1.89%)

DNA sequencing giant Illumina just bought rival Pac Bio for $1.2 billion — here's why

- Illumina just paid $1.2 billion for Pacific Biosciences, to help it retain its dominant position in the DNA sequencing space, biotech experts say.
- Illumina, which is valued at more than $45 billion, makes the machines that companies from 23andMe to Ancestry rely on for their sequencing.

Christina Farr | @chrissyfarr

Published 5:13 PM ET Thu, 1 Nov 2018

CNBC

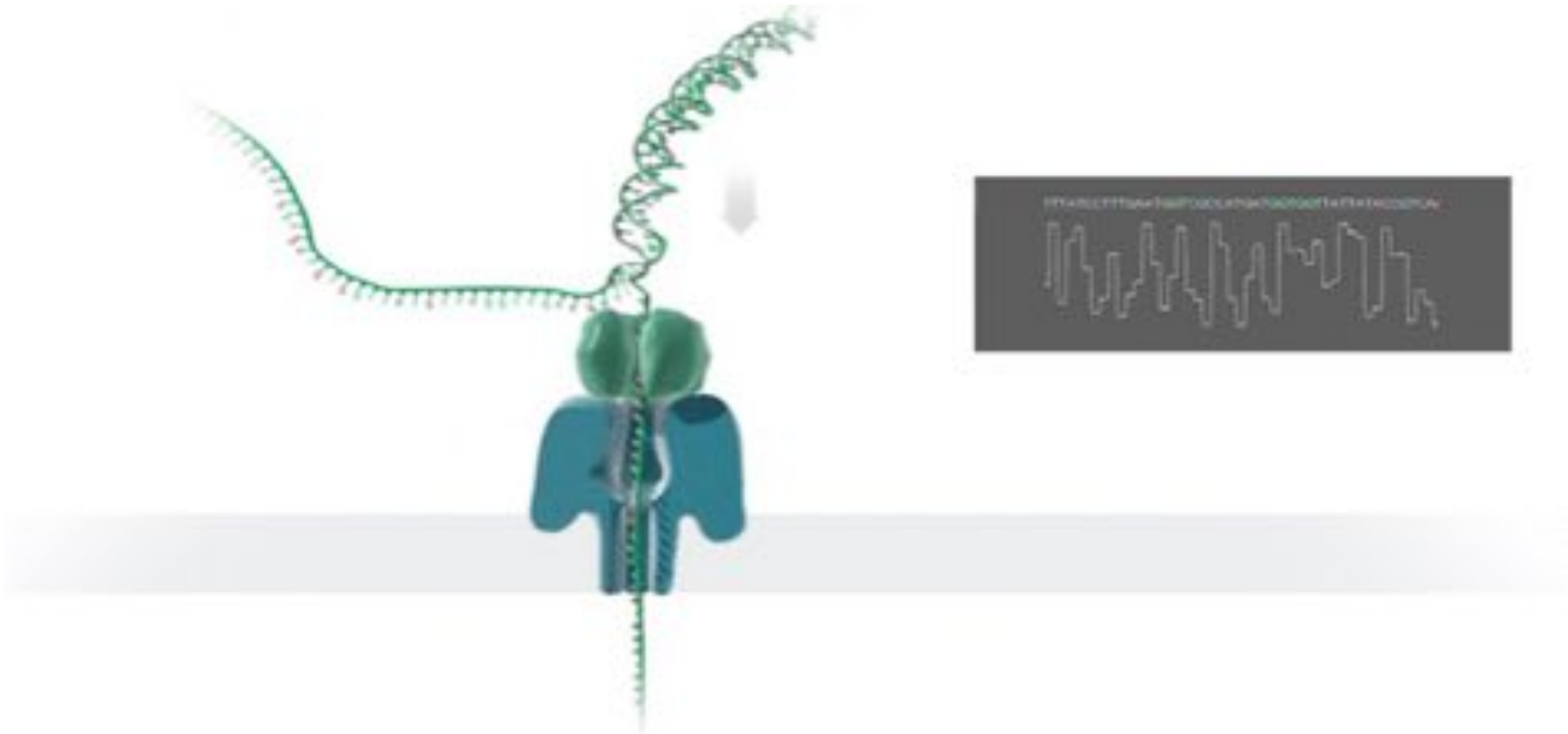
P/E ratio

→ Financial news, comparisons and more

Oxford Nanopore Technologies (ONT)

