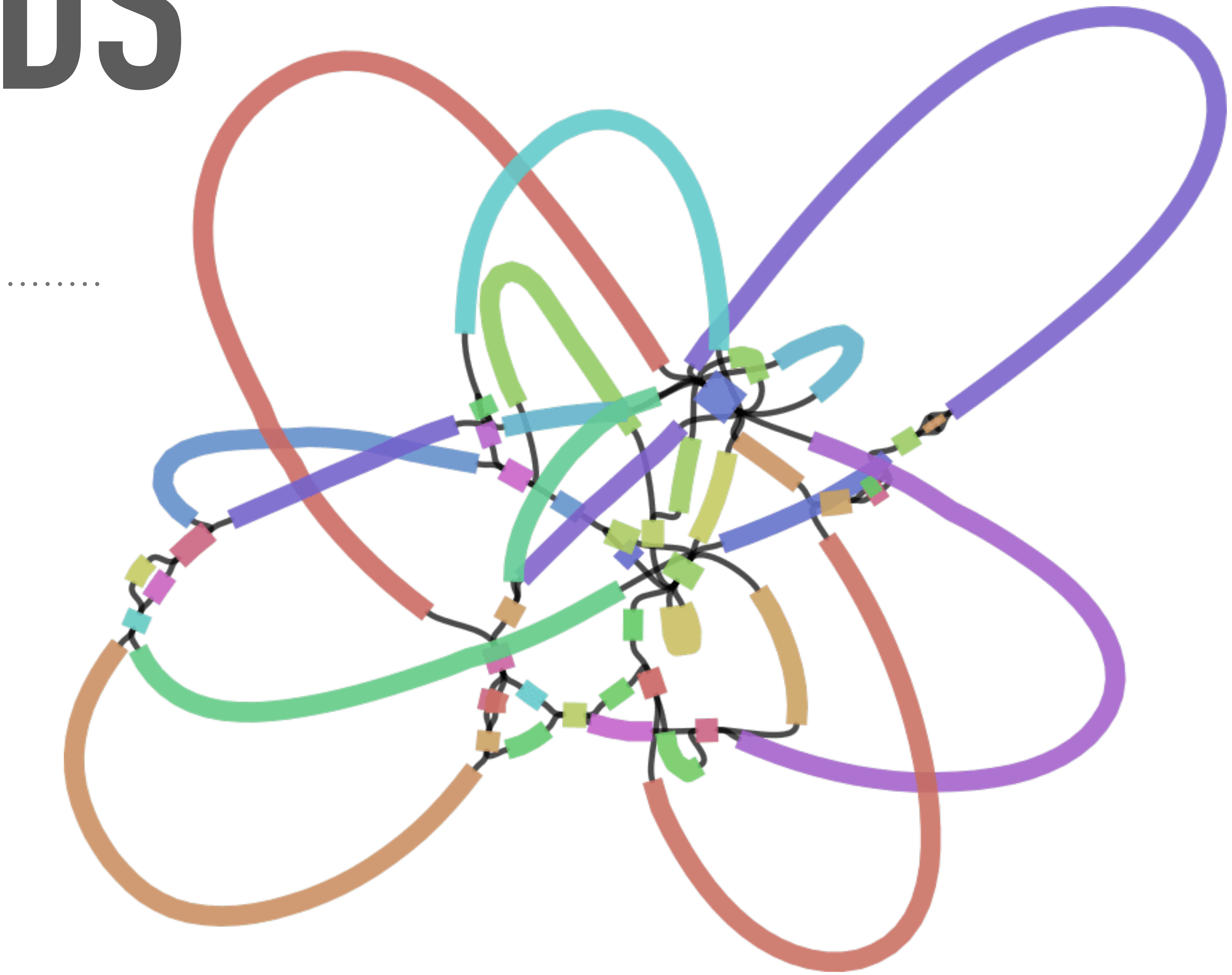
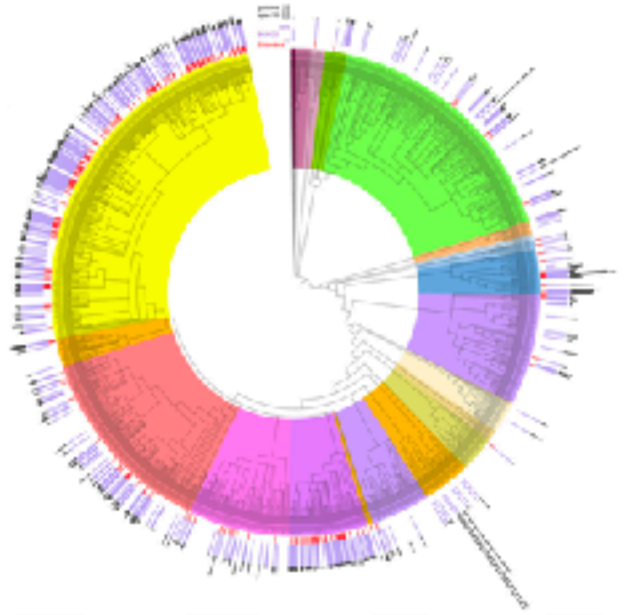
GENOME ASSEMBLY WITH LONG READS

Marcela Uliano-Silva



WHO AM I?

- **Senior Bioinformatician Wellcome Sanger Institute - Darwin Tree of Life Project. Tree of Life Assembly Team (ToLA)**
- **Churchill College Postdoctoral By-Fellow, University of Cambridge**



➤ BSc in Biology (2010) - UFSC, Brazil

➤ MSc in Biophysics (2013) - IBCCF UFRJ, Brazil

➤ PhD in Biophysics (2017) - IBCCF UFRJ, Brazil

➤ Horizon2020 Marie Curie PostDoc Fellow (2017-2019), IZW BeGenDiv (Germany)

➤ TED Fellow

wellcome sanger institute

Darwin

TREE

of

LIFE

Tree of Life: Major Projects

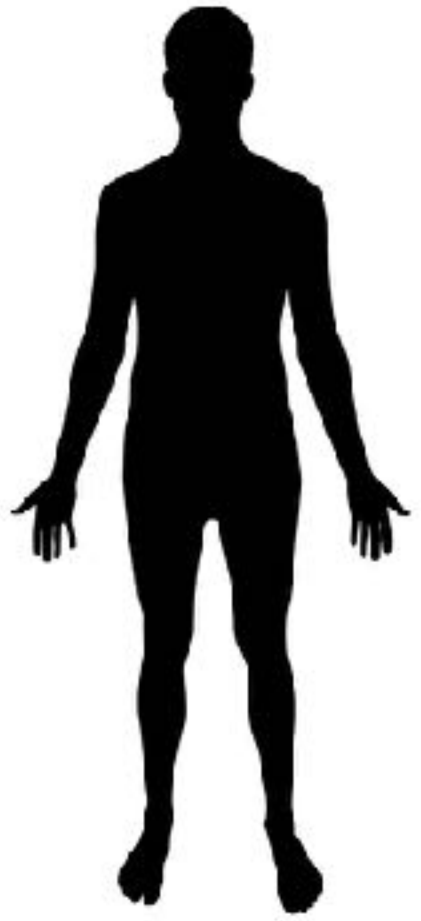
Collaborating widely to deliver across diversity

- ★ **Darwin Tree of Life Project**
    - 70,000 species from Britain and Ireland [Phase 1: 2,000 species]
- ★ **Aquatic Symbiosis Genomics**
    - 1,000 species (500 symbiotic systems) from marine and freshwater
- ★ **Vertebrate Genomes Project**
    - Realising VGP Phase 1 (ordinal - 260 species) and Phase 2 (family) goals
- ★ **European Reference Genome Atlas**
    - Sequencing the genomes of all species in the European continent - Pilot 25 species
- ★ **Earth BioGenome Project**
    - Working to deliver Phase 1 (family) goals, and to "*sequence all life for the future of life*"

wellcome sanger institute

Genome assembly: what is my goal?

- Understand variation in populations (disease-related SNPs etc…)



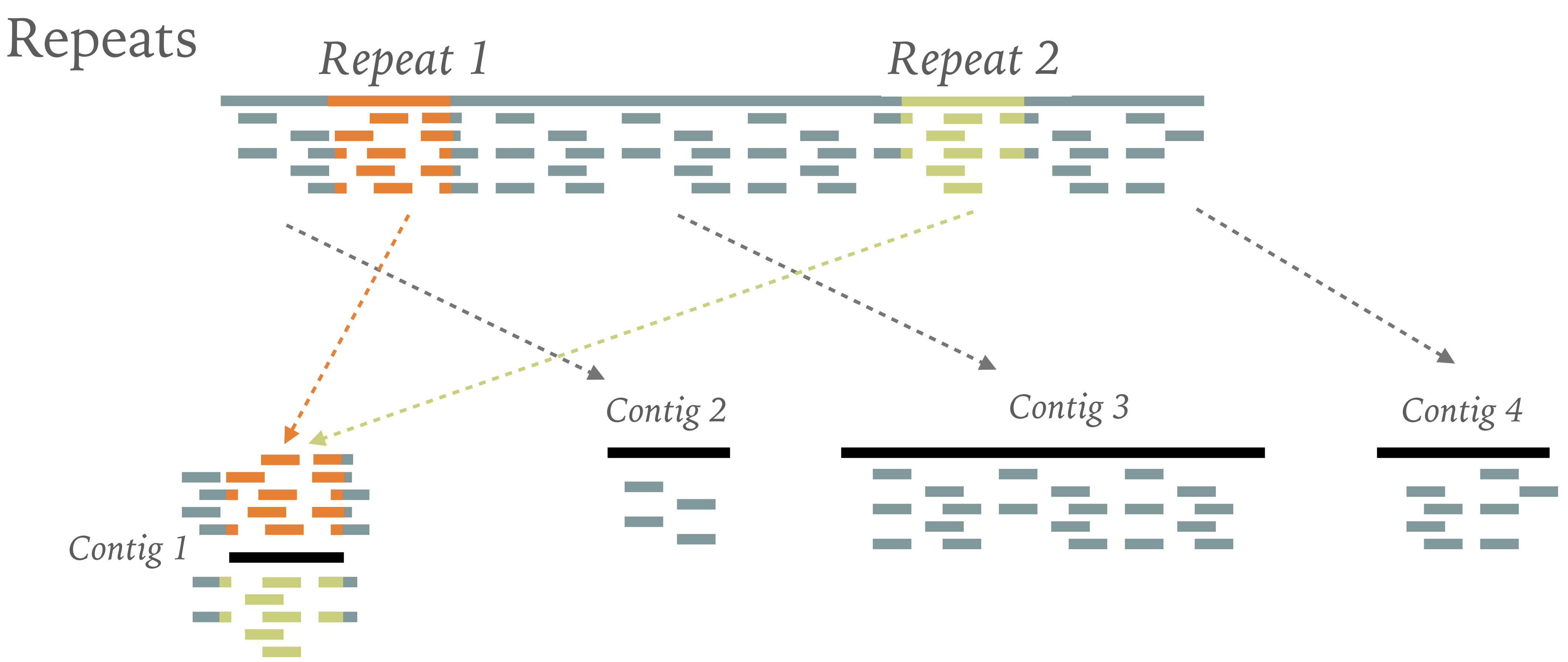
Genome re-sequencing
Assembly by mapping to a reference

- Study the molecular profile of a species never before sequenced (evolutionary studies etc..)

De novo assembly

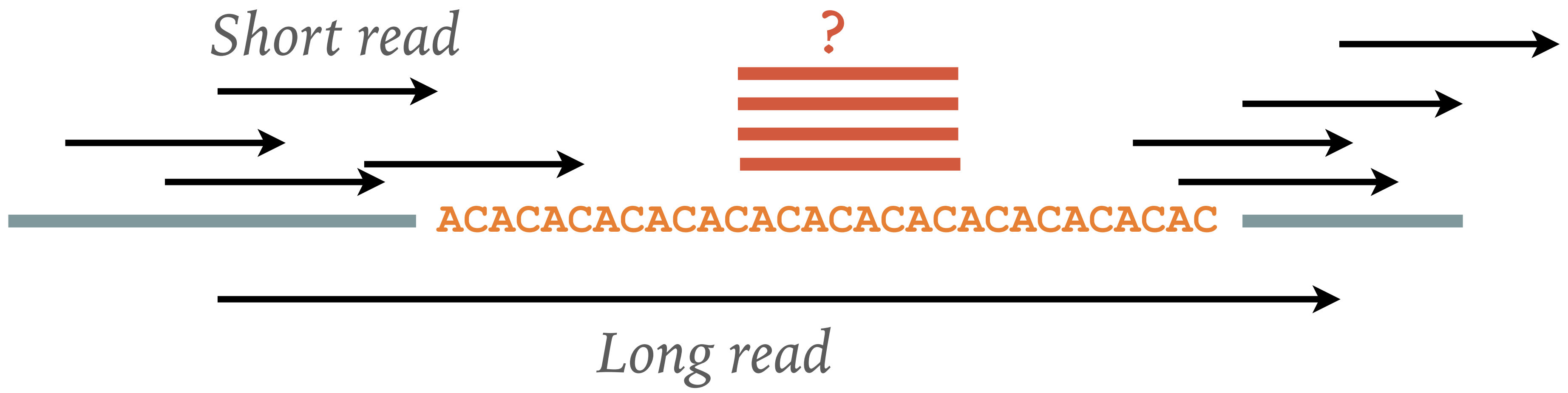
Hurdles

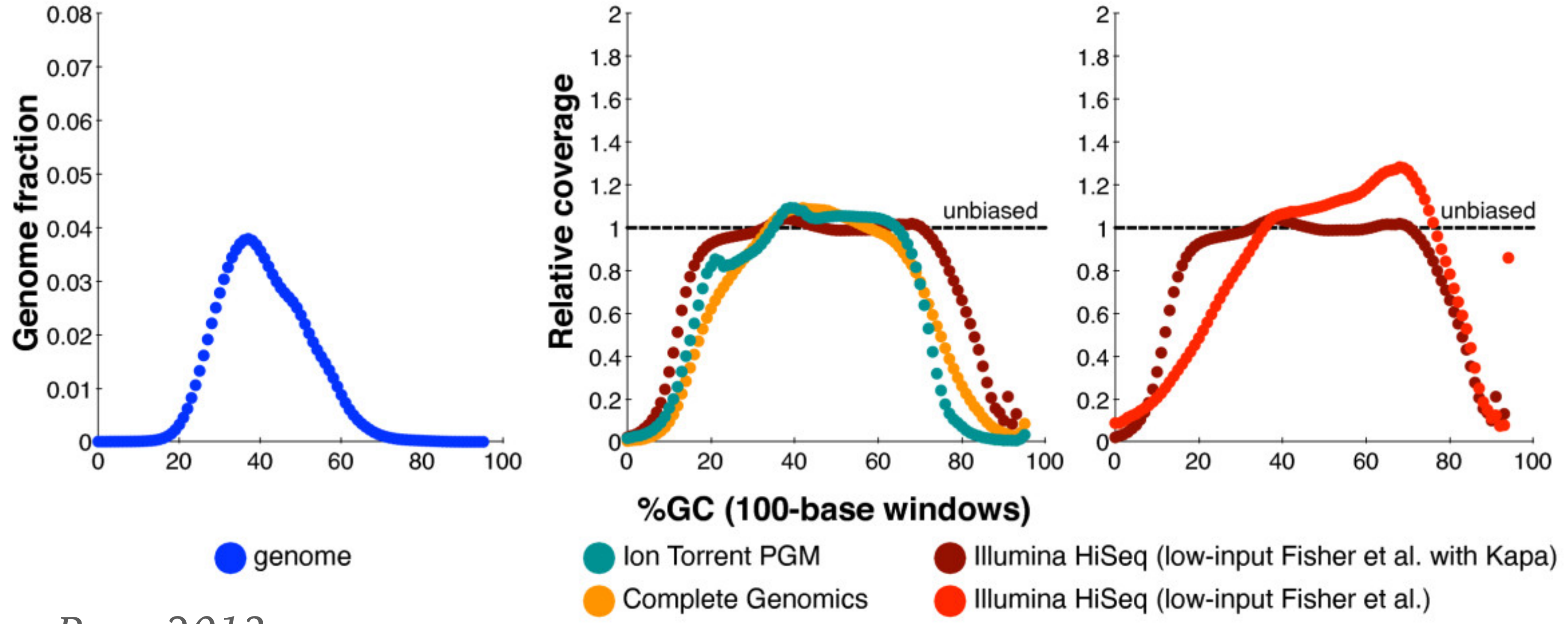
- Heterozygosity
- Sequencing errors
- Repeats
- Low complexity genomic regions
- Base composition and sequencing bias



The repeated element is collapsed into a single contig

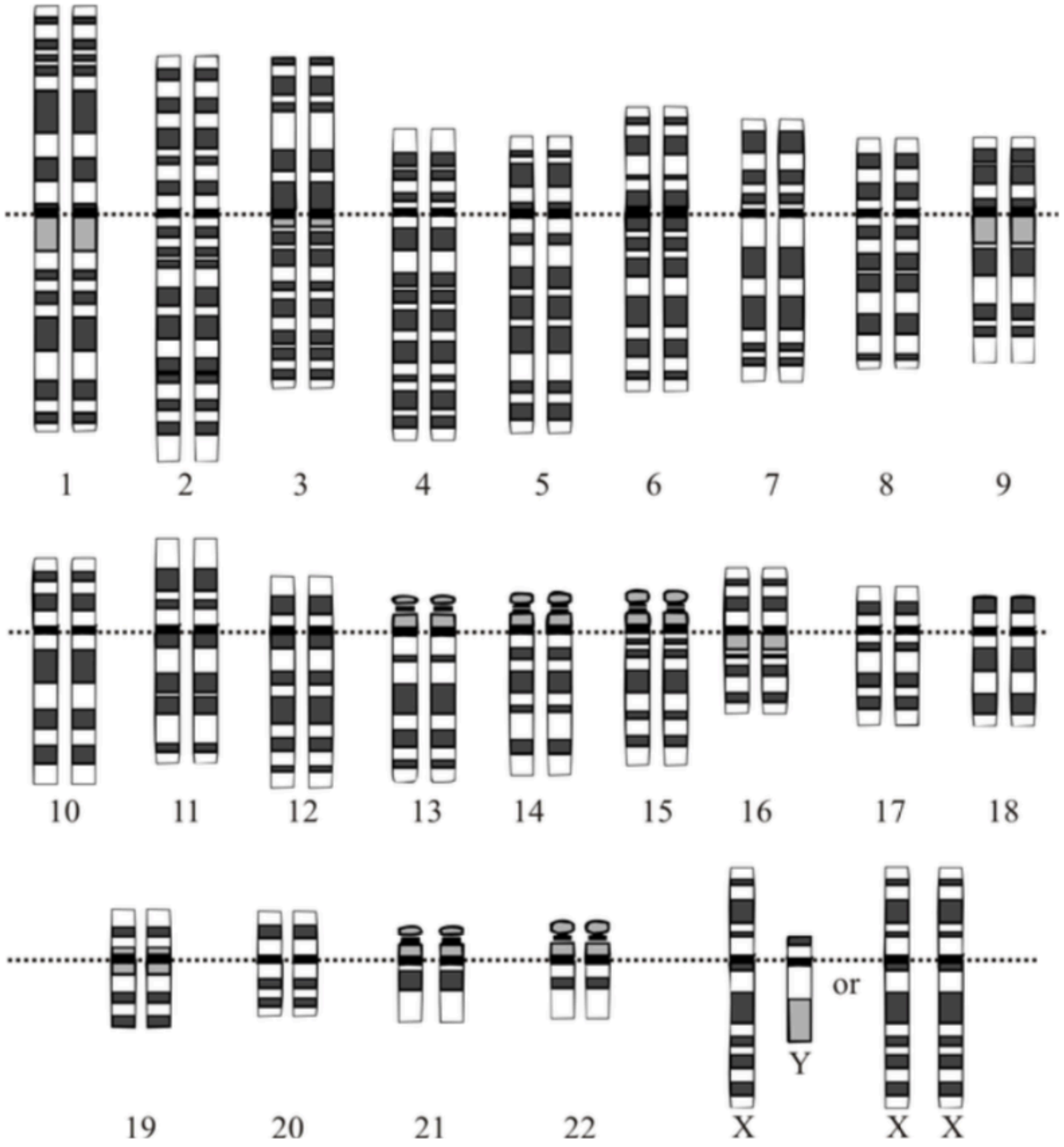
Adapted from Torsten Seemann (2014 talk)





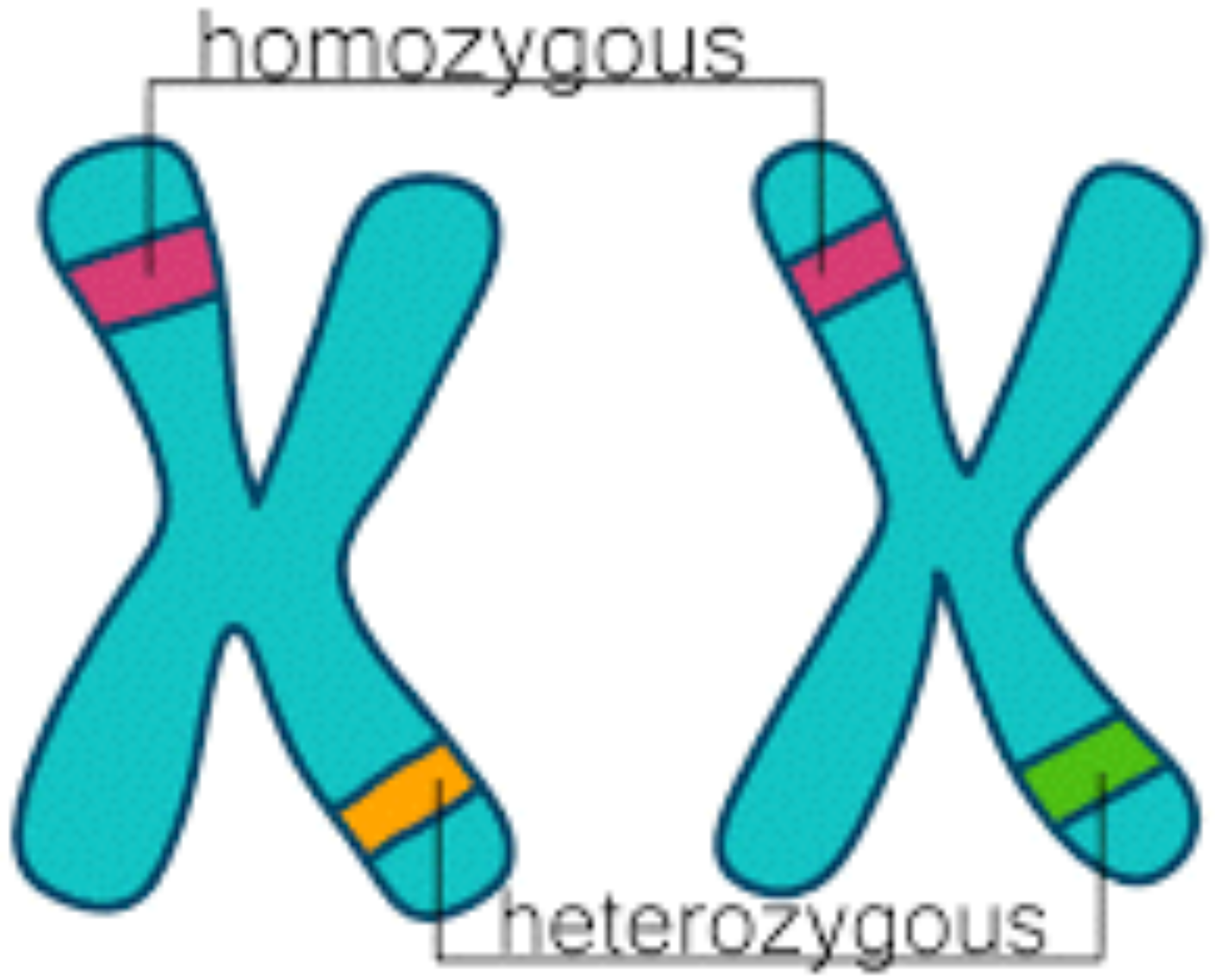
Ross, 2013

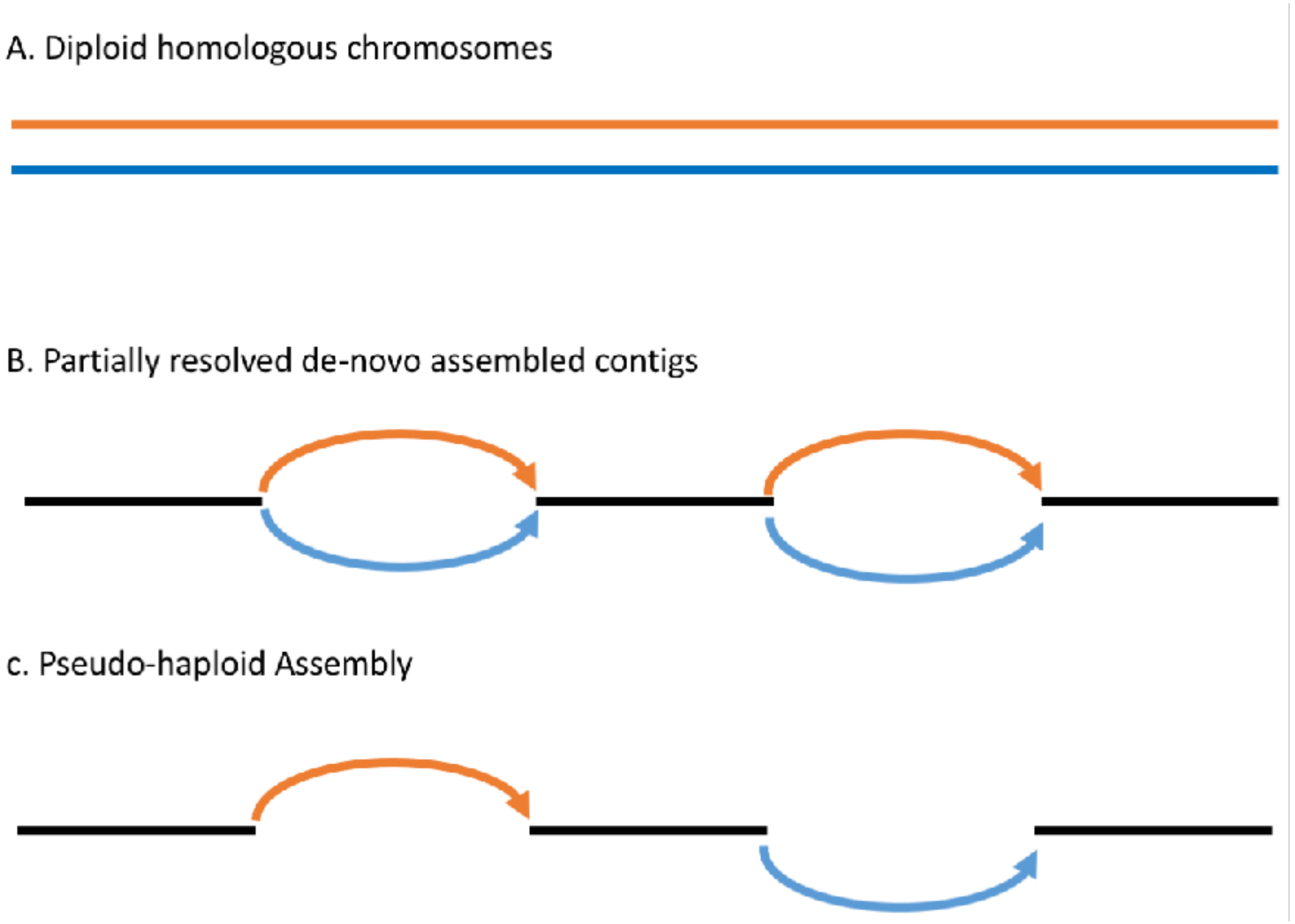
THE HUMAN GENOME – AGAIN



Inside the nucleus of a somatic cell, we will have 6 billion bases of DNA from our genome, as we are ***diploid*** *organisms. The reference human genome is the representation of* ***one*** *copy of each chromosome* ***allele,*** *thus 3 billion bases on average.*

