



Genome assembly: where do I start?

And where do I go once I have contigs and scaffolds

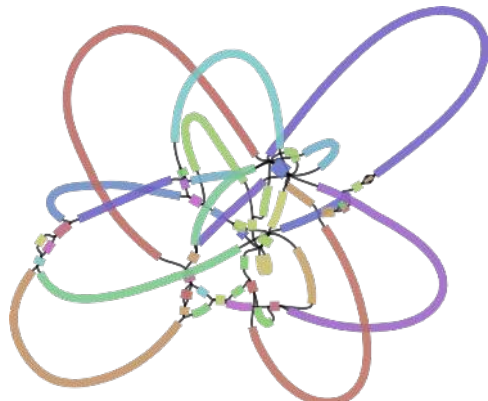
Prof Marcela Uliano-Silva



wellcome
sanger
institute

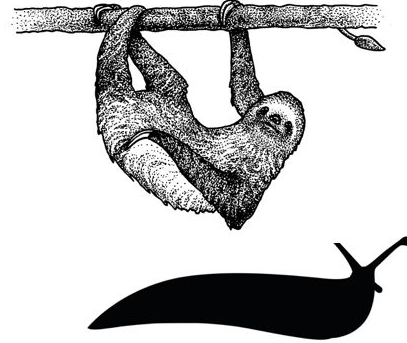


NORD
University



Who am I?

- Senior Bioinformatician - Tree of Life, Wellcome Sanger Institute, Cambridge, UK
 - Associate Professor (Prof II) - Nord University, Bodø, Norway
-
- 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



My two main areas of research

Software development for high-quality genome assembly

Comparative genomics: origins (ancestral linkage groups) and diversification of phenotypes

Software | [Open Access](#) | Published: 18 July 2023

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

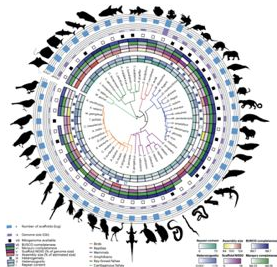
Marcela Uliano-Silva , João Gabriel R. N. Ferreira, Ksenia Krasheninnikova, Darwin Tree of Life Consortium, Giulio Formenti, Linelle Abueg, James Torrance, Eugene W. Myers, Richard Durbin, Mark Blaxter & Shane A. McCarthy

BMC Bioinformatics 24, Article number: 288 (2023) | [Cite this article](#)

Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy

Delphine Larivière, Linelle Abueg, Nadolina Brajuka, Cristóbal Gallardo-Alba, Björn Grüning, Byung June Ko, Alex Ostrovsky, Marc Palmada-Flores, Brandon D. Pickett, Keon Rabbani, Agostinho Antunes, Jennifer R. Balacco, Mark J. P. Chaisson, Haoyu Cheng, Joanna Collins, Melanie Couture, Alexandra Denisova, Olivier Fedrigo, Guido Roberto Gallo, Alice Maria Giani, Grenville MacDonald Gooder, Kathleen Horan, Nivesh Jain, Cassidy Johnson, Heebal Kim, Chul Lee, Tomas Marques-Bonet, Bri Arang Rhie, Simona Secomandi, Marcella Sozzoni, Tatiana Tilley, Marcela Uliano-Silva, M Beek, Robert W. Williams, Robert M. Waterhouse, Adam M. Phillippy, Erich D. Jarvis , V Schatz , Anton Nekrutenko  & Giulio Formenti 

Nature Biotechnology 42, 367–370 (2024) | [Cite this article](#)



Caecilian Genomes Reveal the Molecular Basis of Adaptation and Convergent Evolution of Limblessness in Snakes and Caecilians

Vladimir Ovchinnikov,^{1,†} Marcela Uliano-Silva ^{2,†} Mark Wilkinson,³ Jonathan Wood,² Michelle Smith,⁴ Karen Oliver,⁴ Ying Sims,² James Torrance,² Alexander Suh,^{5,6} Shane A. McCarthy ^{2,7} Richard Durbin ^{2,7,*} and Mary J. O'Connell ^{1,*}

JOURNAL ARTICLE

A chromosome-level assembly supports genome-wide investigation of the DMRT gene family in the golden mussel (*Limnoperna fortunei*)