Nanopore Sequencing

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



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

nanoporetech.com/applications/dna-nanopore-sequencing

More Throughput

MinION

Quick Mobile Sequencing
$1k / instrument
5-10 GB / day

PromethION

High Throughput Desktop Sequencer
$75k / instrument
>>1TB / day

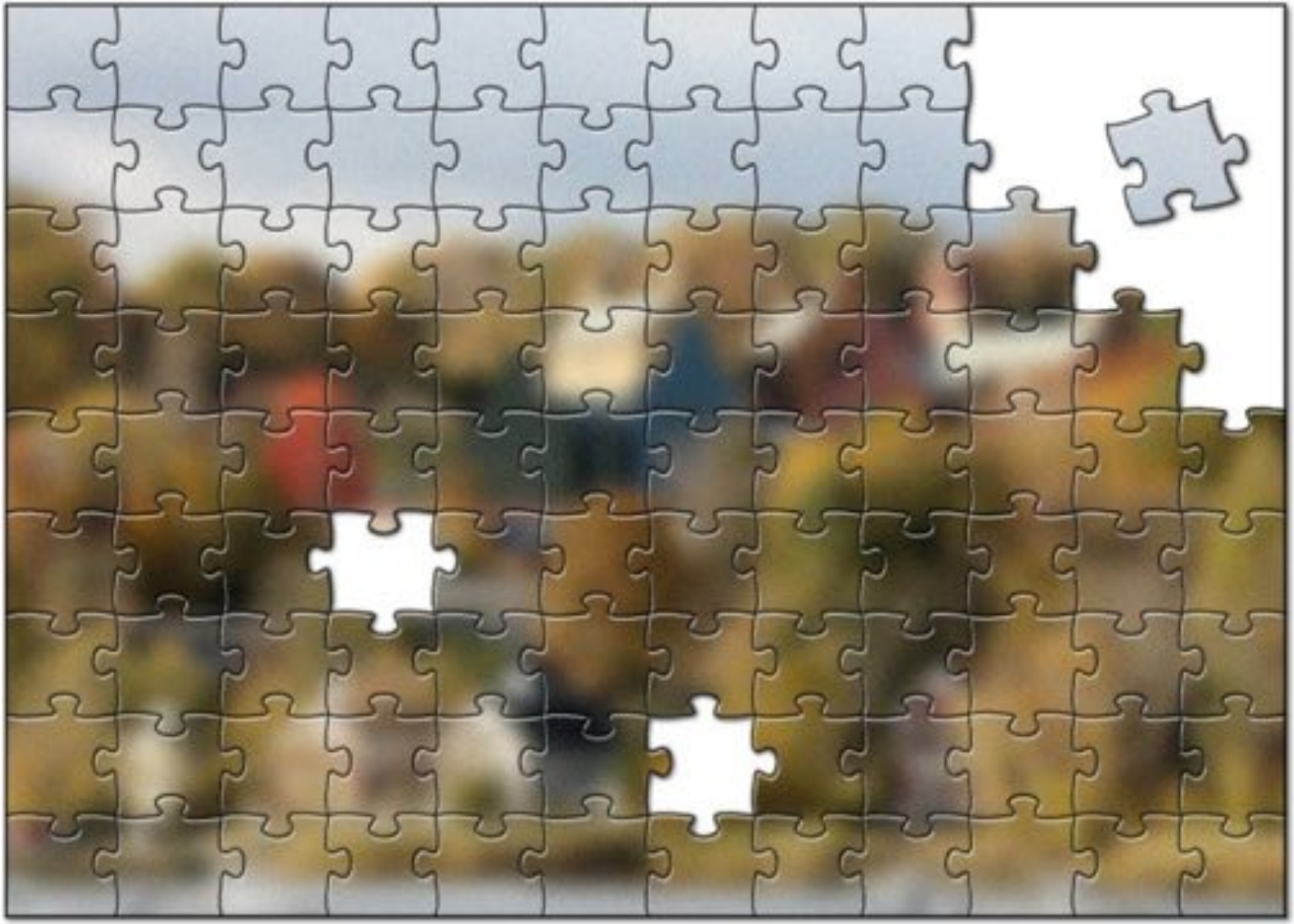
Oxford Nanopore sets sights on IPO

4th April 2019 Callum Cyrus

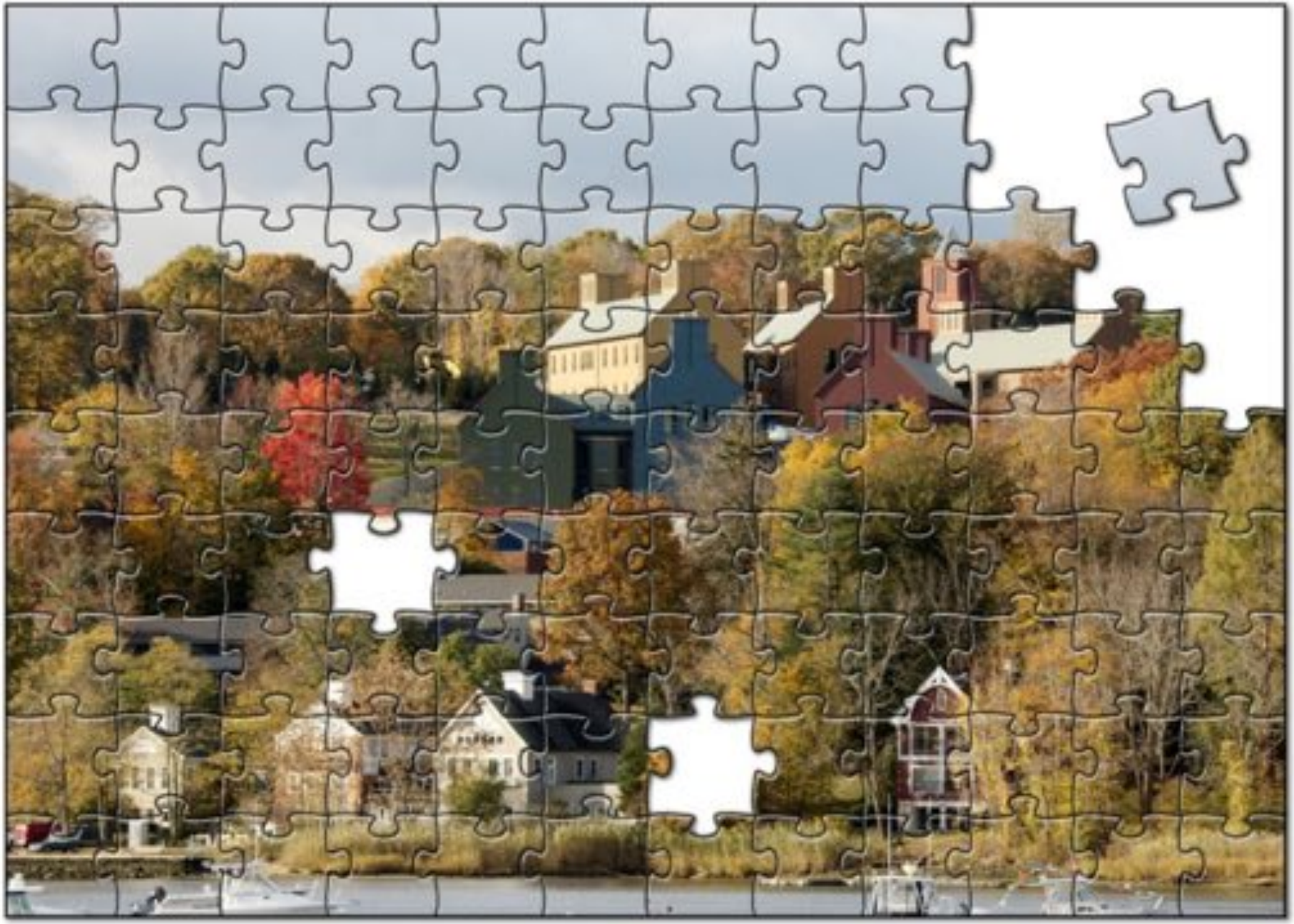
The Oxford University genetic sequencing spinout is reportedly mulling an IPO that would provide exits to investors including commercialisation firm IP Group.