ANOTHER PROBLEM – GENOME HETEROZYGOSITY





A. Diploid homologous chromosomes

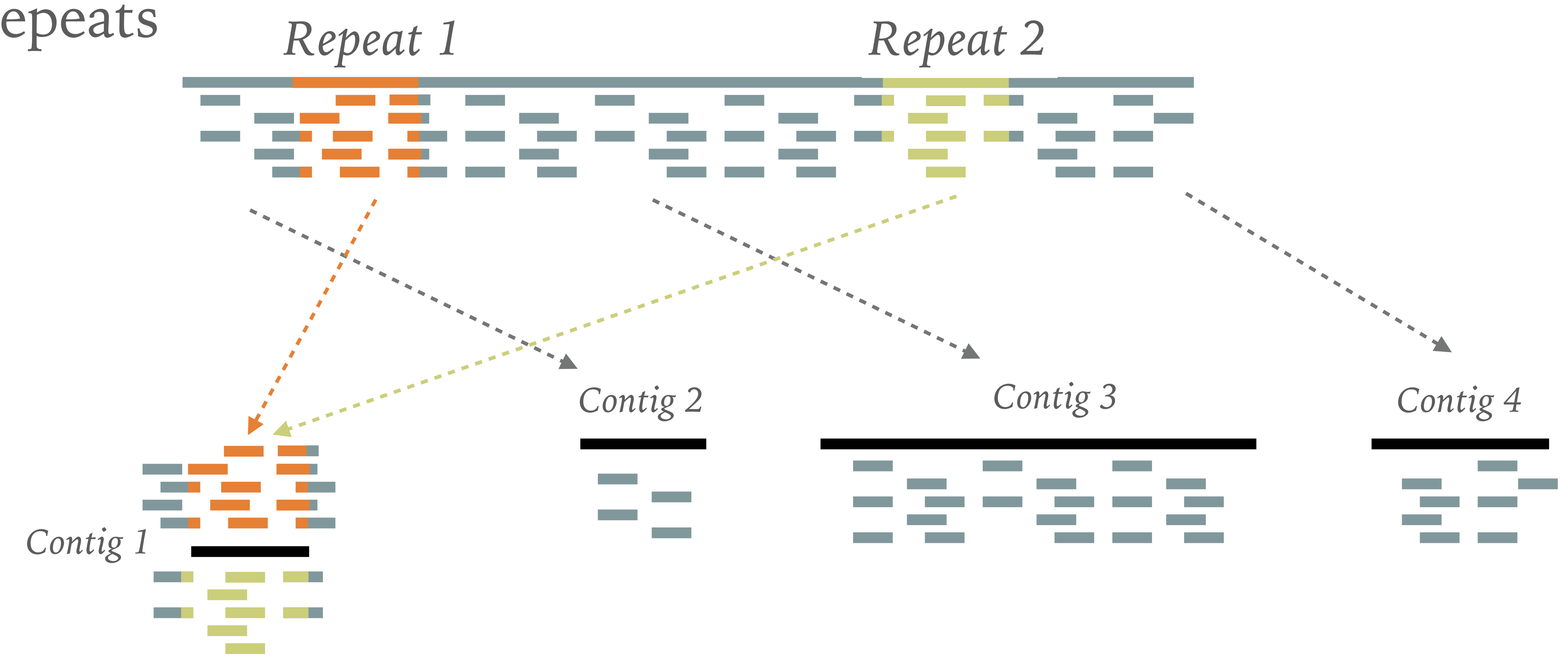
B. Partially resolved de-novo assembled contigs

c. Pseudo-haploid Assembly

Hurdles

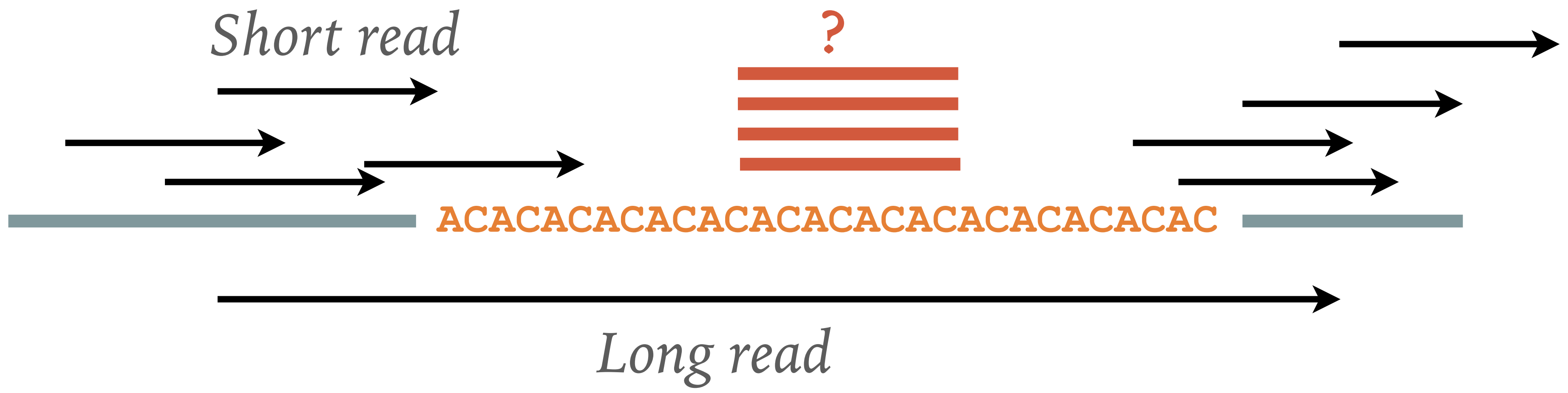
- Heterozygosity
- Sequencing errors
- Repeats
- Low complexity genomic regions
- Base composition and sequencing bias

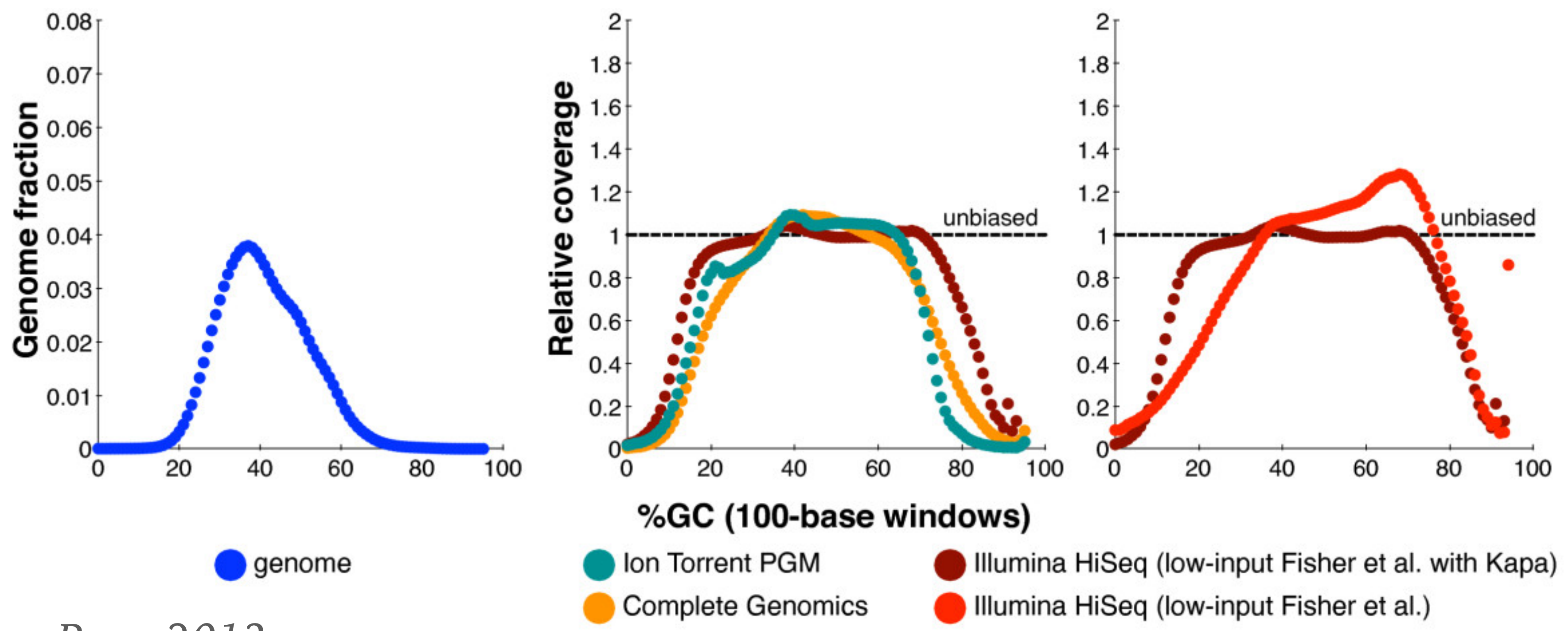
Repeats



The repeated element is collapsed into a single contig

Adapted from Torsten Seemann (2014 talk)

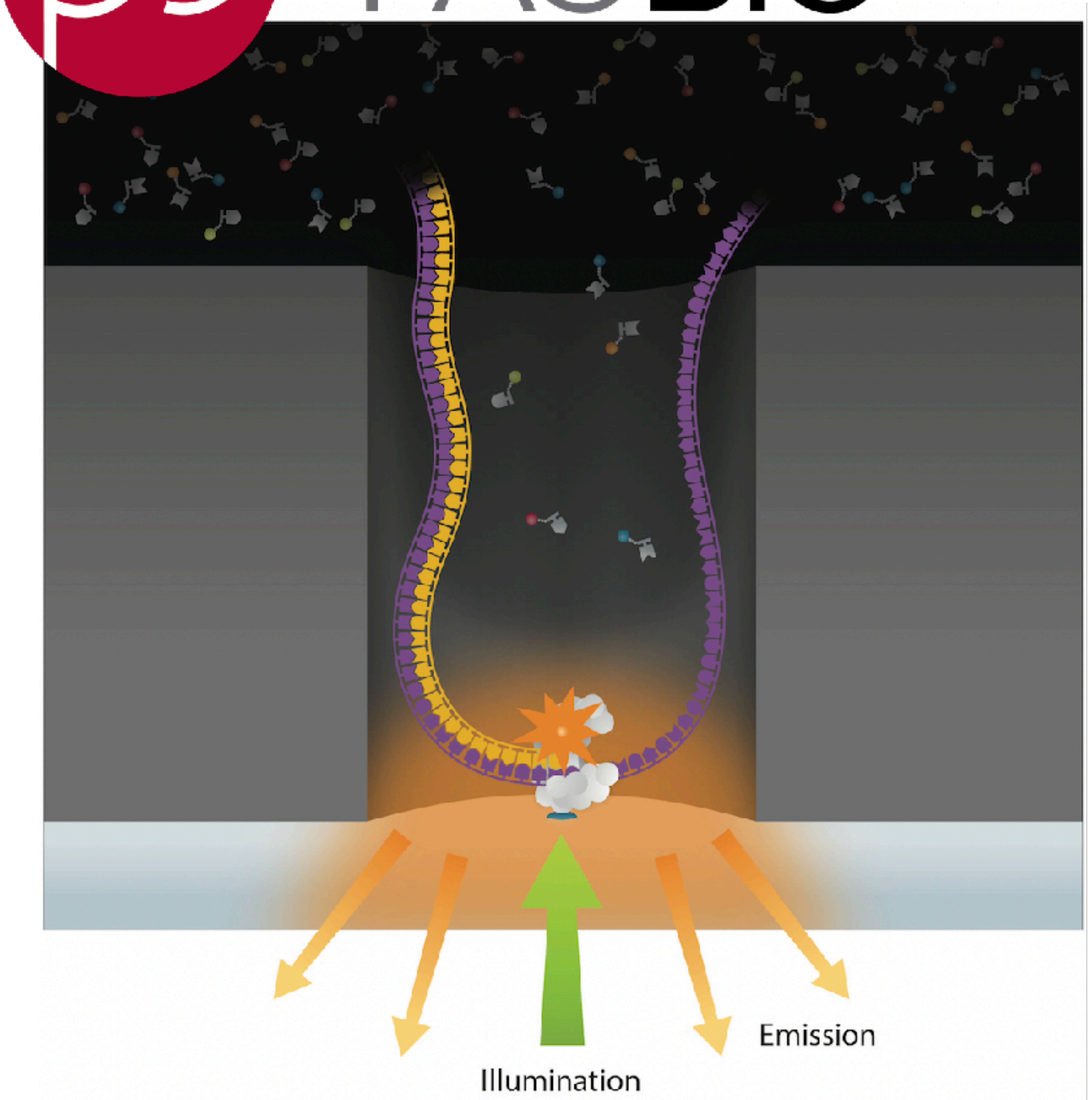




Ross, 2013

LONG READS

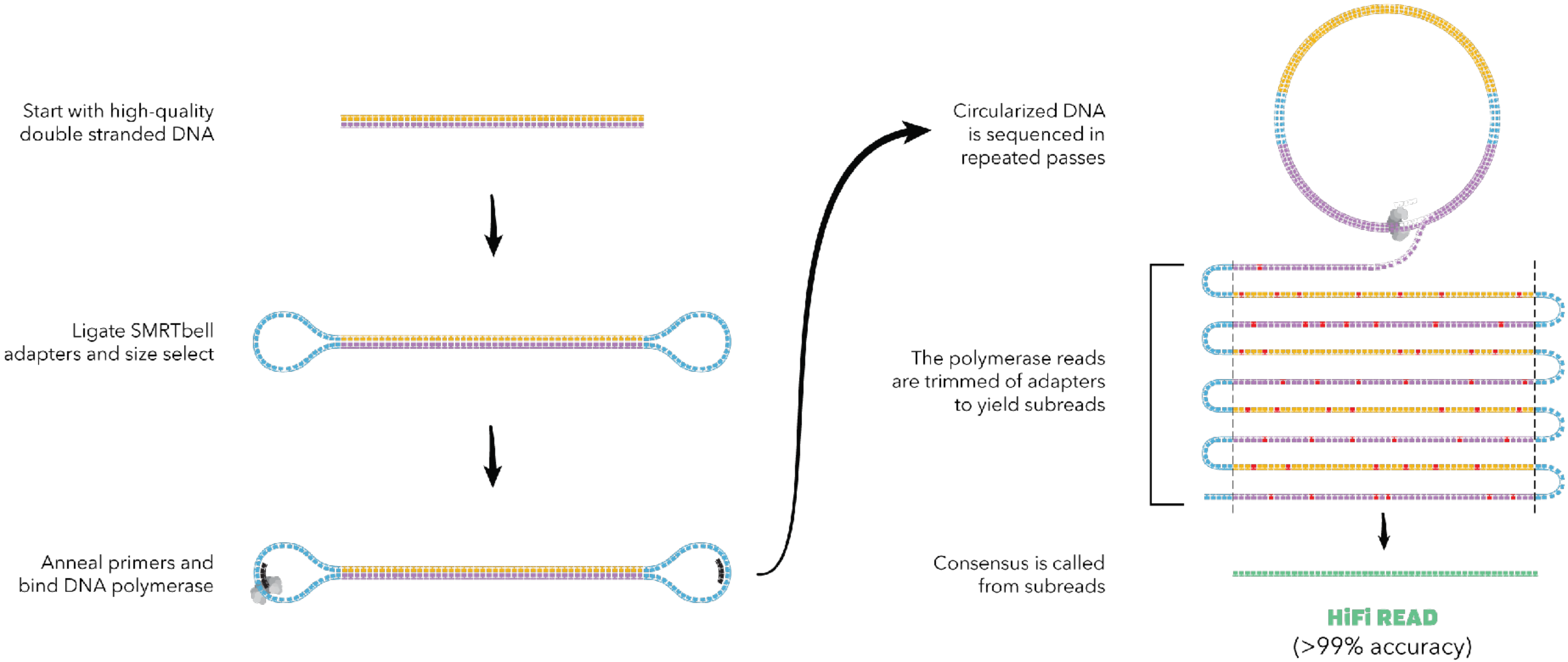
PACBIO®



Single molecule sequencing

DNA Polimerase: 1000bp/s

PACBIO READS



Start with high-quality double stranded DNA

Ligate SMRTbell adapters and size select

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

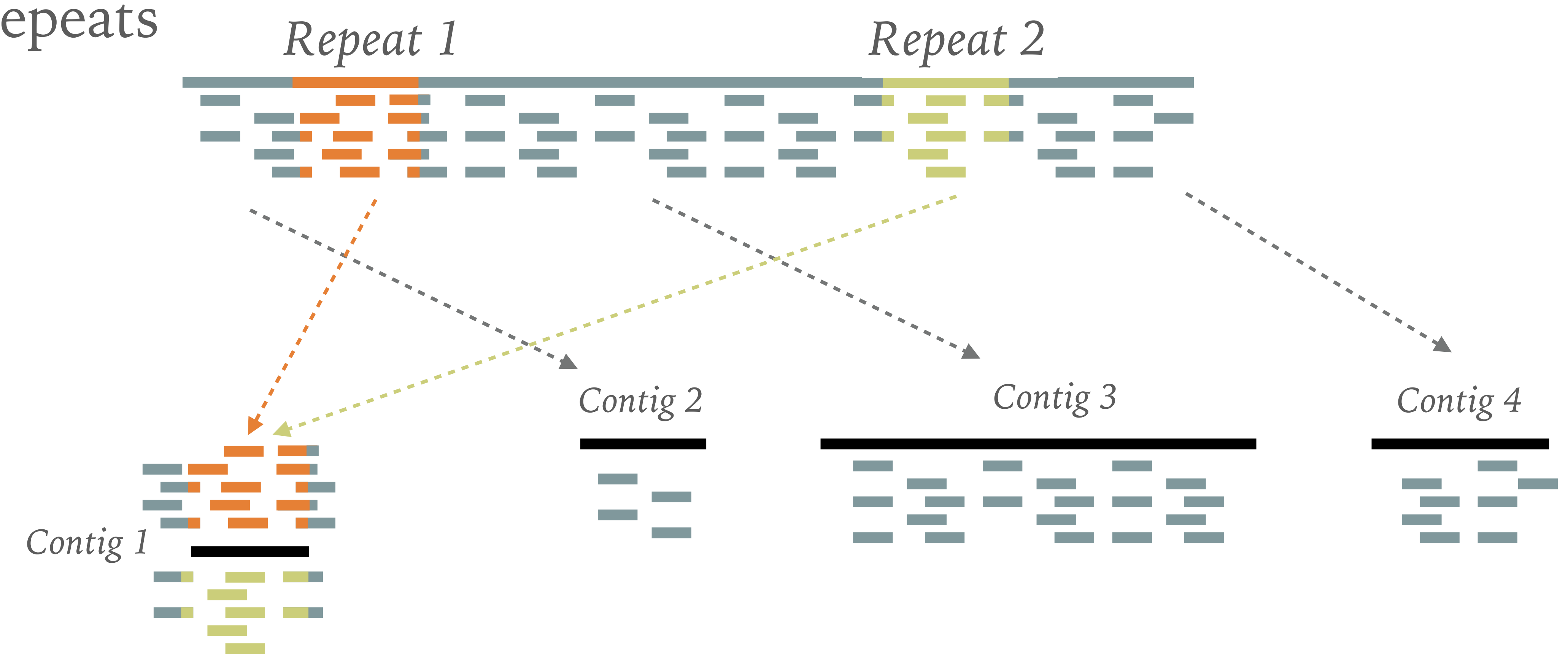
Consensus is called from subreads

HiFi READ
(>99% accuracy)

Hurdles

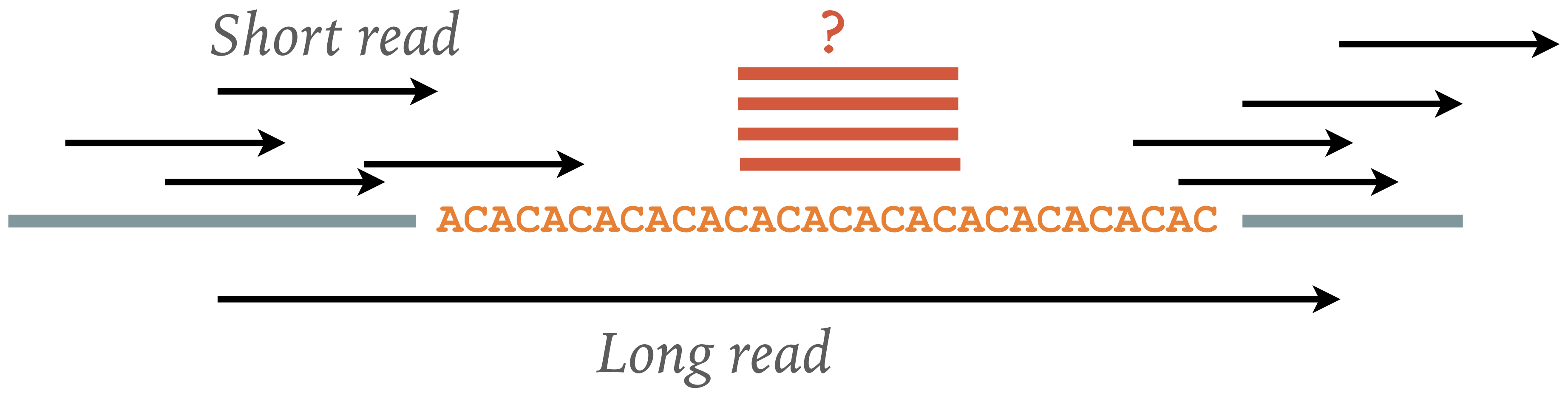
- Heterozygosity
- Sequencing errors
- Repeats
- Low complexity genomic regions
- Base composition and sequencing bias

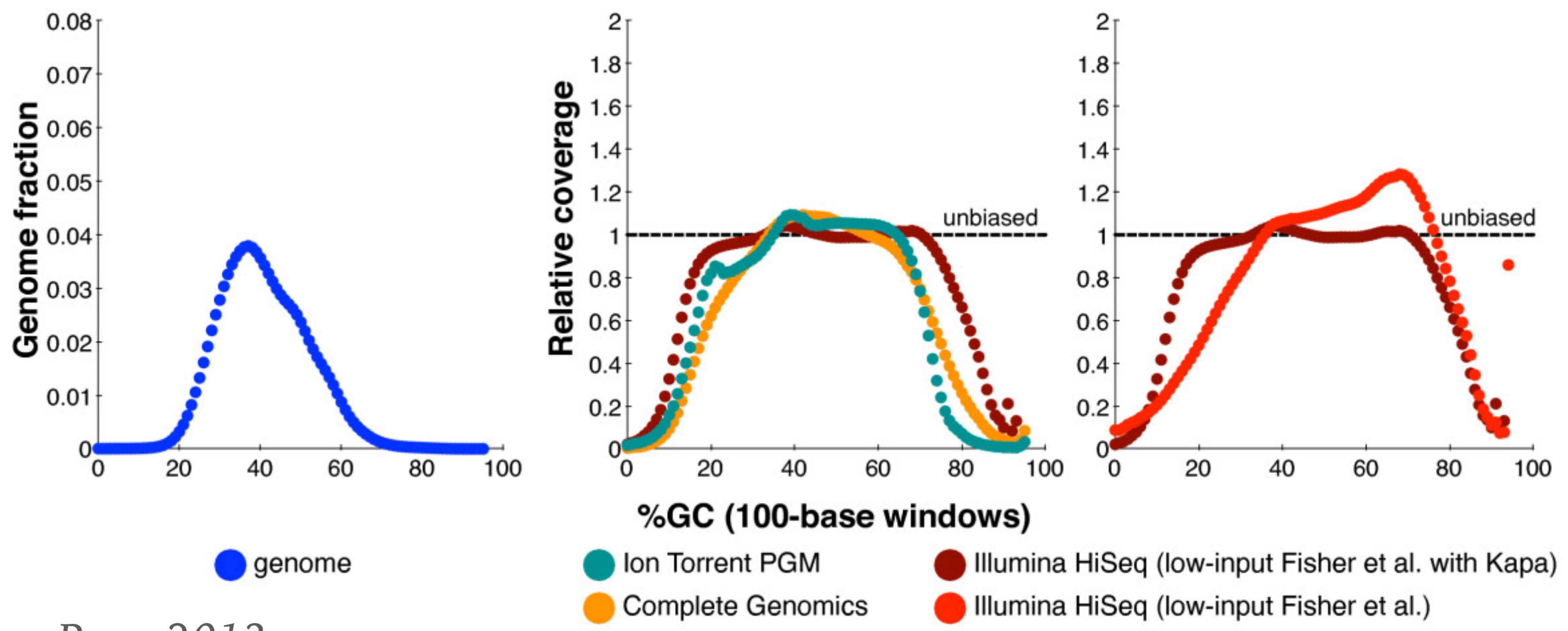
Repeats



The repeated element is collapsed into a single contig

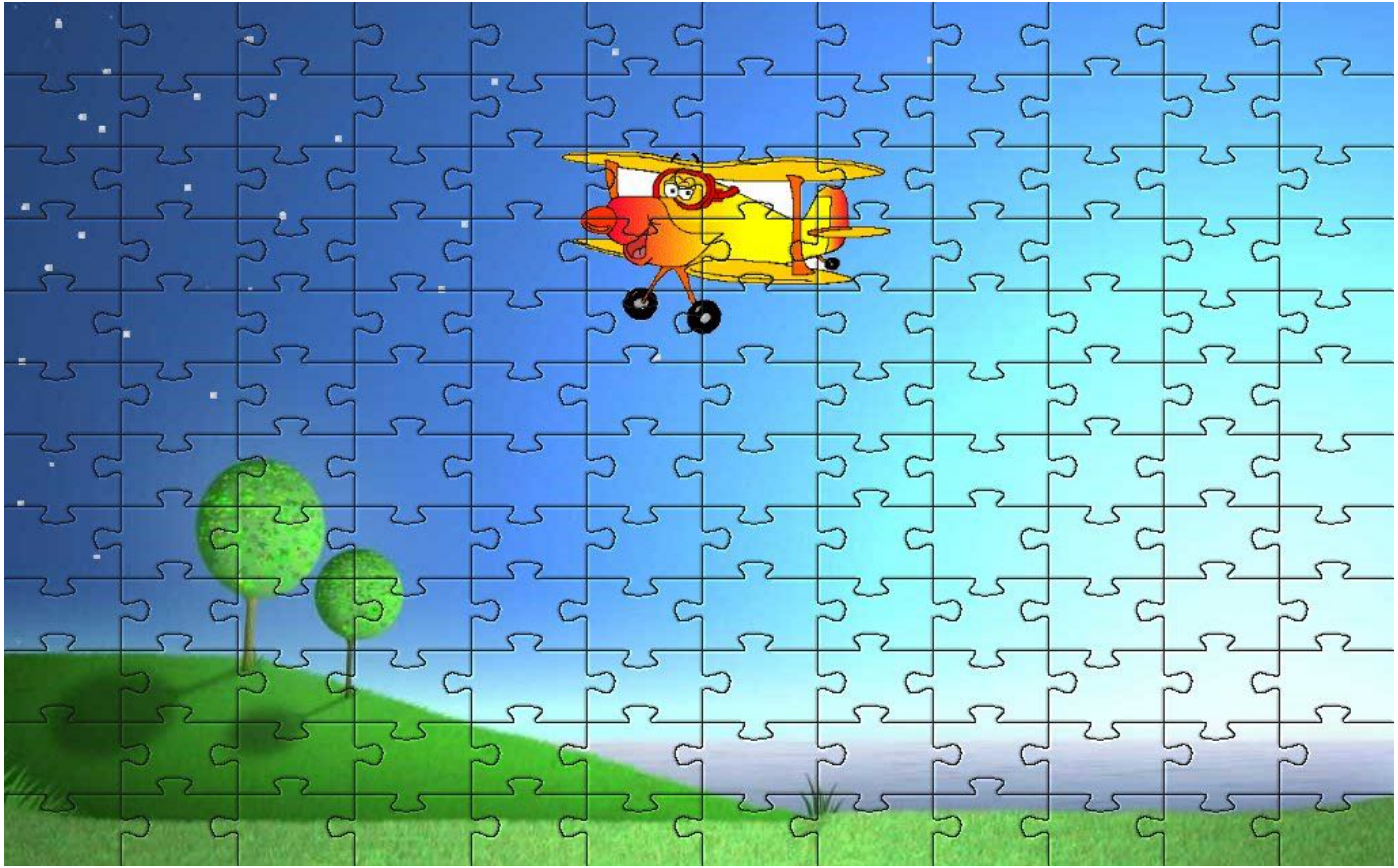
Adapted from Torsten Seemann (2014 talk)