João Gabriel R. N. Ferreira, Juliana A. Americo , Danielle L. A. S. do Amaral, Fábio Sendim, Yasmin R. da Cunha, Tree of Life Programme, Mark Blaxter, Marcela Uliano-Silva, Mauro de F. Rebelo Author Notes

GigaScience, Volume 12, 2023, giad072, <https://doi.org/10.1093/gigascience/giad072>

Published: 30 September 2023 Article history ▼





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



Sequencing

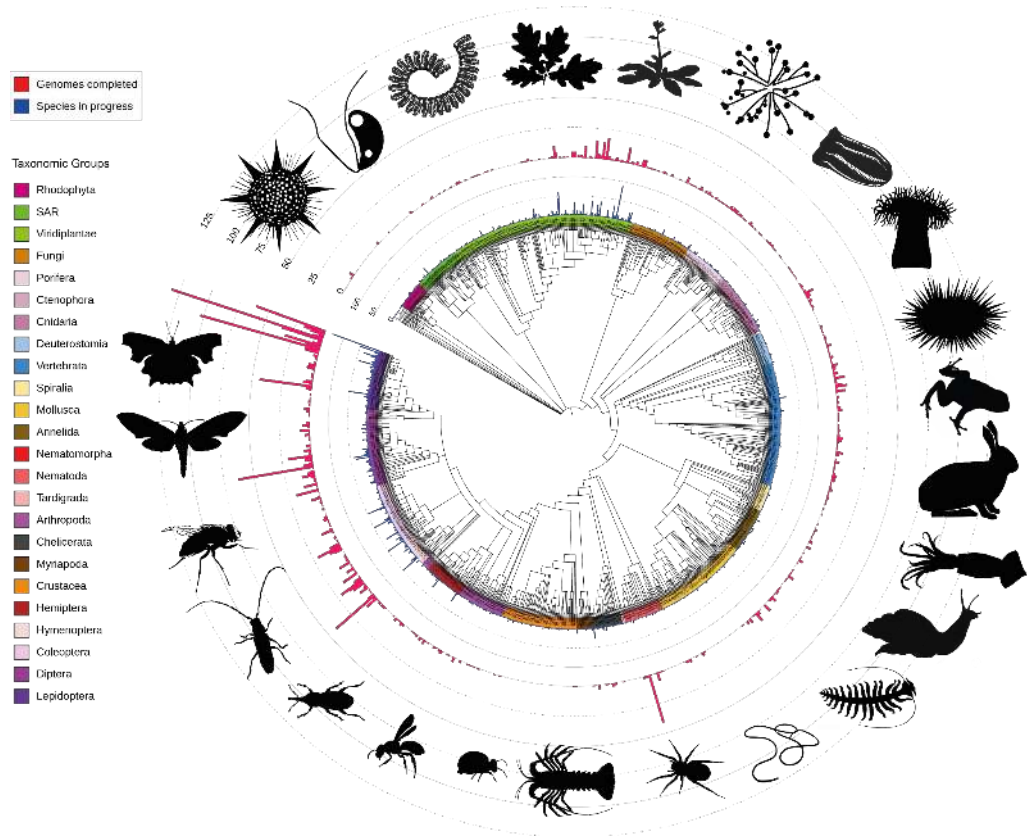
- **6500** species with sequencing data
- 260.3 Tb PacBio data
- 1269.5 Tb HiC data
- 113.4 Tb linked reads
- 33.1 Tb RNAseq data (2099 species)

Assembly

- **2711** species assembled

Public releases

- **1996** assemblies released to INSDC
 - 1509 DToL
 - 201 VGP
 - 151 ASG
- **953** genome notes published





Sequencing

- **6500** species with sequencing data
- 260.3 Tb PacBio data
- 1269.5 Tb HiC data
- 112.4 Tb linked reads
- 3

