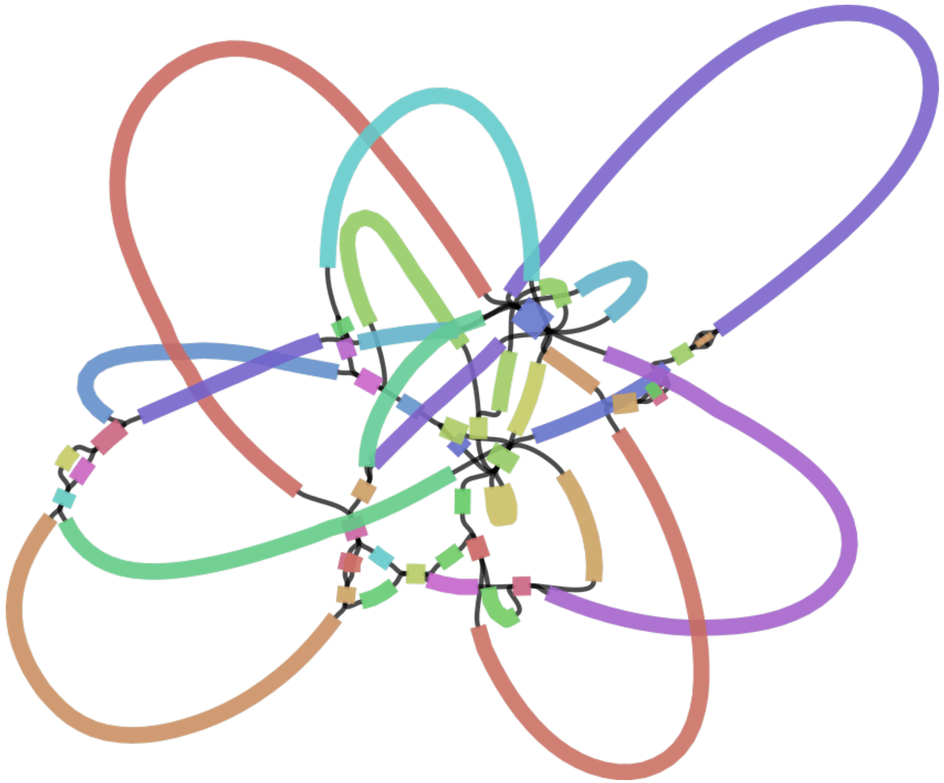
GENOME ASSEMBLY: WHERE DO I START

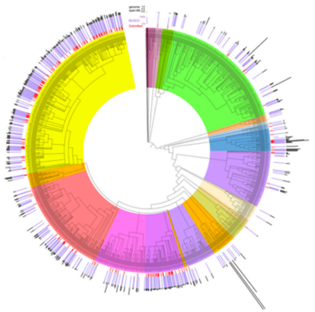


Marcela Uliano-Silva, PhD

Darwin Tree of Life

wellcome sanger institute

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**

- Horizon2020 Marie Curie PostDoc Fellow (2017-2019), IZW, BenGenDiv, Germany
- PhD in Biophysics (2017) - IBCCF UFRJ, Brazil
- MSc in Biophysics (2013) - IBCCF UFRJ, Brazil
- BSc in Biology (2010) - UFSC, Brazil
- 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

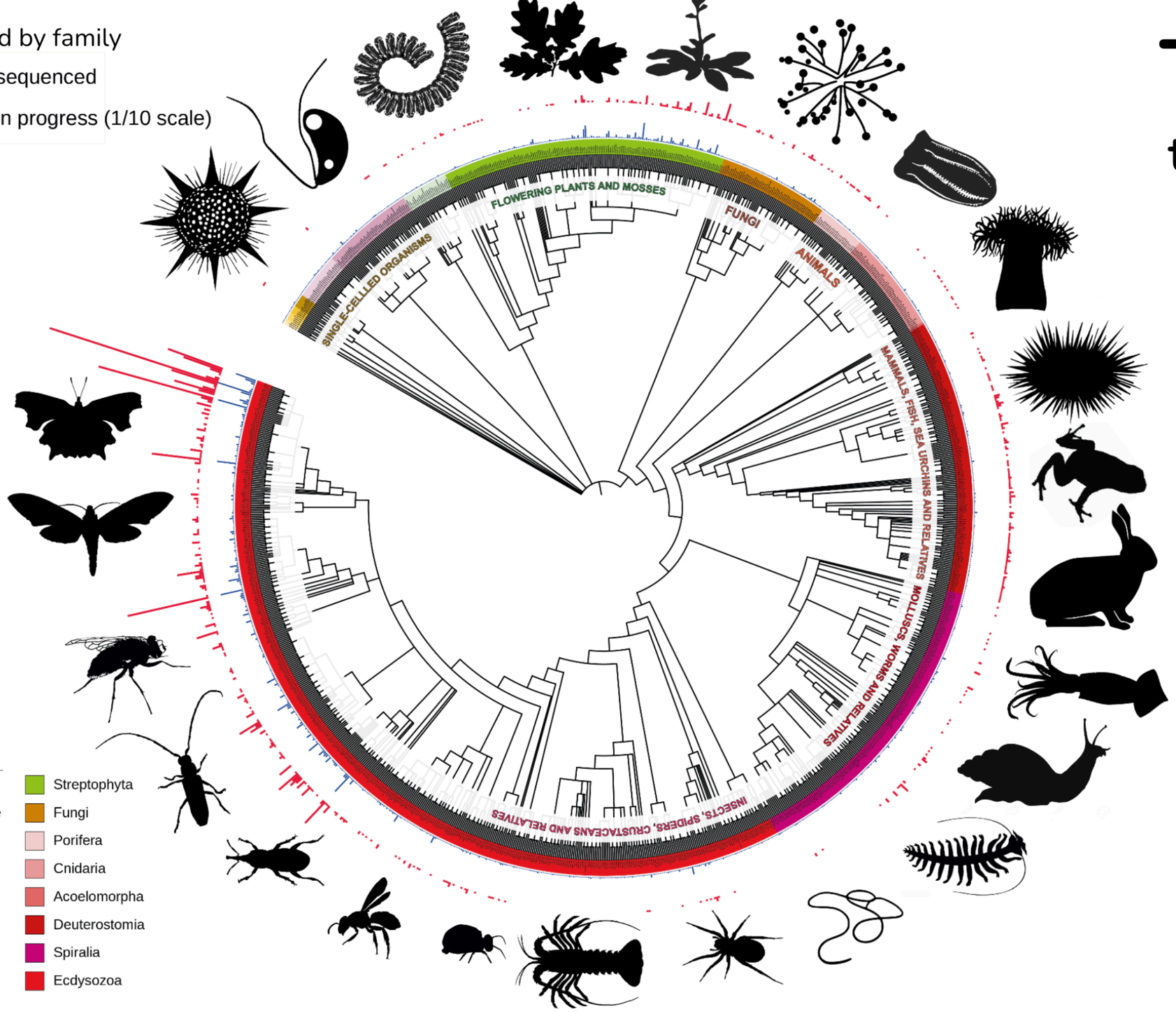
★ **European Reference Genome Atlas**
- Sequencing the genomes of all species in the European continent - Pilot 25 species

★ **Earth BioGenome Project**
- Working to deliver Phase 1 (family) goals, and to "*sequence all life for the future of life*"

wellcome sanger institute

summarised by family

- Species sequenced
- Species in progress (1/10 scale)



Major Taxa

- Apusomonads
- Cryptophyceae
- Haptophyta
- Amoebozoa
- Discoba
- Rhodophyta
- SAR
- Chlorophyta
- Streptophyta
- Fungi
- Porifera
- Cnidaria
- Acoelomorpha
- Deuterostomia
- Spiralia
- Ecdysozoa

Tree of Life

to 01 January 2024

4688 species
from **1194** families
and **48** phyla

1288 species
complete

of which **1103** DToL

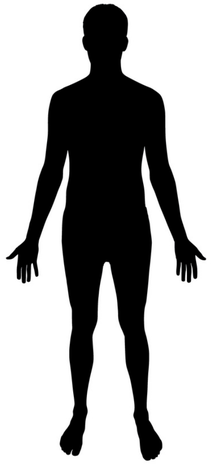
working on
an additional
3400 species

Tree 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

ASSEMBLY GRAPHS

Overlap Layout Consensus

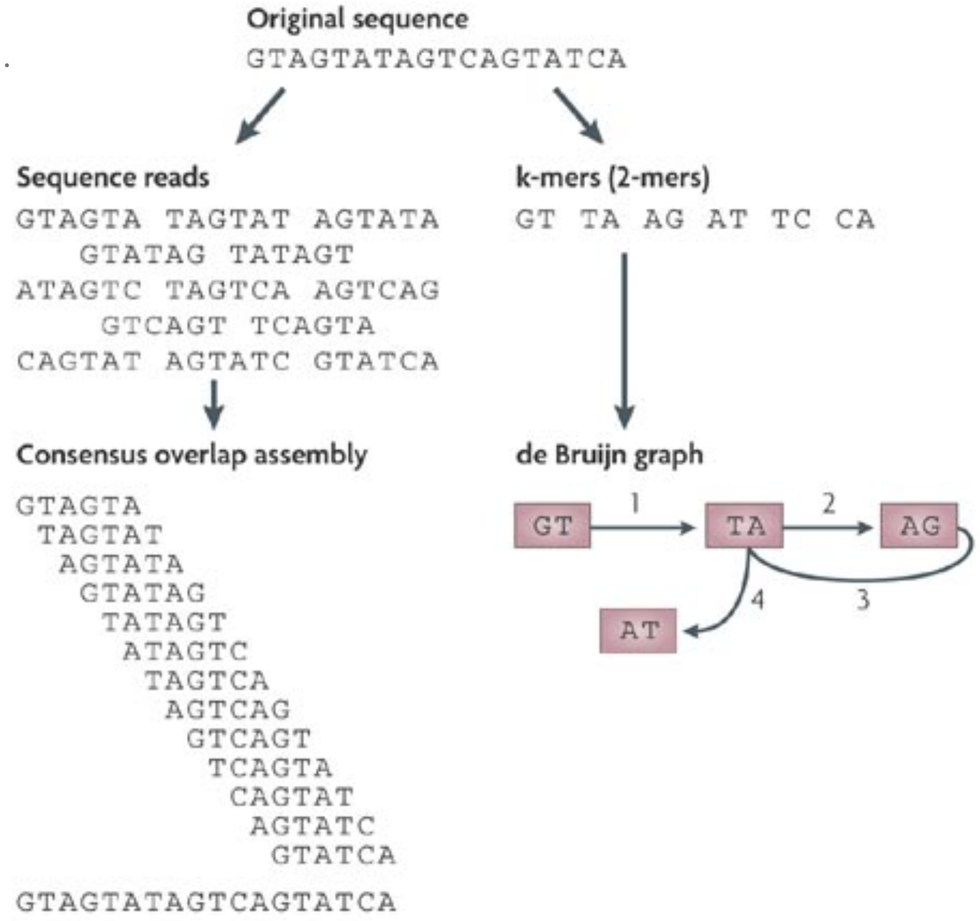
De Brujin

"Both representations share the idea that a **genome assembly corresponds to a path in the graph**: for this reason, the step following the construction of such a graph is the extraction of relevant paths. Under ideal conditions, such as the absence of errors and repeats, we can reconstruct only one relevant path in such graph (that is, there is only one possible assembly). " (Rizzi *et al,* 2019)

ASSEMBLY GRAPHS

Overlap Layout Consensus

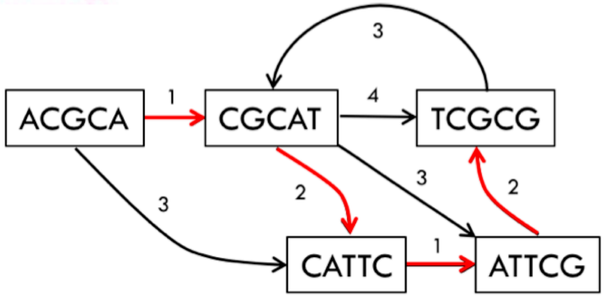
- *Reads of Length L and overlap cutoff*
- *LongReads*



De Brujin

- *Kmers*
- *Short reads*

Nature Reviews | Microbiology

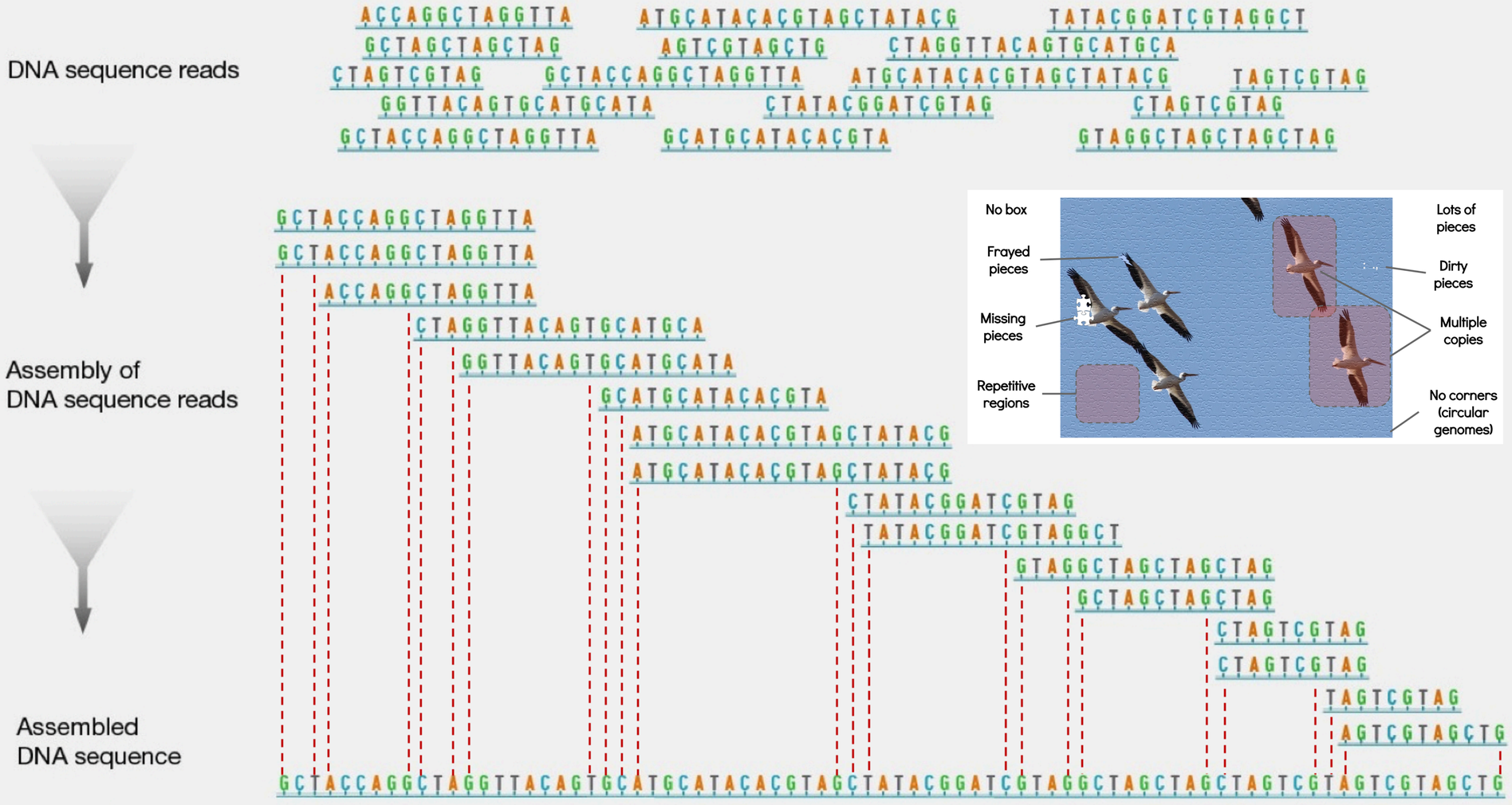


ACGCATTCGCG

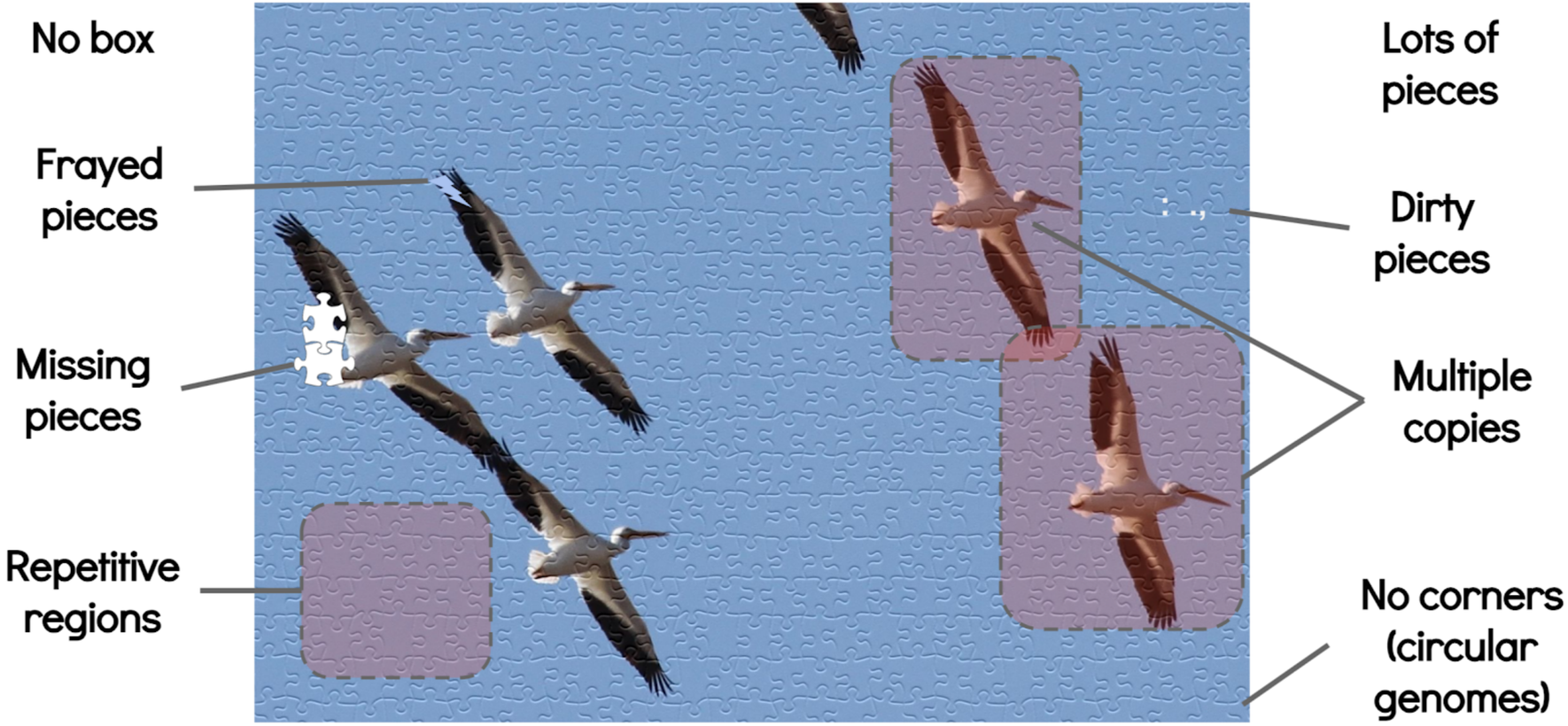
"Both representations share the idea that a **genome assembly corresponds to a path in the graph**: for this reason, the step following the construction of such a graph is the extraction of relevant paths. Under ideal conditions, such as the absence of errors and repeats, we can reconstruct only one relevant path in such graph (that is, there is only one possible assembly). " (Rizzi *et al*, 2019)

I WANT TO TALK TO YOU ABOUT LONG READ SEQUENCING

The Naïve Genome Assembly Approach



What makes a jigsaw puzzle hard?