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...

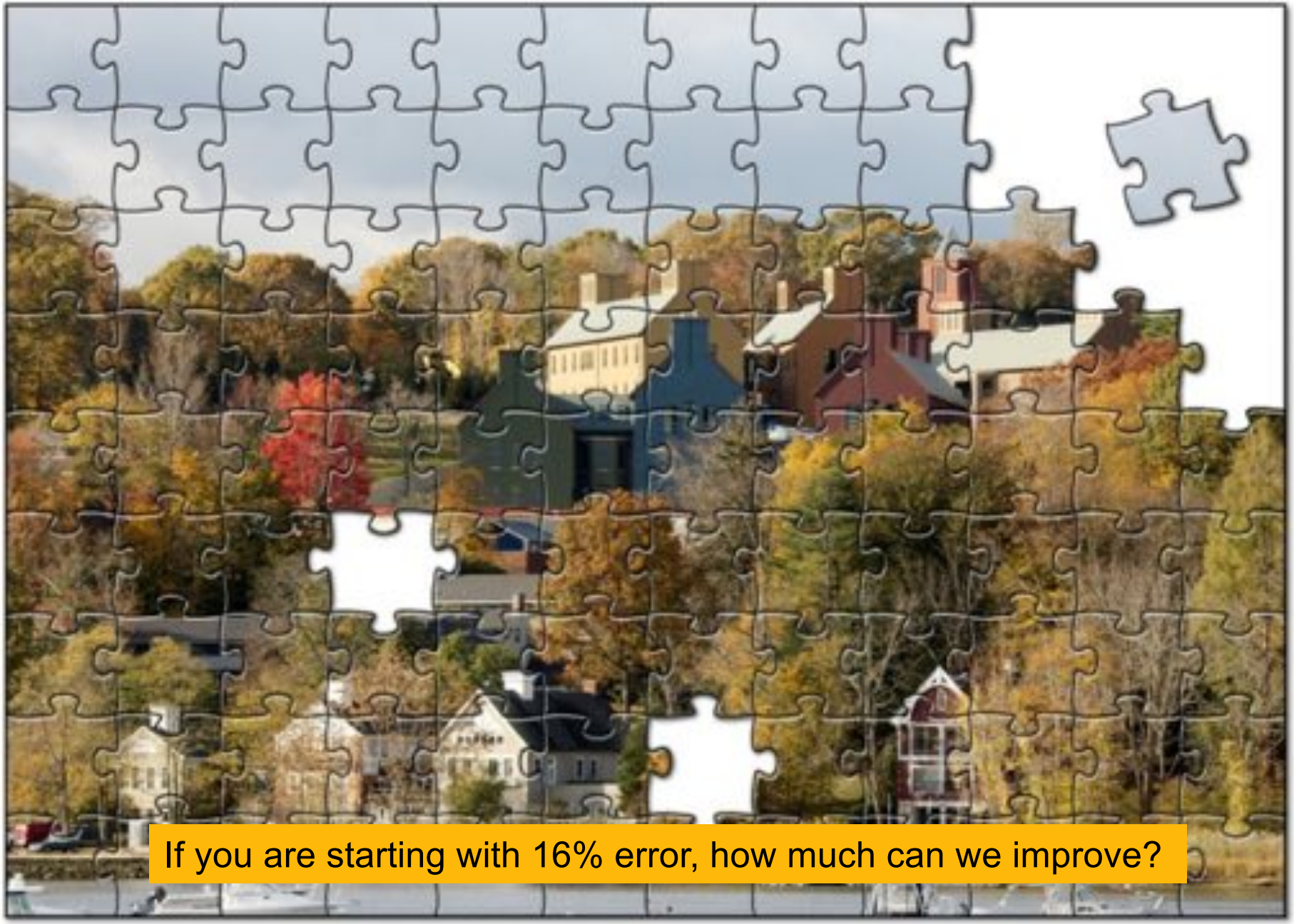
Single Molecule Sequences



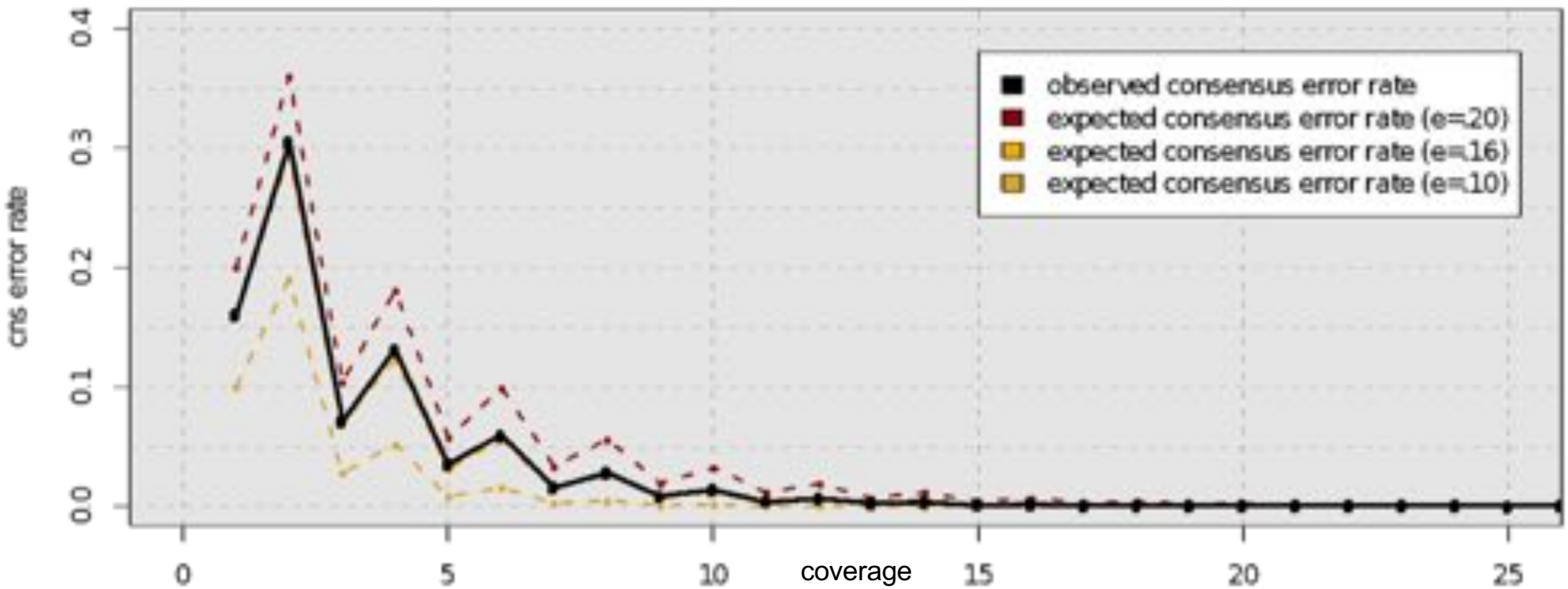
“Corrective Lens” for Sequencing



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage

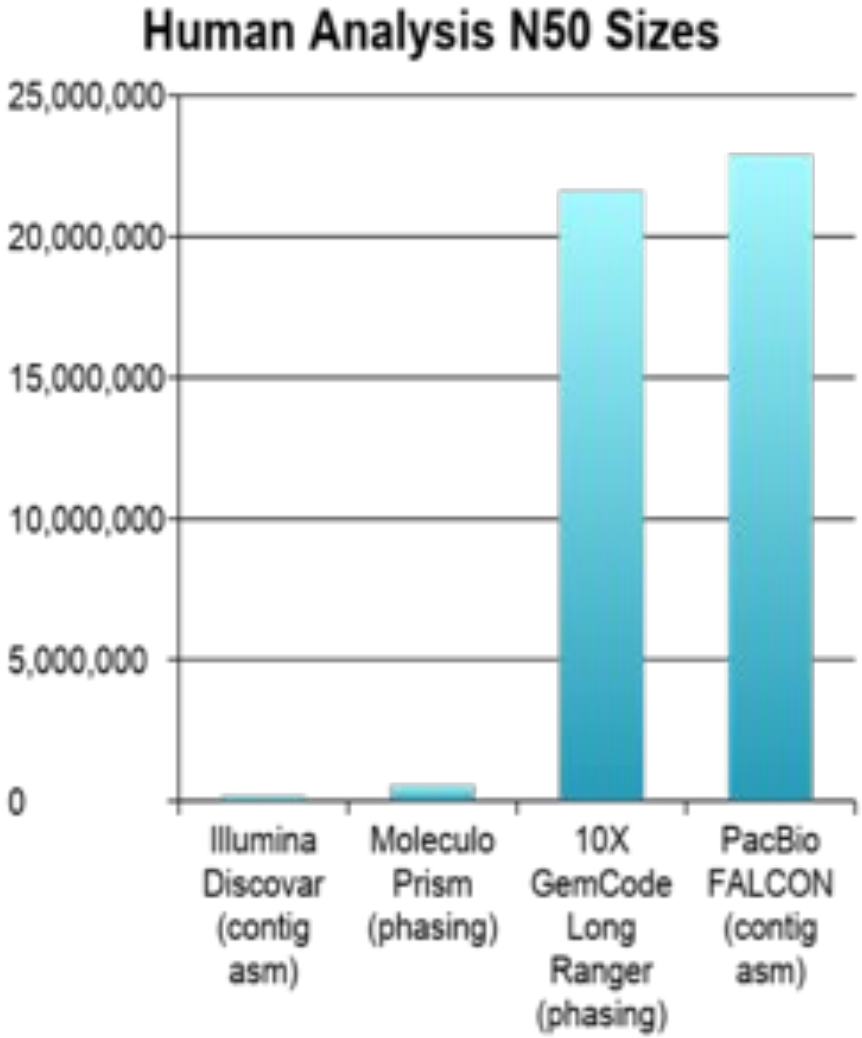


Coverage can overcome random errors