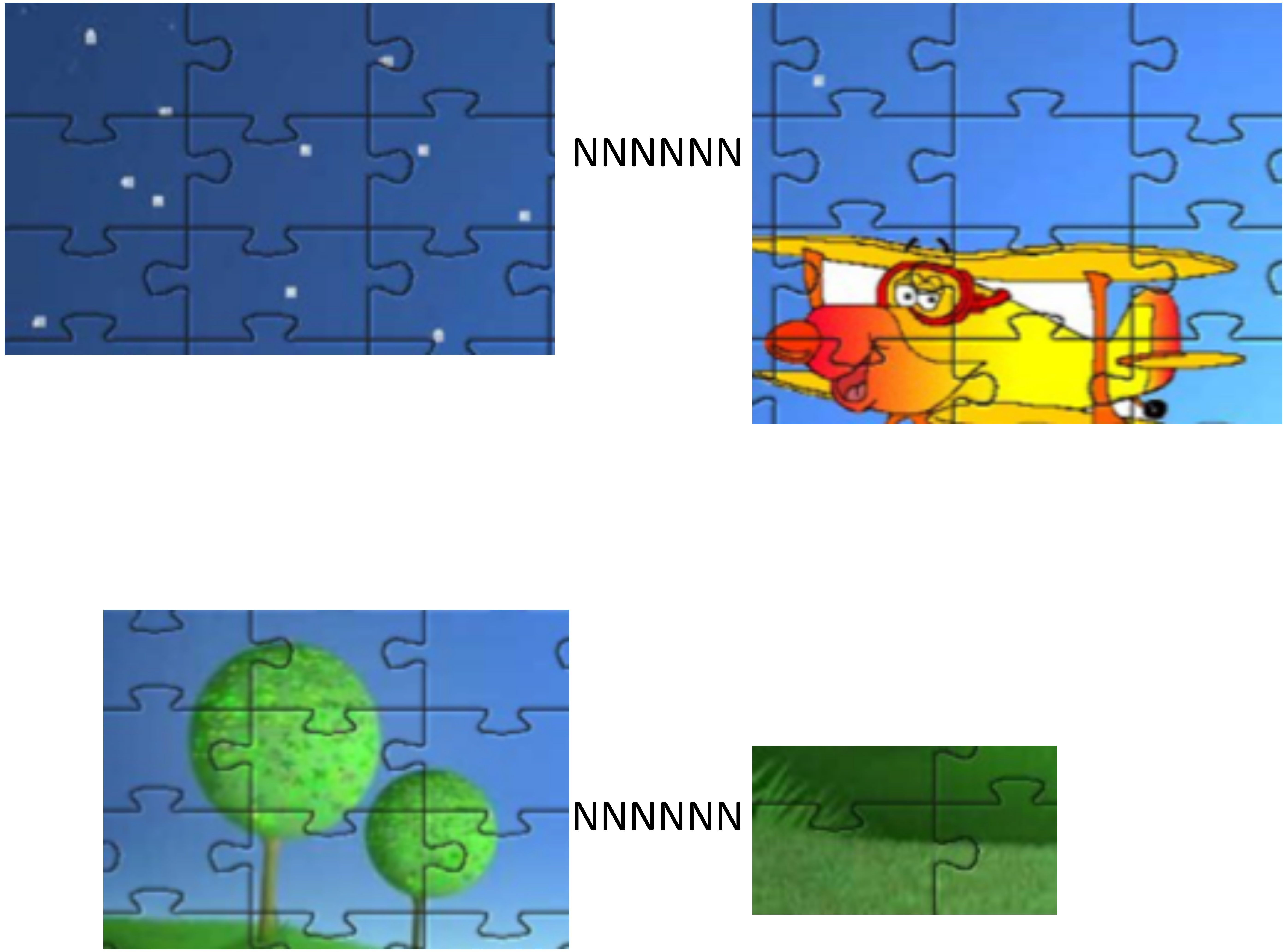


Ross, 2013

I'M A EUKARYOTIC GENOME – THE BLUE AND GREEN ARE MY REPEATS

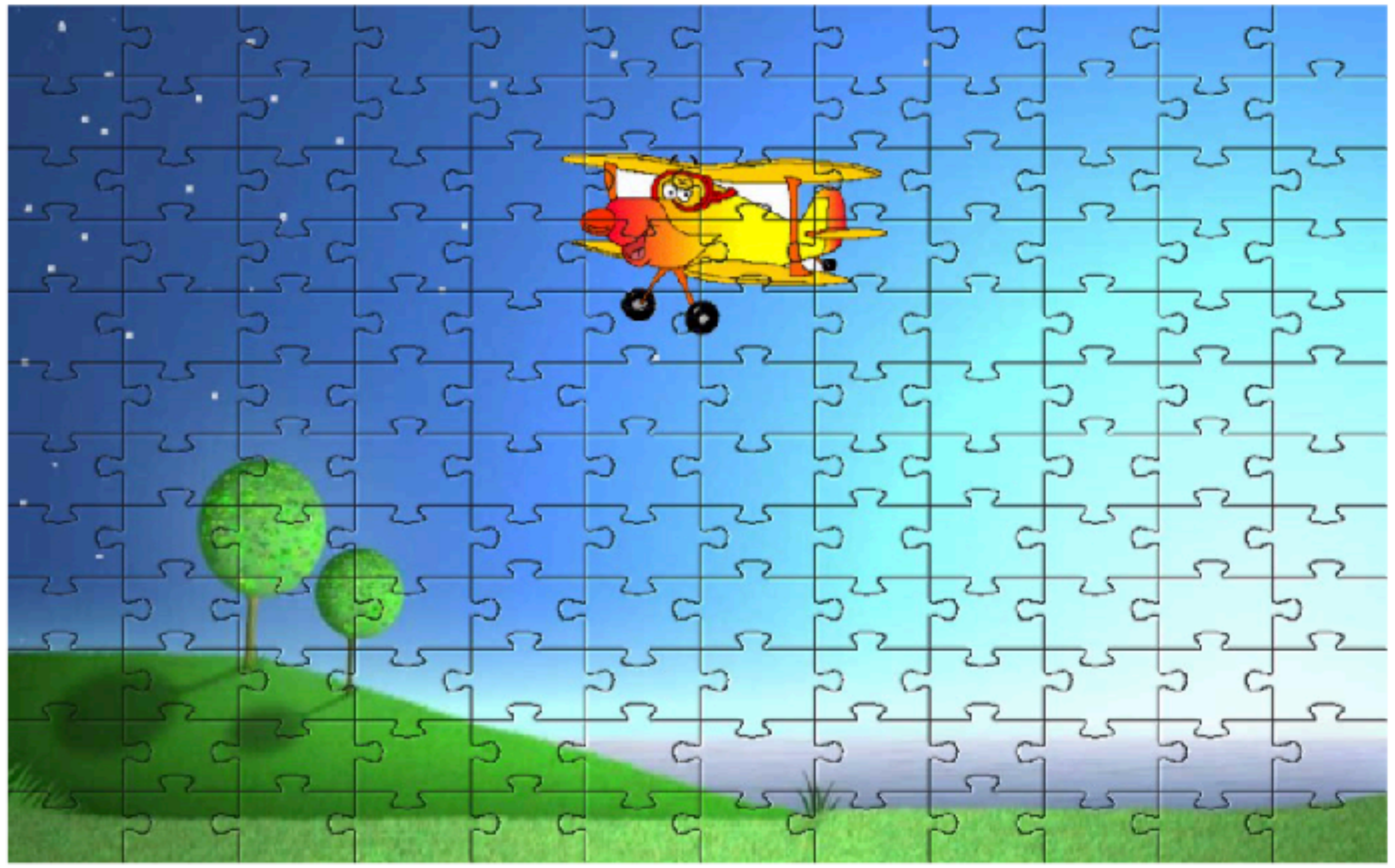


THIS IS A SHORT-READS GENOME ASSEMBLY OF ME

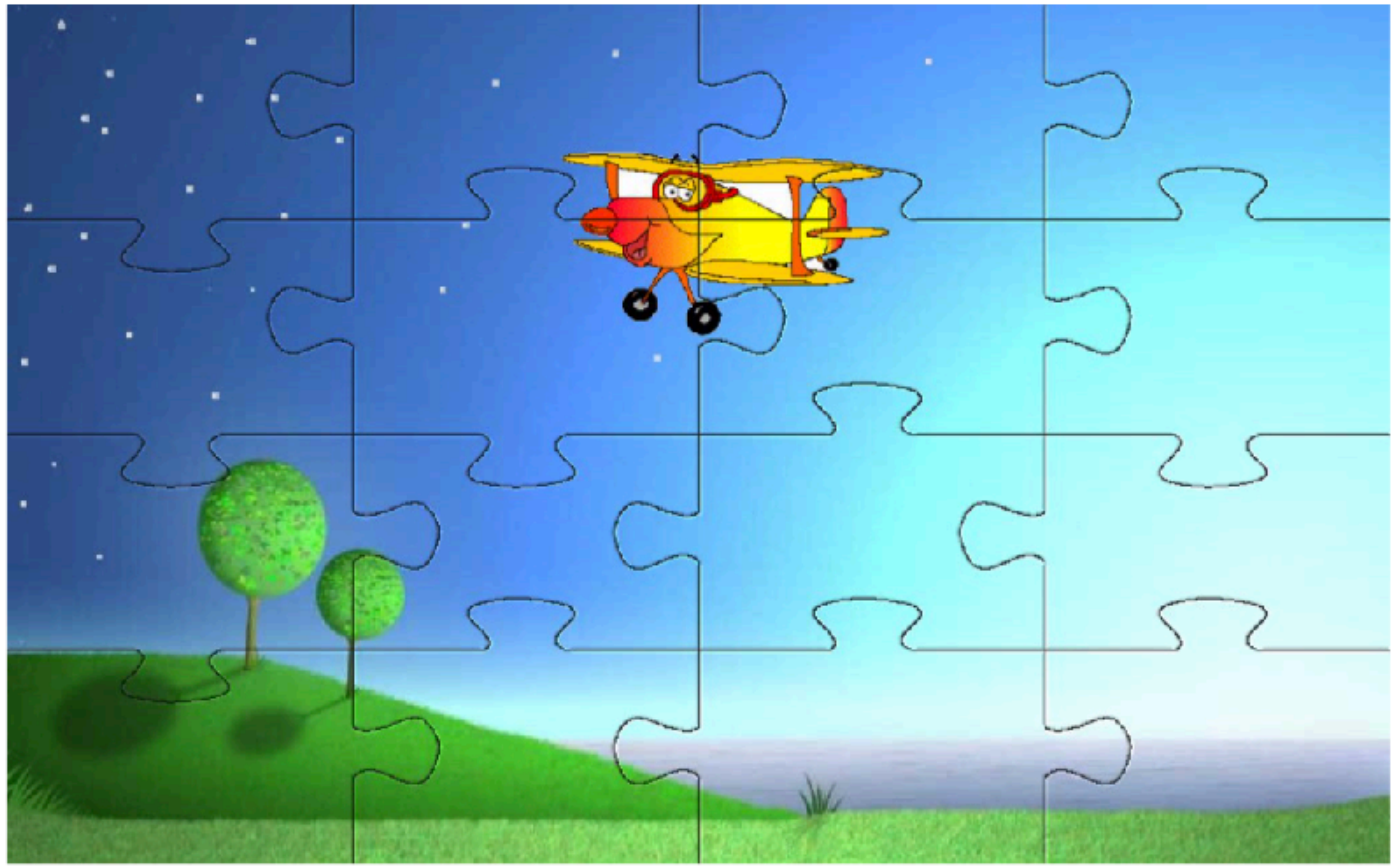


NNNNNN

NNNNNN



Assembling with short reads



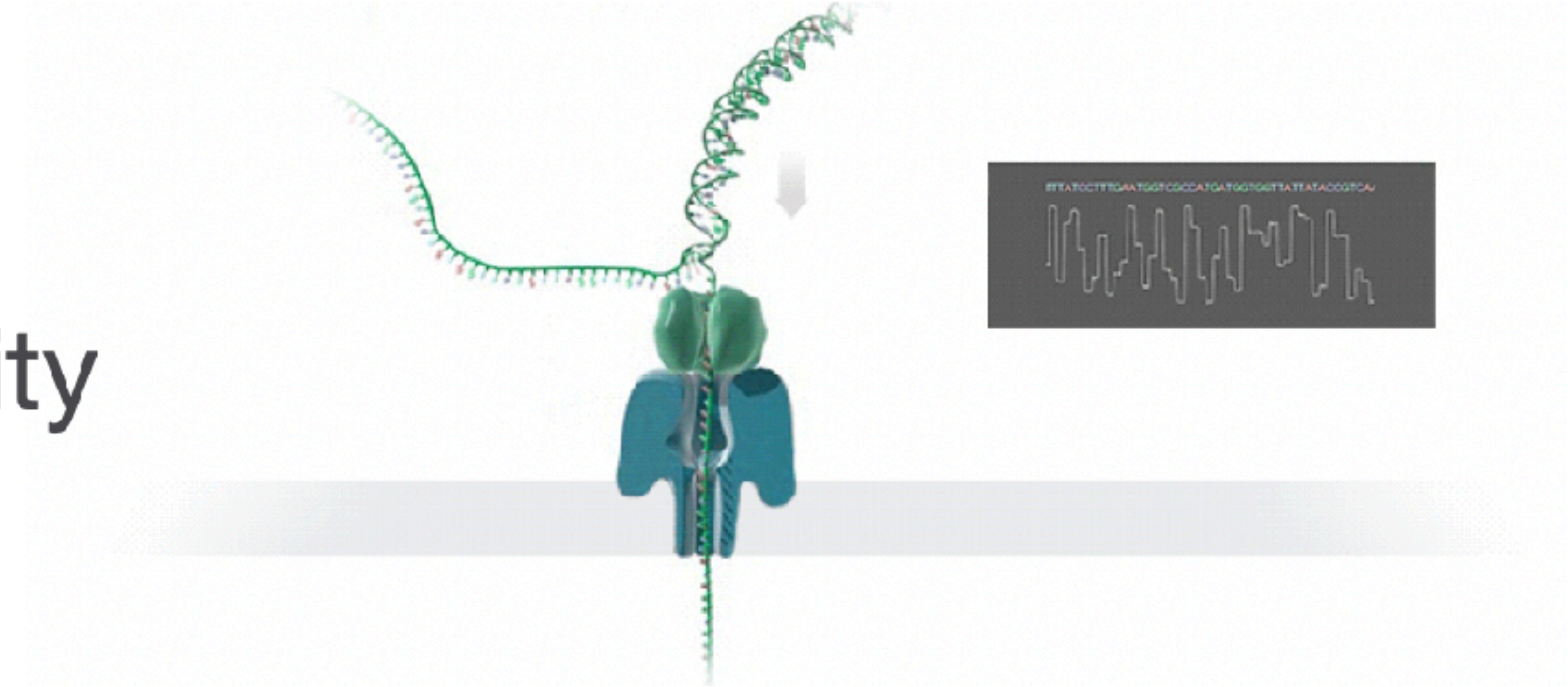
Assembling with long reads

WHAT SEQUENCING STRATEGY TO CHOOSE?

Nanopore ultra-long sequencing

Slide Sergey Koren

- Nanopore UL
  - >100 kb reads, up to 1 Mb
  - 97% (Q15) read quality
  - 99.99% (Q40+) assembly quality
- Pros
  - Length and throughput
  - Reads *span* repeats
- Cons
  - Lower base quality



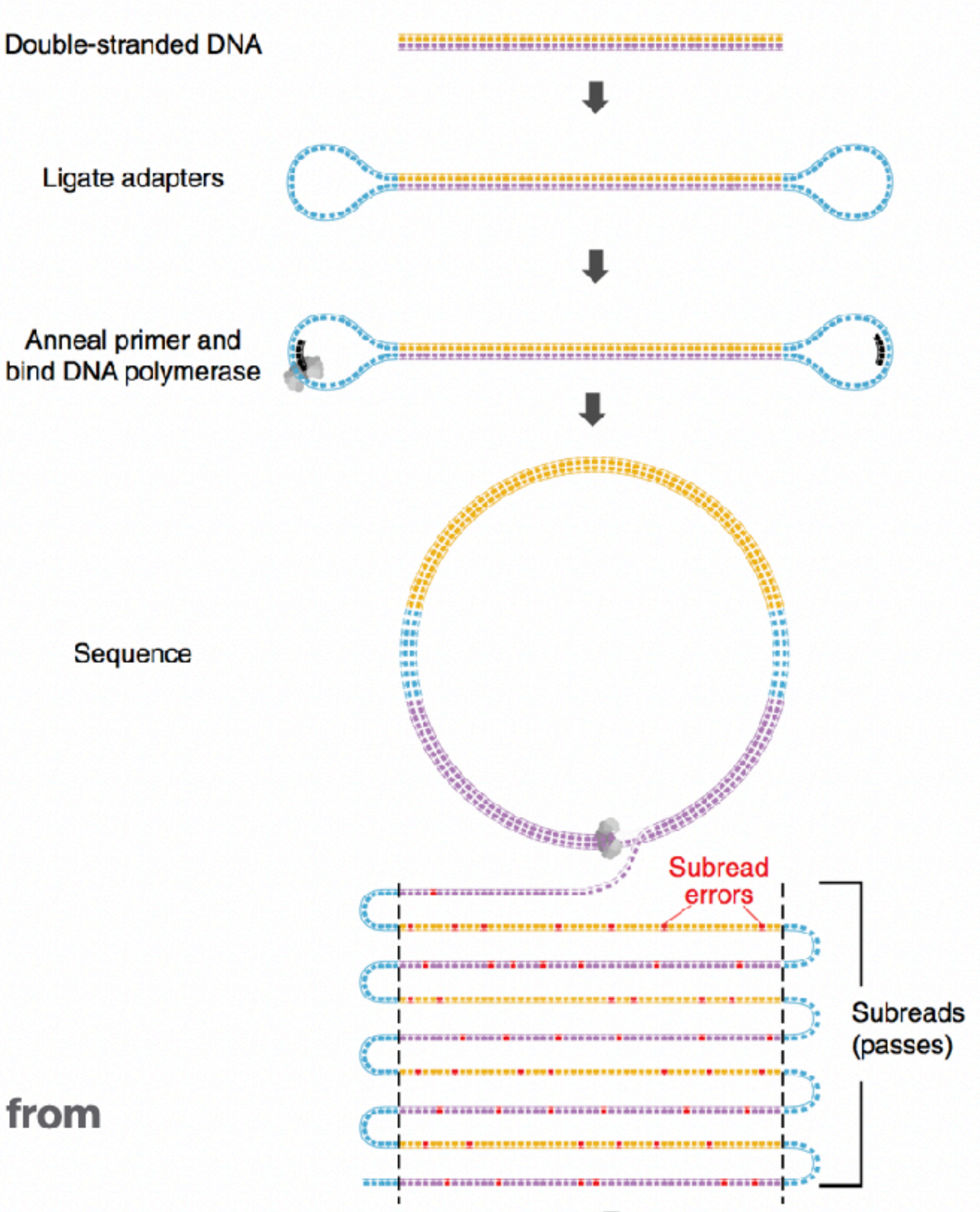
Nanopore sequencing and assembly of a human genome with ultra-long reads.
Jain et al. *Nature Biotechnology* (2018)

Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Shafin et al. *Nature Biotechnology* (2020)

PacBio circular consensus sequencing

Slide Sergey Koren

- PacBio HiFi
  - 20 kb reads
  - 99.9% (Q30) read quality
  - 99.9999% (Q60+) assembly quality
- Pros
  - Near-perfect accuracy
  - Reads *distinguish* repeats
- Cons
  - Limited length and coverage



Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. *Nature Biotechnology* (2019)

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Nurk et al. *Genome Research* (2020)

The best of both worlds

Telomere-to-Telomere

- The human genome is finally finished!
- 8% was left after HGP
- Solved with combination of HiFi + ultra-long ONT

The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

Earth's heart of iron begins to yield its secrets p. 18

Microglia in chronic pain recovery and relapse pp. 33 & 86

Particle acceleration in a nova explosion p

Science

$15
1 APRIL 2022
SPECIAL ISSUE
science.org

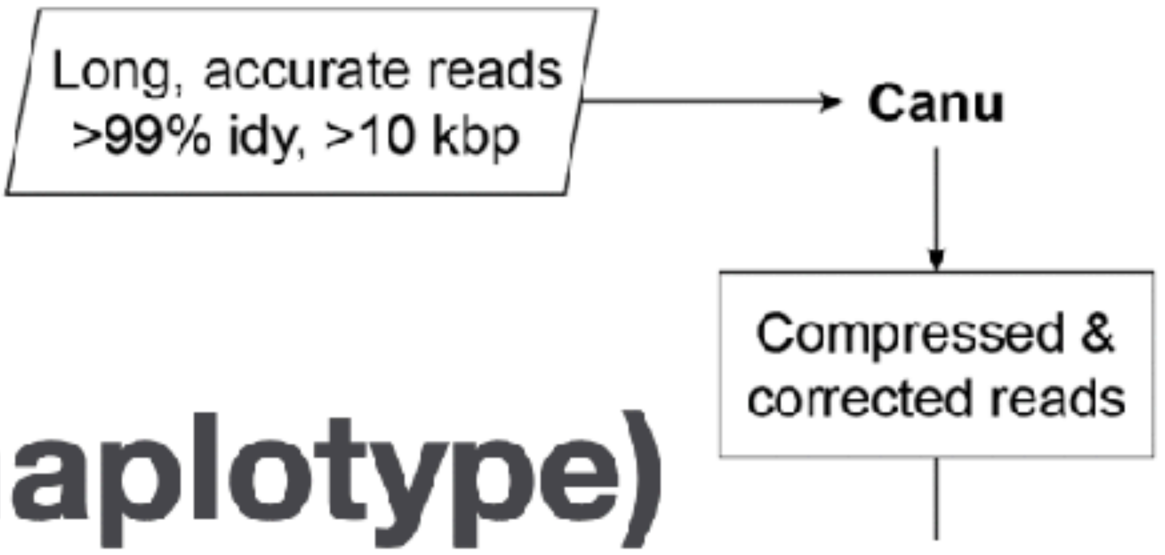
AAAS

FILLING THE GAPS

Closing in on a complete human genome p. 42

Slide Sergey Koren

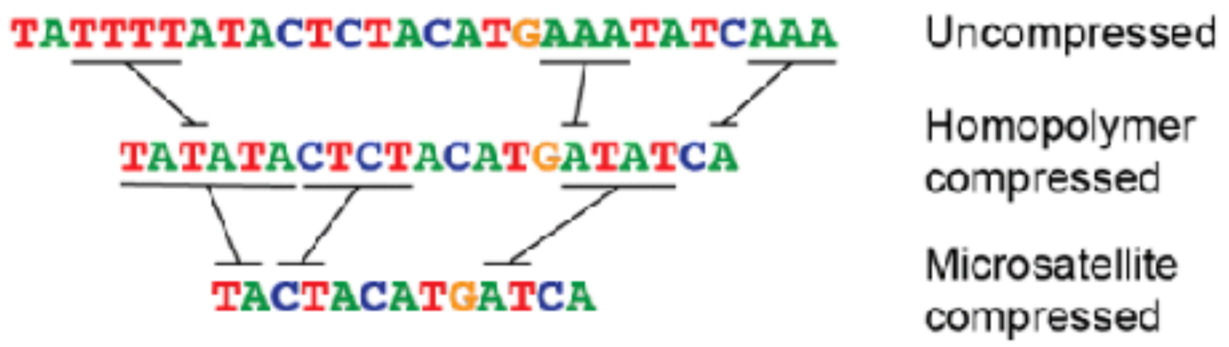
Verkko!