- What helps? Larger pieces (read length); fewer dirty or frayed pieces (errors in reads). fewer repeats and copies…

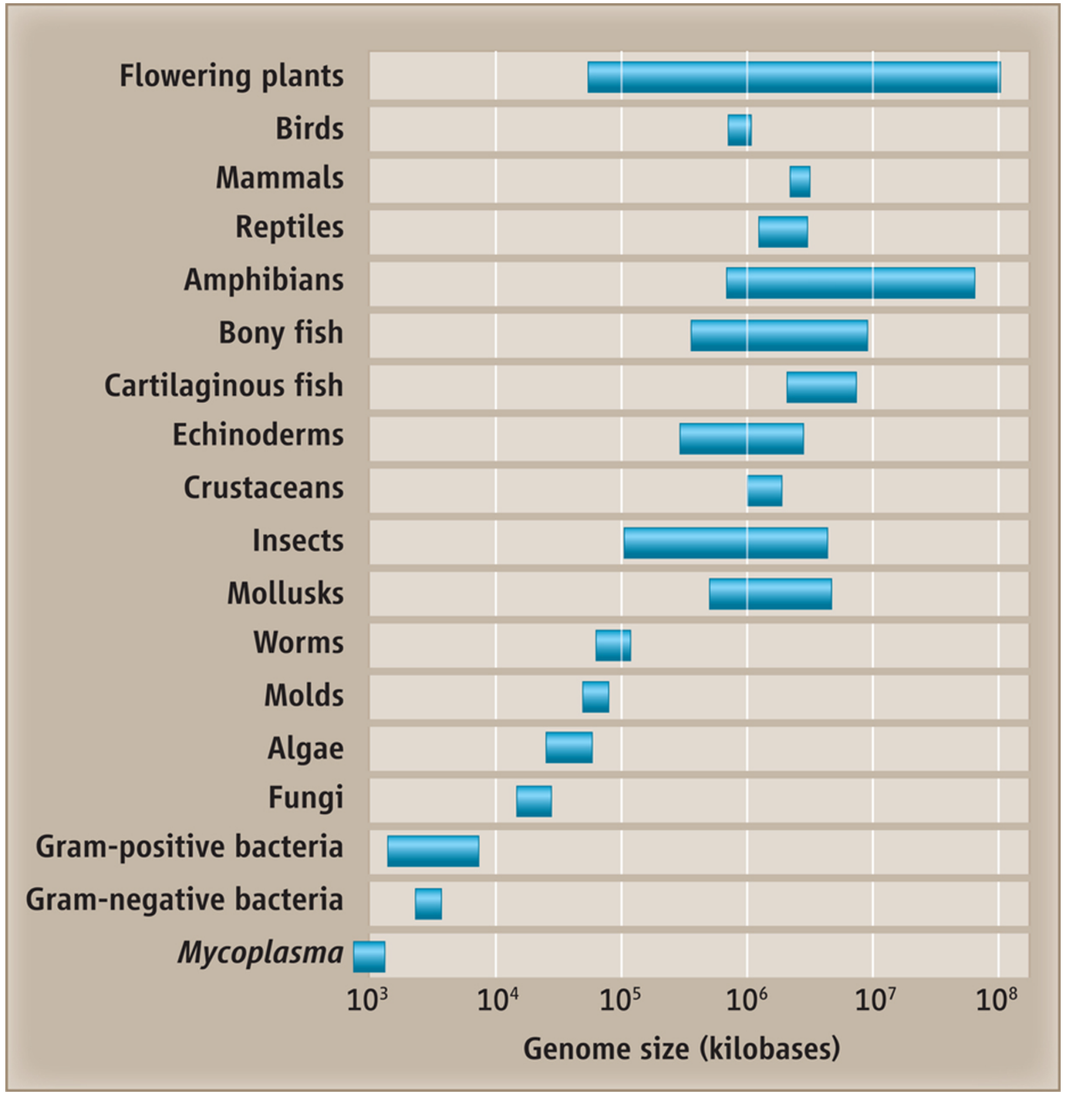
WHAT IS THE PROBLEM WITH SHORT READS?

➤ What are eukaryotic genomes made of?

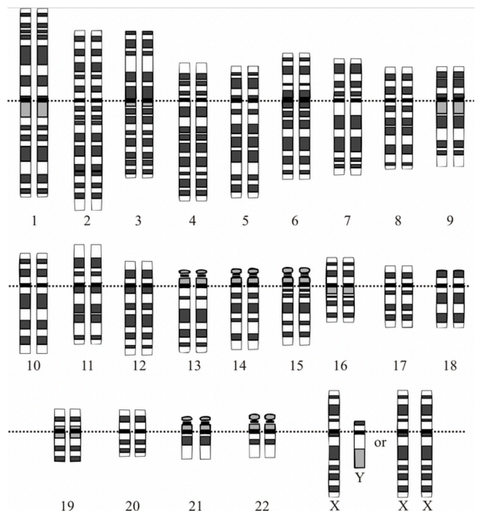
Repeats:

- *Centromeres (Tandem arrays of repeated sequence studded with transposable elements (plants, humans)*
- *Telomeres (tandem arrays of simple repeats)*
- *Mobile elements*
- *Segmental duplications*
- *rRNAs*

But why? (Fedoroff, 2012)

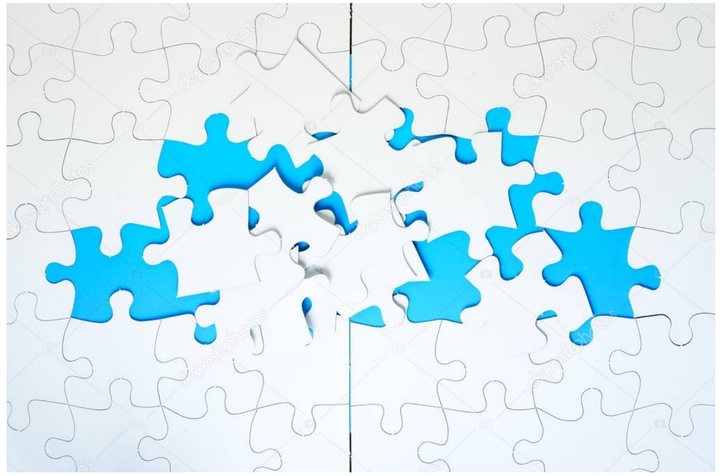


WHEN WE ASSEMBLE A GENOME . . .



What we would like to have

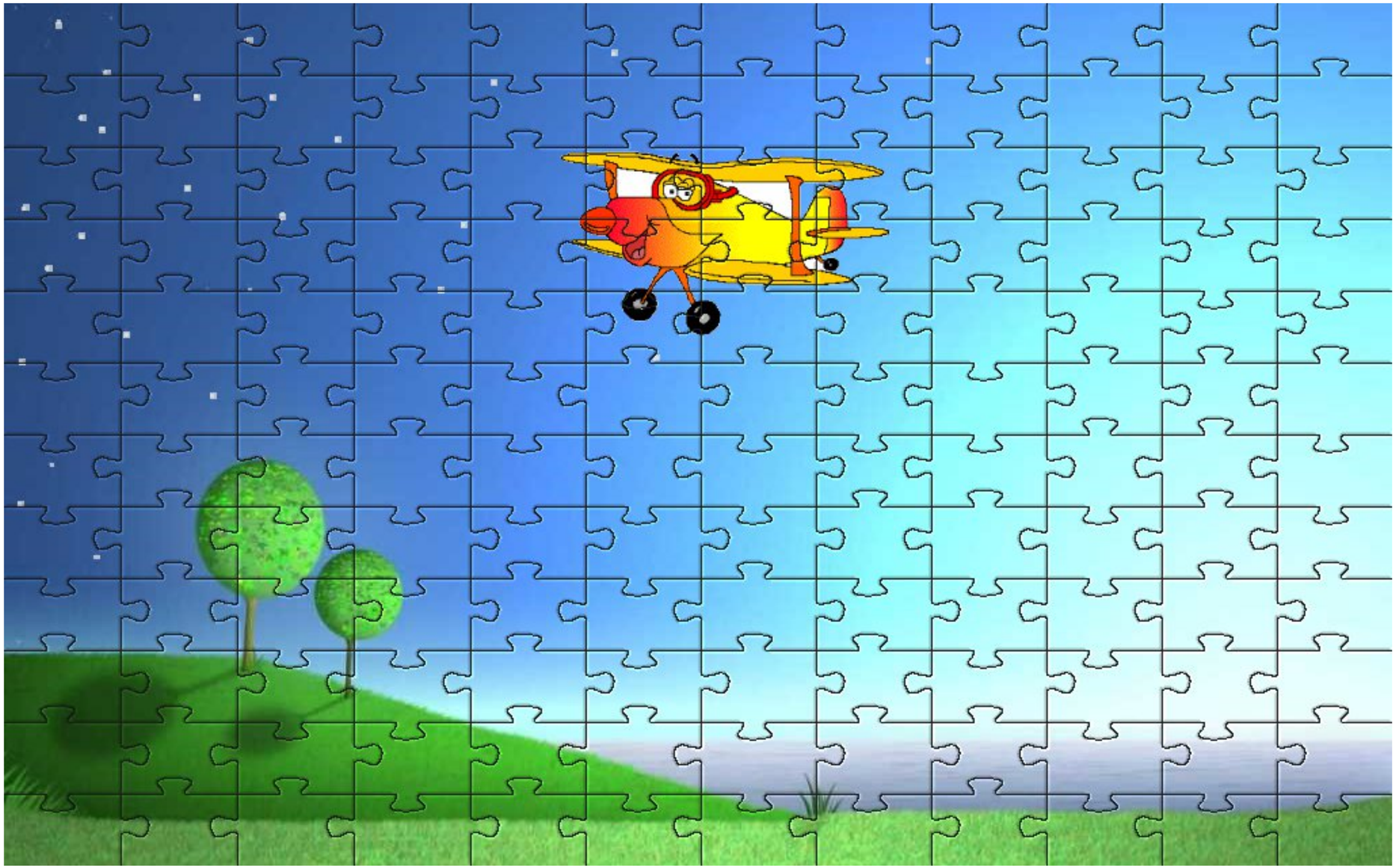
- One DNA sequence for each chromosome



What we really have

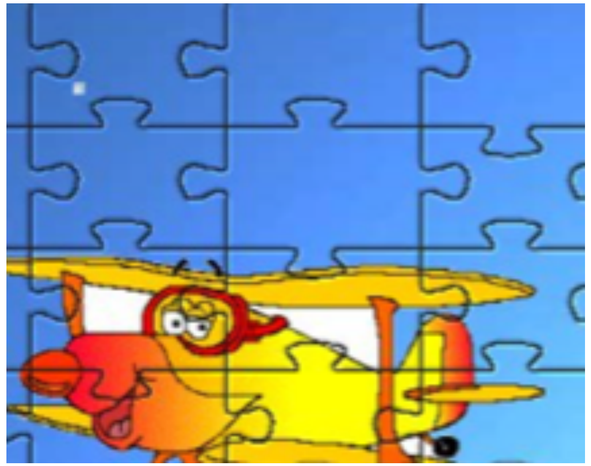
- Contigs, scaffolds, gaps, N50s

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

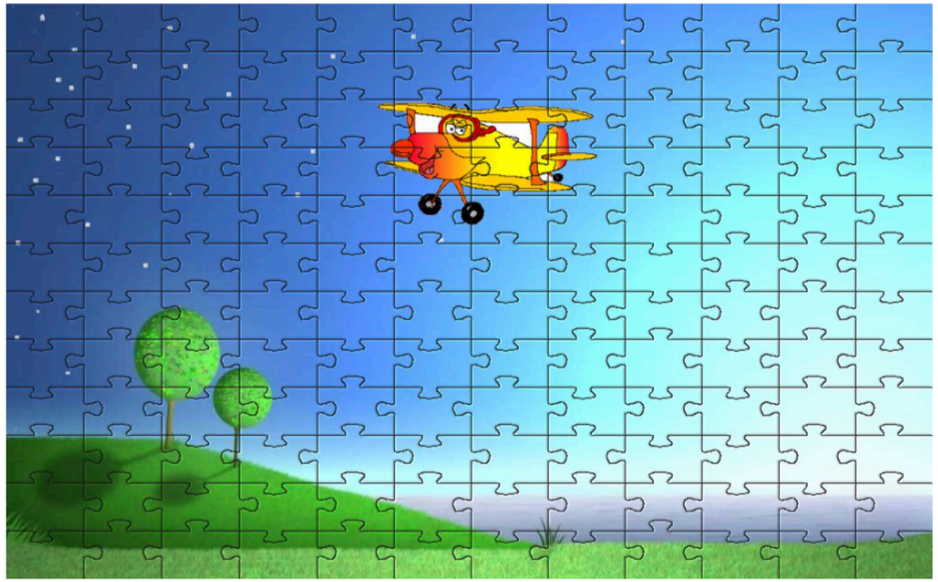


THIS IS A SHORT-READS GENOME ASSEMBLY OF ME

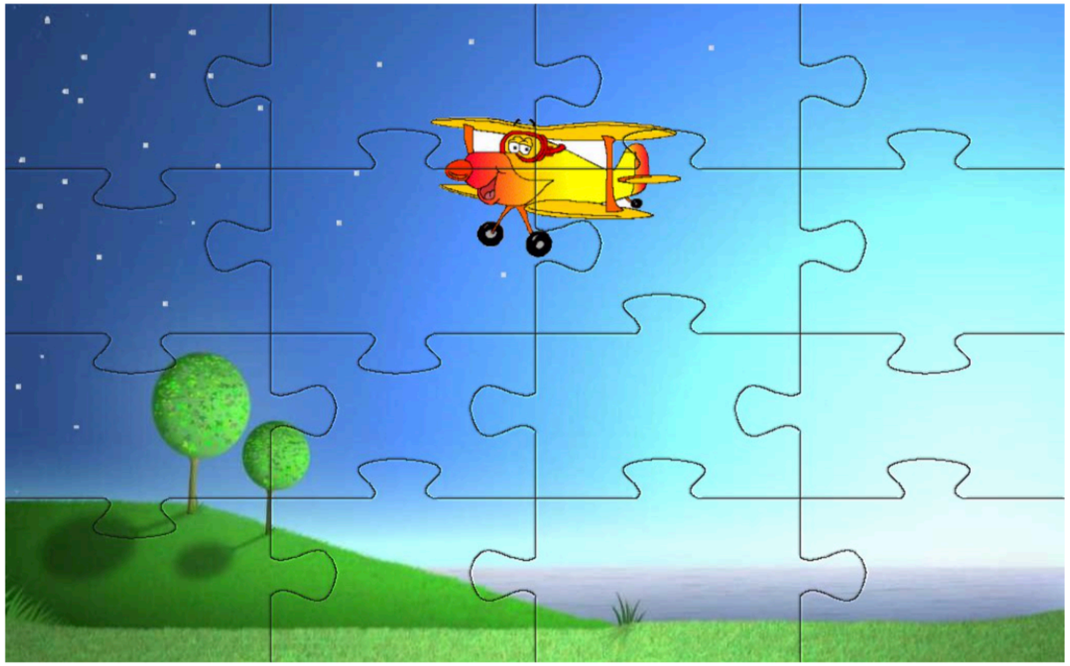
NNNNNN



NNNNNN



Assembling with short reads



Assembling with long reads

YOUR GENOME ASSEMBLY PROJECT STARTS IN THE LAB

High Molecular Weight DNA extraction is key

No one-size-fits-all protocol!

protocols.io FEATURES PLANS BLOG CASE STUDY

Scan me!

OCT 02, 2023

SHARE

WORKS FOR ME 1

Sanger Tree of Life Wet Laboratory Protocol Collection

DOI
dx.doi.org/10.17504/protocols.io.8epv5xxy6g1b/v1

Amy Denton[1], Halyna Yatsenko[1], Jessie Jay[1], kh[1], Caroline Howard[1]

[1]Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA

Tree of Life at the Wellcome Sanger Institute

Tree of Life Genome Note Editor

COPY / FORK

MORE ↓

OXFORD (GIGA)n SCIENCE

GigaScience, 2022, **11**, 1–13
DOI: 10.1093/gigascience/giac068
RESEARCH

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

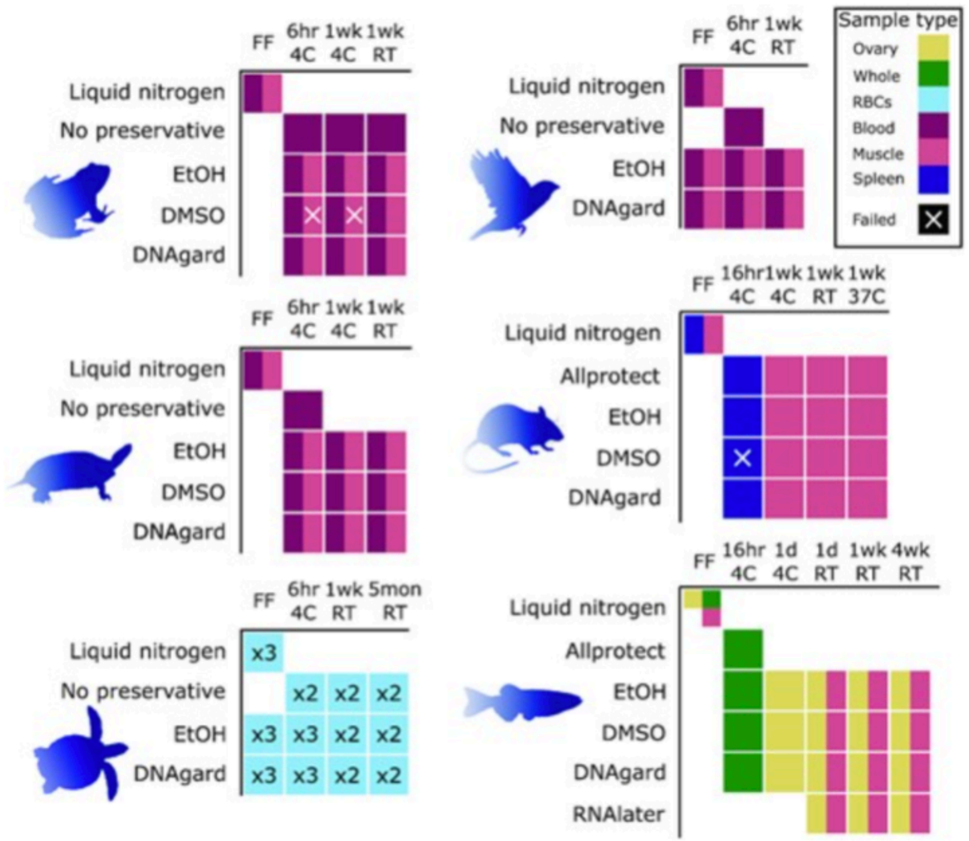
Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

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,*]

No one-size-fits-all protocol!