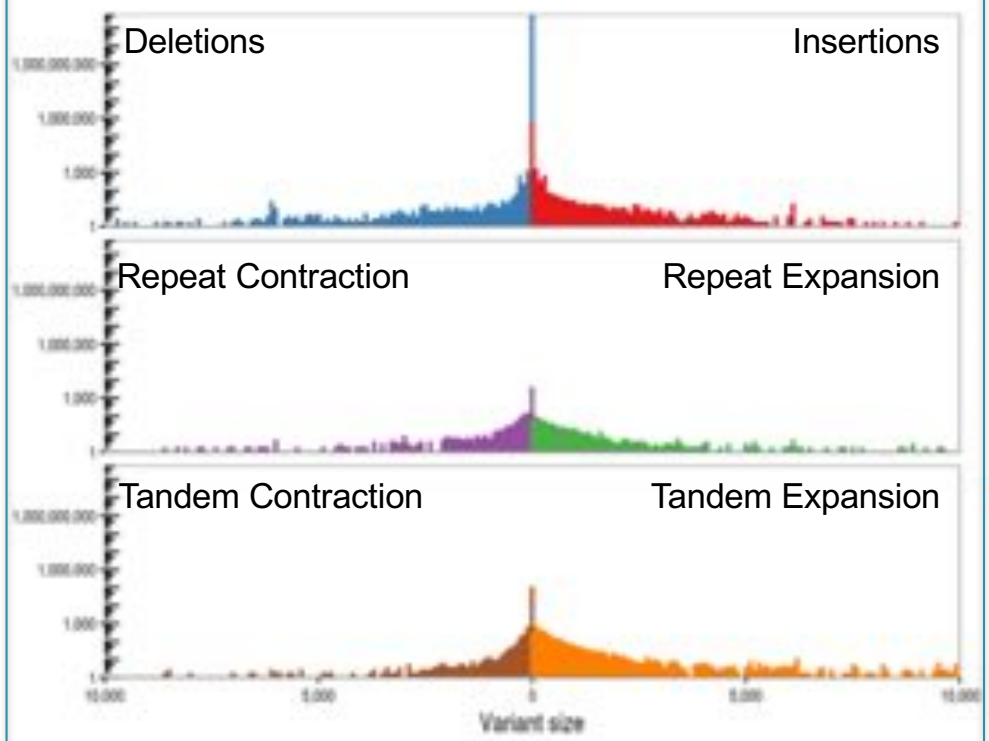
- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

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

Recent Long Read Assemblies

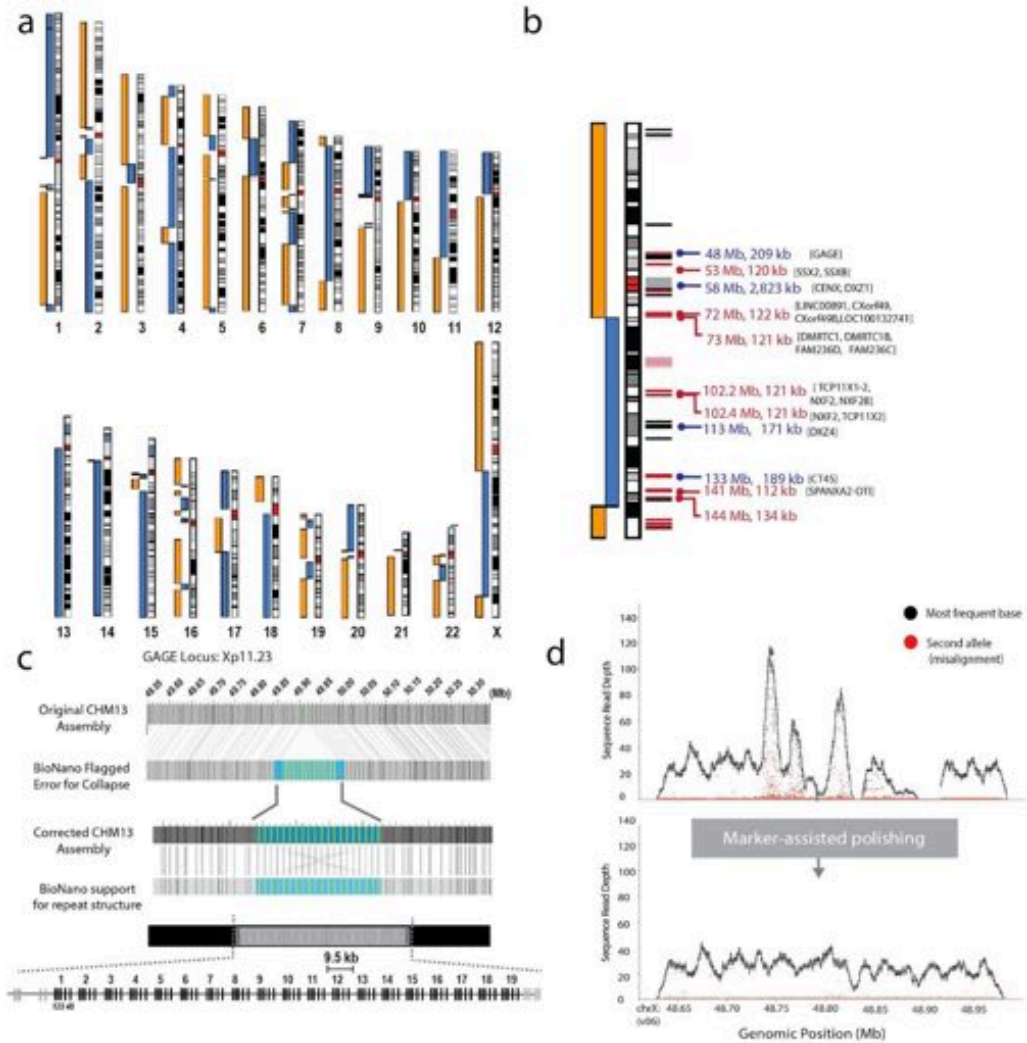


Third-generation sequencing and the future of genomics
Lee et al (2016) *bioRxiv*
doi: http://dx.doi.org/10.1101/048603



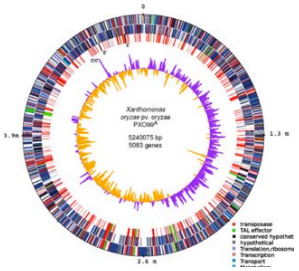
Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad & Schatz (2016) *Bioinformatics*.
doi: 10.1093/bioinformatics/btw369

First Telomere-to-Telomere Human Chromosome



Telomere-to-telomere assembly of a complete human X chromosome
Miga et al. (2019) bioRxiv. https://doi.org/10.1101/735928