Long, accurate reads >99% idy, >10 kbp

Canu

Compressed & corrected reads



- **Sequencing recipe (per haplotype)**
  - 25 PacBio HiFi (20 kb)
  - 25x ONT ultra-long (>100 kb)
  - 30x Illumina Trio or Hi-C

Telomere-to-telomere assembly of diploid chromosomes with Verkko
Rautiainen, *et al.* Nat Biotech (2023)

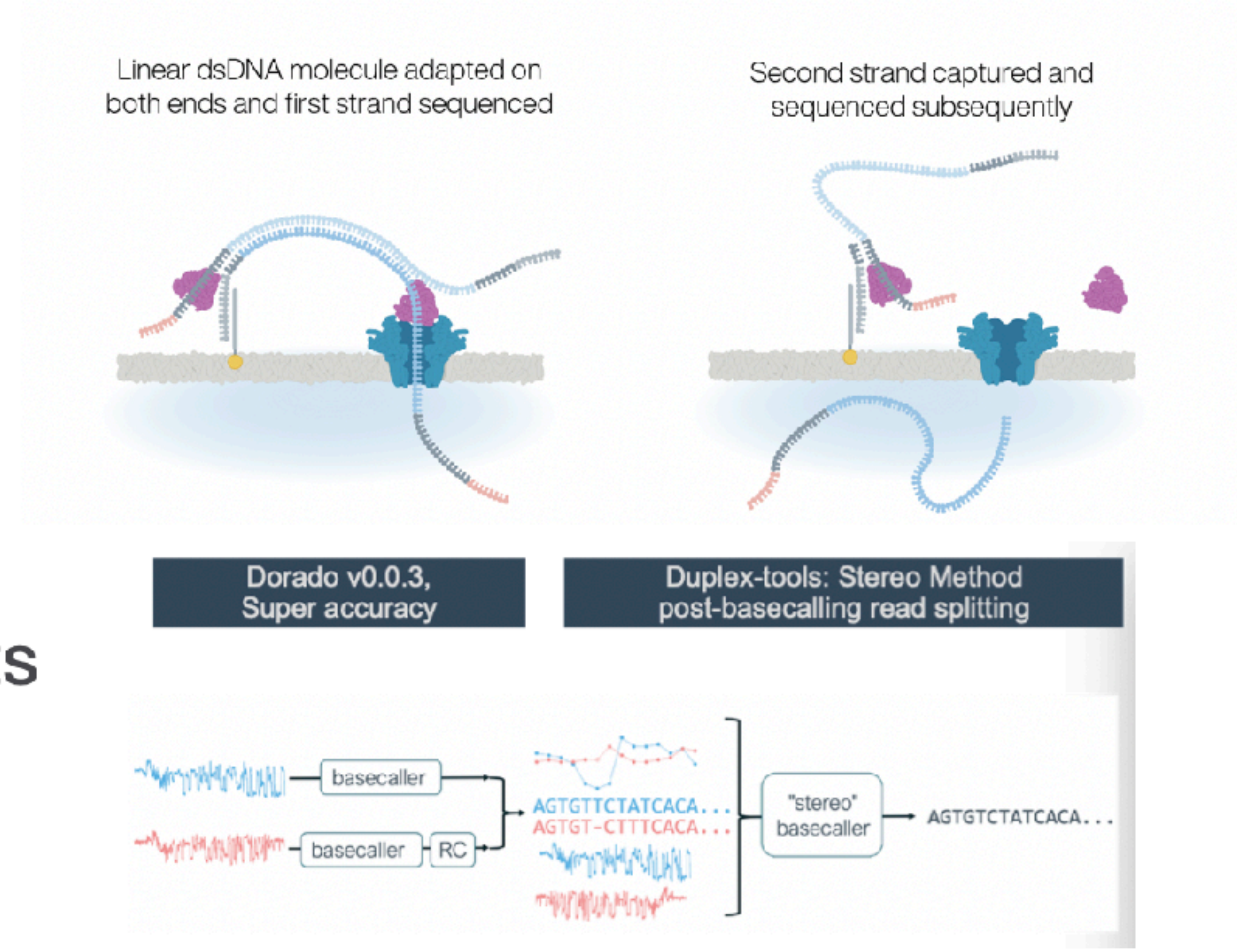
LJA: Assembling Long and Accurate Reads Using Multiplex de Bruijn Graphs
Bankevich, *et al.* Nat Biotech (2021)

Slide Sergey Koren

Nanopore duplex sequencing

Slide Sergey Koren

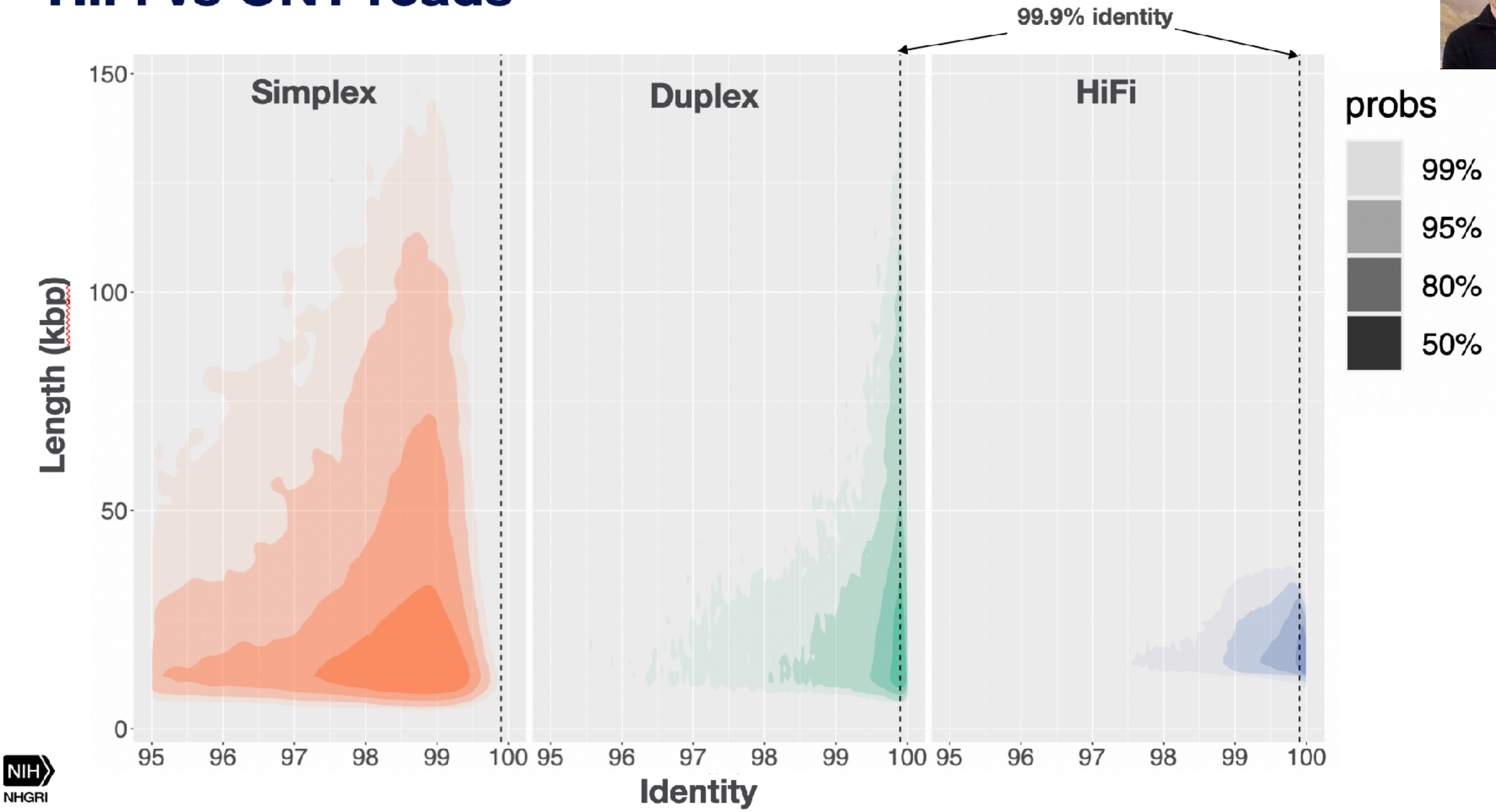
- **Nanopore Duplex**
  - >10 kb reads
  - 99.9% (Q30) read quality
  - 99.999% (Q50+) assembly quality
- **Pros**
  - Near-perfect accuracy
  - No size selection to limit length
  - Reads *distinguish* and *span* repeats
- **Cons**
  - Low throughput


Linear dsDNA molecule adapted on both ends and first strand sequenced
Second strand captured and sequenced subsequently
Dorado v0.0.3, Super accuracy
Duplex-tools: Stereo Method post-basecalling read splitting
basecaller
basecaller
RC
AGTGTTCTATCACA...
AGTGT-CTTTCACA...
"stereo" basecaller
AGTGTCTATCACA...

NIH NHGRI

Slide Sergey Koren

HiFi vs ONT reads



PacBio HiFi data at the core of DToL strategy

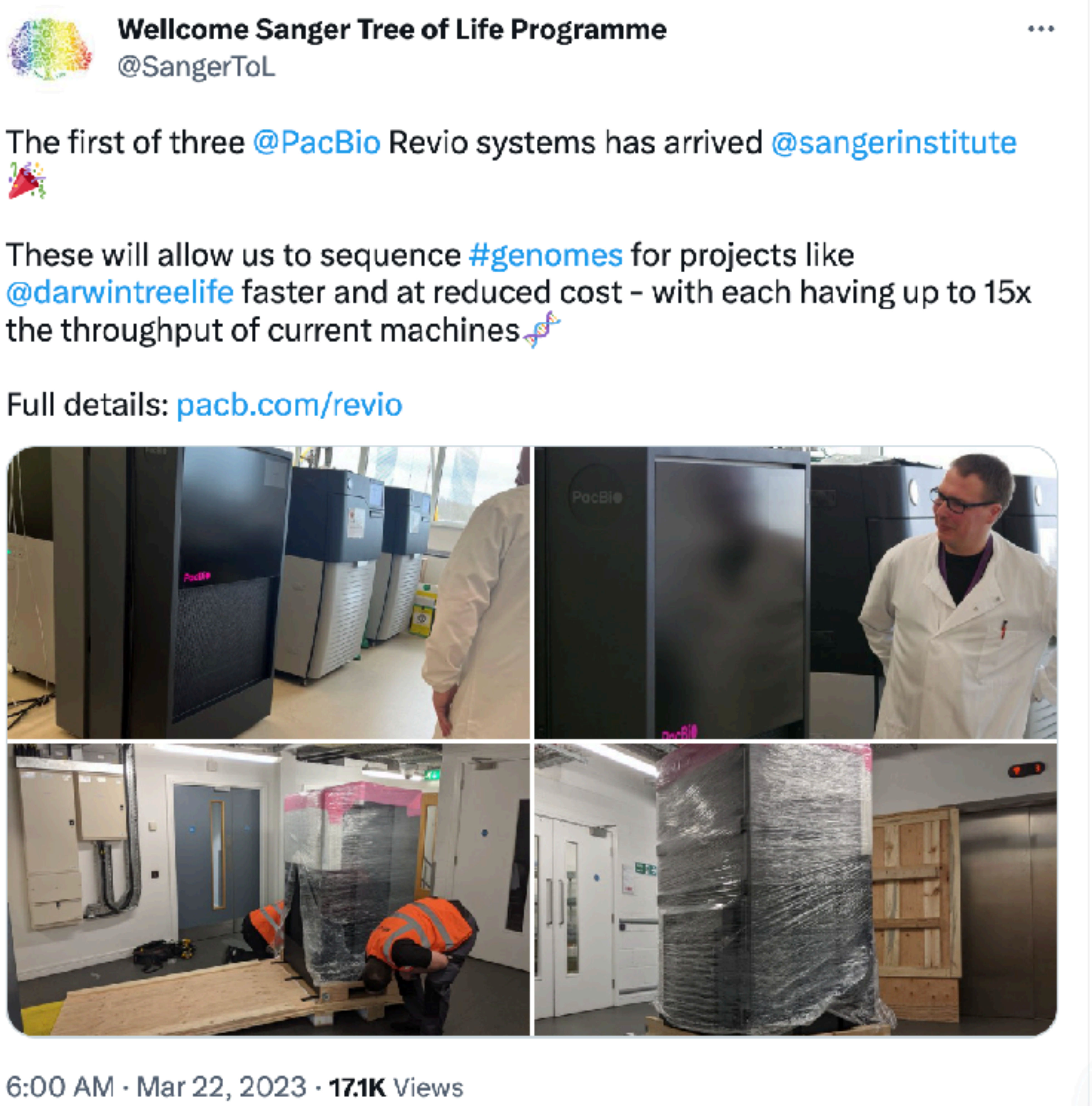
Sequel IIe installed in Sanger SciOps

We generate

- HiFi CCS reads to ~25x
- Hi-C (Illumina) reads to ~100x
- linked reads (Illumina) to ~50x
- RNASeq (30 M Illumina read pairs)

also

- ONT long reads *(for large, repetitive genomes*
- isoSEQ RNASeq *(for gene finding in Family representatives)*



Wellcome Sanger Tree of Life Programme
@SangerToL

The first of three @PacBio Revio systems has arrived @sangerinstitute 🎉

These will allow us to sequence #genomes for projects like @darwintreelife faster and at reduced cost - with each having up to 15x the throughput of current machines 🧬

Full details: pacb.com/revio

6:00 AM · Mar 22, 2023 · **17.1K** Views

wellcome sanger institute

ASSEMBLY

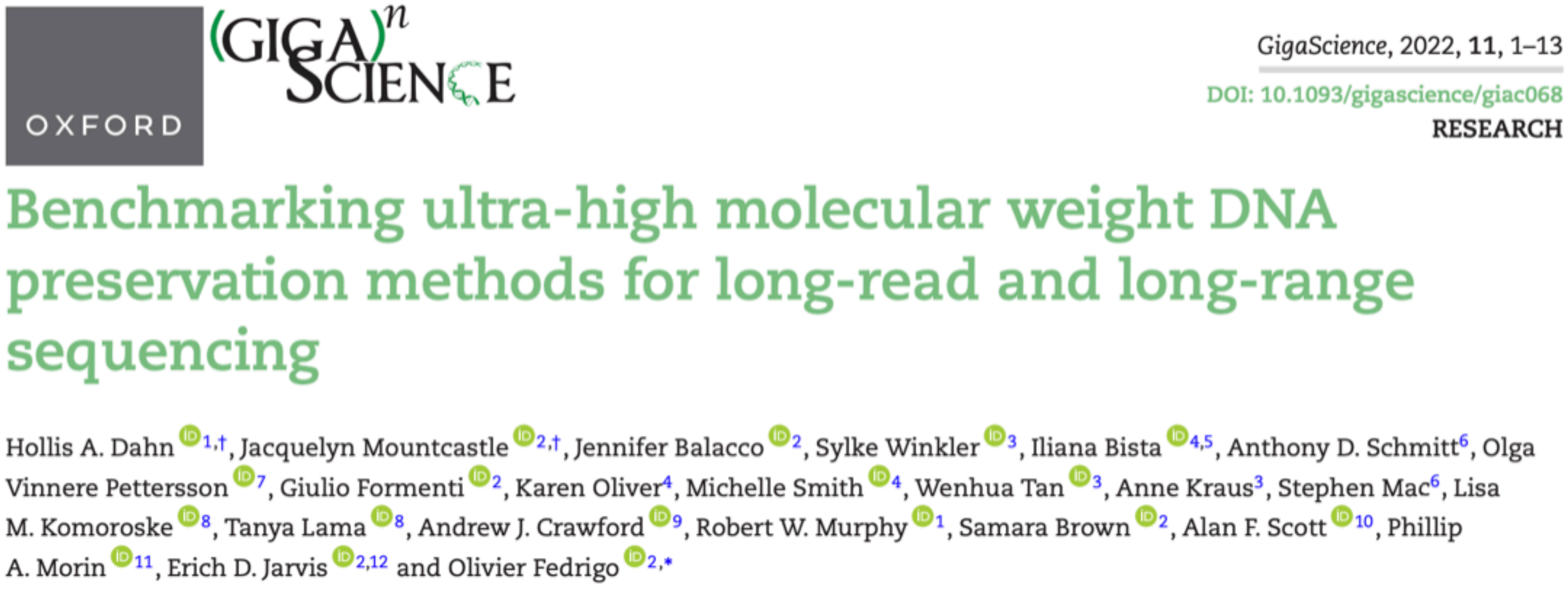
Darwin Tree of Life

ToLa - Tree of Life Assembly team

Band names? AC/GC?

Shane, Ksenia, Chenxi, Marcela, Eerik, Noah, James, Yumi, Willian

BREATHE



OXFORD (GIGA)^n SCIENCE

GigaScience, 2022, **11**, 1–13

DOI: 10.1093/gigascience/giac068

RESEARCH

Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

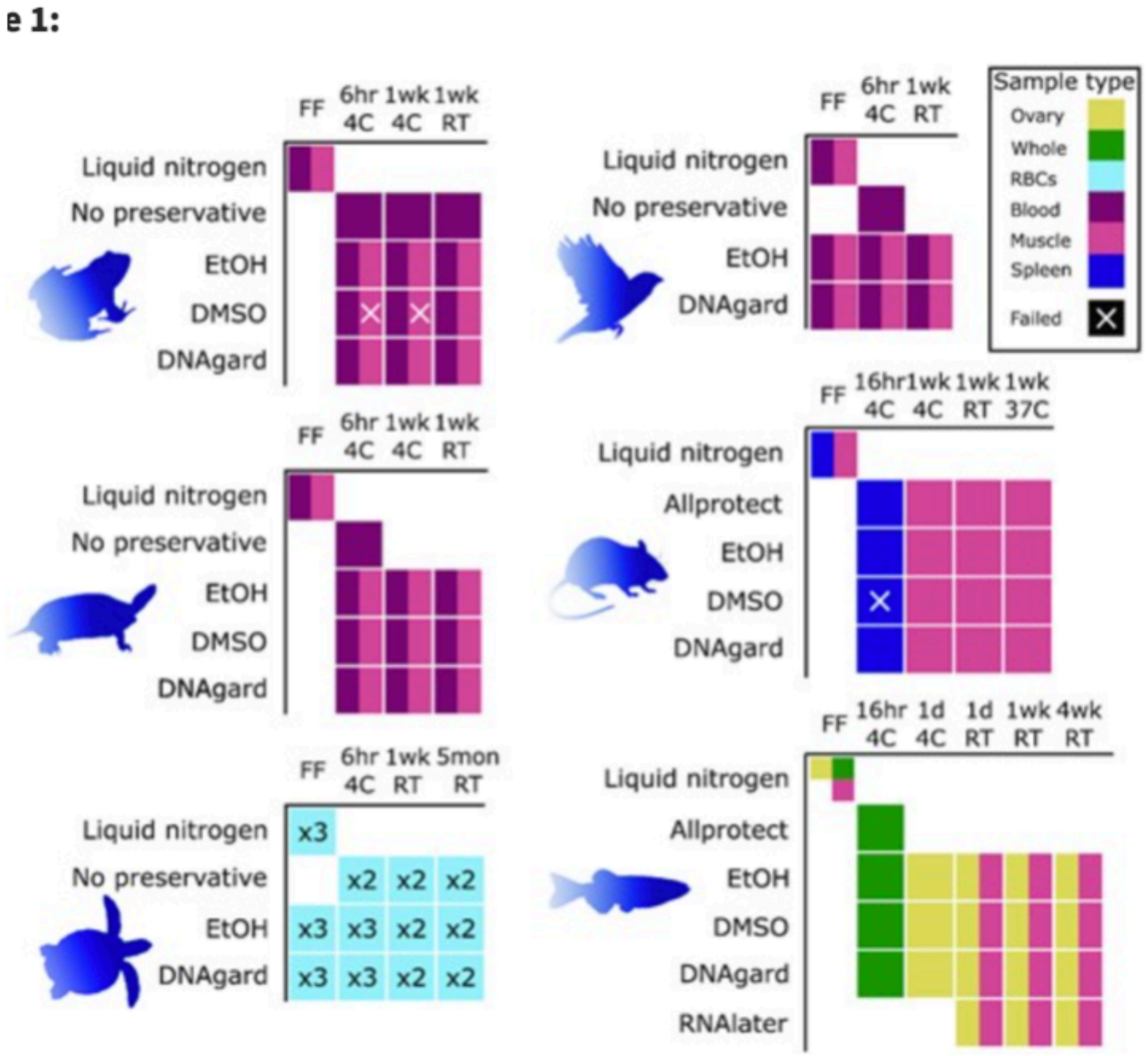
Hollis A. Dahn[1,†], Jacquelyn Mountcastle[2,†], Jennifer Balacco[2], Sylke Winkler[3], Iliana Bista[4,5], Anthony D. Schmitt[6], Olga Vinnere Pettersson[7], Giulio Formenti[2], Karen Oliver[4], Michelle Smith[4], Wenhua Tan[3], Anne Kraus[3], Stephen Mac[6], Lisa M. Komoroske[8], Tanya Lama[8], Andrew J. Crawford[9], Robert W. Murphy[1], Samara Brown[2], Alan F. Scott[10], Phillip A. Morin[11], Erich D. Jarvis[2,12] and Olivier Fedrigo[2,*]

VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

No one-size-fits-all protocol!

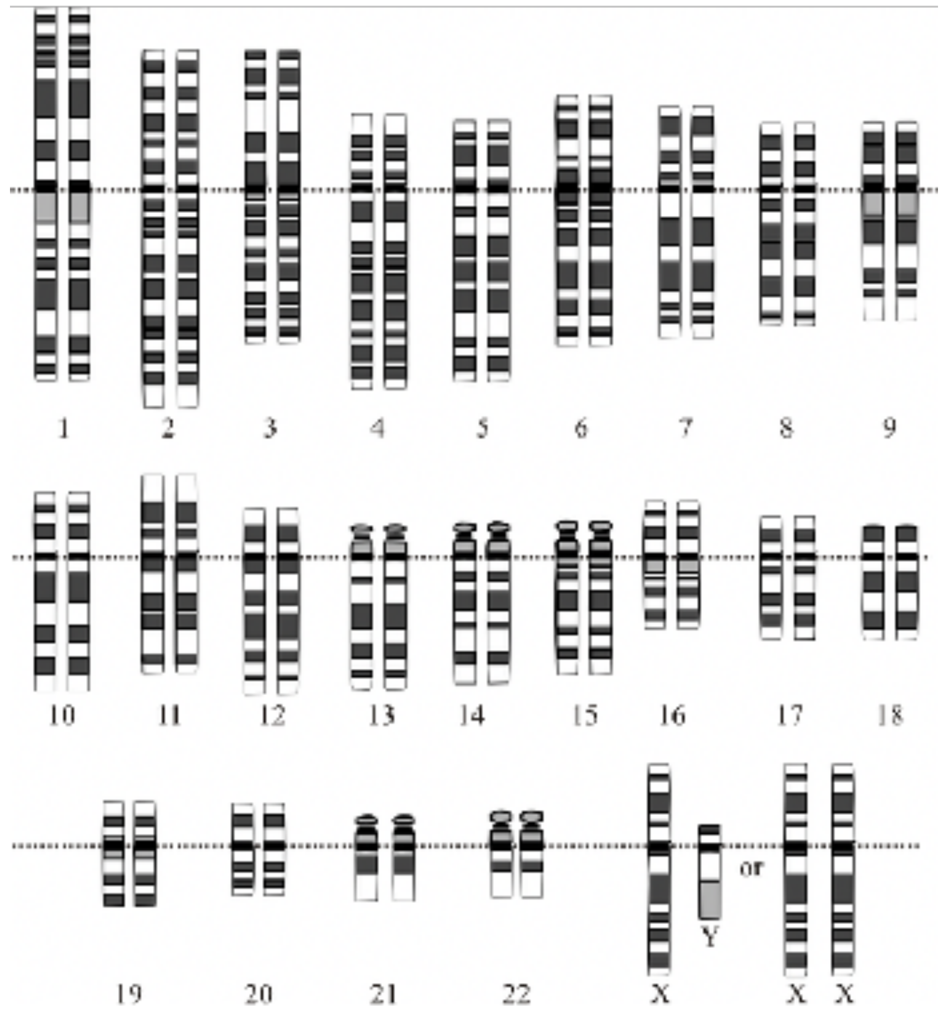
slack Channel: all.things.up.to.assembly



TAKE HOME MESSAGE

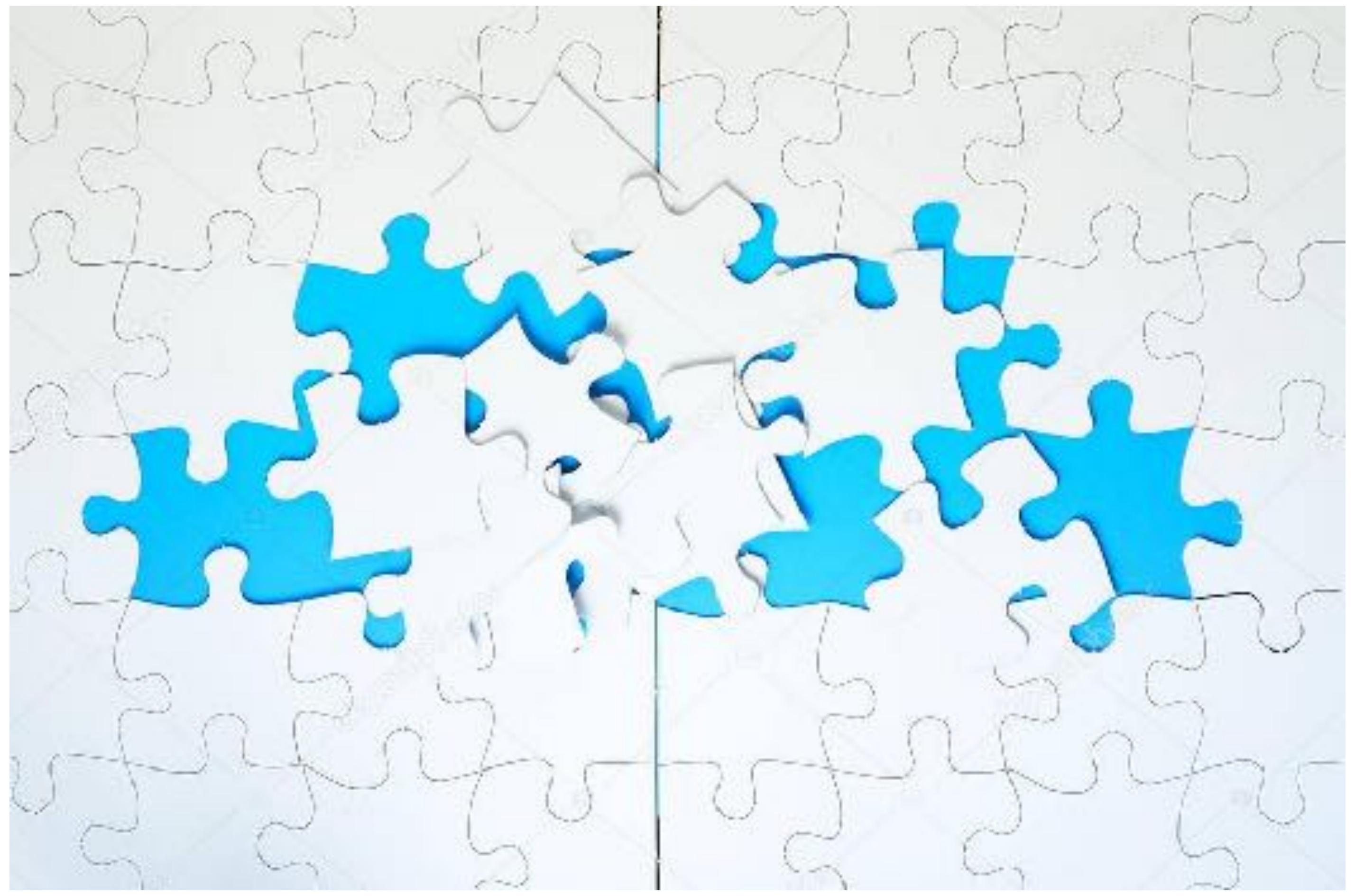
- Know your species! Do your research prior to sequencing. Try to have the best sample you can (fresh and immediately flash-frozen).
- **Estimated genome size, repeat content, heterozygosity**
- From our experience at the Darwin Tree of Life
- 25x coverage of PacBio HiFi (for both haplotypes) + 100x coverage of Hi-C is yielding high-quality assemblies
- Nanopre duplex is promising: not available to the public yet