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

e 1:



I EXTRACTED HMW DNA: WHAT DO I DO NOW?

Darwin Tree of Life

Our recipe working across the Tree of Life:

Chromosome level genomes

- 25x Pacbio HiFi
- 100x Hi-C (Arima/Qiagen)

T2T (Telomere to Telomere) genomes

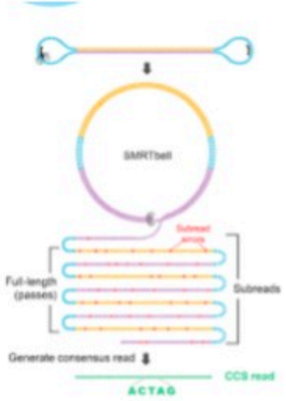
- The above plus 25x ONT Ultra Long (>100Kb reads)

DToL Current Pipeline

MitoHiFi

For mitochondria genome assembly

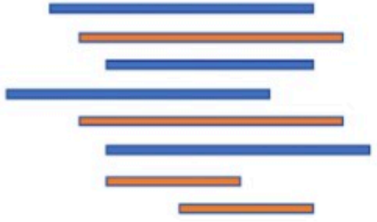
- Sequencing technologies: PacBio HiFi + HiC (Arima or Qiagen)



1- Kmer Jellyfish/ GenomeScope, asmstats, smudgeplot (se possível poliploide)

Assembly

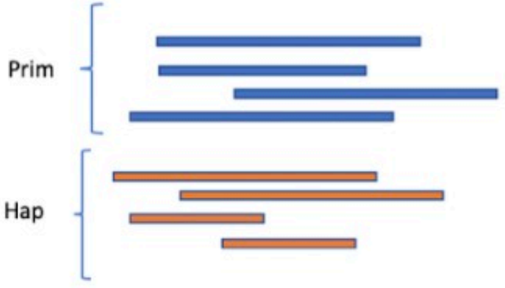
Hicanu or Hifiasm



2 - asmstats, BUSCO, merqury

Haplotype separation

Purge dups



3 - asmstats, BUSCO, merqury

Scaffolding

Yahs scaffolding (Arima or Qiagen HiC)

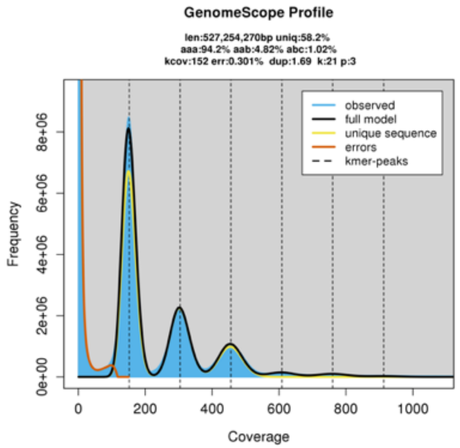
4 - asmstats, BUSCO, merqury, HiC heatmap

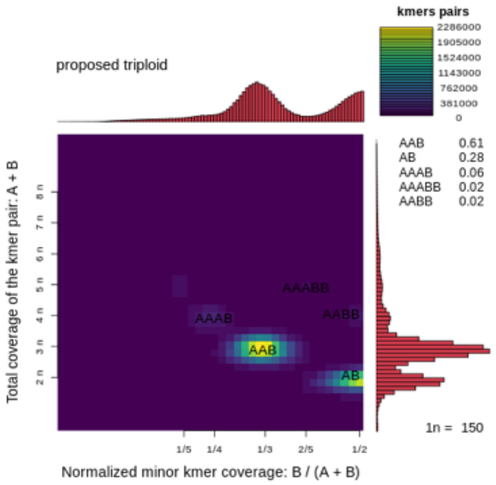
Curated assembly

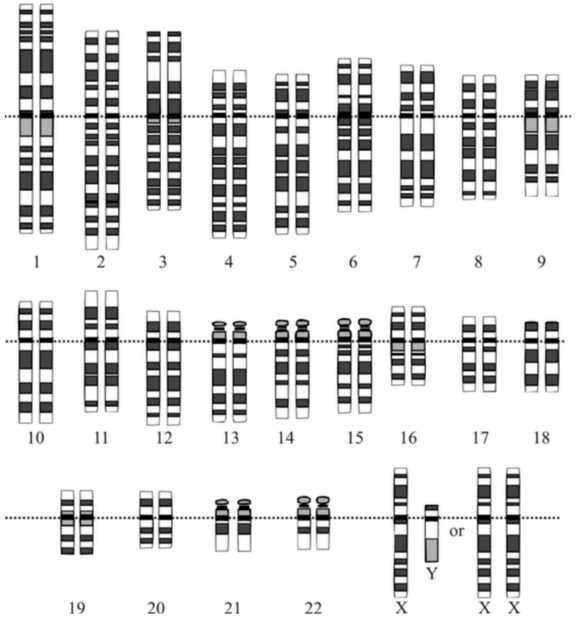
5 - asmstats, BUSCO, merqury, HiC heatmap

Key considerations to start your genome assembly project

- Genome size (flow cytometry, Kmer analysis, GoaT) **https://goat.genomehubs.org/**
- Heterozygosity (kmer analyses: jellyfish, genomescope)
- Repetitive content (kmer analyses: jellyfish, genomescope)
- Ploidy (kmer analyses: smudgeplots)



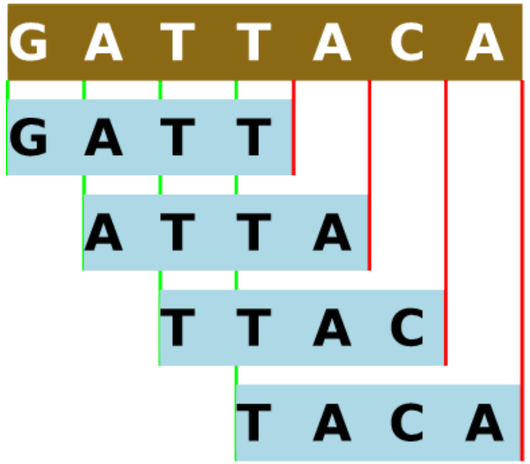




I HAVE HIGH-QUALITY DATA (ILLUMINA, PACBIO HIFI, DUPLEX NANOPORE)

I WILL do a kmer analysis first thing

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)

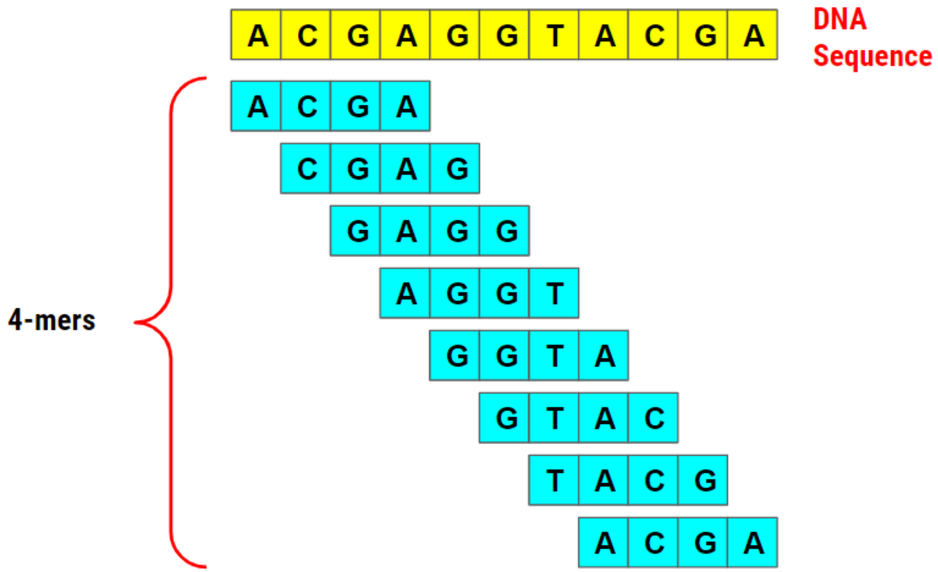
https://bioinfologics.github.io/post/2018/09/17/k-mer-counting-part-i-introduction/

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
```

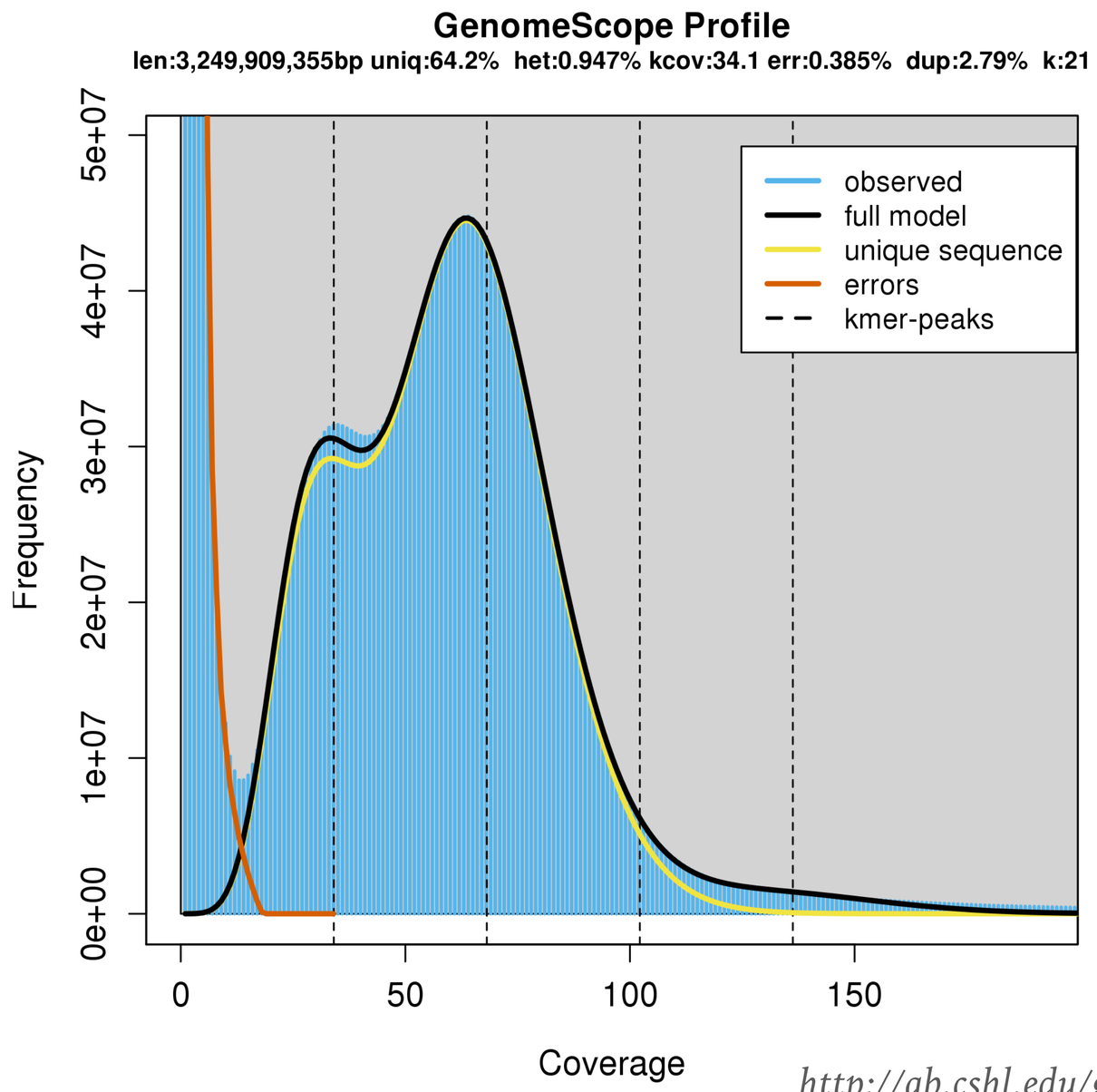
Now the histogram file looks like this:

```
# 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
```

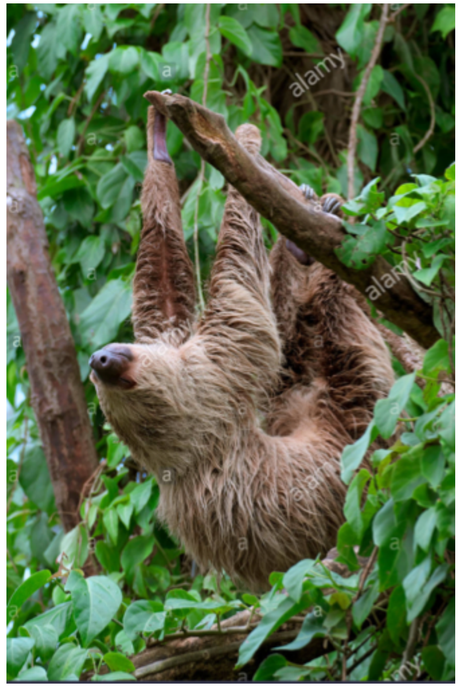
Here, only **4159** *8-mers* are *unique*, out of **4726** *distinct 8-mers*, that are present in the genome's **5377** *total 8-mers*.

Bernardo Cavijo's post

A TYPICAL KMER PLOT FOR A DIPLOID SPECIES



Choloepus didactylus (VGP)

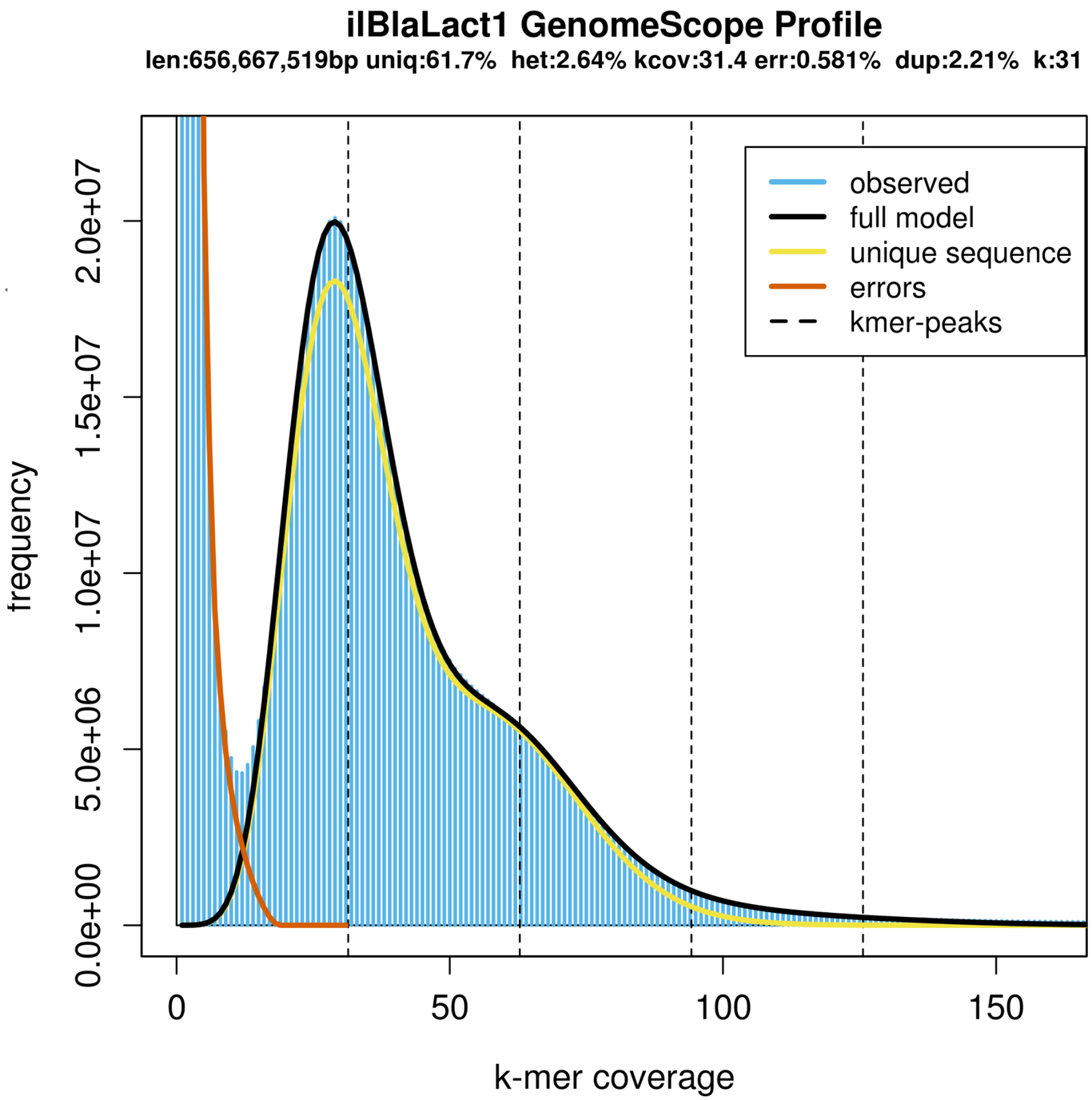


http://qb.cshl.edu/genomescope/analysis.php?code=bVuZNlhwn2tVCHhRN71I

A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH HIGH HETEROZYGOSITY

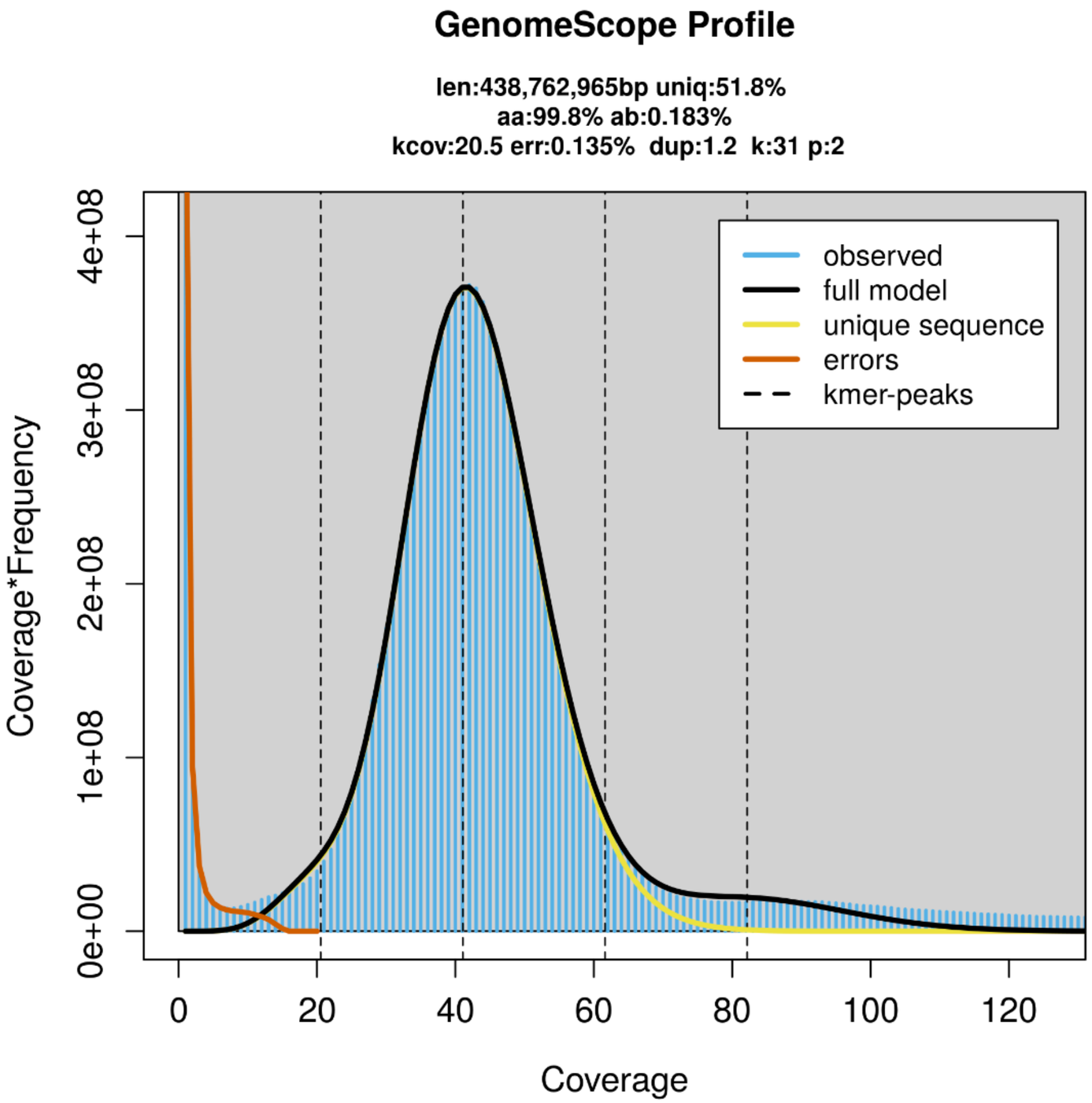
Blastobasis lacticolella (DToL)

Wakely's dowd



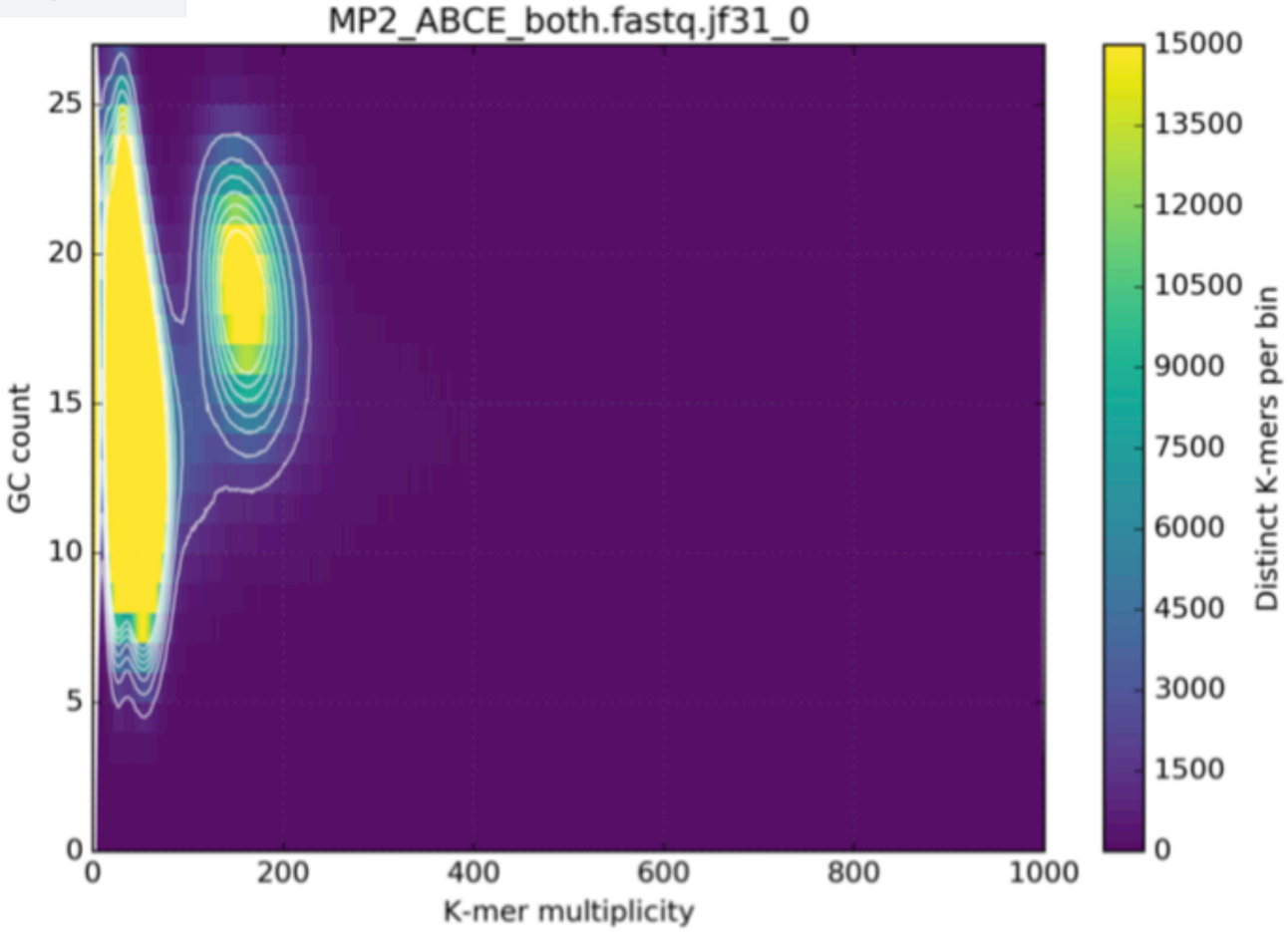
A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH LOW HETEROZYGOSITY

Urtica urens



SPOTTING BACTERIAL CONTAMINATION: KMER AND ITS GC CONTENT

github.com/TGAC/KAT



You can use KAT to plot this!

README.md

KMER ANALYSIS TOOLKIT

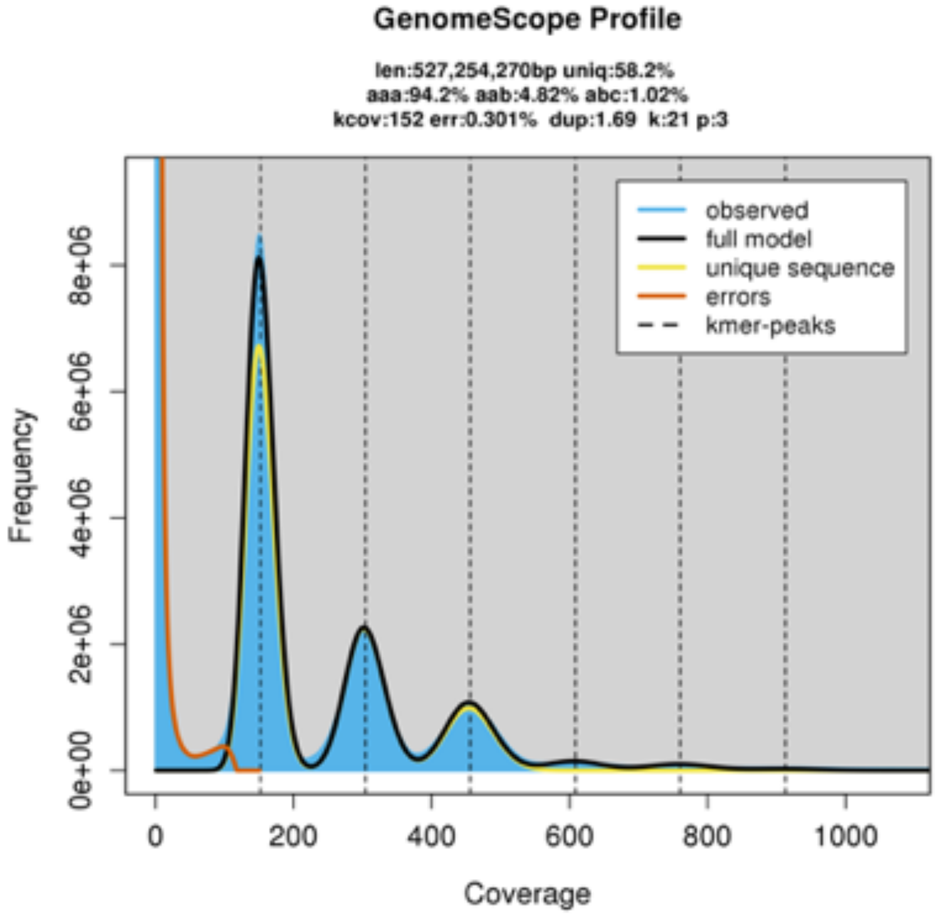
KAT - The K-mer Analysis Toolkit

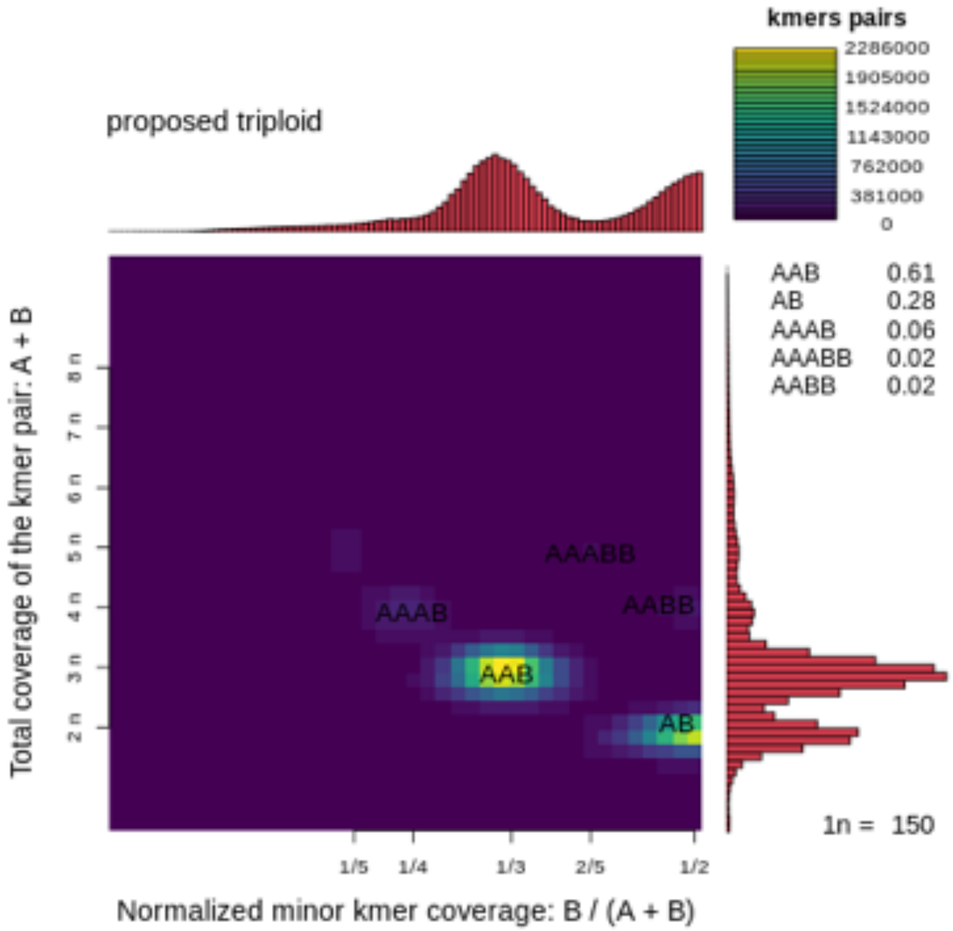
KAT is a suite of tools that analyse jellyfish hashes or sequence files (fasta or fastq) using kmer counts. The following tools are currently available in KAT:

- **hist**: Create an histogram of k-mer occurrences from a sequence file. Adds metadata in output for easy plotting.
- **gcp**: K-mer GC Processor. Creates a matrix of the number of K-mers found given a GC count and a K-mer count.
- **comp**: K-mer comparison tool. Creates a matrix of shared K-mers between two (or three) sequence files or hashes.
- **sect**: SEquence Coverage estimator Tool. Estimates the coverage of each sequence in a file using K-mers from another sequence file.

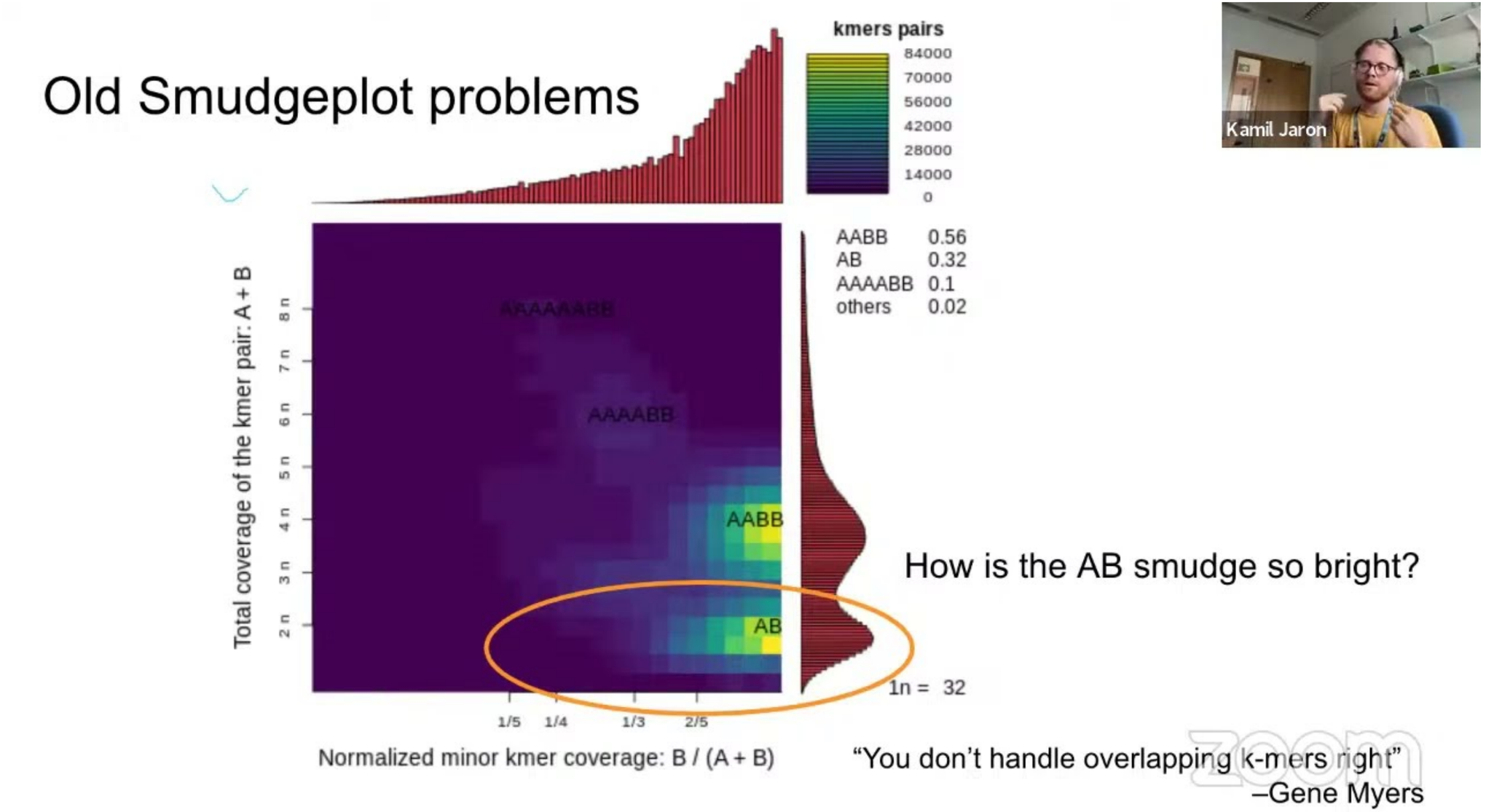
Tubastraea tagusensis

KMER PROFILE FOR A TRIPLOID SPECIES





MORE ON SMUDGEPLOTS



github.com/thegenemyers/MERQURY.FK

README.md

MerquryFK & KatFK: Fast & Simple

Authors: Gene Myers & Arang Rhie
First: Feb 24, 2021
Current: Aug 11, 2021

Introduction

The original Merqury is a collection of R, Java, and shell scripts for producing k-mer analysis plots of genomic sequence data and assemblies with **meryl** as its core k-mer counter infra-structure. **MerquryFK** replaces meryl with the **FastK** k-mer counter suite to considerably speed up analyses. Moreover, all the R, Java, and shell scripts have been refactored into a typical collection of UNIX command line tools that the user will hopefully experience as easier to comprehend and invoke. In addition, we have realized some analyses, **KatComp** and **KatGC**, that one finds only in the somewhat similar KAT k-mer suite developed at the Earlham Institute. Lastly, we include in this collection, **PloidyPlot** which is an improved version of the ploidy plotting tool SmudgePlot.

There are some general conventions for our tools programmed for your convenience. First, suffix extensions need not be given for arguments of a known type. For example, if an argument is a fasta or fastq with root name "foo" without extensions, then our commands will look for `foo.fasta, foo.fa, foo.fastq, and foo.fq` if you specify `foo` as the argument. Second, option arguments (those that begin with a '-') can be in any order and in any position relative to the non-optional primary arguments (which must be given in the order specified). We find this pretty convenient when for example you have typed out an entire CNplot command (2. below) but forgot that you wanted .pdf's. All you do is append -pdf to what you've already typed and then hit return. So for example, `CNplot -w4 -h3 Assembly -ls Reads -pdf` is acceptable input.

For the tools that take a FastK k-mer table as an input, we use the syntax <name>[.ktab], to describe it on the command line indicating that the .ktab extension is optional as per the convention above. Regardless of whether the extension is given, it is expected that the associated histogram file <name>.hist is also present (this file is always produced by a run of FastK that produces a k-mer table). Also note carefully that these tables must be produced with the option -t or -t1 set so that all k-mers that occur 1 or more times in the underlying data set are in the table.

KNOWING THE CHALLENGE, YOU GO AND BUILD CONTIGS WITH ASSEMBLERS

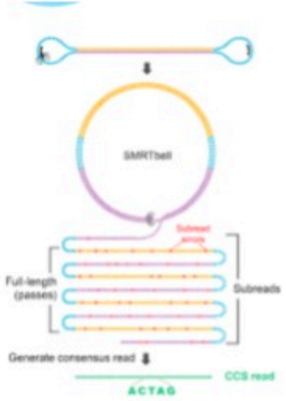
CONTIG

DToL Current Pipeline