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

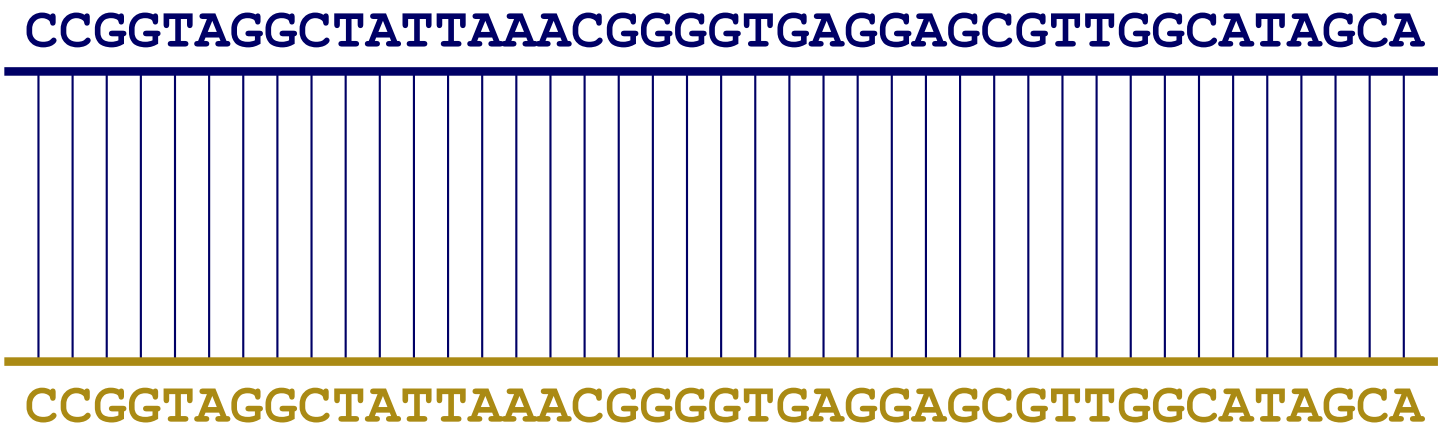
Part 2: Whole Genome Alignment

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
NHGRI

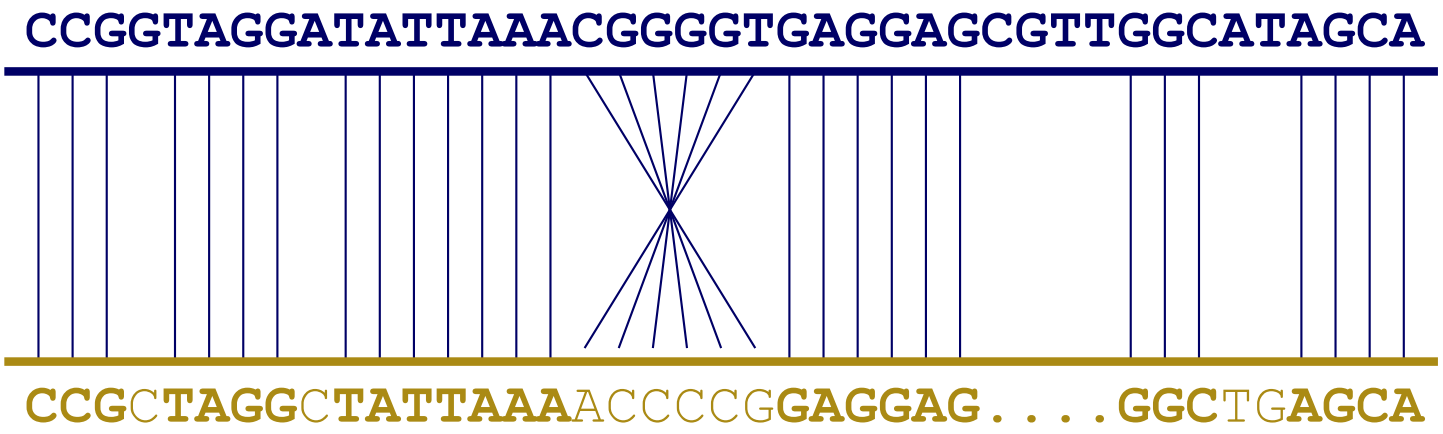
Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



Not so fast...

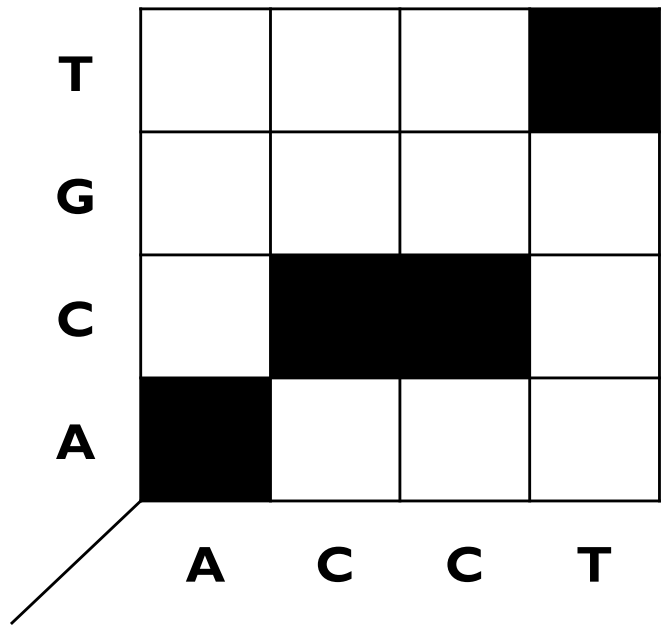
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