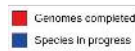
Assembly

- **27**

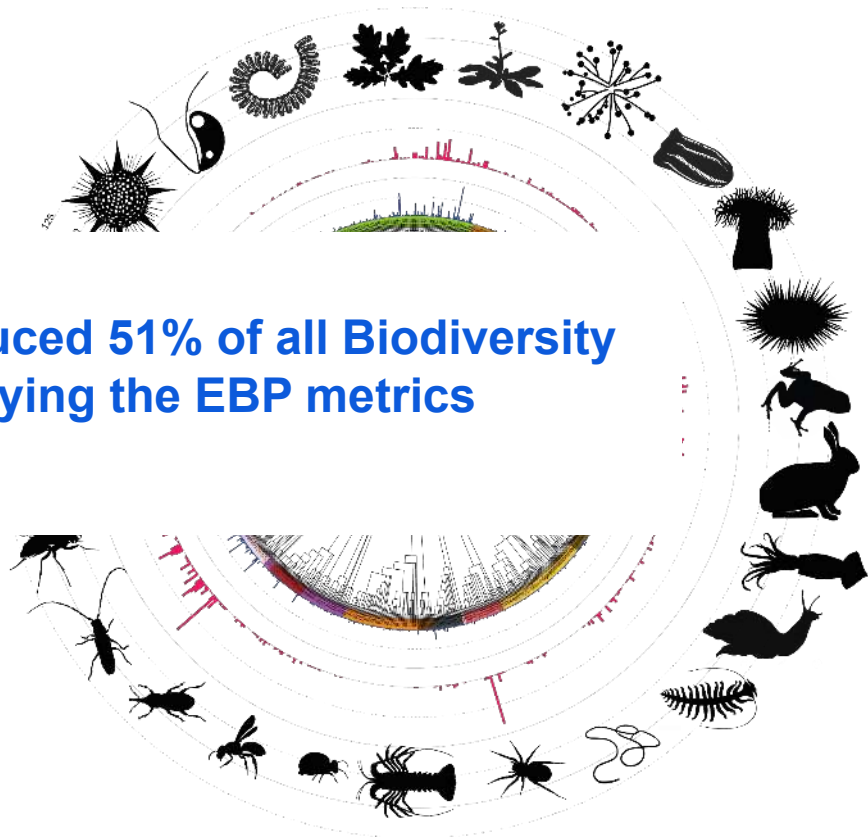
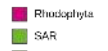
Public

- **1996** assemblies released to INSDC
 - 1509 DToL
 - 201 VGP
 - 151 ASG
- **953** genome notes published

As of October 2024, we have produced 51% of all Biodiversity assemblies worldwide satisfying the EBP metrics

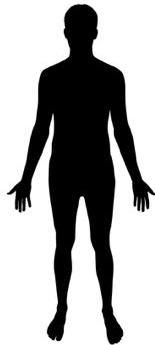


Taxonomic Groups



Genome assembly: what is my goal?

- Understand variation in populations (disease-related SNPs etc...)
- Study the molecular profile of a species never before sequenced (evolutionary studies etc..)



Genome re-sequencing
Assembly by mapping to a reference



De novo assembly

Genome assembly

Let's try to reconstruct the sentence bellow (our genome) from some fragments (reads):

- It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, ...*

It was the best of	times, it was the worst	of times, it was the	age of wisdom, it was	the age of foolishness, ...
It was the best	of times, it was the	worst of times, it was	the age of wisdom, it	was the age of foolishness,
It was the	best of times, it was	the worst of times, it	was the age of wisdom,	it was the age of foolishness, ...
It was	the best of times, it	was the worst of times,	it was the age of	wisdom, it was the age of foolishness, ...
It	was the best of times,	it was the worst of	times, it was the age	of wisdom, it was the age of foolishness, ...