MitoHiFi

For mitochondria genome assembly

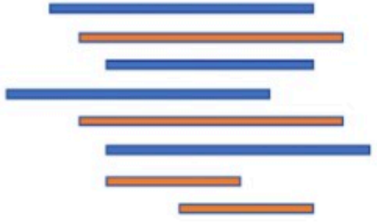
- Sequencing technologies: PacBio HiFi + HiC (Arima or Qiagen)



1- Kmer Jellyfish/ GenomeScope, asmstats, smudgeplot (se possível poliploide)

Assembly

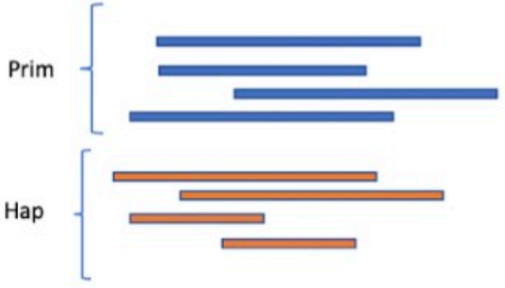
Hicanu or Hifiasm



2 - asmstats, BUSCO, merqury

Haplotype separation

Purge dups



3 - asmstats, BUSCO, merqury

Scaffolding

Yahs scaffolding (Arima or Qiagen HiC)

4 - asmstats, BUSCO, merqury, HiC heatmap

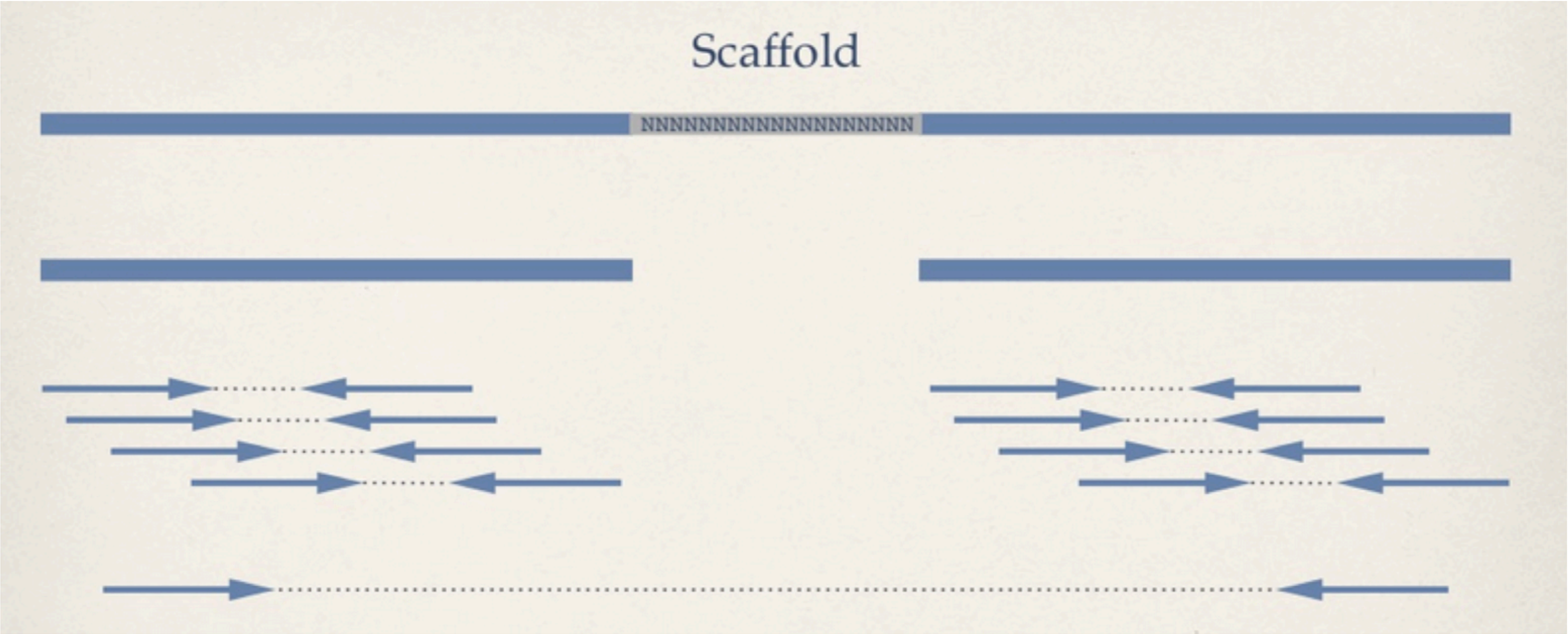
Curated assembly

5 - asmstats, BUSCO, merqury, HiC heatmap

DOES MY ASSEMBLED SIZE CORRESPONDS WITH MY ESTIMATED GENOME SIZE?

Genomics is a game of going back and forth

Scaffolding methods



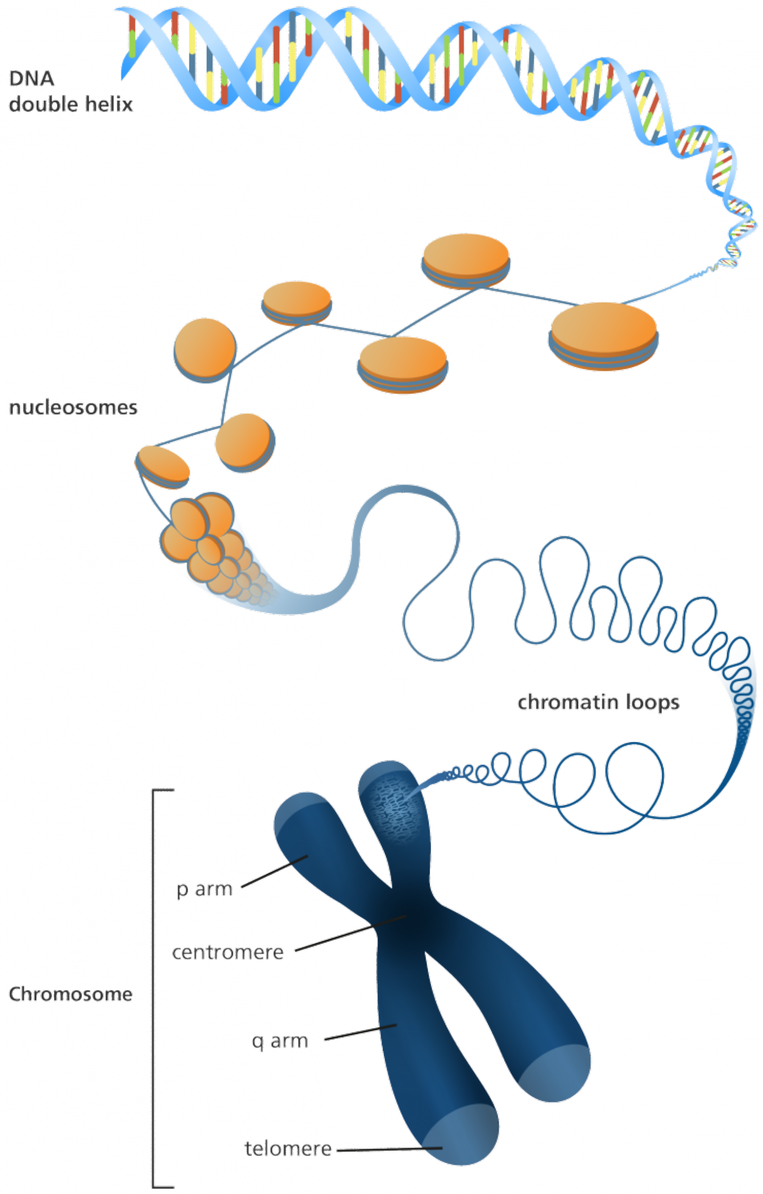
Scaffold: joining and orienting contigs

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

HOW DO I BUILD UP SCAFFOLDS AND CHROMOSOMES?

Hi-C and Ultralong Nanopore

The human genome consists of over 3 billion nucleotides and is contained within 23 pairs of chromosomes. ***If the chromosomes were aligned end to end and the DNA stretched, the genome would measure roughly 2 meters long. Yet the genome functions within a sphere smaller than a tenth of the thickness of a human hair (10 micron)***. … the genome does not exist as a simple one-dimensional polymer; instead the genome folds into a complex compact three-dimensional structure. (Lajoie et al 2015)

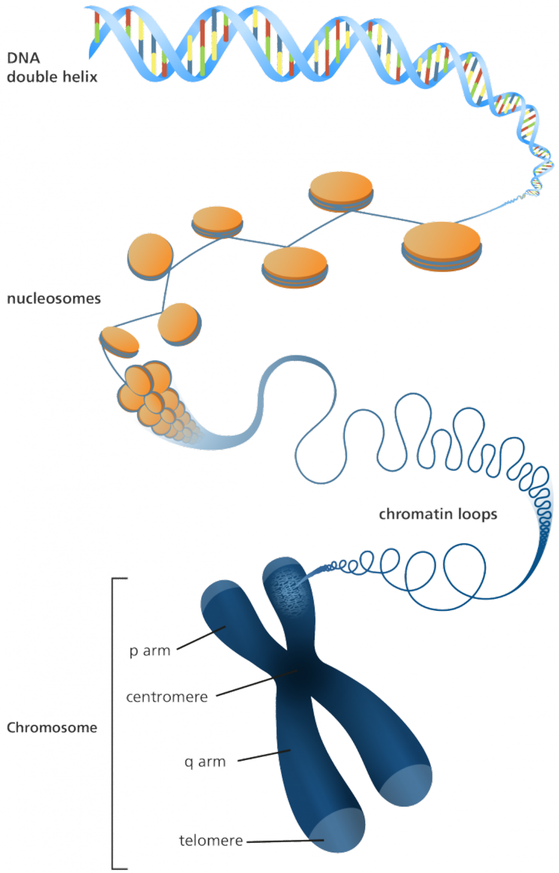


Chromosome conformation

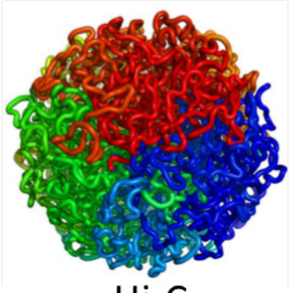
➤*The organisation of the chromatin in the nucleus is extremely relevant to biological function at the gene level as well as the global nuclear level.*

➤*The study of the packaging and organisation of chromatin in the nucleus will shed light on:*

- *the spatial aspects of gene regulation*
- *chromosome morphogenesis*
- *genome stability*
- *genome transmission*
- *biophysics of chromatin*
- *pathologies related to genome instability or nuclear morphology*



Published in final edited form as:
Science. 2009 October 9; 326(5950): 289–293. doi:10.1126/science.1181369.



Hi-C

Comprehensive mapping of long range interactions reveals folding principles of the human genome

Erez Lieberman-Aiden[1,2,3,4,*], **Nynke L. van Berkum**[5,*], **Louise Williams**[1], **Maxim Imakaev**[2], **Tobias Ragoczy**[6,7], **Agnes Telling**[6,7], **Ido Amit**[1], **Bryan R. Lajoie**[5], **Peter J. Sabo**[8], **Michael O. Dorschner**[8], **Richard Sandstrom**[8], **Bradley Bernstein**[1,9], **M. A. Bender**[10], **Mark Groudine**[6,7], **Andreas Gnirke**[1], **John Stamatoyannopoulos**[8], **Leonid A. Mirny**[2,11], **Eric S. Lander**[1,12,13,†], and **Job Dekker**[5,†]

[1] Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02139, USA.

[2] Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts 02139, USA.

[3] Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.

[4] Department of Applied Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.

[5] Program in Gene Function and Expression and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA.

[6] Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.

[7] Department of Radiation Oncology, University of Washington School of Medicine, University of Washington, Seattle, Washington 98195, USA.

[8] Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.

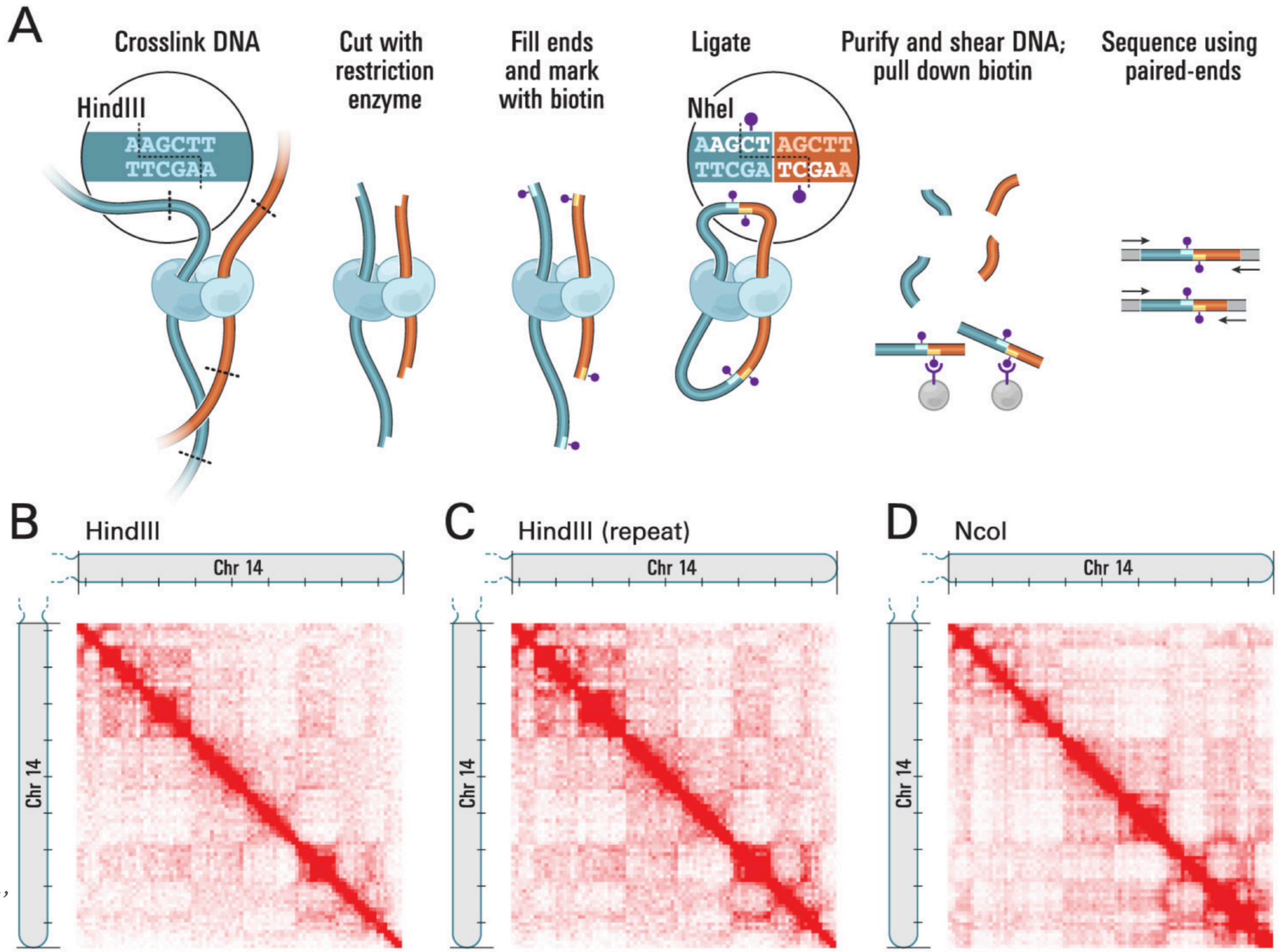
[9] Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.

[10] Department of Pediatrics, University of Washington, Seattle, Washington 98195, USA.

[11] Department of Physics, MIT, Cambridge, Massachusetts 02139, USA.

[12] Department of Biology, MIT, Cambridge, Massachusetts 02139, USA.

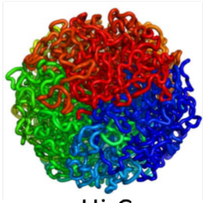
[13] Department of Systems Biology, Harvard Medical School, Boston MA 02115.



Lieberman-Aiden et al., 2010

HI-C

- Intrachromosomal contact probability is on average much higher than interchromosomal.
- Interaction probability rapidly decays with increasing genomic distance.



Hi-C

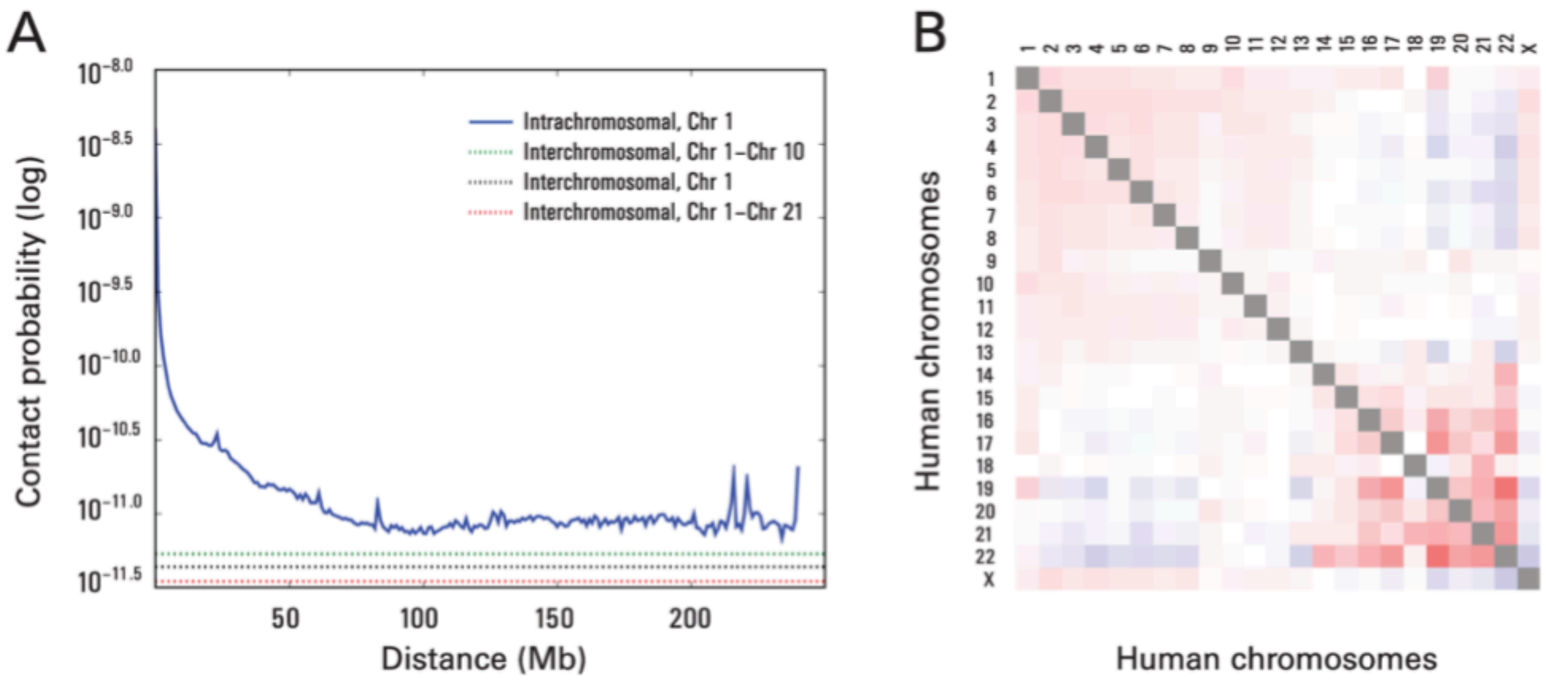


Fig. 2.