WHEN WE ASSEMBLE A GENOME …



What we would like to have

- One DNA sequence for each chromosome



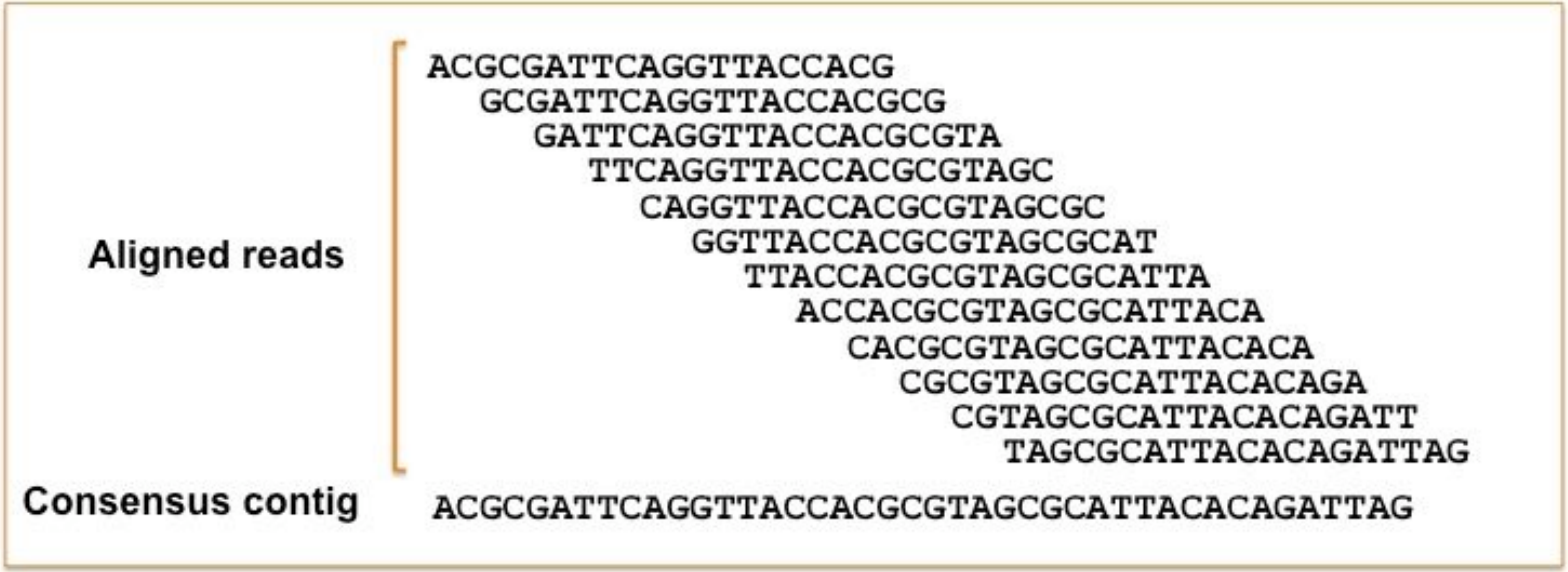
What we really have

- Contigs, scaffolds, gaps, N50s

GENOME ASSEMBLY METRICS

A DNA sequence with gaps

CONTIG



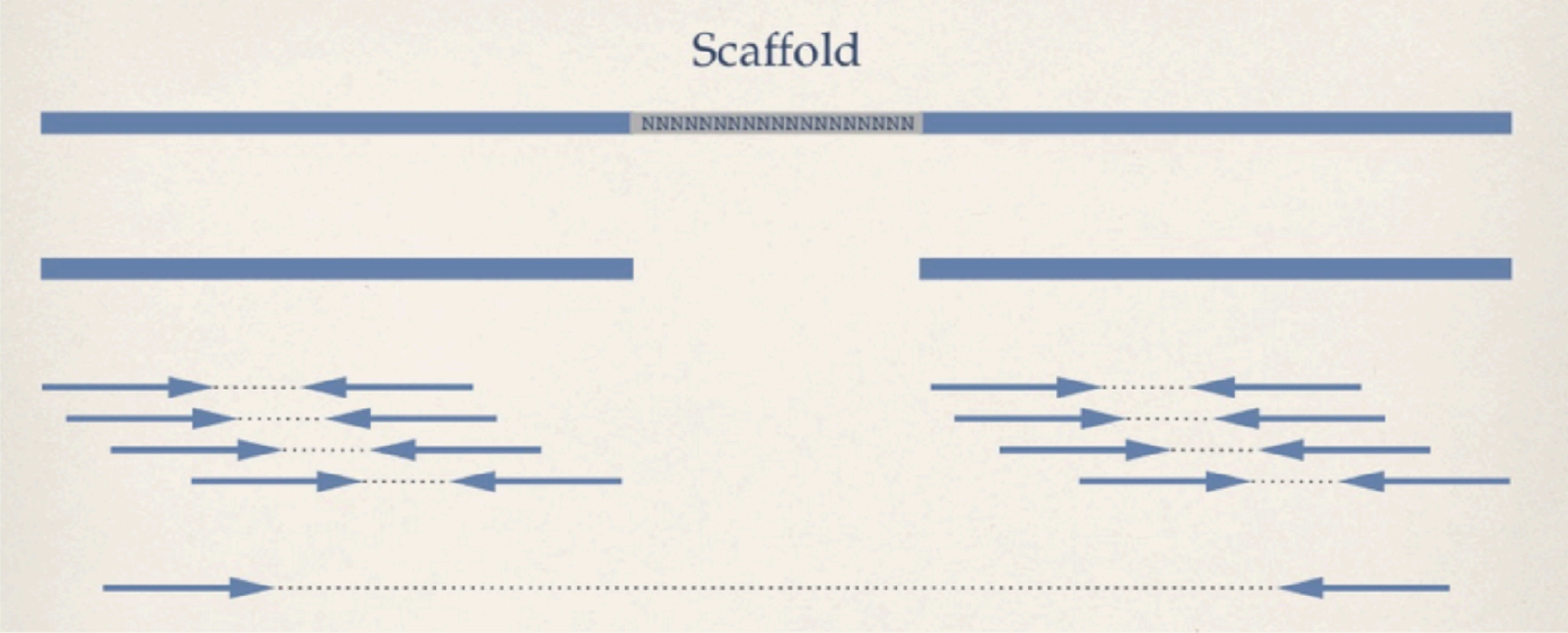
Aligned reads

```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA
                    CGTAGCGCATTACACAGATT
                      TAGCGCATTACACAGATTAG
```

Consensus contig

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

Scaffolding methods

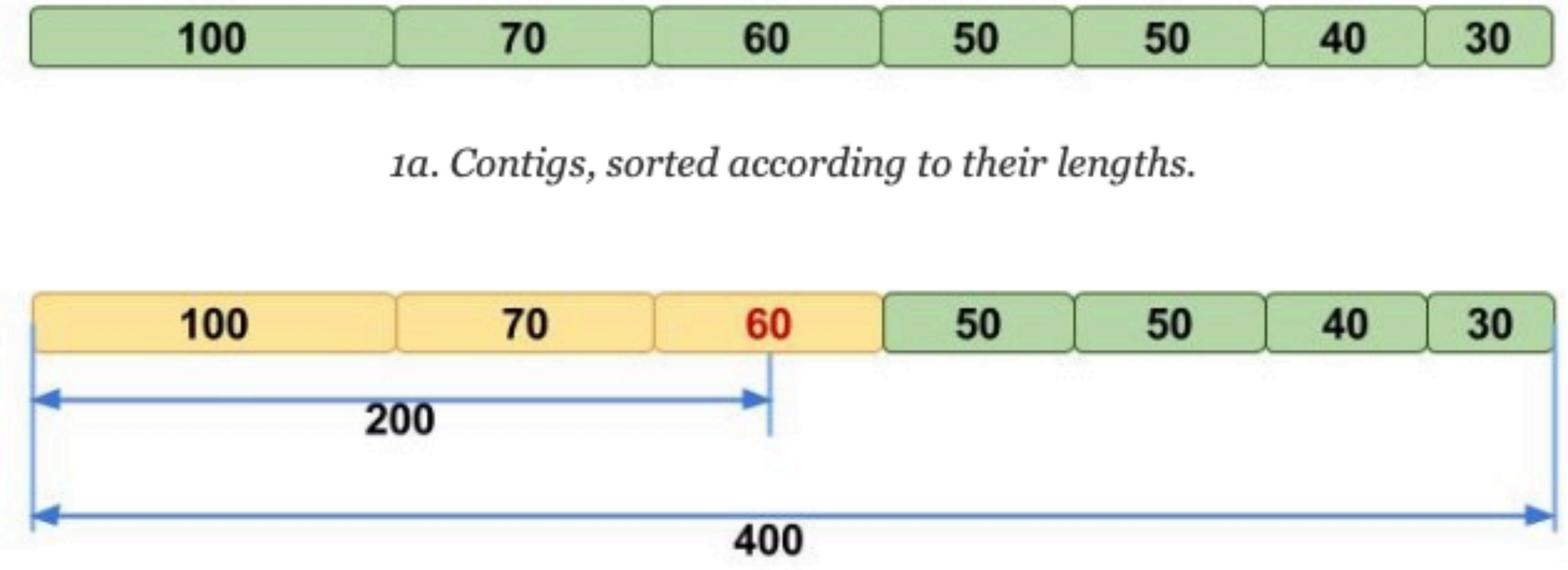


Scaffold: joining and orienting contigs

*Scaffolding methods: mate-pairs (blerg), optical maps (bionano), **Hi-C**, Nanopore UltraLong reads*

N50

- N50: half of the genome is assembled in scaffolds that are the N50 size, or larger



1a. Contigs, sorted according to their lengths.

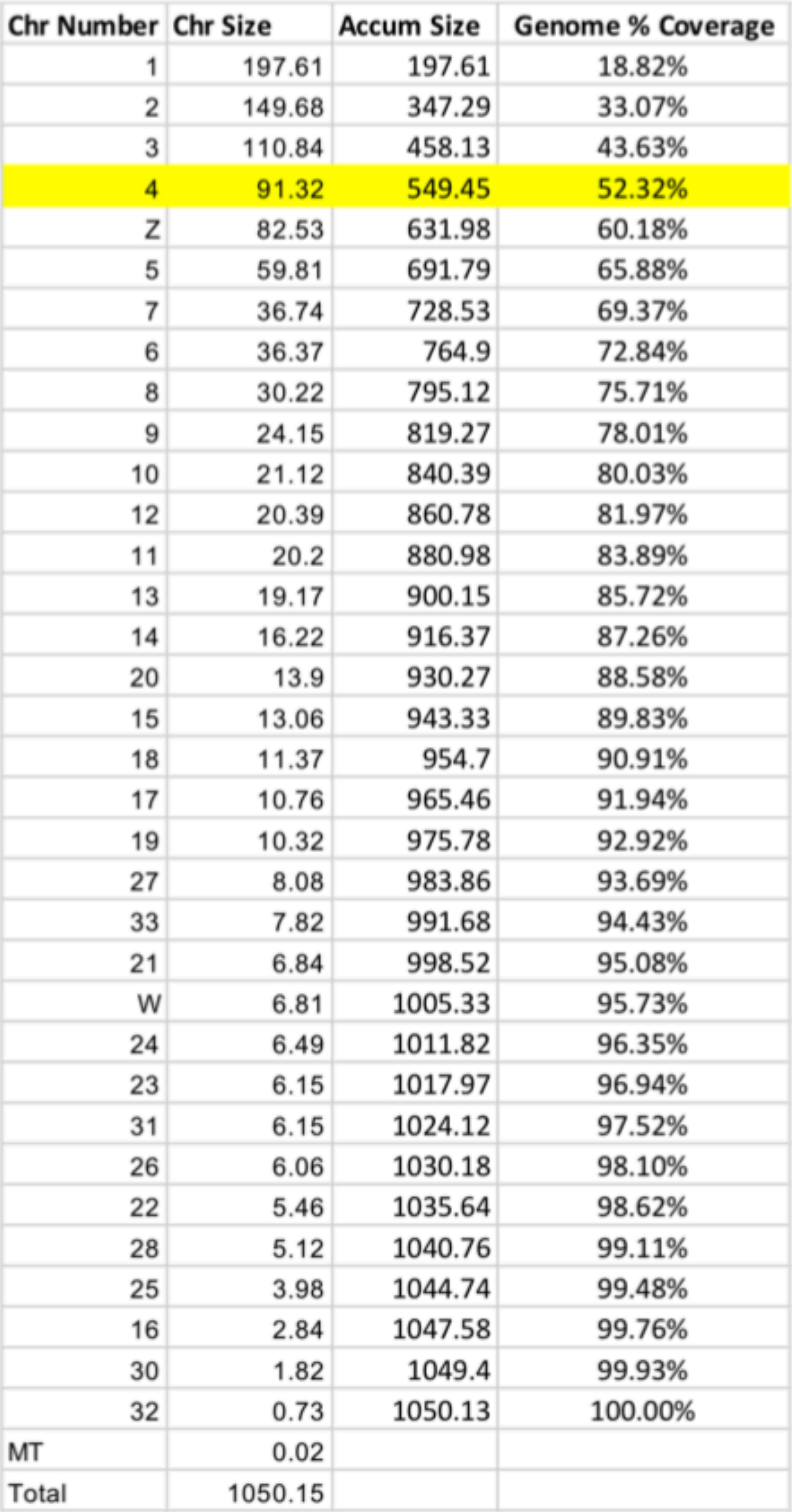
1b. Calculation of N50 using sorted contigs.

Fig. 1. Example of calculating N50 for a set of seven contigs. Here N50 equals 60 kbp.

molecularecologist.com

Quality metrics in genomics

- **N50: half of the genome is assembled in scaffolds that are the N50 size, or larger**



Chr Number	Chr Size	Accum Size	Genome % Coverage
1	197.61	197.61	18.82%
2	149.68	347.29	33.07%
3	110.84	458.13	43.63%
4	91.32	549.45	52.32%
Z	82.53	631.98	60.18%
5	59.81	691.79	65.88%
7	36.74	728.53	69.37%
6	36.37	764.9	72.84%
8	30.22	795.12	75.71%
9	24.15	819.27	78.01%
10	21.12	840.39	80.03%
12	20.39	860.78	81.97%
11	20.2	880.98	83.89%
13	19.17	900.15	85.72%
14	16.22	916.37	87.26%
20	13.9	930.27	88.58%
15	13.06	943.33	89.83%
18	11.37	954.7	90.91%
17	10.76	965.46	91.94%
19	10.32	975.78	92.92%
27	8.08	983.86	93.69%
33	7.82	991.68	94.43%
21	6.84	998.52	95.08%
W	6.81	1005.33	95.73%
24	6.49	1011.82	96.35%
23	6.15	1017.97	96.94%
31	6.15	1024.12	97.52%
26	6.06	1030.18	98.10%
22	5.46	1035.64	98.62%
28	5.12	1040.76	99.11%
25	3.98	1044.74	99.48%
16	2.84	1047.58	99.76%
30	1.82	1049.4	99.93%
32	0.73	1050.13	100.00%
MT	0.02		
Total	1050.15		

Scaffold N50

@ Chromosome level

N50 = 91Mb

Assembled size= 1Gb

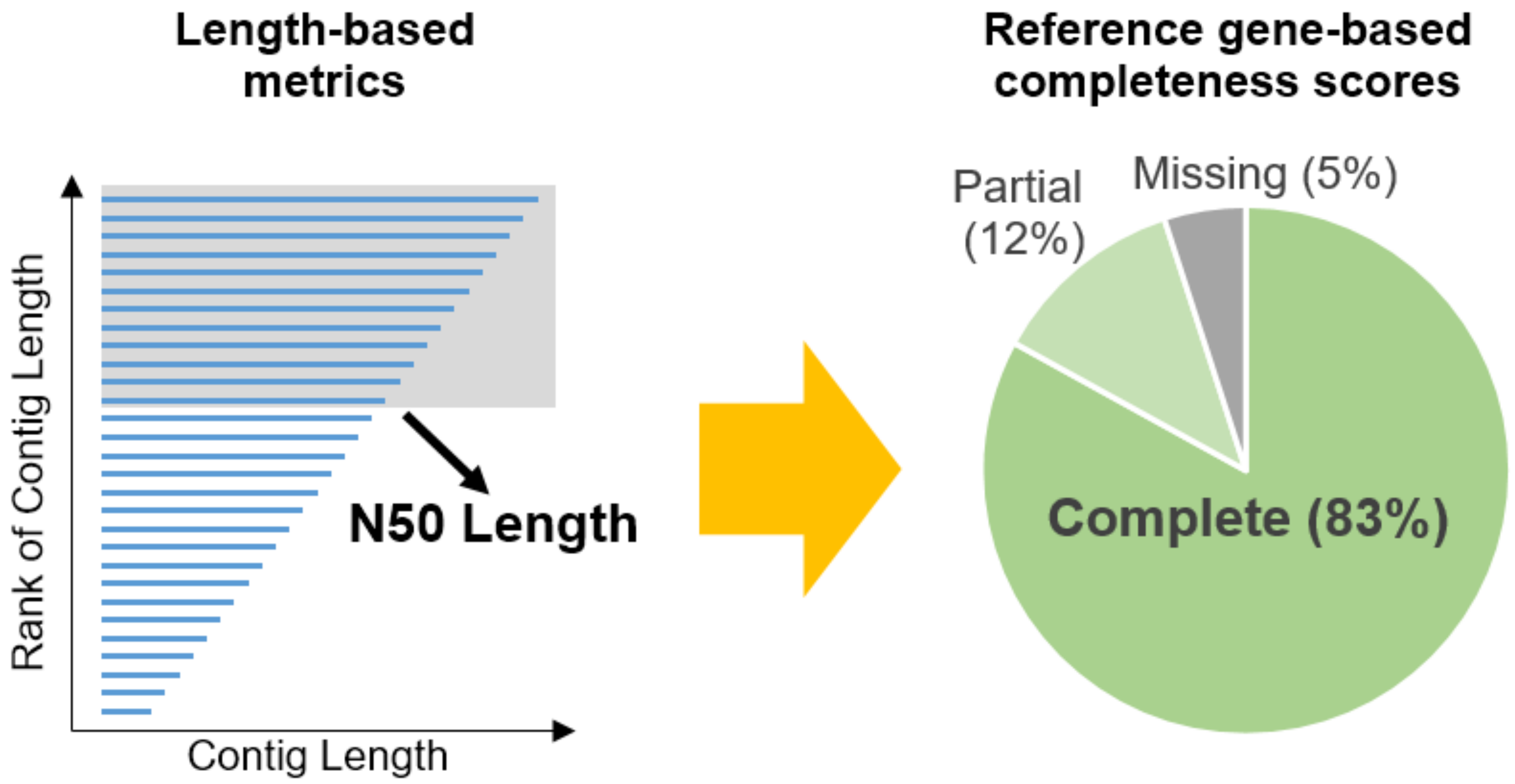
How many scaffolds= 32

BUSCO v3

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

- The quality metrics for genome assembly should not be only the ones related to contiguity, rather, the composition of the genes present in the assembly is also crucial

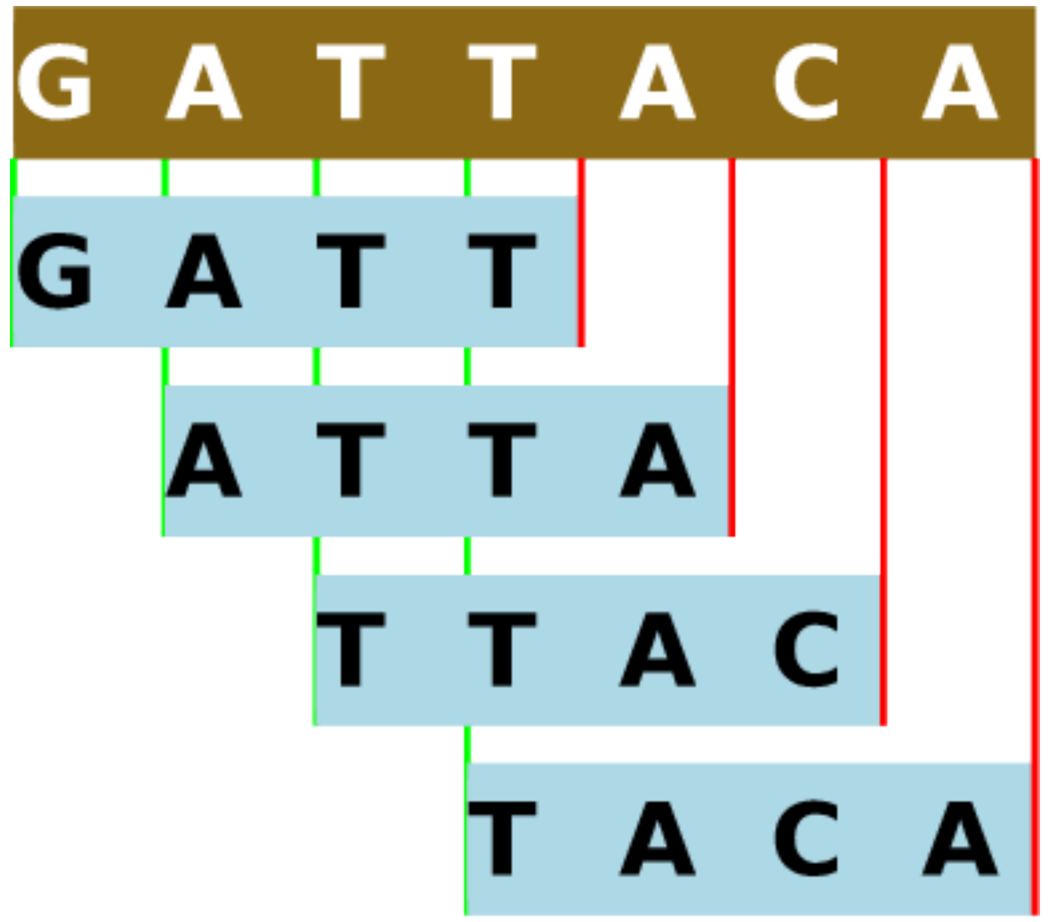
More accurate assessment for genome assembly!



OUR LAB TODAY!

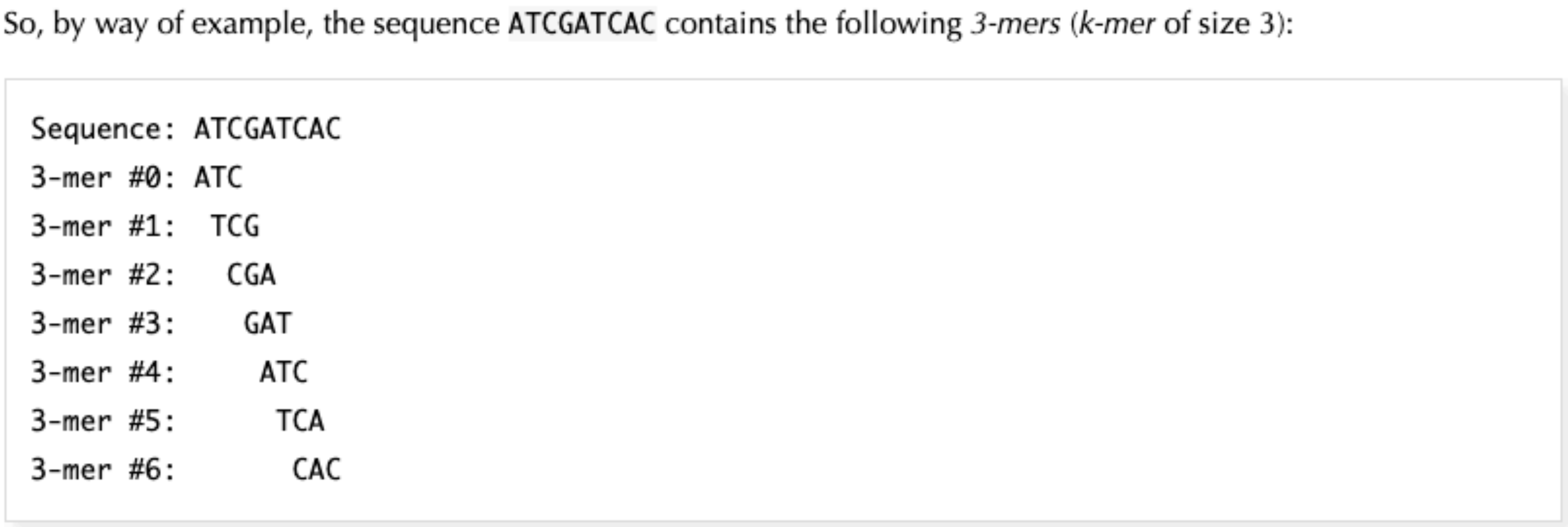
- Count Kmers
- Run general reads statistics
- Run genome assembly with Hifiasm
- Run MitoHiFi to assemble and annotate a mitochondrial genome!

KMER ANALYSIS



WHAT ARE K-MERS ?

- In biology, k-mers are unique subsequences of a sequence of length k



So, by way of example, the sequence ATCGATCAC contains the following *3-mers* (*k-mer* of size 3):

```
Sequence: ATCGATCAC
3-mer #0: ATC
3-mer #1:  TCG
3-mer #2:   CGA
3-mer #3:    GAT
3-mer #4:     ATC
3-mer #5:      TCA
3-mer #6:       CAC
```

APPLICATIONS OF K-MER ANALYSIS

- Genome assembly: K-mers used to construct De Brujin graphs
- Detect bacterial contamination on eukaryotic genome assembly (CG content discrepancies)
- Correcting NSG data
- Detect horizontal gene transfers
- Identification of CpG Islands
- **Estimation of genome size and heterozygosity**
- Genome assembly k-mer completeness

WHY ARE K-MERS SO POPULAR?

"Decomposing a sequence into its *k-mers* for analysis allows this set of fixed-size chunks to be analysed rather than the sequence, and this can be more efficient." (Bernardo Cavijo)

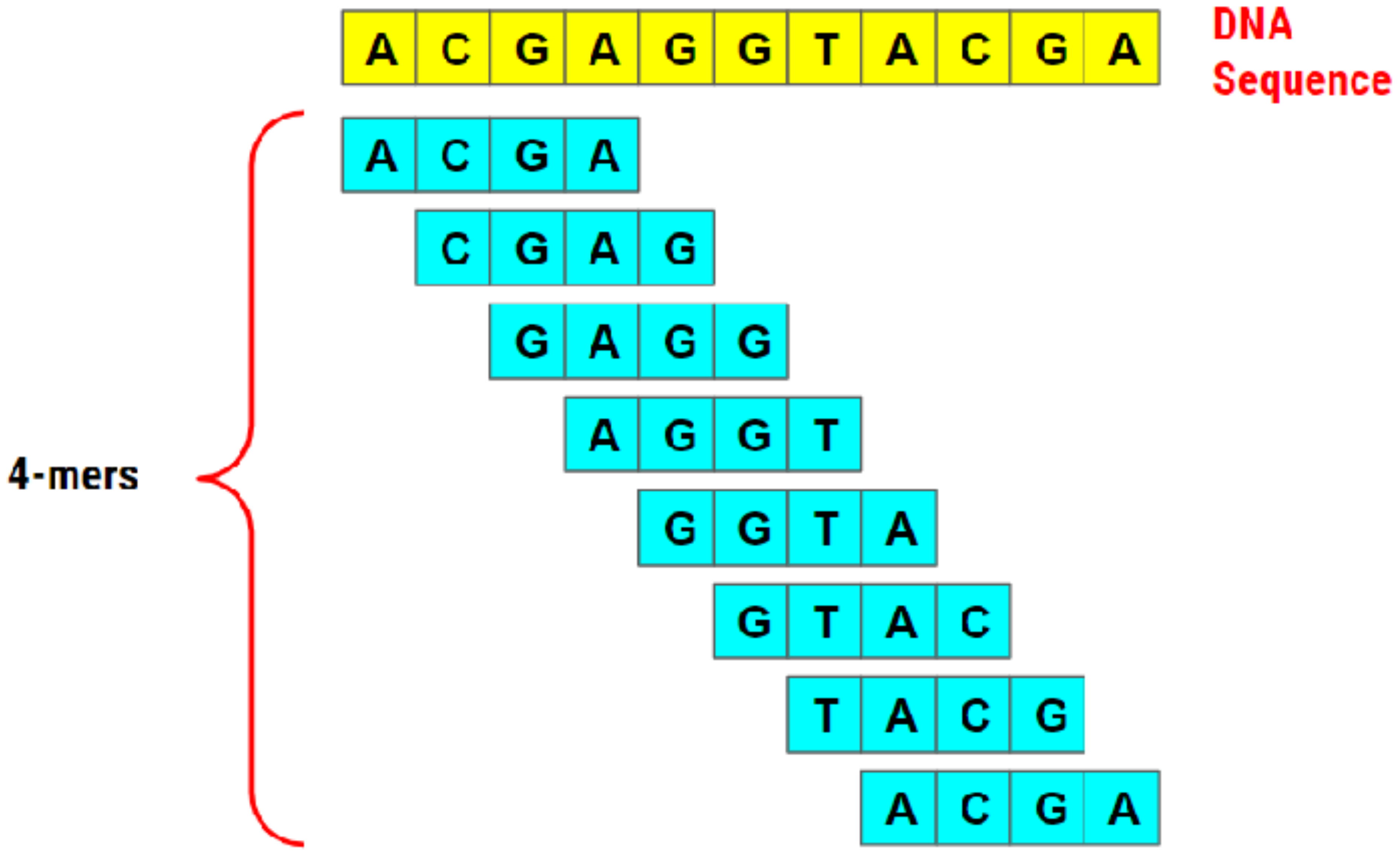
https://bioinfologics.github.io/post/2018/09/17/k-mer-counting-part-i-introduction/

KMER SIZE

Choosing *k*: specificity vs. Sensitivity

- Using a *k* that is too small will result in many unrelated sequences being composed of the same *k-mers*, in a textbook case of specificity loss because there being very few possible *k-mers* of that size. In the extreme of the small *k*, k=1 only distinguishes two *canonical k-mers*: A and C. 1-mer analysis is incredibly popular in biology, but it is best known by the name of *GC content analysis*.