Genome assembly

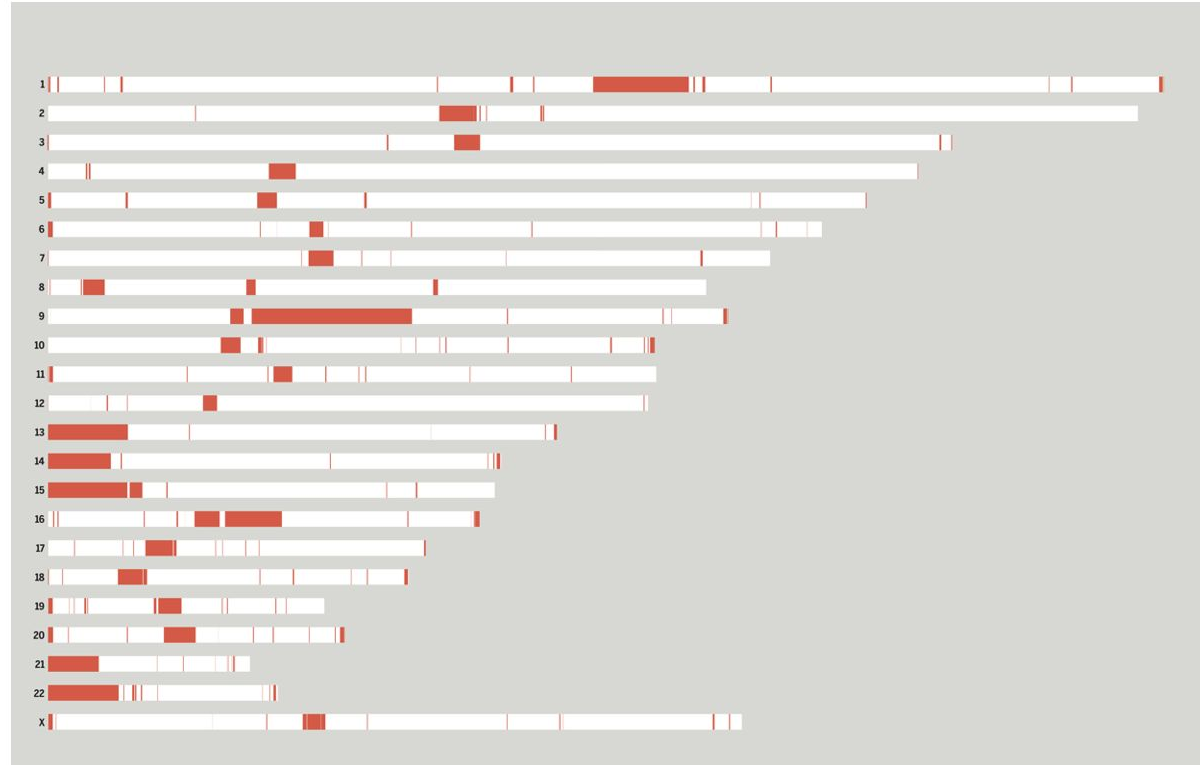
It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
of times, it was the
times, it was the worst
times, it was the age

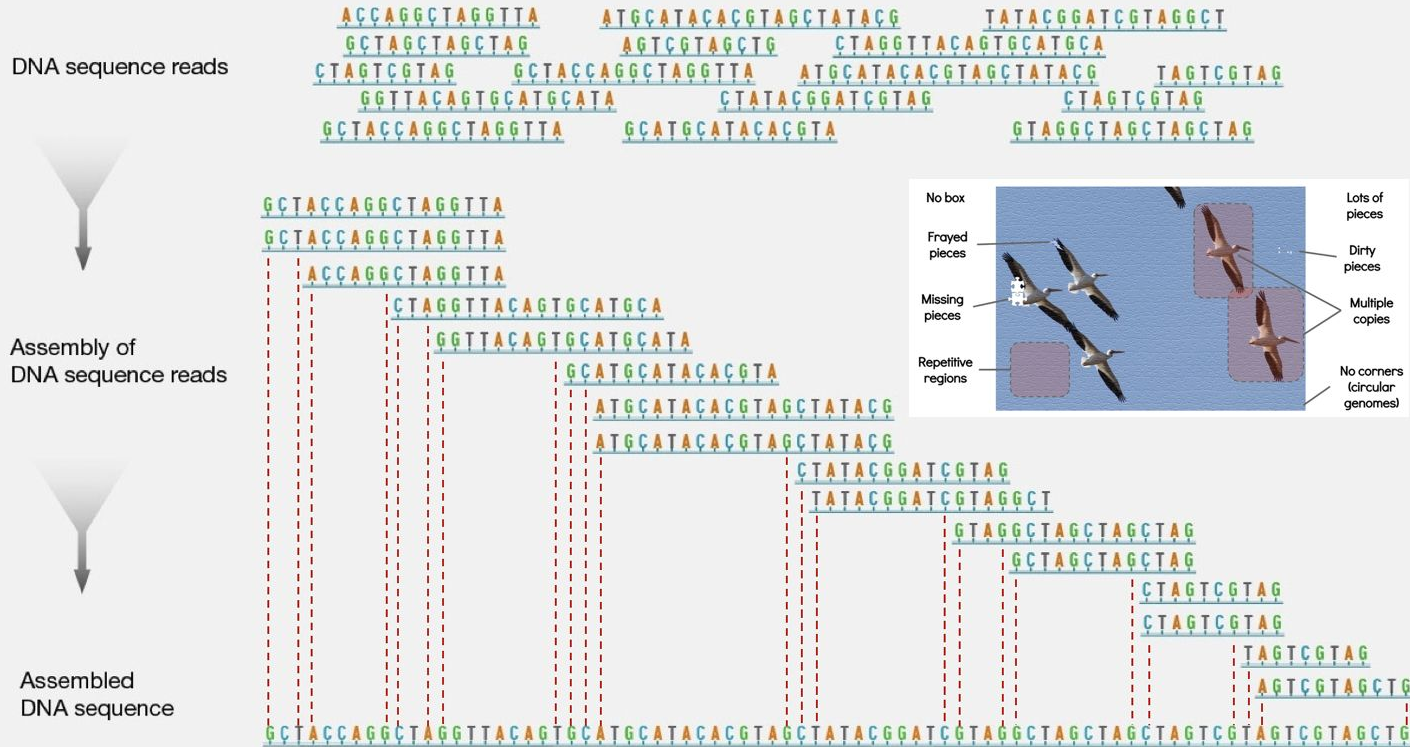
What is the next word, 'world' or 'age'?