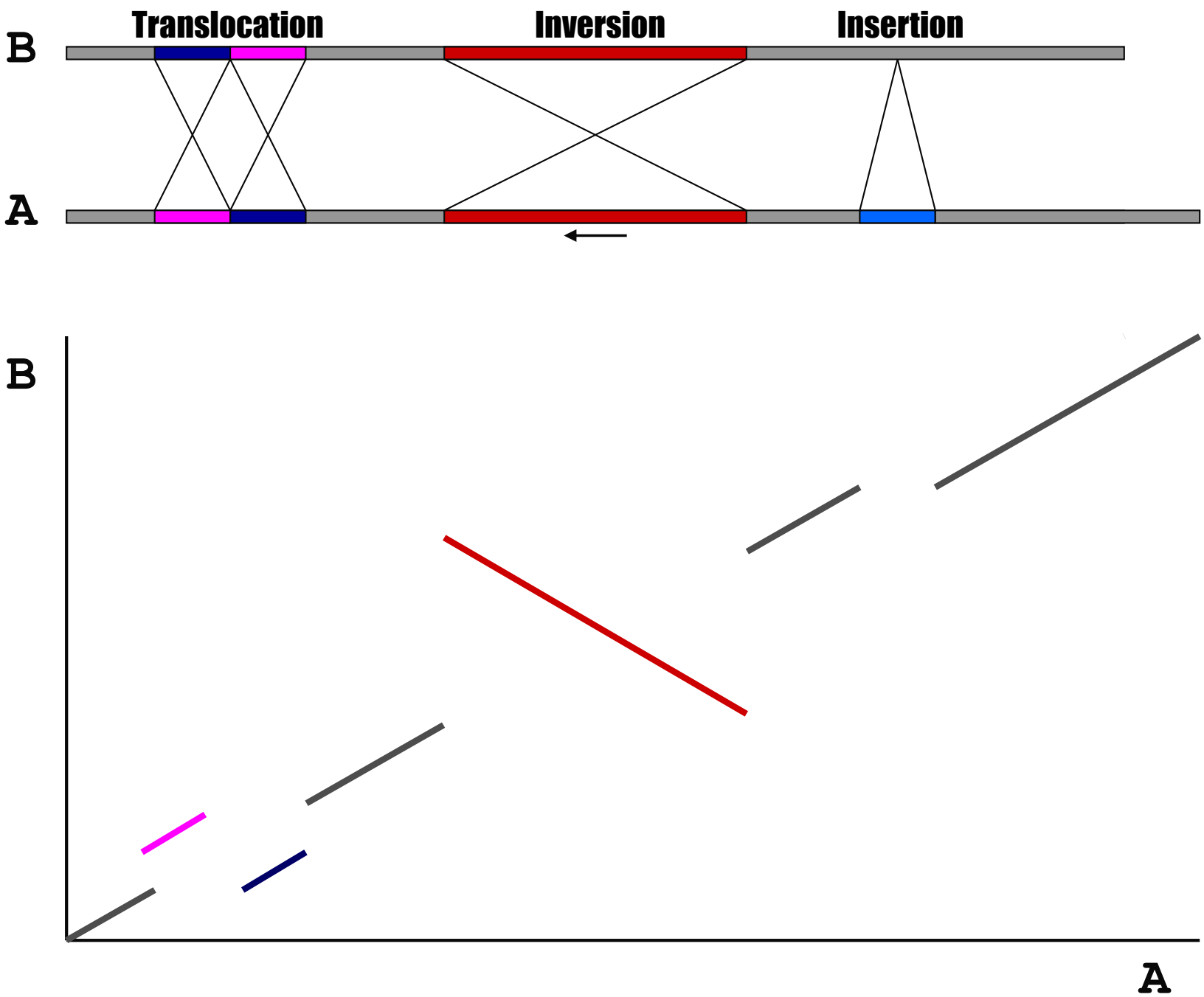
WGA visualization

- How can we visualize *whole* genome alignments?

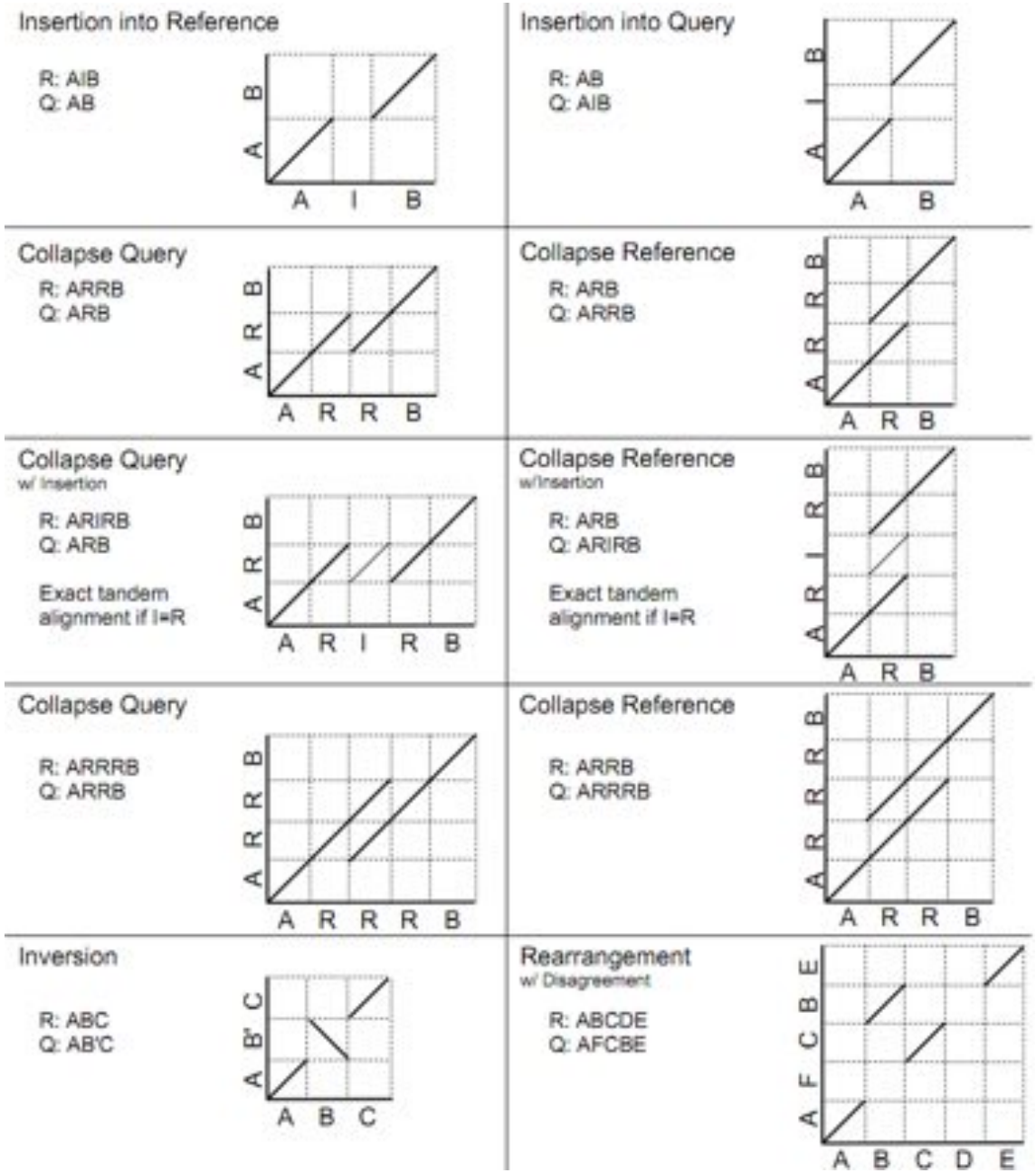
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - A perfect alignment between *A* and *B* would completely fill the positive diagonal