The presence and organization of chromosome territories. (**A**) Probability of contact decreases as a function of genomic distance on chromosome 1, eventually reaching a plateau at ~90M (blue). The level of interchromosomal contact (black dashes) differs for different pairs of chromosomes; loci on chromosome 1 are most likely to interact with loci on chromosome 10 (green dashes) and least likely to interact with loci on chromosome 21 (red dashes). Interchromosomal interactions are depleted relative to intrachromosomal interactions. (**B**) Observed/expected number of interchromosomal contacts between all pairs of chromosomes. Red indicates enrichment, and blue indicates depletion (up to twofold). Small, gene-rich chromosomes tend to interact more with one another.

HOW TO DO HI-C SEQUENCING

- You have a protocol for Hi-C extraction
- This is sequenced as short Illumina reads
- You map the Hi-C data to your built contigs (Arima Mapping pipeline or BWA men -5SP)
- Ran YaHS and/or Salsa for scaffolding
- Build and look at Hi-C HeatMaps

NNN

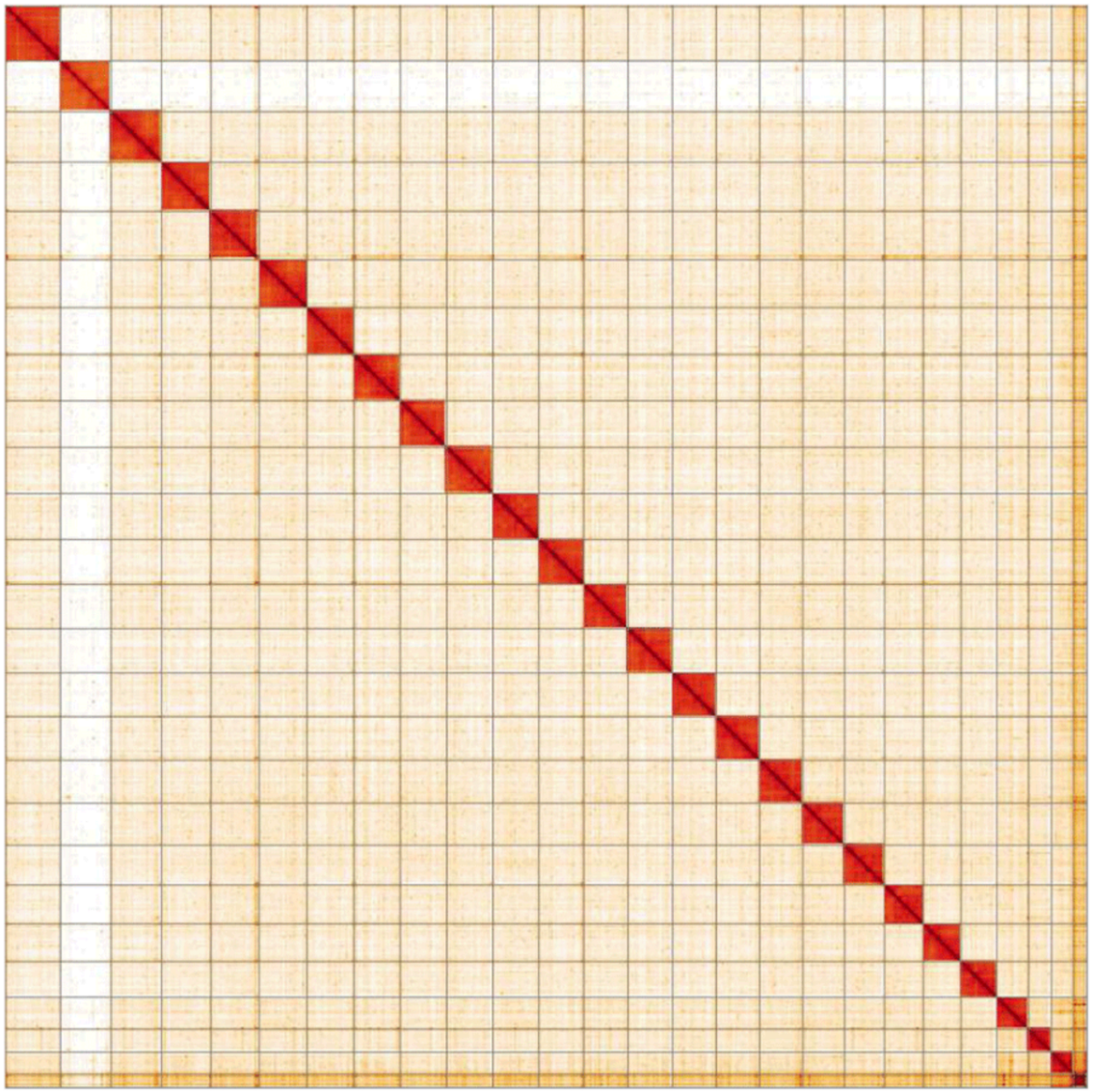
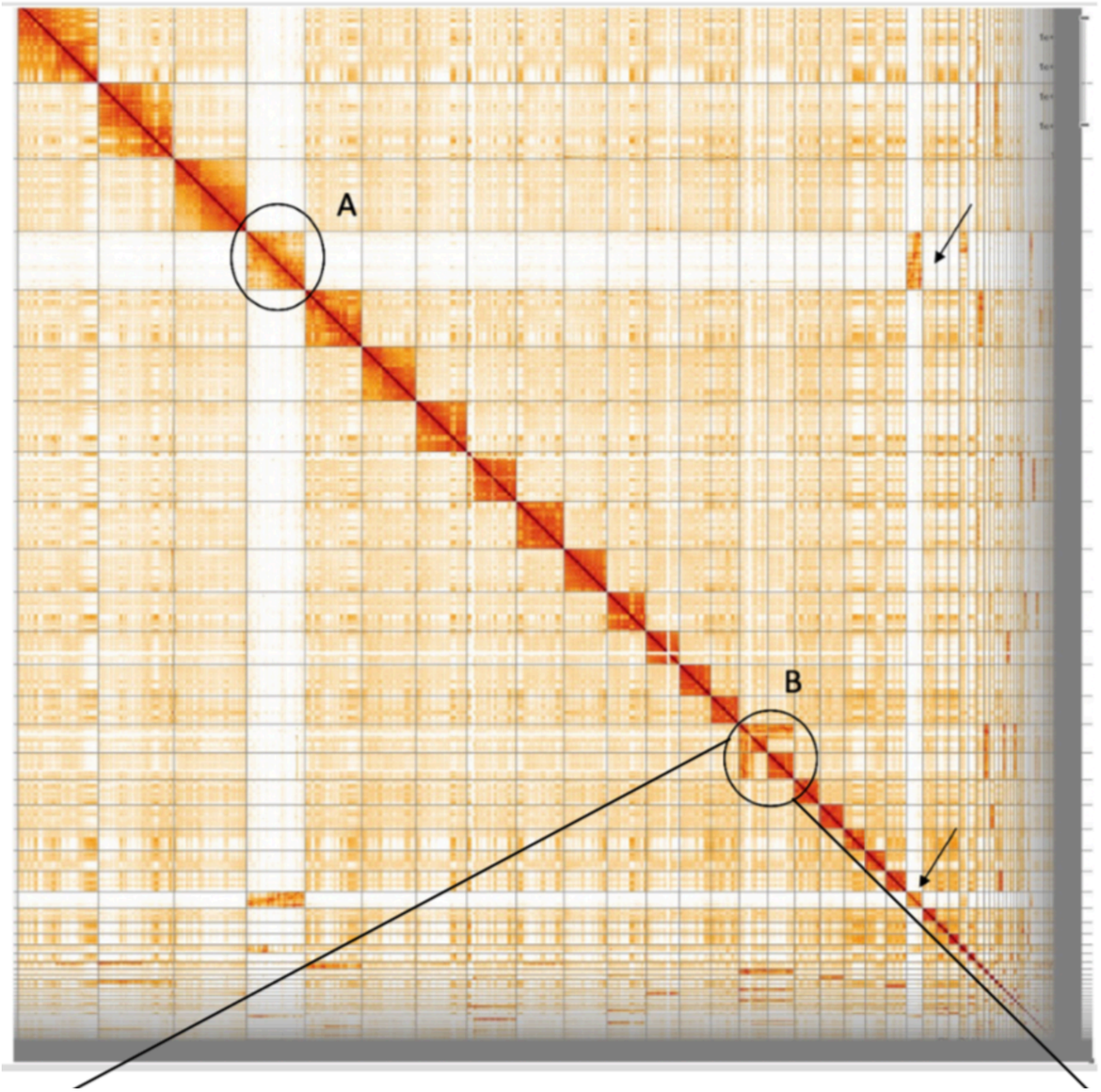


Figure 5. Genome assembly of *Pieris rapae*, ilPieRapa1.1: Hi-C contact map.

Hi-C contact map of the ilPieRapa1.1 assembly, visualised in HiGlass. Chromosomes are given in size order from left to right and top to bottom.

YOU DO MORE THAN SCAFFOLDING WITH HI-C: YOU SEE BIOLOGY



Choloepus didactylus VGP

Non-curated output

3.2 Gb, 281 scaffolds, N50 = 161 Mb

SHARED SEQUENCES: CENTROMERES

wellcomeopenresearch.org

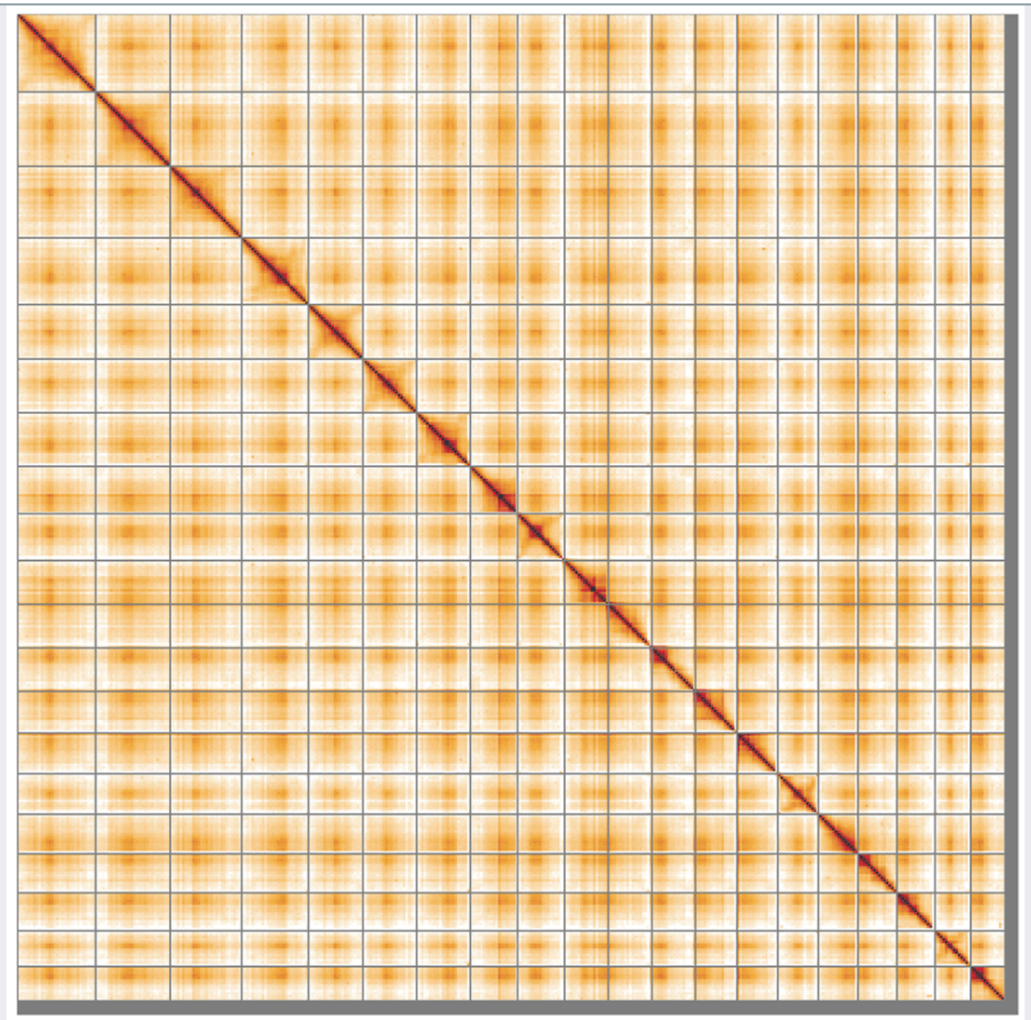


Figure 5. Genome assembly of *Ilex aquifolium*, drIleAqui2.1: Hi-C contact map of the drIleAqui2.1 assembly, visualised using HiGlass.

YaHS: yet another Hi-C scaffolding tool

Chenxi Zhou[1,2], Shane A. McCarthy[1,2], and Richard Durbin[1,2,*]

[1] Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK
[2] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
[*] Correspondence: rd109@cam.ac.uk

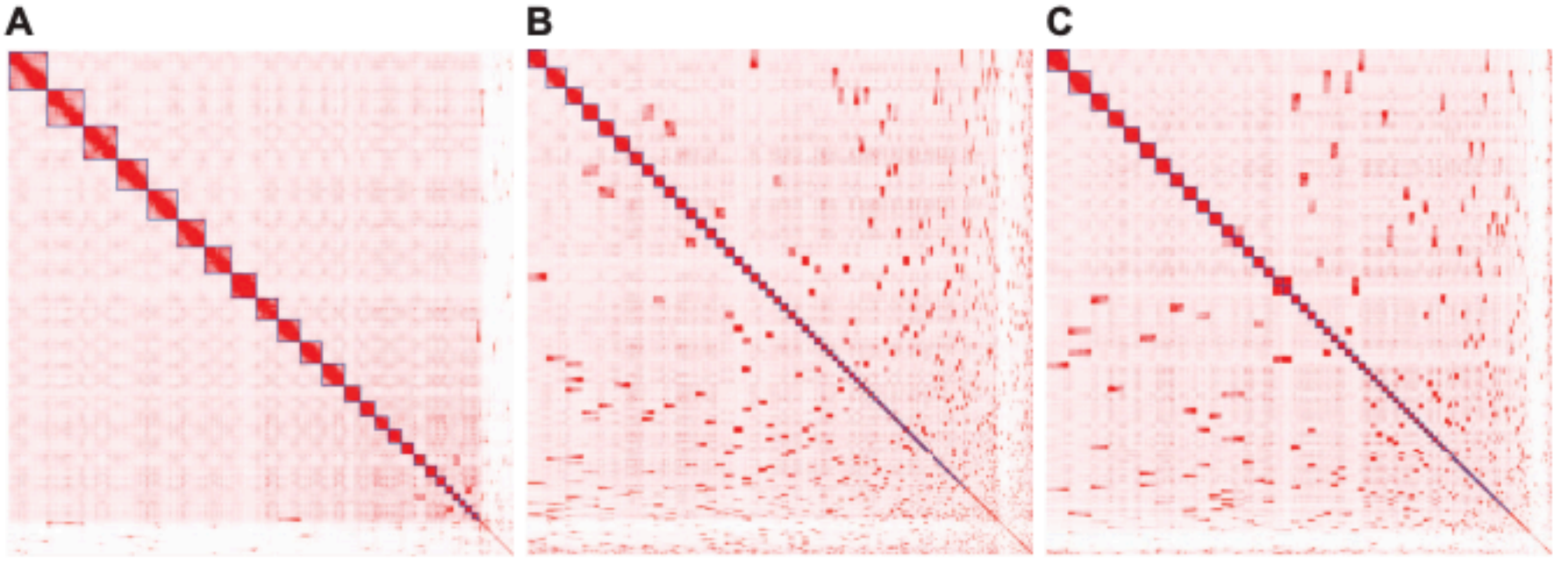
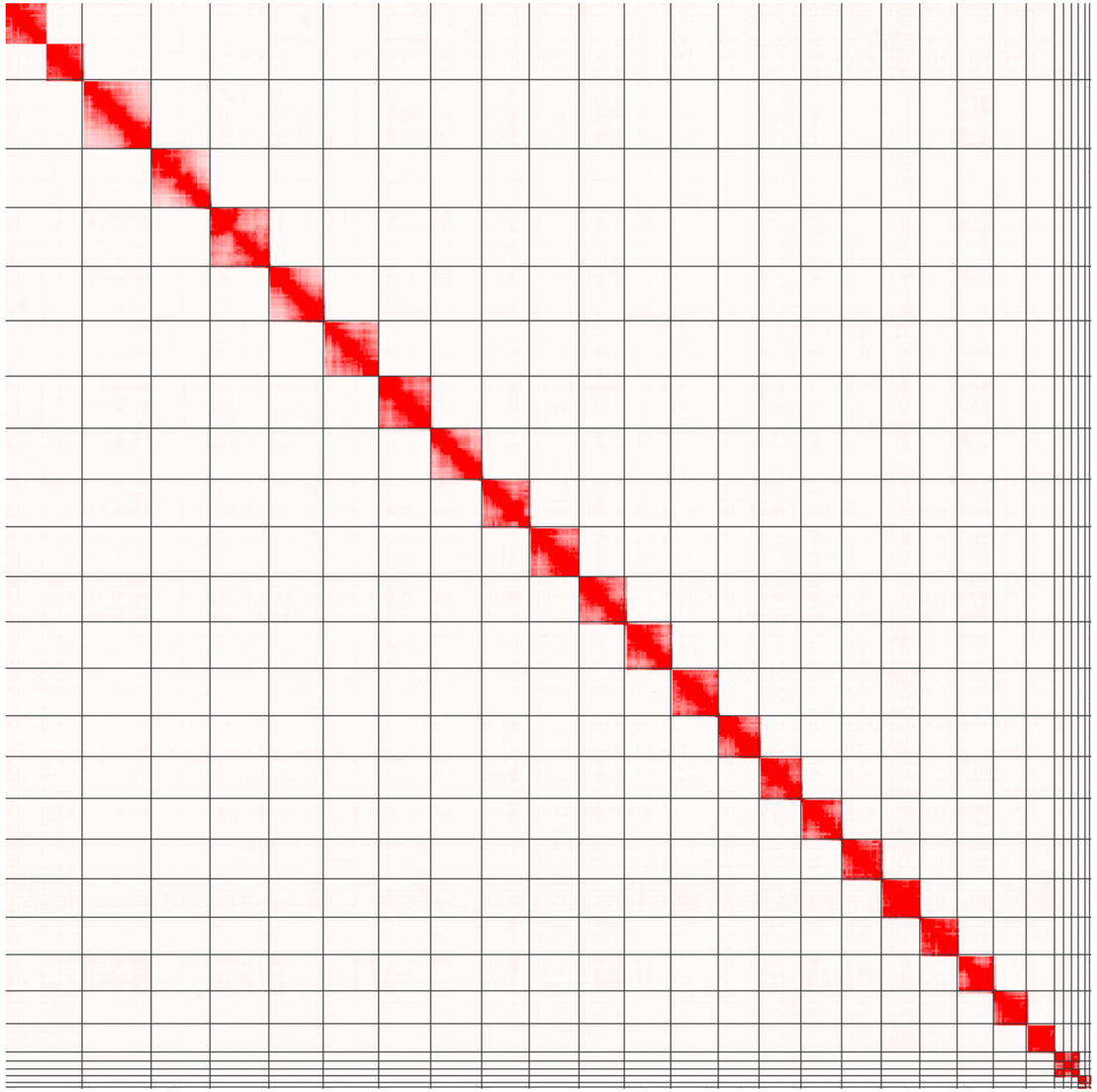


Figure 1. Hi-C contact maps of genome assemblies constructed with YaHS (**A**), SALSA2 (**B**) and pin_hic (**C**) for the simulated T2T data without contig errors. The blocks highlighted with blue squares in diagonal line are scaffolds. The contact maps were plotted with Juicebox (Durand *et al.*, 2016).

HI-C: DETECTING MISASSEMBLES

Look at me!!!!

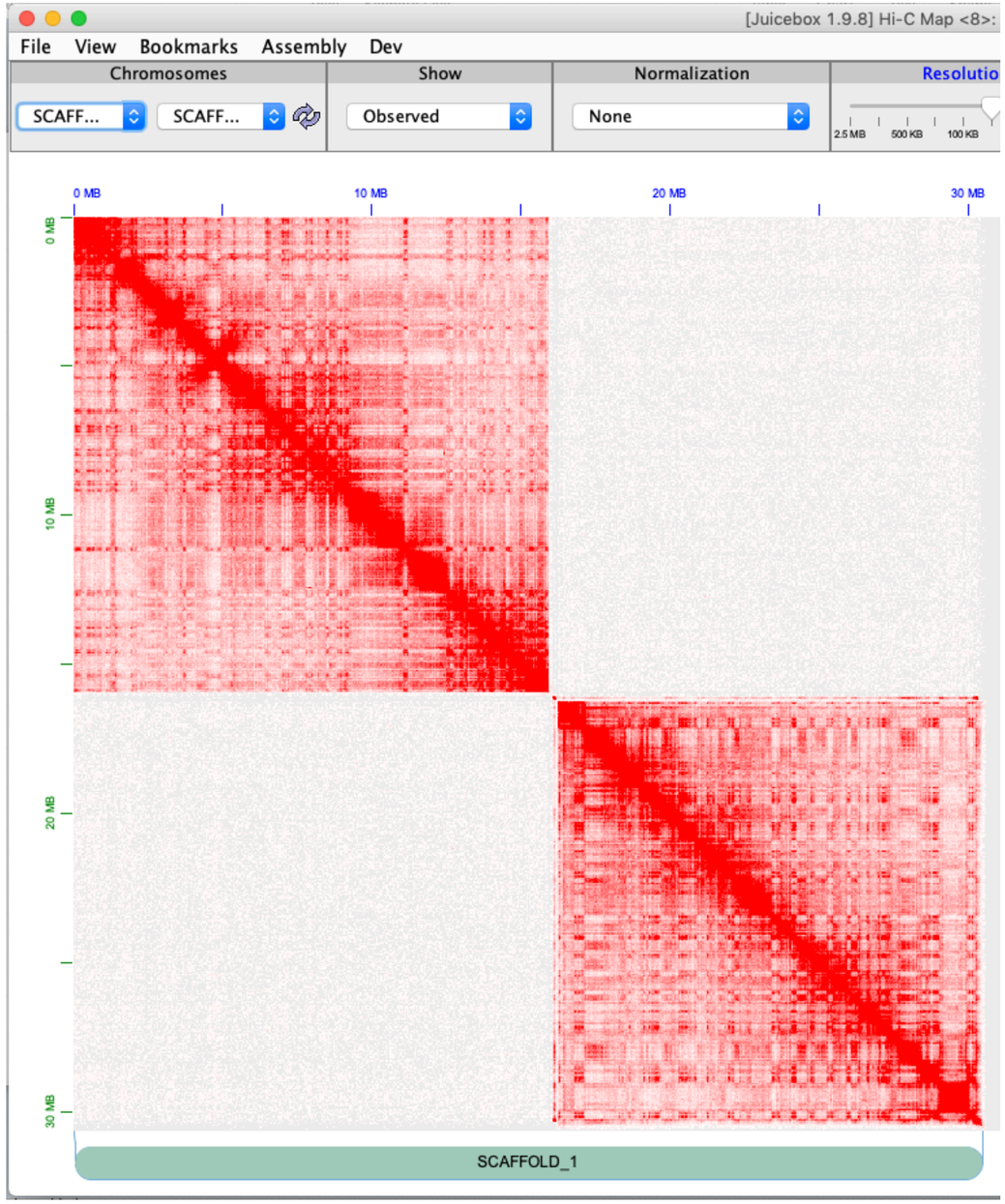
Lycaena phlaeas - ilLycPhla1



Darwin TREE of LIFE

HI-C: DETECTING MISASSEMBLES

Lycaena phlaeas - ilLycPhla1



OXFORD (GIGA)n SCIENCE

GigaScience, 10, 2021, 1–9

doi: 10.1093/gigascience/giaa153
Review

REVIEW

Significantly improving the quality of genome assemblies through curation

Kerstin Howe*, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey and Jonathan Wood

Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK

*Correspondence address. Kerstin Howe, Tree of Life, Wellcome Sanger Institute, Cambridge CB10 1SA, UK.
E-mail: kerstin@sanger.ac.uk http://orcid.org/0000-0003-2237-513X

Abstract

Genome sequence assemblies provide the basis for our understanding of biology. Generating error-free assemblies is therefore the ultimate, but sadly still unachieved goal of a multitude of research projects. Despite the ever-advancing improvements in data generation, assembly algorithms and pipelines, no automated approach has so far reliably generated near error-free genome assemblies for eukaryotes. Whilst working towards improved datasets and fully automated pipelines, assembly evaluation and curation is actively used to bridge this shortcoming and significantly reduce the number of assembly errors. In addition to this increase in product value, the insights gained from assembly curation are fed back into the automated assembly strategy and contribute to notable improvements in genome assembly quality. We describe our tried and tested approach for assembly curation using gEVAL, the genome evaluation browser. We outline the procedures applied to genome curation using gEVAL and also our recommendations for assembly curation in a gEVAL-independent context to facilitate the uptake of genome curation in the wider community.

Keywords: genome; assembly; curation; gEVAL

Raw data

- ❑ fits expected species
- ❑ all raw data groups belong to same project

Automated assembly

- ❑ false duplications removed
- ❑ contaminants removed
- ❑ trailing Ns removed
- ❑ kmer spectrum fits expectations