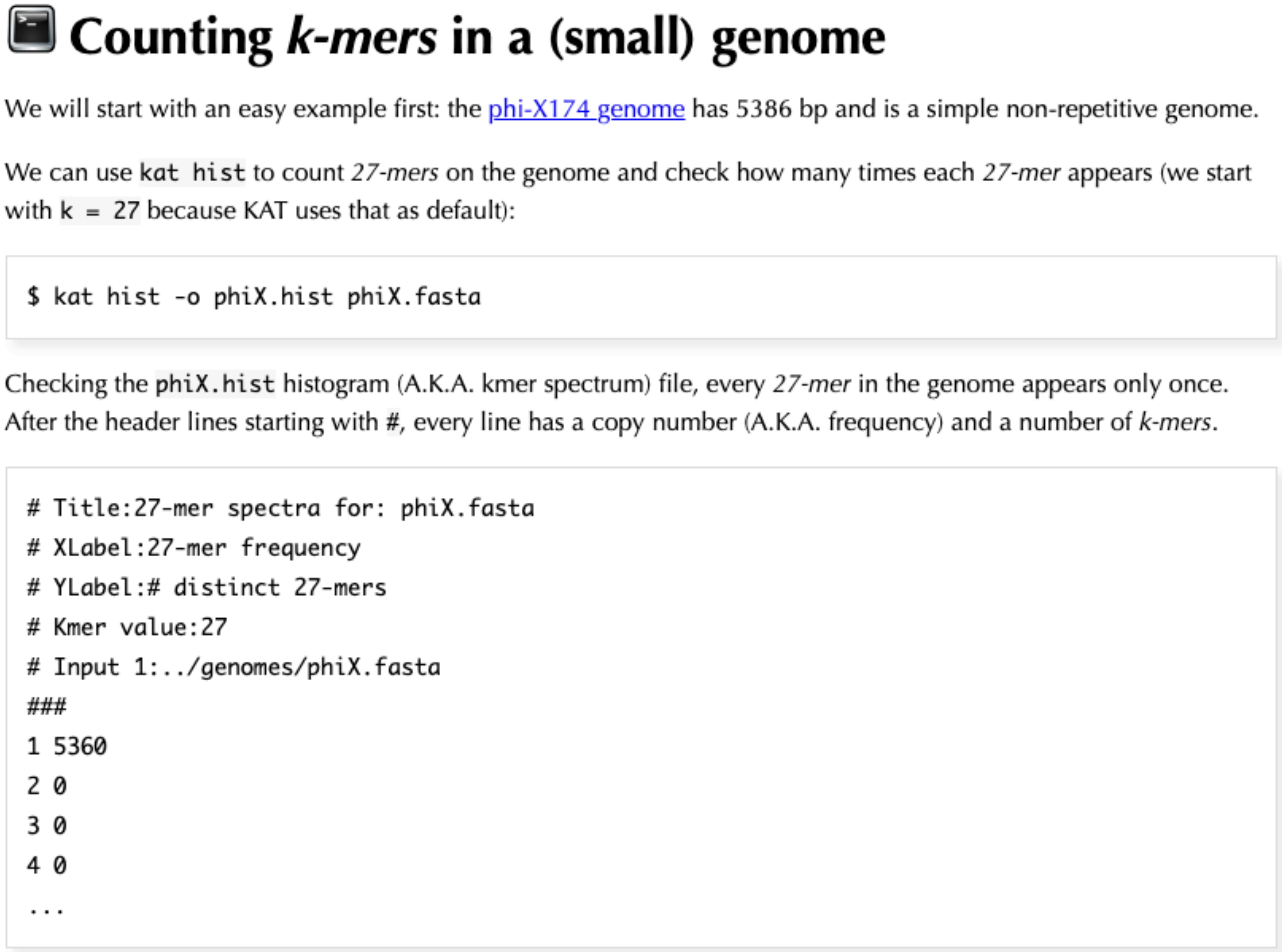
- Using extremely large *k* values would sacrifice many of the benefits and sensitivity of *k-mer* analyses in the first place. (Bernado Cavijo's post)



Why do we chose k=31 so often?

One reason is: it is specific enough that a large number of them are unique both in mammalian-sized genomes and in bacterial genome databases.

COUNT AND HISTO



Counting *k-mers* in a (small) genome

We will start with an easy example first: the phi-X174 genome has 5386 bp and is a simple non-repetitive genome.

We can use `kat hist` to count *27-mers* on the genome and check how many times each *27-mer* appears (we start with `k = 27` because KAT uses that as default):

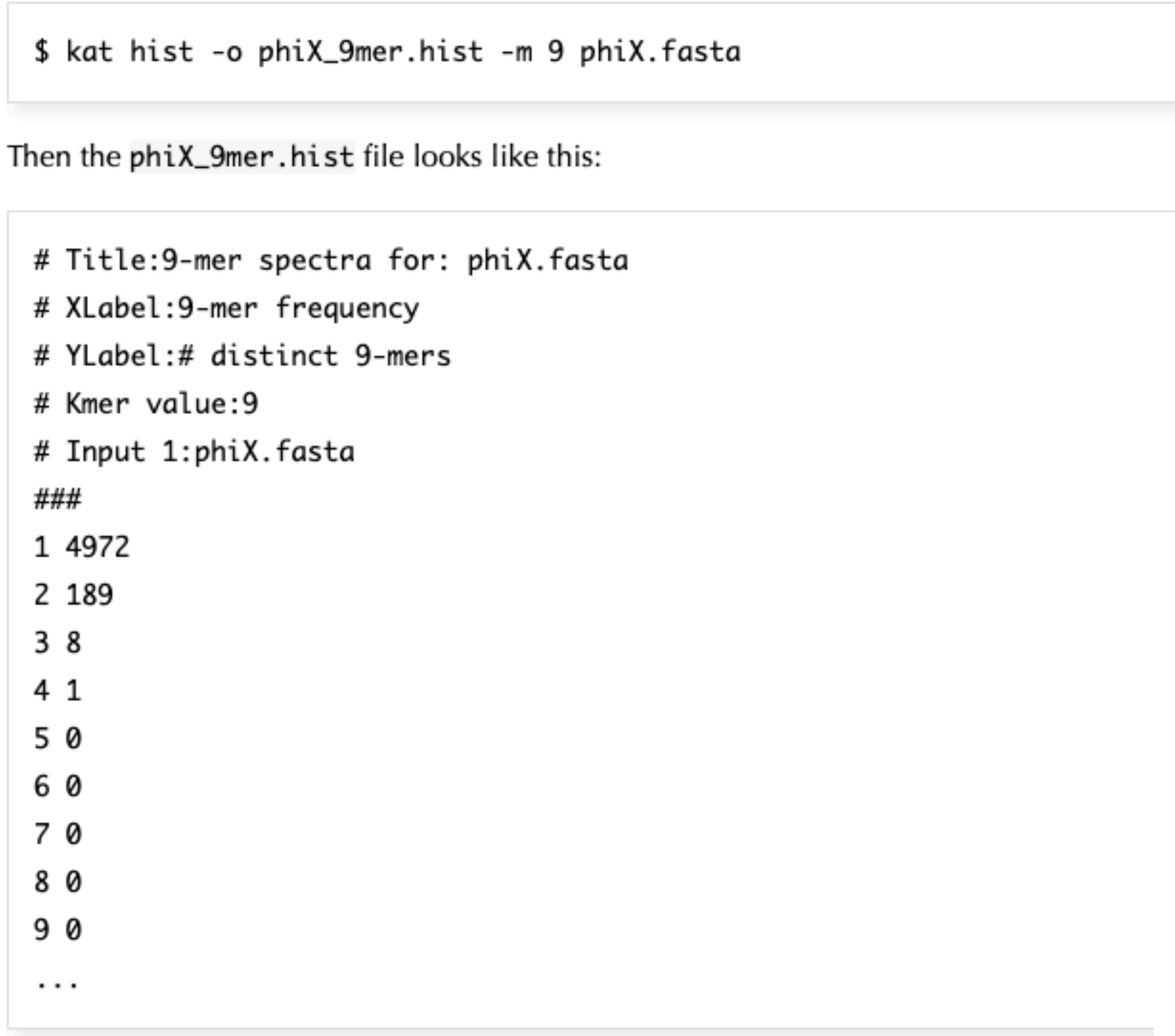
```
$ kat hist -o phiX.hist phiX.fasta
```

Checking the `phiX.hist` histogram (A.K.A. kmer spectrum) file, every *27-mer* in the genome appears only once. After the header lines starting with `#`, every line has a copy number (A.K.A. frequency) and a number of *k-mers*.

```
# Title:27-mer spectra for: phiX.fasta
# XLabel:27-mer frequency
# YLabel:# distinct 27-mers
# Kmer value:27
# Input 1:../genomes/phiX.fasta
###
1 5360
2 0
3 0
4 0
...
```

Bernardo Cavijo's post

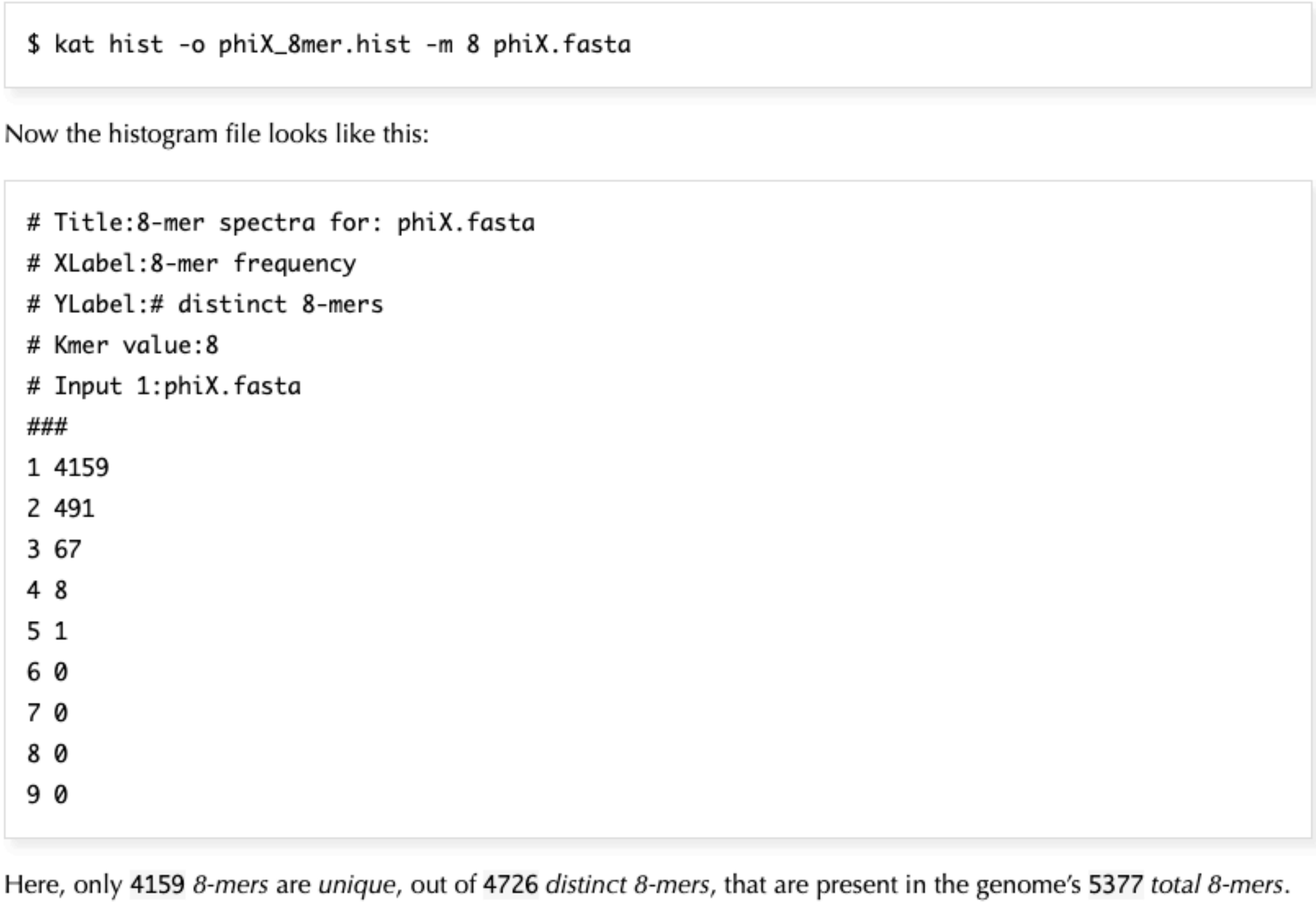
COUNT AND HISTO



```
$ kat hist -o phiX_9mer.hist -m 9 phiX.fasta
```

Then the `phiX_9mer.hist` file looks like this:

```
# Title:9-mer spectra for: phiX.fasta
# XLabel:9-mer frequency
# YLabel:# distinct 9-mers
# Kmer value:9
# Input 1:phiX.fasta
###
1 4972
2 189
3 8
4 1
5 0
6 0
7 0
8 0
9 0
...
```



```
$ kat hist -o phiX_8mer.hist -m 8 phiX.fasta
```

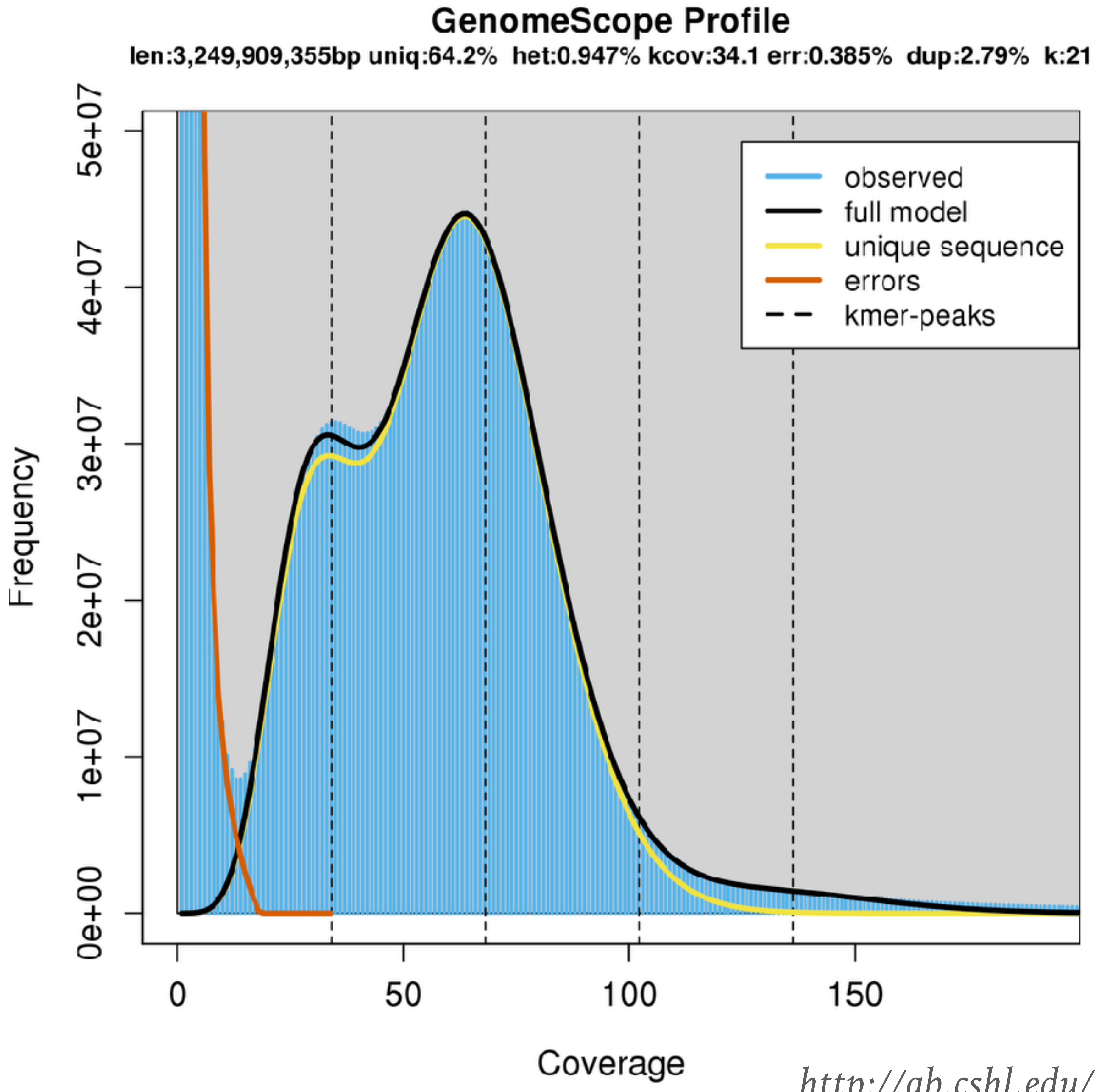
Now the histogram file looks like this:

```
# Title:8-mer spectra for: phiX.fasta
# XLabel:8-mer frequency
# YLabel:# distinct 8-mers
# Kmer value:8
# Input 1:phiX.fasta
###
1 4159
2 491
3 67
4 8
5 1
6 0
7 0
8 0
9 0
```

Here, only `4159` *8-mers* are *unique*, out of `4726` *distinct 8-mers*, that are present in the genome's `5377` *total 8-mers*.

Bernardo Cavijo's post

A TYPICAL KMER PLOT FOR A DIPLOID SPECIES



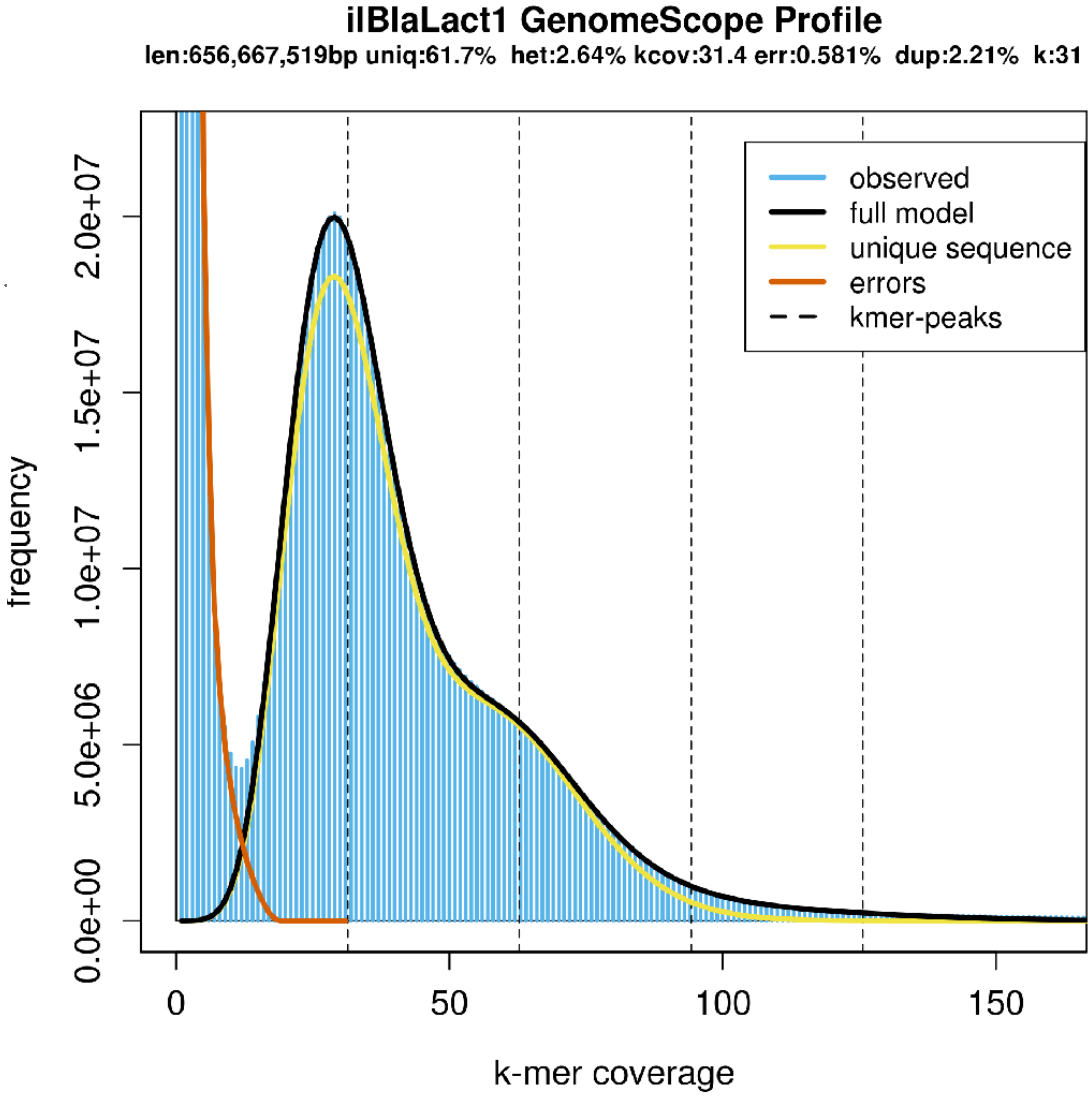
Choloepus didactylus (VGP)

http://qb.cshl.edu/genomescope/analysis.php?code=bVuZNlhwn2tVCHhRN71I

A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH HIGH HETEROZYGOSITY

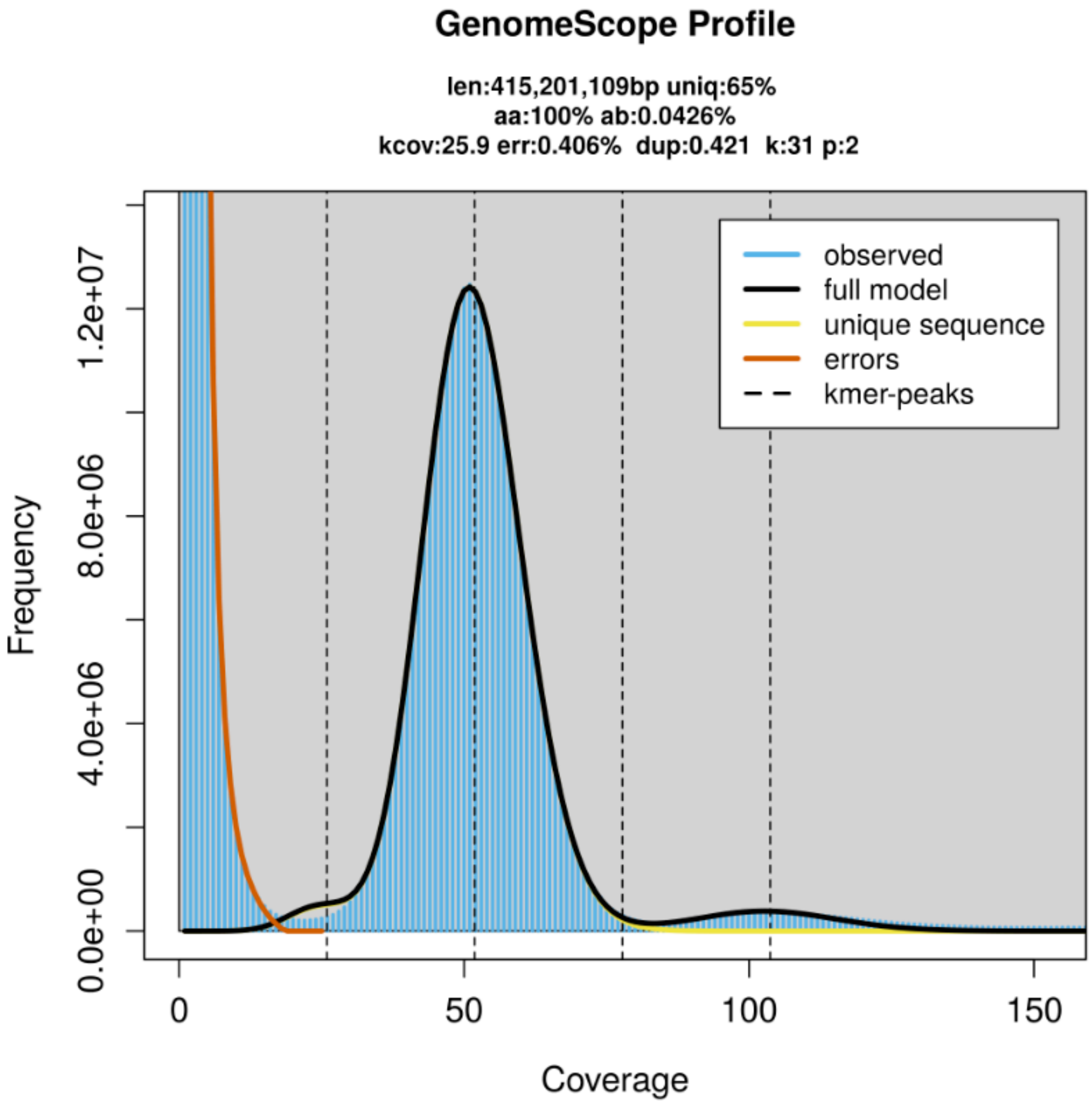
Blastobasis lacticolella (DToL)

Wakely's dowd



A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH LOW HETEROZYGOSITY

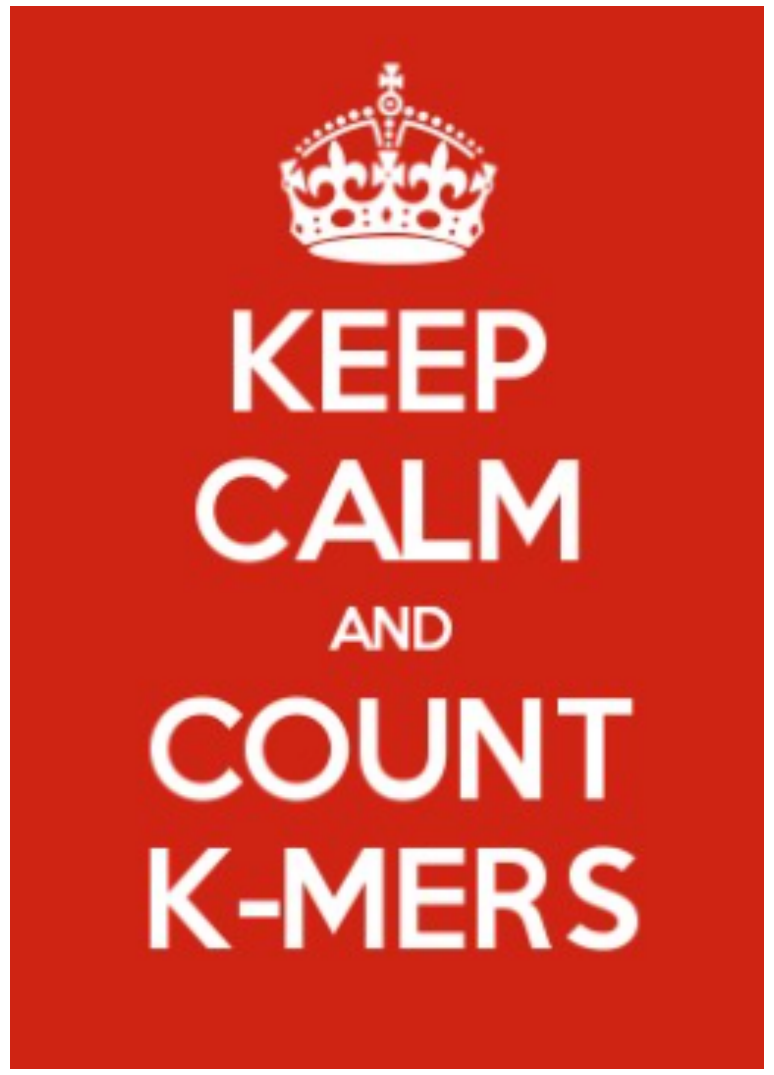
Rhytidiadelphus loreus
Little Shaggy-moss