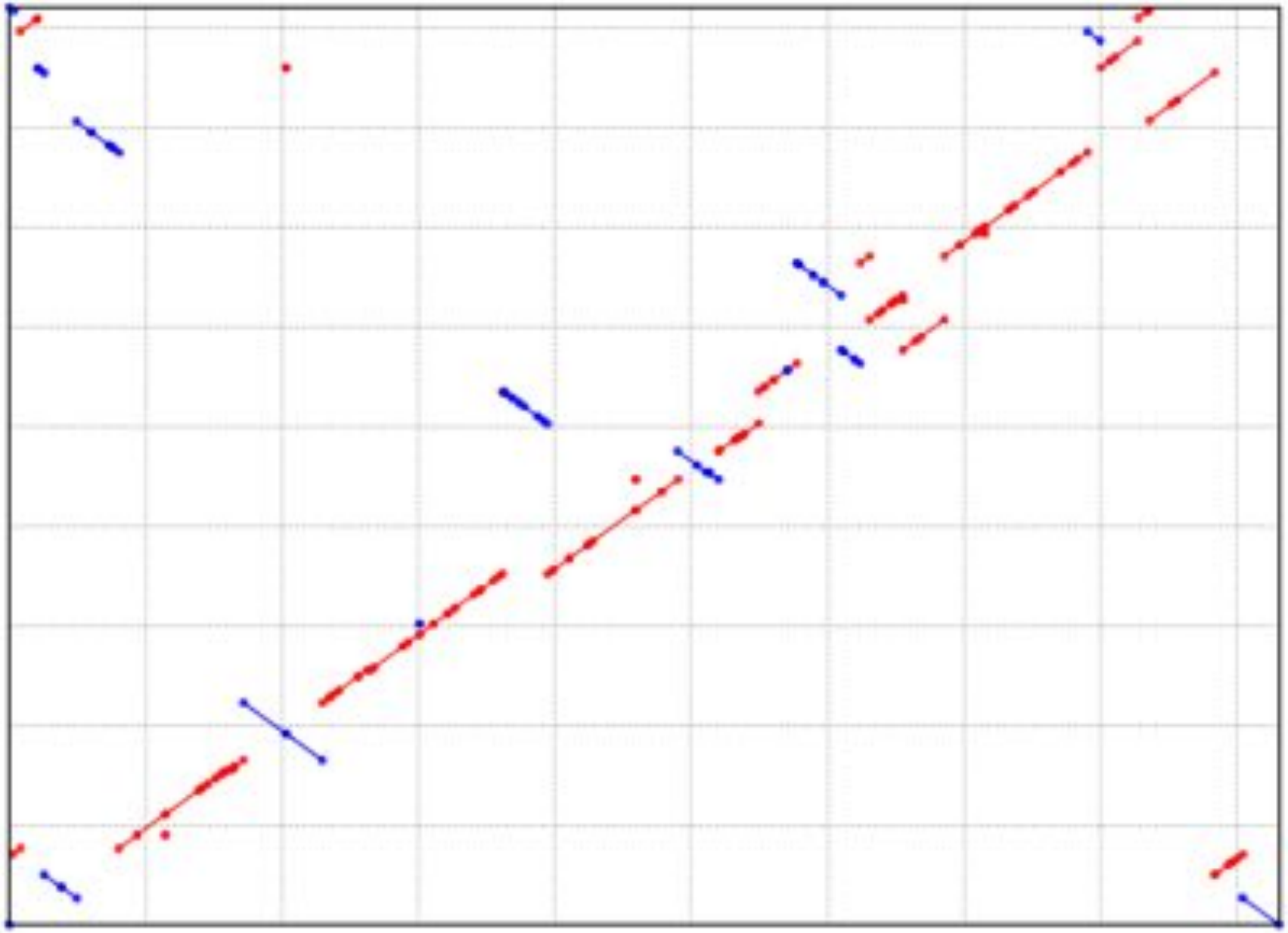

Translocation
Inversion
Insertion
B
A
B
A

SV Types



- 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



Alignment of 2 strains of *Y. pestis*

http://mummer.sourceforge.net/manual/

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Work on HW2