- ❑ haplotypes correctly assorted (if fully haplotype-resolved)
- ❑ organelle genomes produced with dedicated pipeline
- ❑ polishing reliable (e.g. NUMT/NUPTs preserved and frameshifts corrected)
- ❑ organelle genomes as unintended byproduct of assembly identified and removed

Cleaned assembly

- ❑ all available raw data aligned, discrepancies marked up
- ❑ all available data from other sources aligned, discrepancies marked up
- ❑ discrepancies assessed and corrected where necessary
  - ❑ misjoins broken
  - ❑ missing joins made
  - ❑ inversions reoriented
  - ❑ false duplications removed

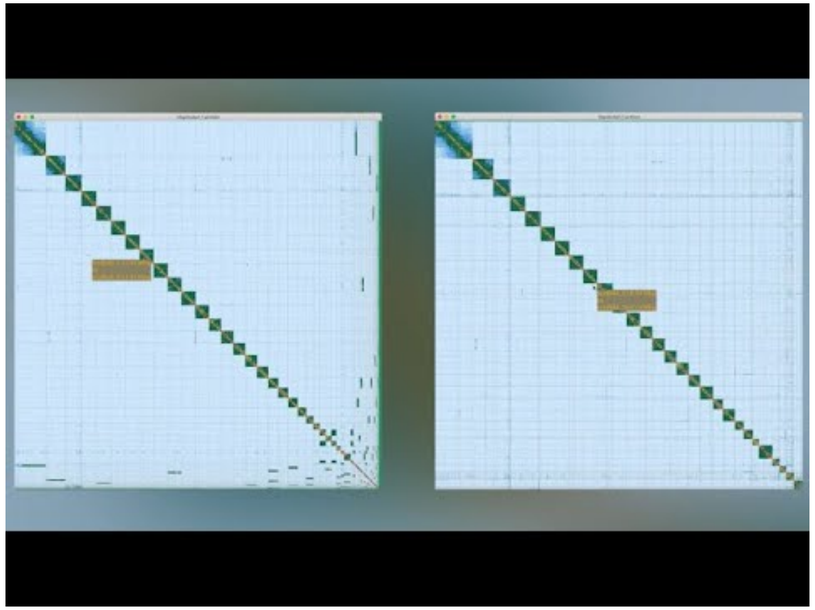
Structurally correct assembly

- ❑ chromosome-level units identified and named
- ❑ sex chromosomes identified and named where possible
- ❑ unlocalised units identified and named
- ❑ organelle genomes named

Submission to INSDC archive

LOTS OF MATERIALS FROM THE CURATION TEAM

https://gitlab.com/wtsi-grit/rapid-curation/-/blob/main/Interpreting_HiC_Maps_guide.pdf



https://bga23.org/manual-curation/

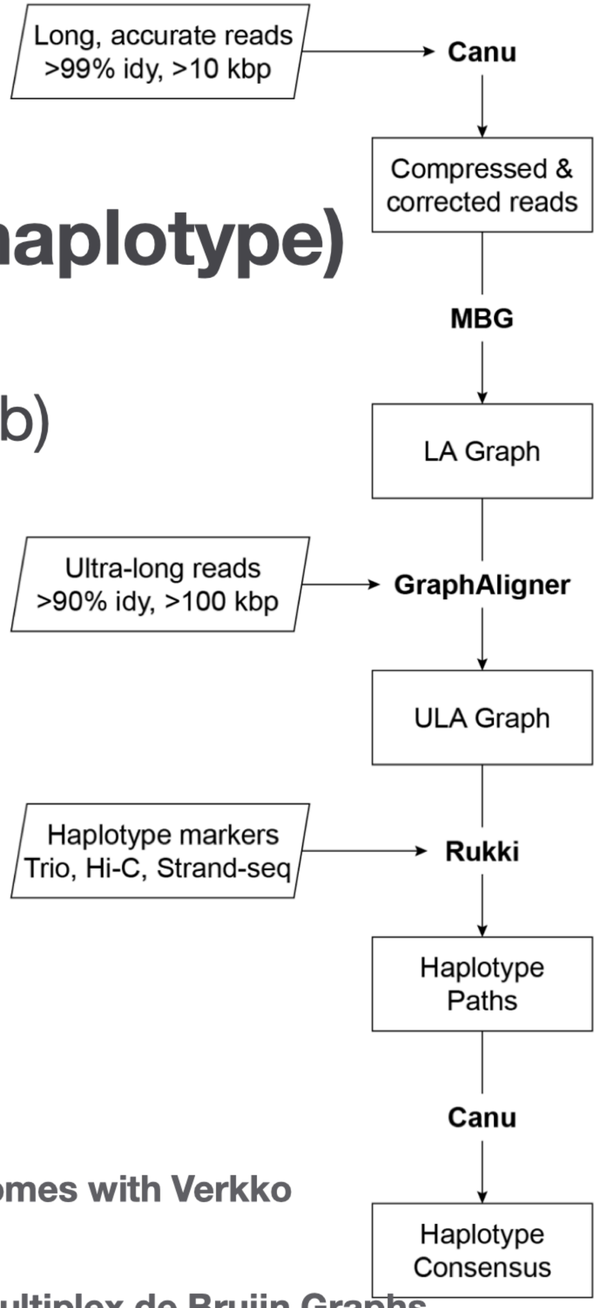
https://docs.google.com/presentation/d/1g9Ubjjjpl4Vxvw-HSOlodJUDJBaicBbU/edit#slide=id.p13

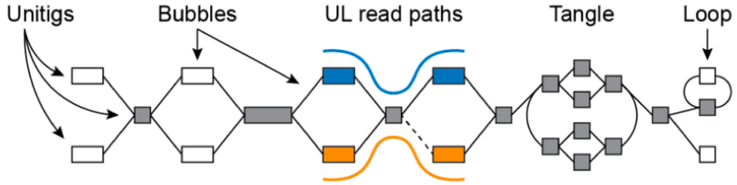
THE ONLY TYPE OF HYBRID ASSEMBLY YOU SHOULD USE IS VERKKO

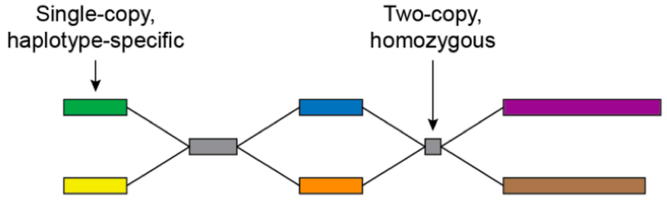
Slide Sergey Koren

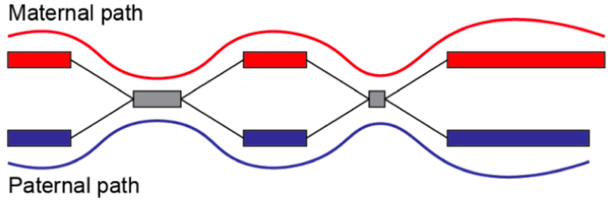
Verkko!

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









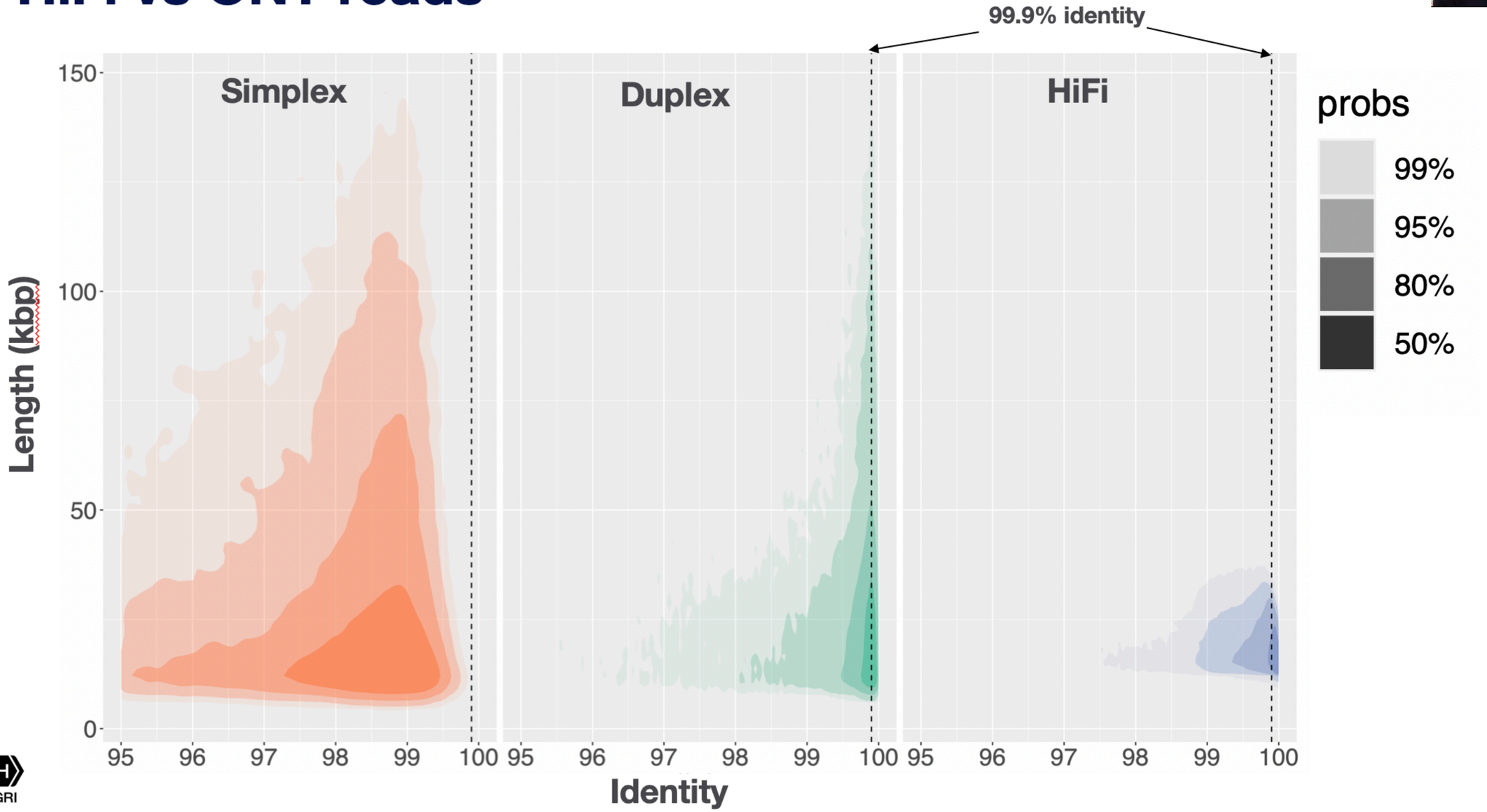
Telomere-to-telomere assembly of diploid chromosomes with Verkko
Rautiainen, *et al.* Nat Biotech (2023)

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

NIH NHGRI

Slide Sergey Koren

HiFi vs ONT reads



NIH
NHGRI

Software | Open Access | Published: 18 July 2023

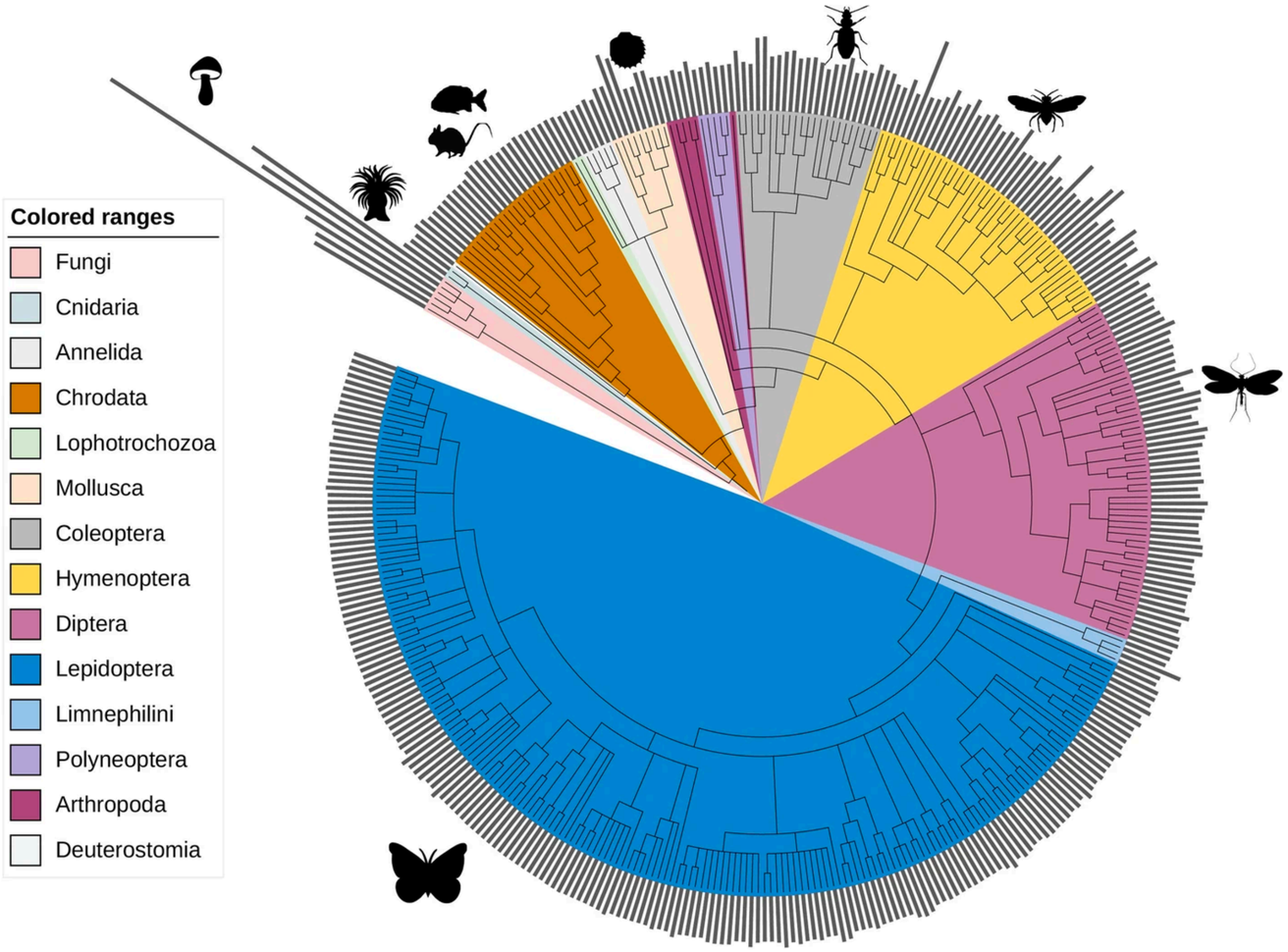
MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads

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

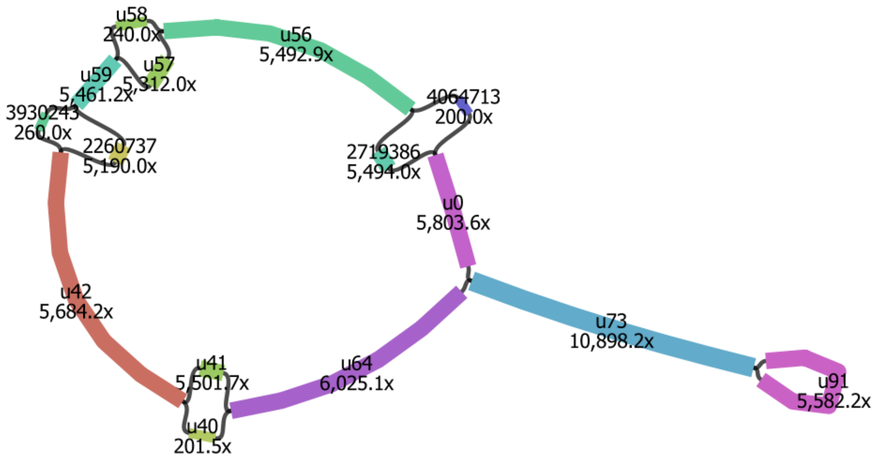
BMC Bioinformatics **24**, Article number: 288 (2023) | Cite this article

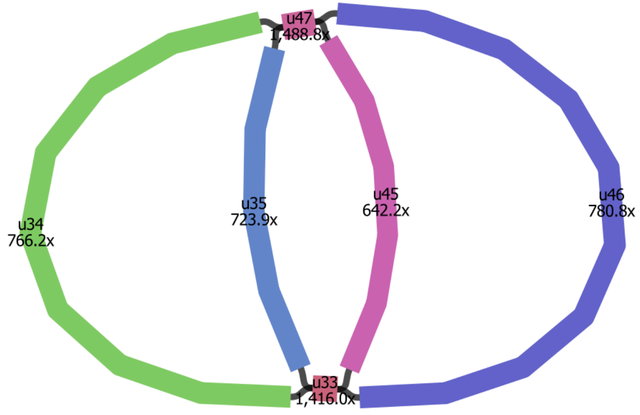
Use the docker container!!!

Pleeeeaseeee….



Oatk: an organelle assembly toolkit

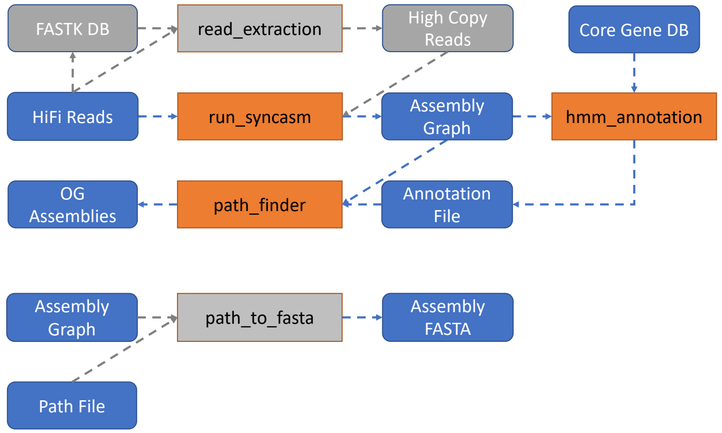




Chenxi Zhou

- Classification of graph components
  - Mitochondria
  - Plastid
  - others
- For each organelle graph component, list all possible paths and do selection
- For circular chloroplast assemblies, do rotation with reference to the conserved gene order

Oatk workflow



Darwin Tree of Life

Mutual aid: How you can join, and how you can benefit.

- We would like to sequence **your** species of interest to support your future plans in genomic understanding of biodiversity - be it population genetics, conservation, ecosystems, evolution, …

Tell us through the form at tinyurl.com/dtol-suggest

https://darwintreeoflife.org

To conclude

- There is no one-size-fits all protocol for DNA extraction
- Long reads will be ideal for chromosome level genome assembly
- You MUST investigate the genome size, heterozygosity and repeat content at the beginning of your genome assembly project
- Hi-C is useful not only for scaffolding, but for genome curation as well
- All Tree of Life is data Open for the use of all!

https://wellcomeopenresearch.org/gateways/treeoflife

Wellcome Open Research / Gateways

535 papers in the collection so far

SUBMIT TO THIS GATEWAY

TRACK

Search this Gateway

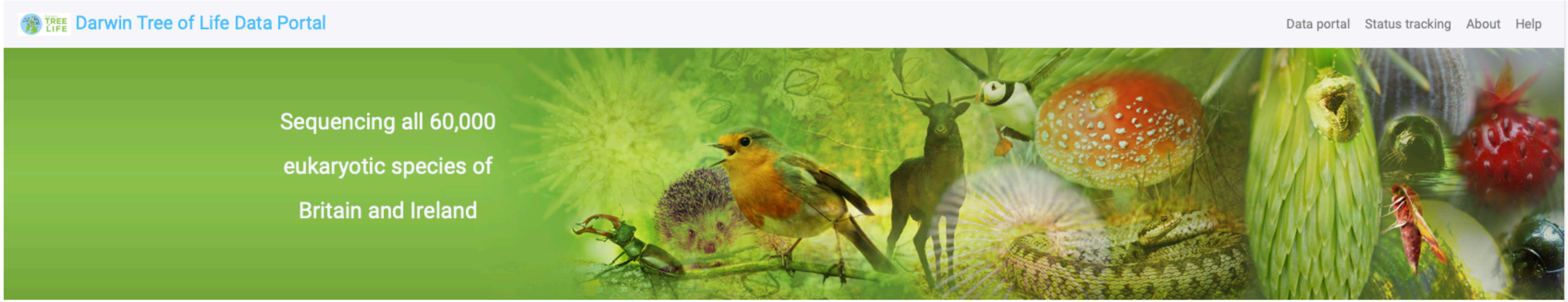
The Tree of Life Programme

This gateway collates genome sequences released by the Wellcome Sanger Institute as part of the Darwin Tree of Life project (sequencing the genomes of all known species of animals, plants, fungi and protists in Britain and Ireland) and other initiatives.

Read more in the blog →

https://wellcomeopenresearch.org/treeoflife

portal.darwintreeoflife.org



THINGS I HAVEN'T MENTIONED THAT YOU MUST LEARN ABOUT

Research | Open access | Published: 27 September 2022

Widespread false gene gains caused by duplication errors in genome assemblies

Byung June Ko, Chul Lee, Juwan Kim, Arang Rhie, Dong Ahn Yoo, Kerstin Howe, Jonathan Wood, Seoae Cho, Samara Brown, Giulio Formenti, Erich D. Jarvis & Heebal Kim

Genome Biology **23**, Article number: 205 (2022) | Cite this article