KMERS CAN BE ANALYSED ONLY FOR HIGH-QUALITY DATA

This means that:

- *If you have sequenced PacBio CLR, you should have short-read sequencing for kmer analysis (and for polishing)*
- *But with* ***Pacbio HiFi that you can count kmers as you do with short-reads!***



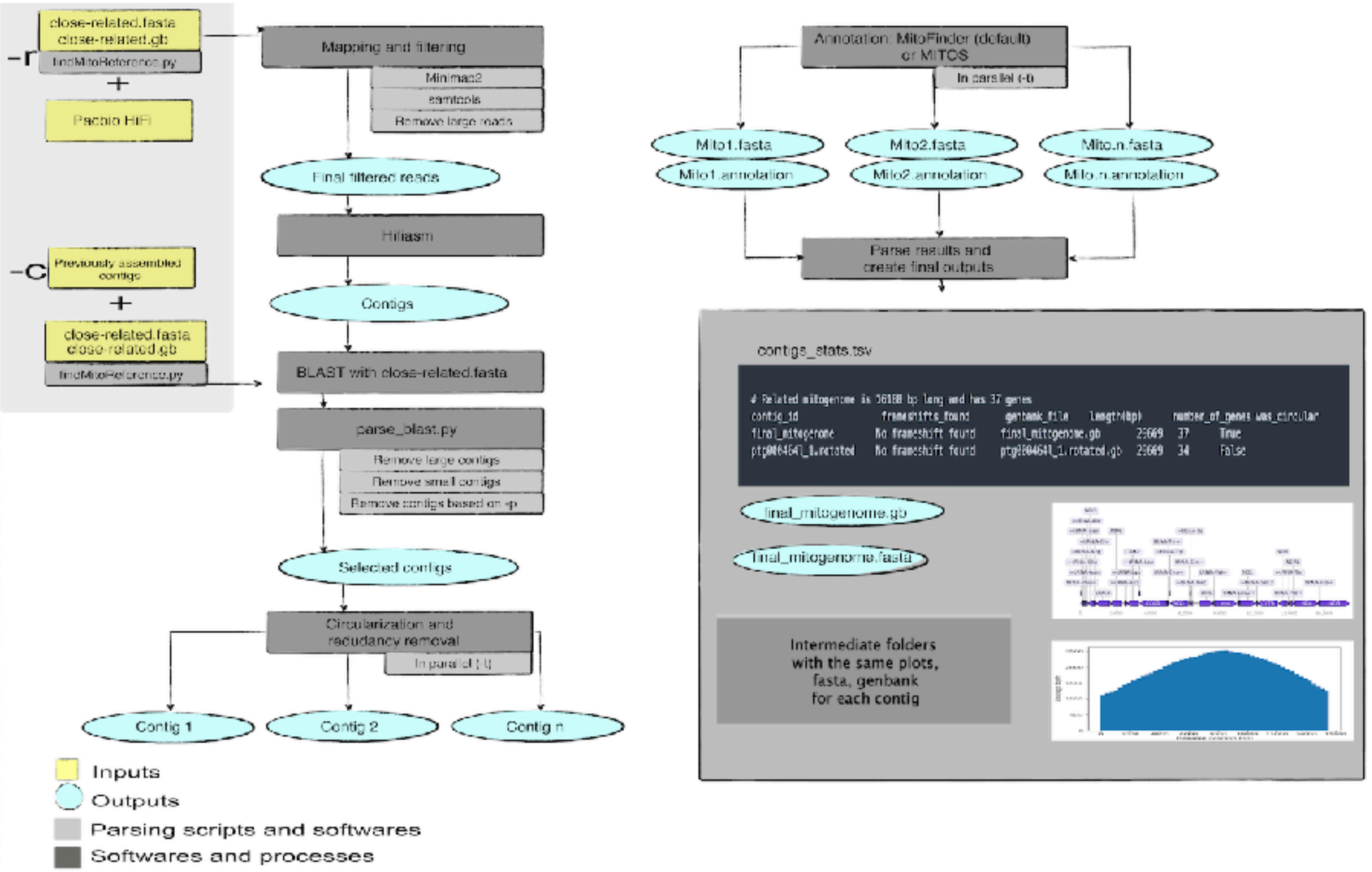
HOW'S THE DEBUGGING GOING?
JORGE CHAM © 2005
www.phdcomics.com

Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio High Fidelity reads

Marcela Uliano-Silva, João Gabriel R. N. Ferreira, Ksenia Krasheninnikova, Darwin Tree of Life Consortium, Giulio Formenti, Linelle Abueg, James Torrance, Eugene W. Myers, Richard Durbin, Mark Blaxter, Shane A. McCarthy

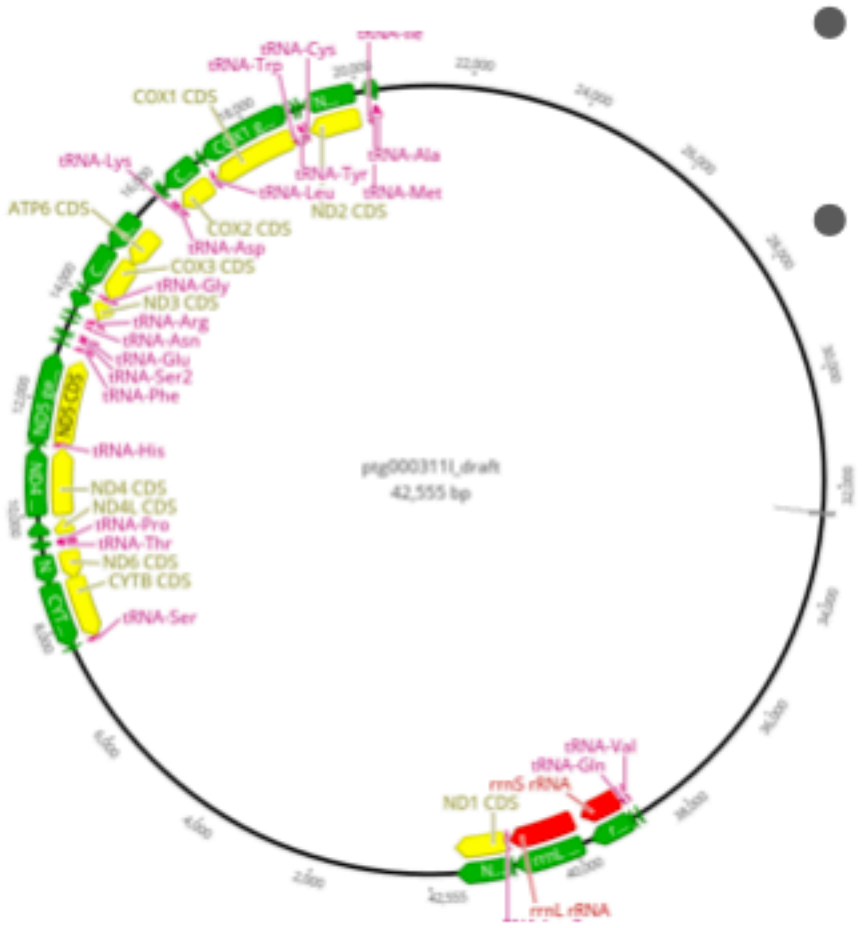
doi: https://doi.org/10.1101/2022.12.23.521667

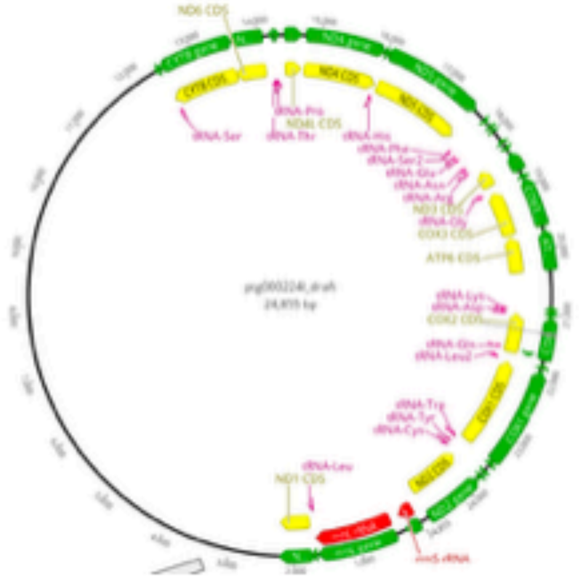


```
# Related mitogenome is 16188 bp long and has 37 genes
contig_id               frameshifts_found       genbank_file    length(bp)      number_of_genes was_circular
final_mitogenome        No frameshift found     final_mitogenome.gb     23669   37      True
ptg000464l_1.rotated    No frameshift found     ptg000464l_1.rotated.gb 23669   34      False
```

MitoHifi GitHub python

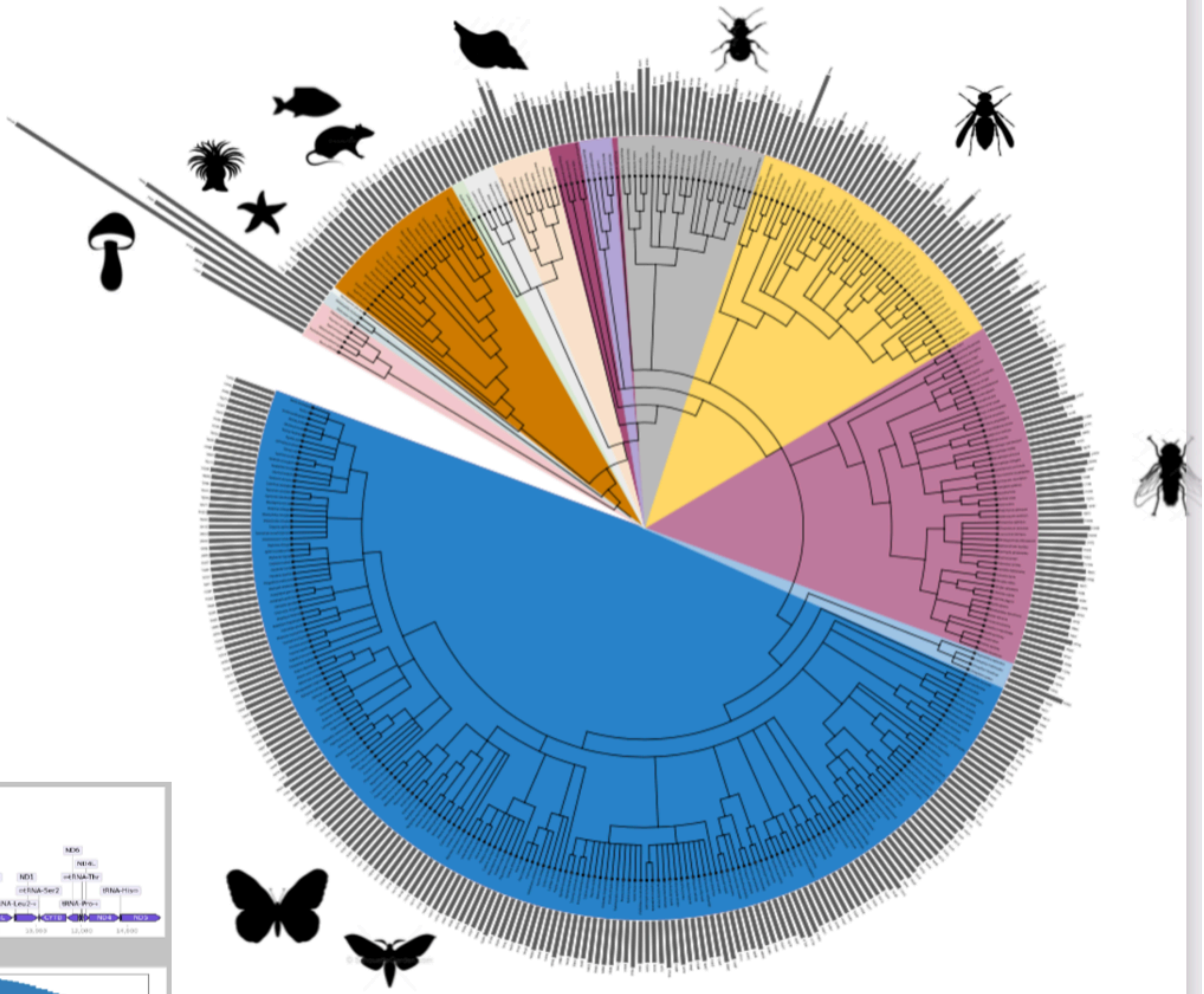
- Assemble
- Circularise
- Annotate
- Heteroplasmy

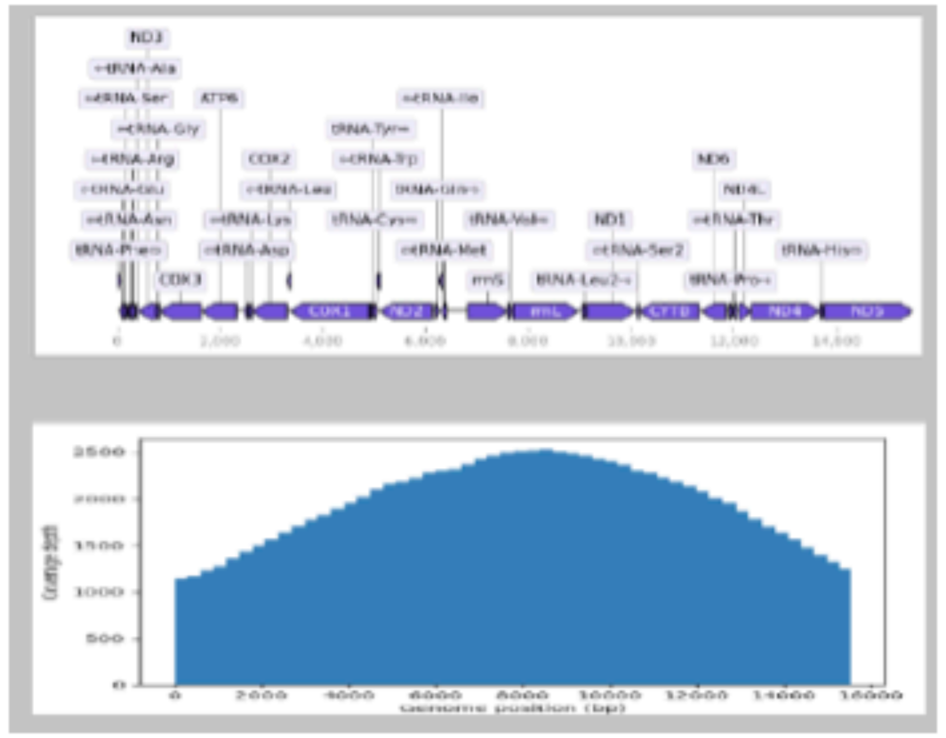




Colored ranges

- ■ Fungi
- ■ Cnidaria
- ■ Annelida
- ■ Chrodata
- ■ Lophotrochozoa
- ■ Mollusca
- ■ Coleoptera
- ■ Hymenoptera
- ■ Diptera
- ■ Lepidoptera
- ■ Limnephilini
- ■ Polyneoptera
- ■ Arthropoda
- ■ Deuterostomia

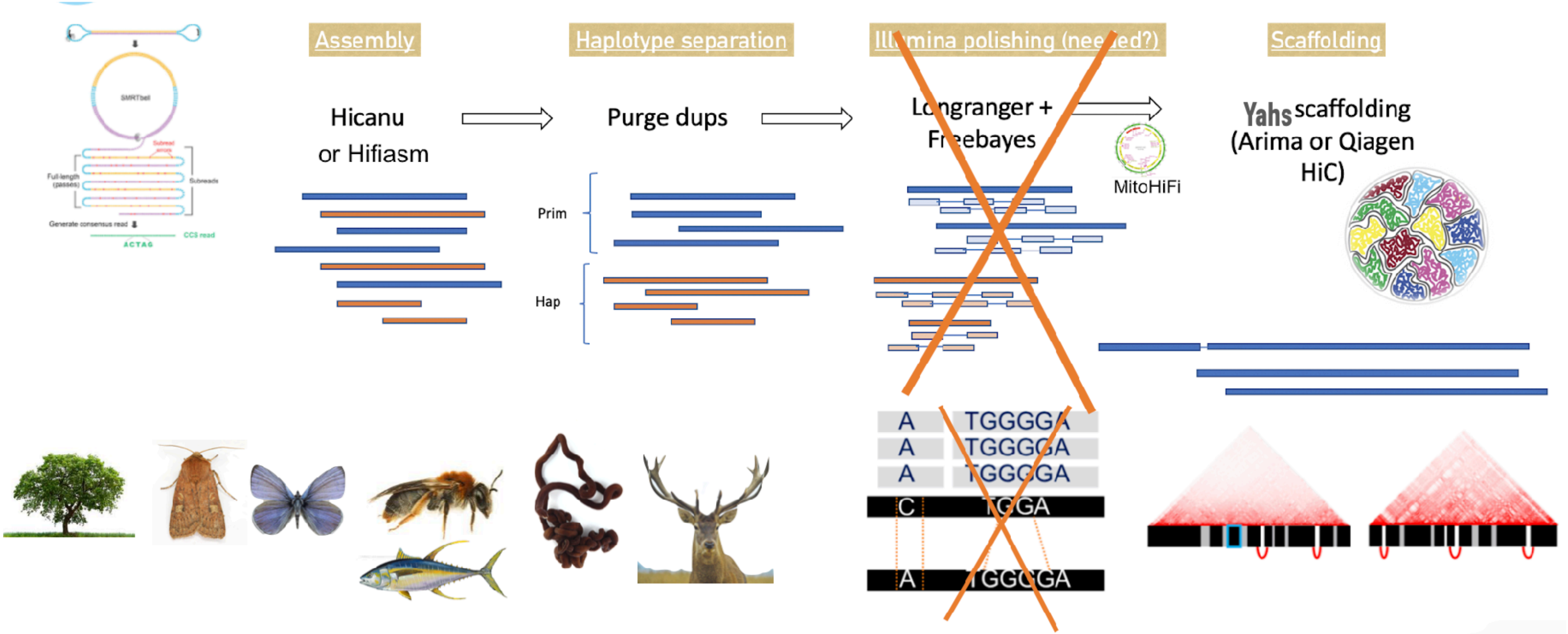




Uliano-Silva et al., submitted

DToL Current Pipeline

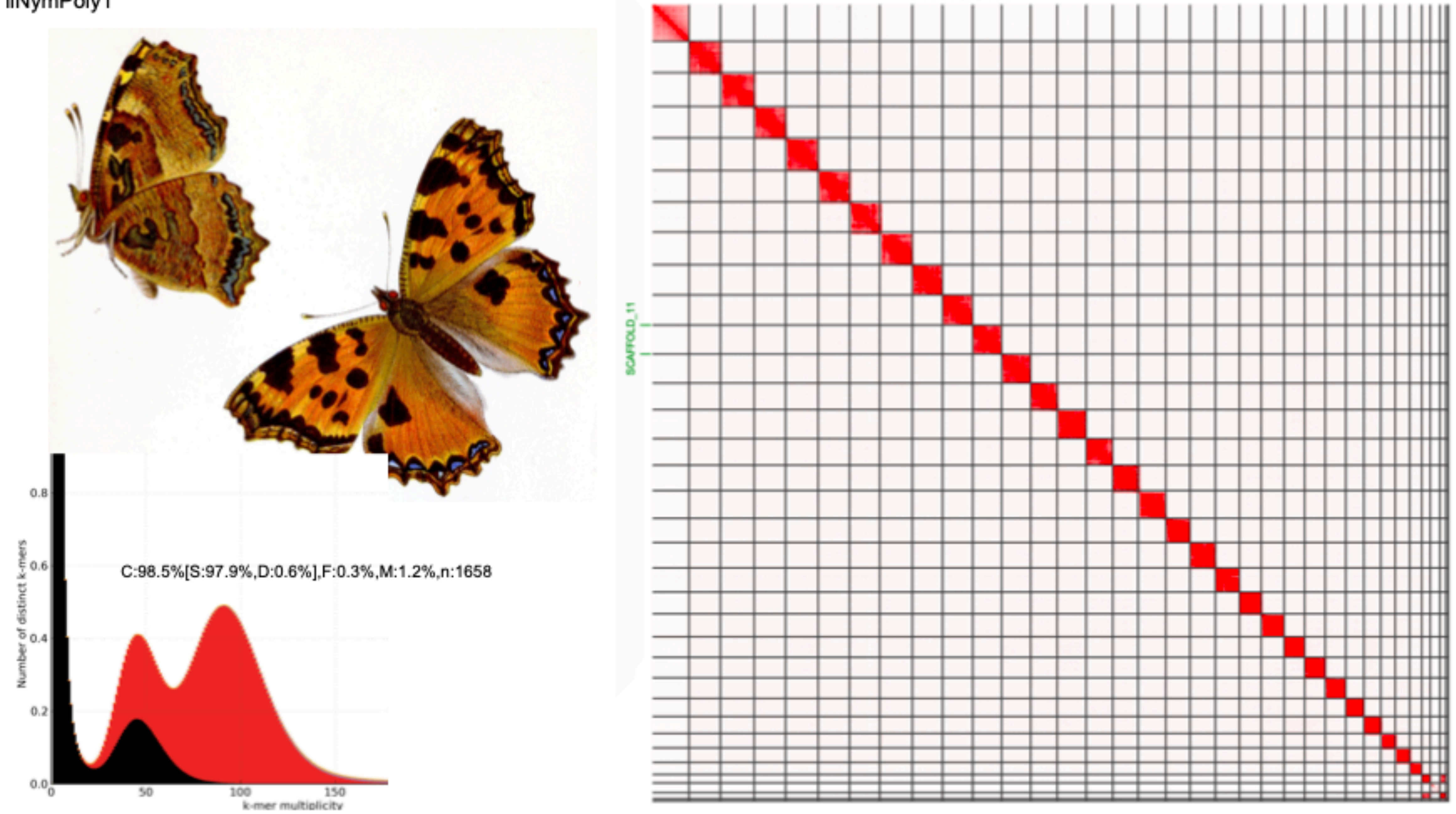
- Sequencing technologies: PacBio HiFi + HiC (Arima or Qiagen)



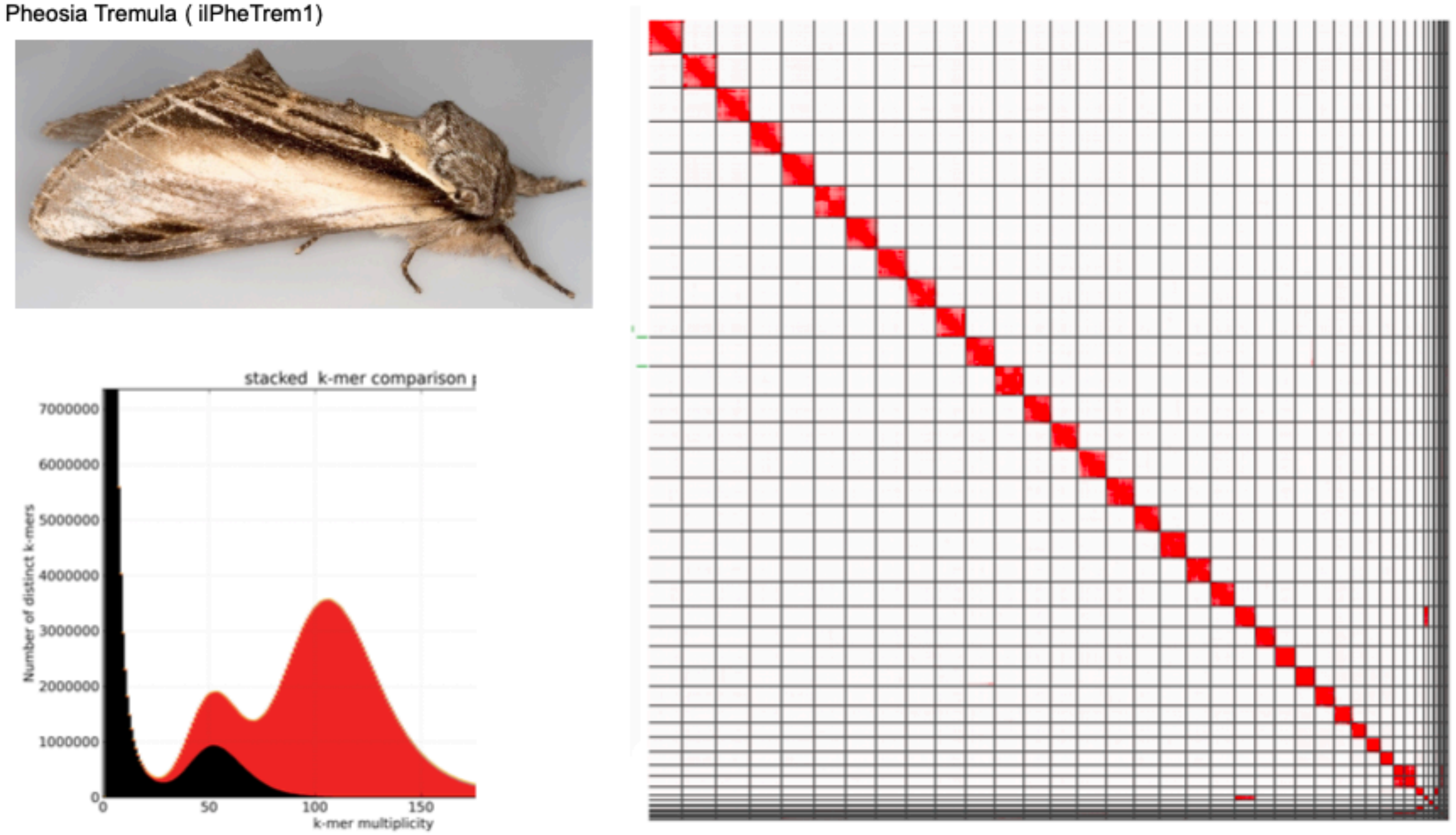
OUT OF THE BOX

ilNymPoly1

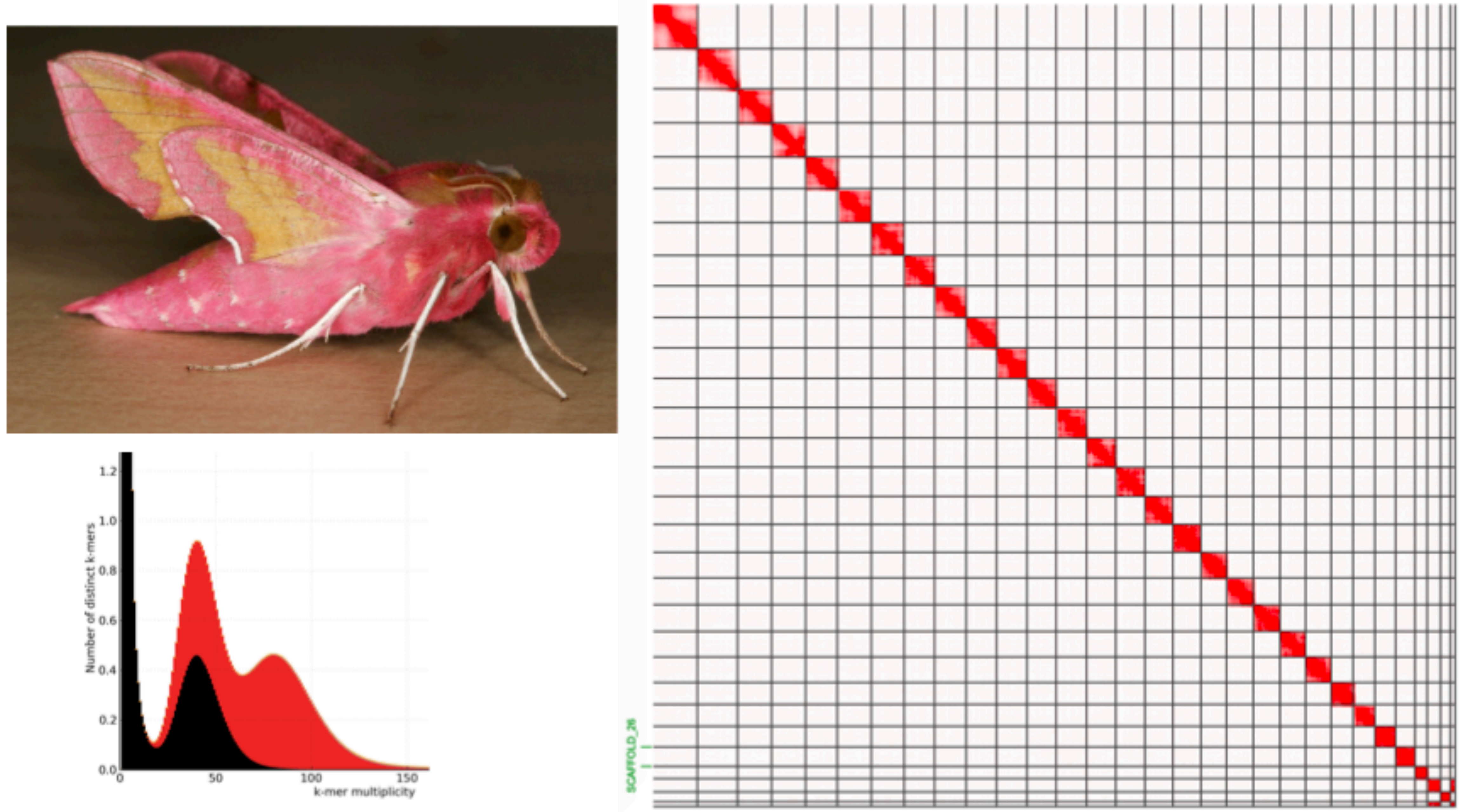
Hifiasm


C:98.5%[S:97.9%,D:0.6%],F:0.3%,M:1.2%,n:1658

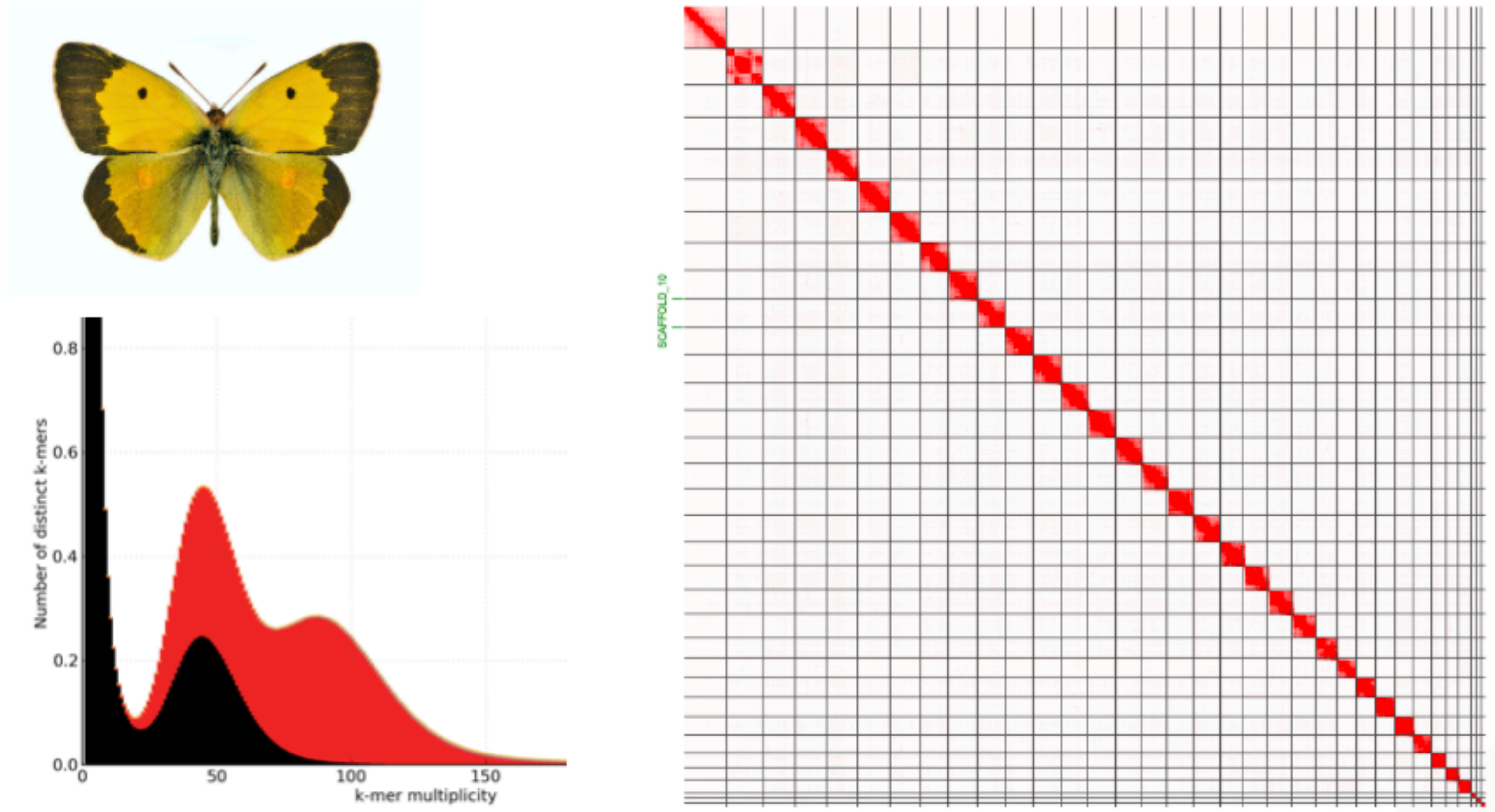
Pheosia Tremula (ilPheTrem1)



ilDeiPorc1 - *Deilephila porcellus*



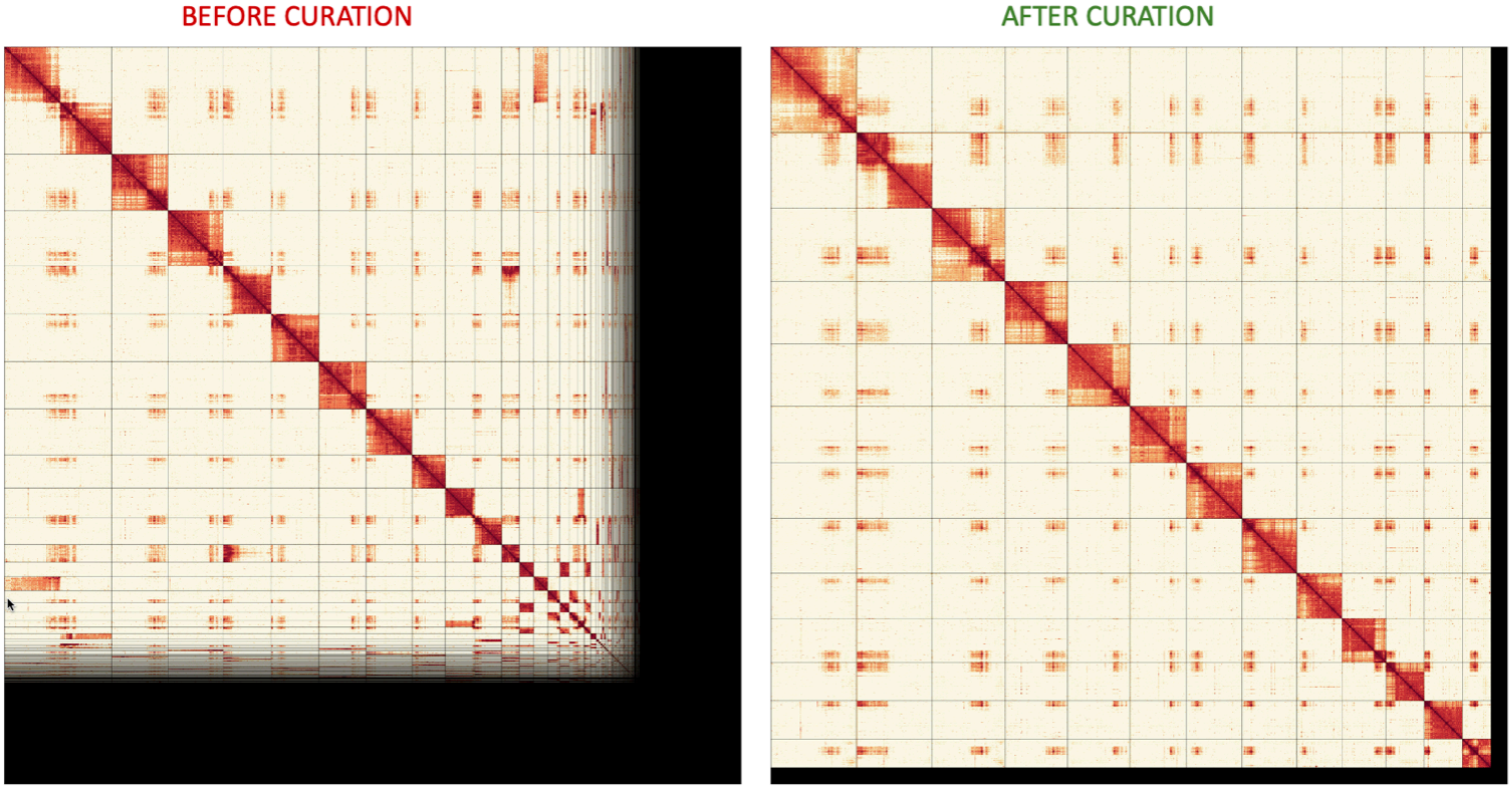
ilColCroc2 - Colias_crocea



BUT! MANUAL CURATION IS ESSENTIAL

Jo Wood's GRIT curation team

iyOphLute1_1 -> 372 breaks, 1,377 joins, 2 haplotig removals

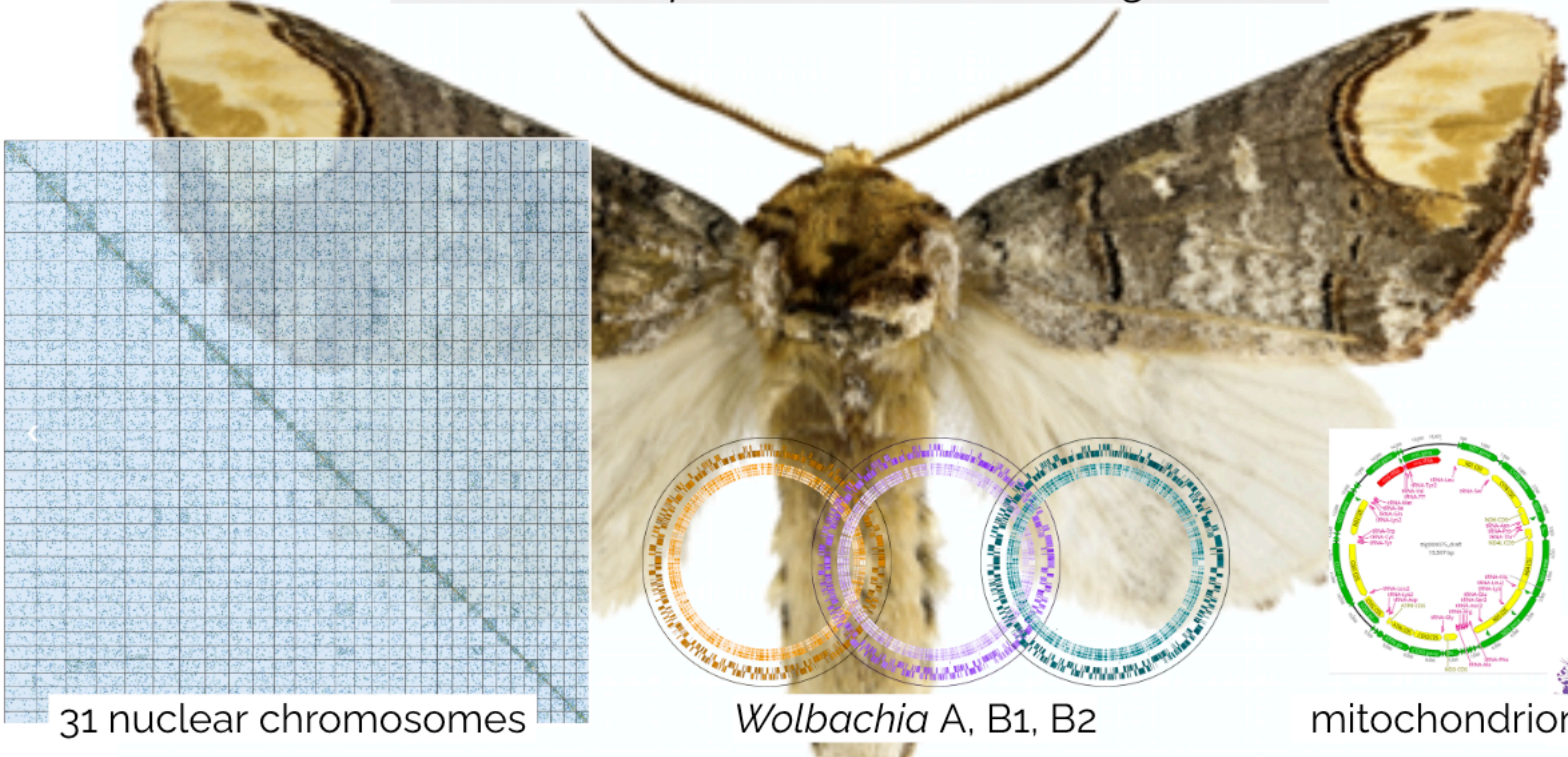


Ophion luteus

Alan Tracey

Assembling genomes of target *and* cobiome

Phalera bucephala: one moth, five genomes



31 nuclear chromosomes

Wolbachia A, B1, B2

mitochondrion

Emmelien Vancaester, Claudia Weber, Marcela Uliano, Alan Tracey, James Torrance, Jonathan Wood

October 2 - 6

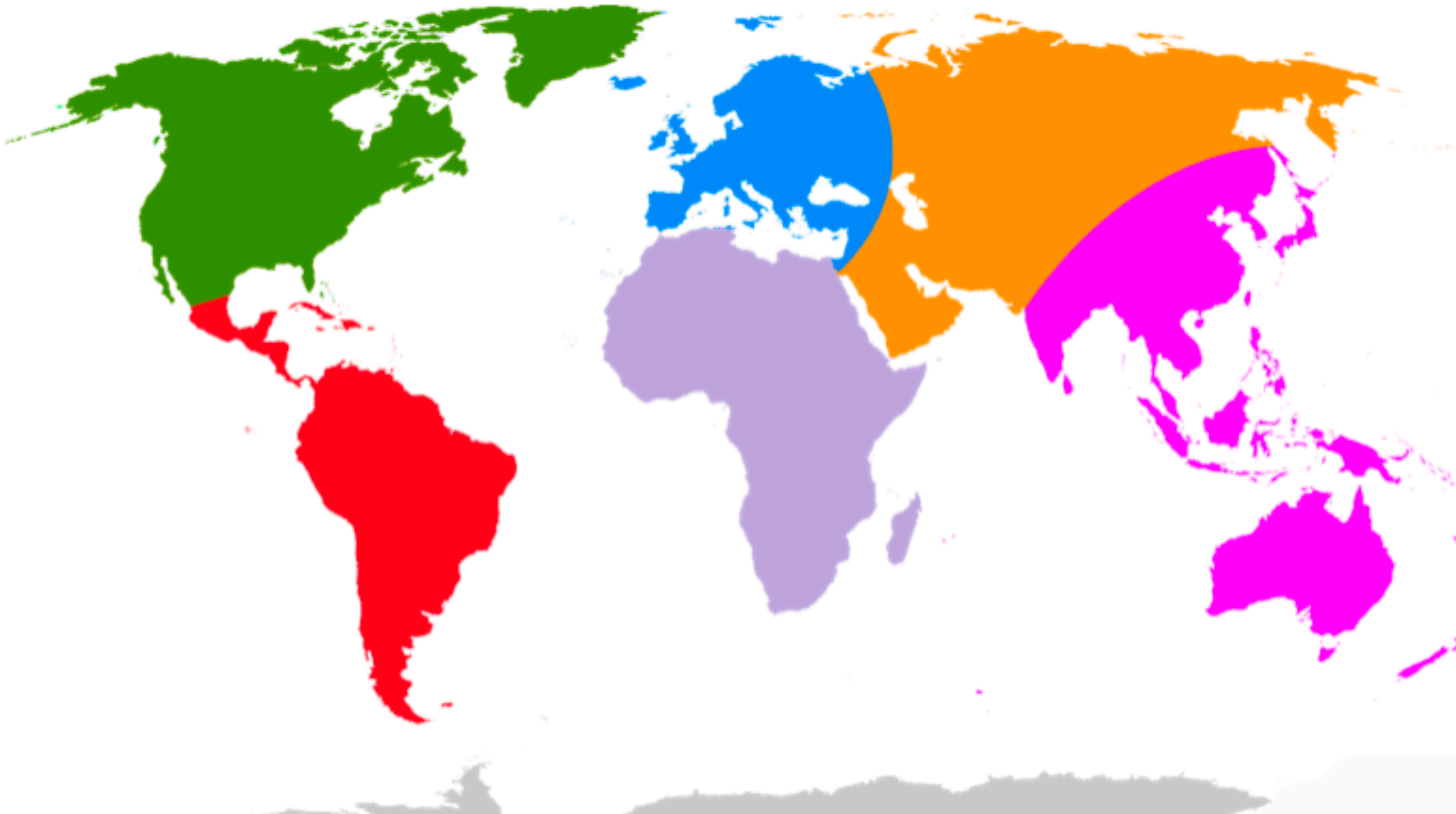
SAVE THE DATE

Biodiversity Genomics 2023

Regional Meetings, Scientific topics, JEDI sessions, keynote speakers.

Free to attend!

@BiodivGenomics



Obrigada! Thank you!

Royal Botanic Gardens Kew

Bill Baker
Ester Gaya
Paul Kersey
Ilia Leitch
Greg Palmer

Royal Botanic Garden Edinburgh

David Bell
David Long
Laura Forrest
Mary Gibby
Michelle Hart
Neil Bell
Pete Hollingsworth
Rebecca Yahr

The Marine Biological Association

Nova Mieszkowska
Willie Wilson
Michael Cunliffe
John Bishop
Helen Jenkins
Robert Mrowicki
Padrick Adkins
Joanna Harley

The University of Edinburgh

Alex Twyford

Earlham Institute

Neil Hall
Iain Macaulay
Karim Gharbi
Jim Lipscombe
David Swarbreck
Ross Lowe
Rob Davey
Felix Shaw
Sally Warring
Jamie McGowan
Alice Minotto
Seanna McTaggart

Natural History Museum

Ian Barnes
Gavin Broad
Jonathan Gabriel
Charlotte Barclay
Andrew Briscoe
Mark Carine
Matt Clark
Gerry Hey
Lauren Hughes
Tim Littlewood
Jacqueline MacKenzie-Dodds
Raju Misra
Ben Price
Chris Raper
Fred Rumsey
John Tweddle
Heather Allen
Darren Chooneea
Lyndall Pereira da Conceicoa
Laura Sivess
Olga Sivell

University of Oxford

University of Oxford and Wytham Woods
Peter Holland
Owen Lewis
Tom Richards
Liam Crowley
Amber Harper
Elisabet Alacid Fernandez
Estelle Kilias
Nigel Fisher
František Sládeček
Lauren Sumner-Rooney
Doug Boyes (CEH)
Alistair McGregor (Brookes Univ)
Karl Wotton (Exeter Univ)
Will [illegible] (Exeter Univ)

University of Cambridge

Richard Durbin
Shane McCarthy
Iliana Bista

EMBL-EBI

Paul Flicek
Suran Jayathilaka
Fergal Martin
David Thybert
Jeena Rajan
Kevin Howe
Guy Cochrane
Peter Harrison
Leanne Haggerty
Jamie Allen
Carlos Garcia Giron
Matthieu Muffato

wellcome sanger institute

Tree of Life
Alan Tracey
Amit Vishwakarma
Andrew Varley
Chloe Leech
Damon Lee Pointon
Emmelien Vancaester
Graeme Oatley
James Torrance
Joanna Collins
Jonathan Wood
Katie Woodcock
Kenneth Haug
Kerstin Howe
Ksenia Krasheninnikova
Maja Todorovic
Manuela Kieninger
Mara Lawniczak
Marcela Uliano da Silva
Mark Blaxter
Matt Berriman
Michelle Strickland
Nancy Holroyd
Nick Salmon
Radka Platte
Raquel Amaral
Robbie Heathcote
Sarah Pelan
Sophie Potter
Victoria Wright
William Chow
Ying Sims

Scientific Operations
Carol Smee
Catherine McCarthy
Elizabeth Cook
Emma Betteridge
Iraad Bronner
Michelle Smith
Mike Quail
Naomi Park
Alex Dove
Barbora Pardubska
Carlos Jimenez Verdejo
Craig Corton
Emily Gallagher
Emma Taluy
Esther Mellado
Harriet Johnson
Hermione Blomfield-Smith
Irene Fabiola
James Uphill
John Tushabe
Karen Oliver
Michelle Smith
Robin Moll
Tracey Chillongworth

Collaborators
Jonas Korlach *et al.*
Pacific Biosciences
Dan Turner *et al.*
Oxford Nanopore

Team301
Chris Laumer
Claudia Weber
Emmelein Vancaester
Erna King
Lewis Stevens
Max Brown
Pablo Gonzalez
Rich Challis

wellcome sanger institute

Obrigada! Thank you!

Royal Botanic Gardens

Alex Twyford

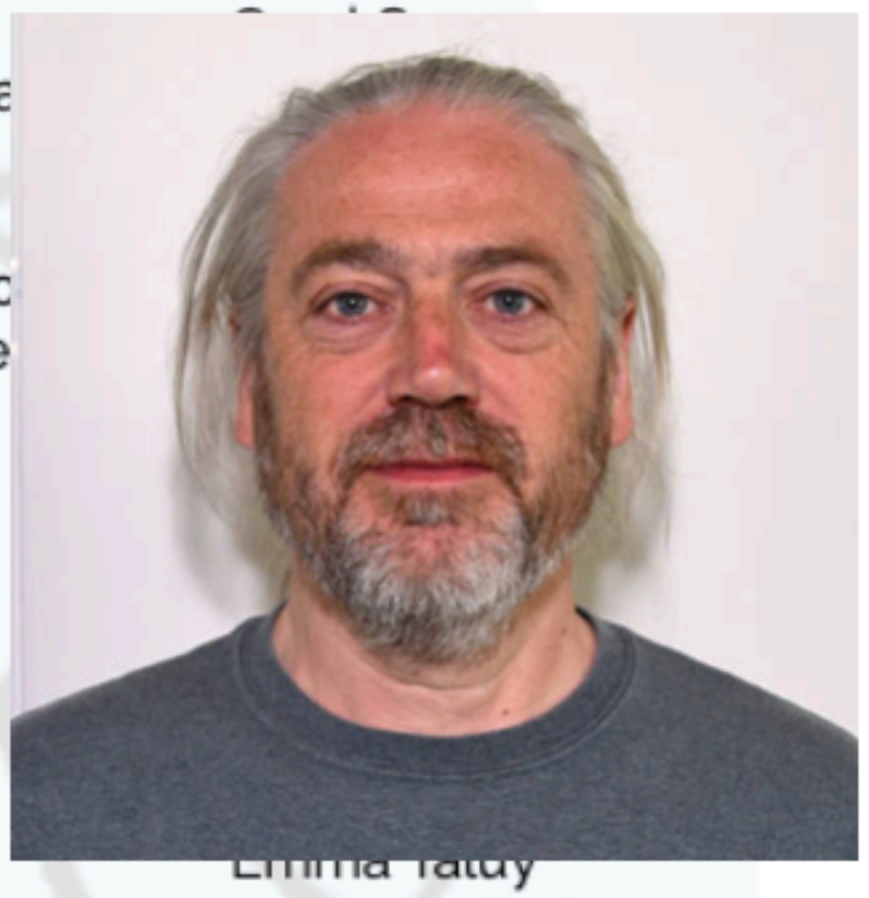
Olga Sivell

Jayathilaka
Martin
Thybert
Rajan
Howe
ochrane
Harrison
e Haggerty
Allen
Garcia Giron
eu Muffato

Matt Berriman
Michelle Strickland
Nancy Holroyd
Nick Salmon
Radka Platte
Raquel Amaral
Robbie Heathcote
Sarah Pelan
Sophie Potter
Victoria Wright
William Chow
Ying Sims

wellcome sanger institute

Scientific Operations

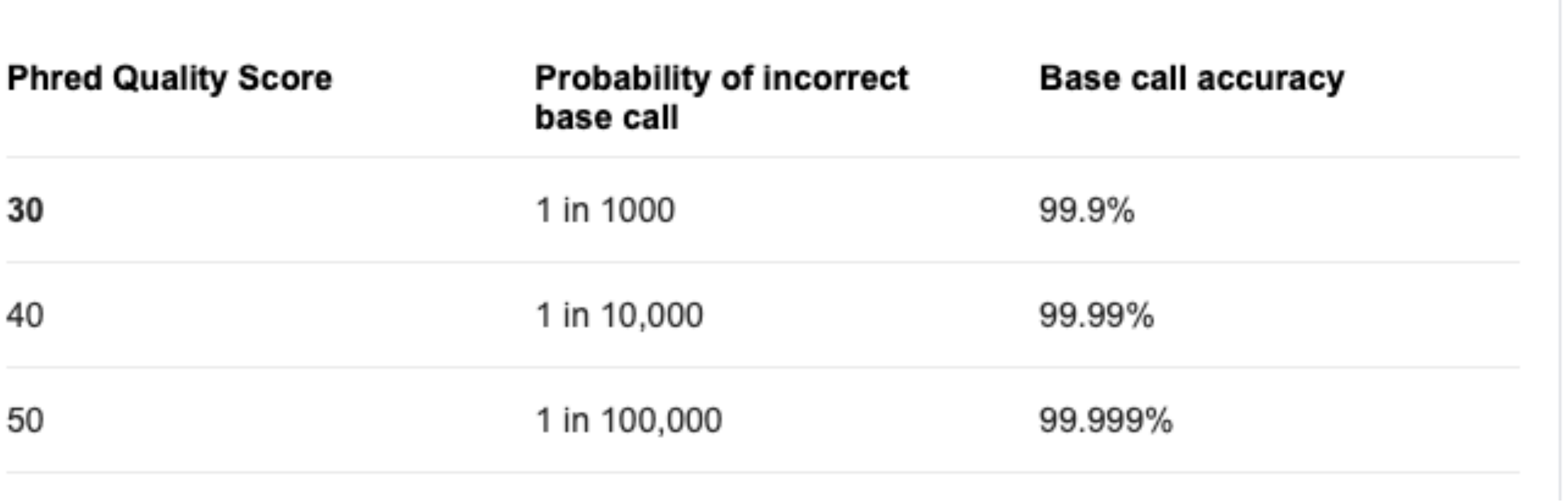


Emma Tatay
Esther Mellado
Harriet Johnson
Hermione Blomfield-Smith
Irene Fabiola
James Uphill
John Tushabe
Karen Oliver
Michelle Smith
Robin Moll
Tracey Chillongworth

Team301

Collaborators
Jonas Korlach *et al.*
Pacific Biosciences
Dan Turner *et al.*
Oxford Nanopore

wellcome sanger institute



Phred Quality Score	Probability of incorrect base call	Base call accuracy
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%