"Whole genome alignments revealed that 4 to 16% of the sequences are falsely duplicated in the previous assemblies, impacting hundreds to thousands of genes. These lead to overestimated gene family expansions.

The main source of the false duplications is heterotype duplications, where the haplotype sequences were relatively more divergent than other parts of the genome leading the assembly algorithms to classify them as separate genes or genomic regions." Kim et al, 2022

HOW TO IDENTIFY RETAINED HAPLOTIGS? PURGING AND MERQURY!!!!!

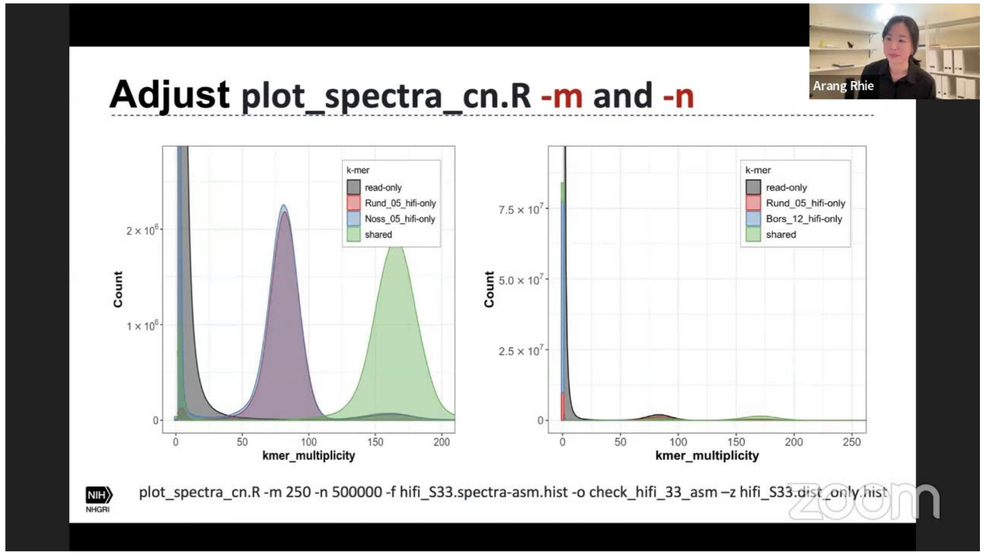
Bioinformatics, 36(9), 2020, 2896–2898
doi: 10.1093/bioinformatics/btaa025
Advance Access Publication Date: 23 January 2020
Applications Note
OXFORD

Genome analysis

Identifying and removing haplotypic duplication in primary genome assemblies

Dengfeng Guan[1,2], Shane A. McCarthy [2], Jonathan Wood[3], Kerstin Howe [3], Yadong Wang[1,*] and Richard Durbin [2,3,*]

[1]Department of Computer Science and Technology, Center for Bioinformatics, Harbin Institute of Technology, Harbin 150001, China, [2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK and [3]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK



Rhie *et al. Genome Biology* (2020) 21:245
https://doi.org/10.1186/s13059-020-02134-9
Genome Biology

METHOD Open Access

Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies

Arang Rhie*, Brian P. Walenz, Sergey Koren and Adam M. Phillippy

* Correspondence: arang.rhie@nih.gov
Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Abstract

Recent long-read assemblies often exceed the quality and completeness of available reference genomes, making validation challenging. Here we present Merqury, a novel tool for reference-free assembly evaluation based on efficient k-mer set operations. By comparing k-mers in a de novo assembly to those found in unassembled high-accuracy reads, Merqury estimates base-level accuracy and completeness. For trios, Merqury can also evaluate haplotype-specific accuracy, completeness, phase block continuity, and switch errors. Multiple visualizations, such as k-mer spectrum plots, can be generated for evaluation. We demonstrate on both human and plant genomes that Merqury is a fast and robust method for assembly validation.

Keywords: Genome assembly, Assembly validation, Benchmarking, K-mers, Haplotype phasing, Trio binning

YouTube CZ

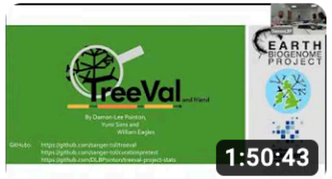
Pesquisar

Biodiversity Genomics Academy 2023

Biodiversity Genomics Academy

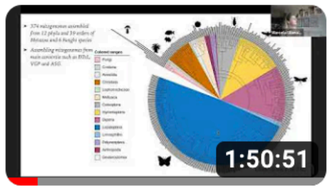
26 vídeos 711 visualizações Última atualização em 1 d...

Reproduzir tu... Ordem aleató...

1 

Fri 8 Sep, 11:00 - The Treeval pipeline: Generating evidence for manual curation

Biodiversity Genomics Academy • 273 visualizações • Transmitido há 4 meses

2 

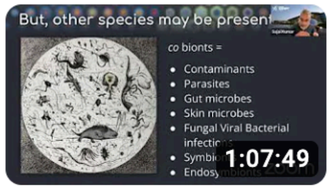
Fri 8 Sep, 13:00 - Assembling Mitogenomes from PacBio HiFi reads using MitoHiFi

Biodiversity Genomics Academy • 257 visualizações • Transmitido há 4 meses

3 

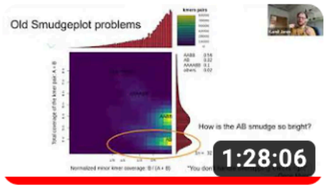
Mon 11 Sep, 08:00 - Visualising genome assembly cobionts by running BlobToolKit locally

Biodiversity Genomics Academy • 150 visualizações • Transmitido há 4 meses

4 

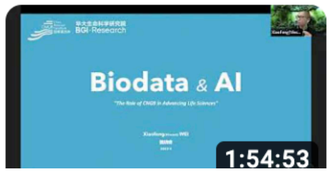
Mon 11 Sep, 17:00 - Checking for cobionts in public genomes using BlobToolKit

Biodiversity Genomics Academy • 90 visualizações • Transmitido há 4 meses

5 

Tue 12 Sep, 14:00 - Understanding k-mers and ploidy using Smudgeplot

Biodiversity Genomics Academy • 274 visualizações • Transmitido há 4 meses

6 

Wed 13 Sep, 09:00 - Starting a comparative genome study from CNGBdb

Biodiversity Genomics Academy • 86 visualizações • Transmitido há 4 meses

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)

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

Team301

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

Collaborators

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

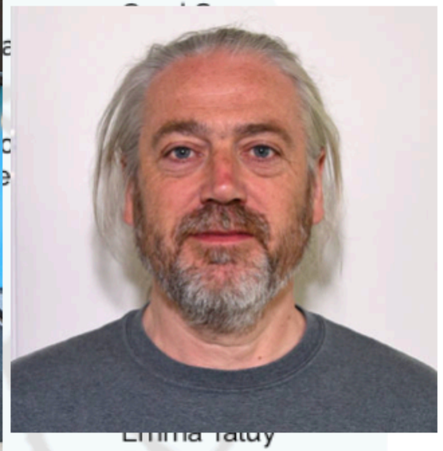
wellcome sanger institute

Obrigada! Thank you!

mu2@sanger.ac.uk

wellcome sanger institute

Scientific Operations



Team301

Jayathilaka
Martin
Thybert
Rajan
Howe
Cochrane
Harrison
Haggerty
Allen
Garcia Giron
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

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

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%