What are eukaryotic genomes made of?

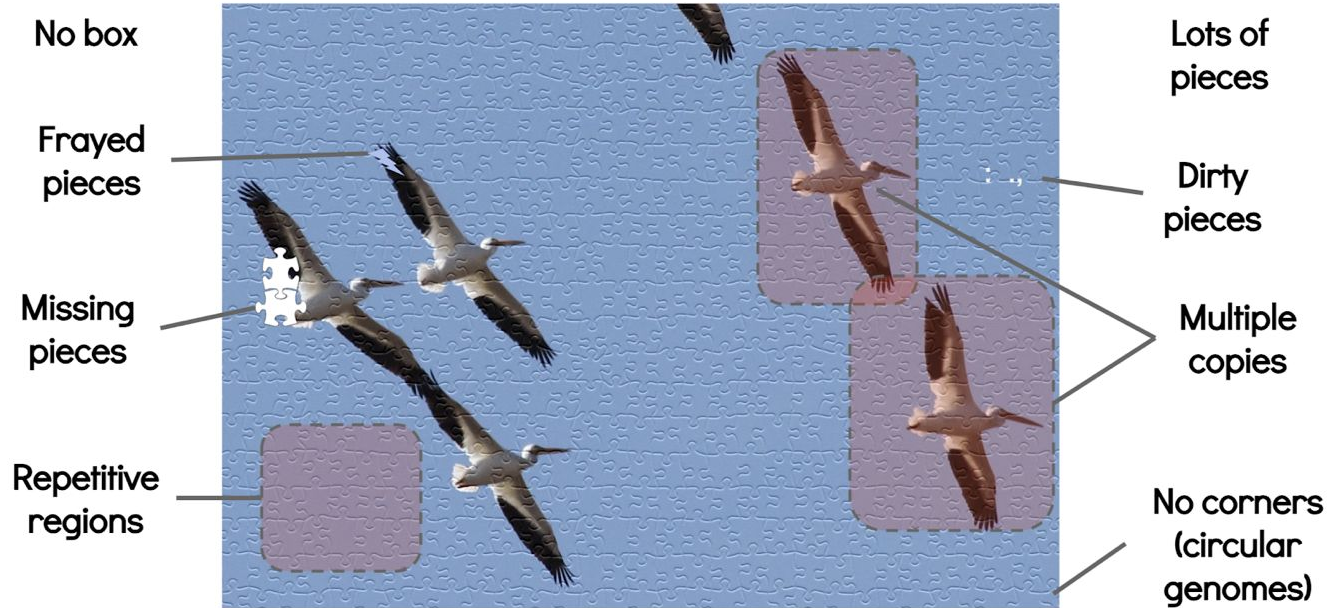
- Genes (exon, introns)
 - Repetitive elements
1. Mobile elements (transposons)
 2. Centromeres (tandem arrays of repeat sequence studded with transposable elements)
 3. Telomeres (tandem arrays of simple repeats)
 4. Segmental duplications
 5. rRNAS



The Naïve Genome Assembly Approach

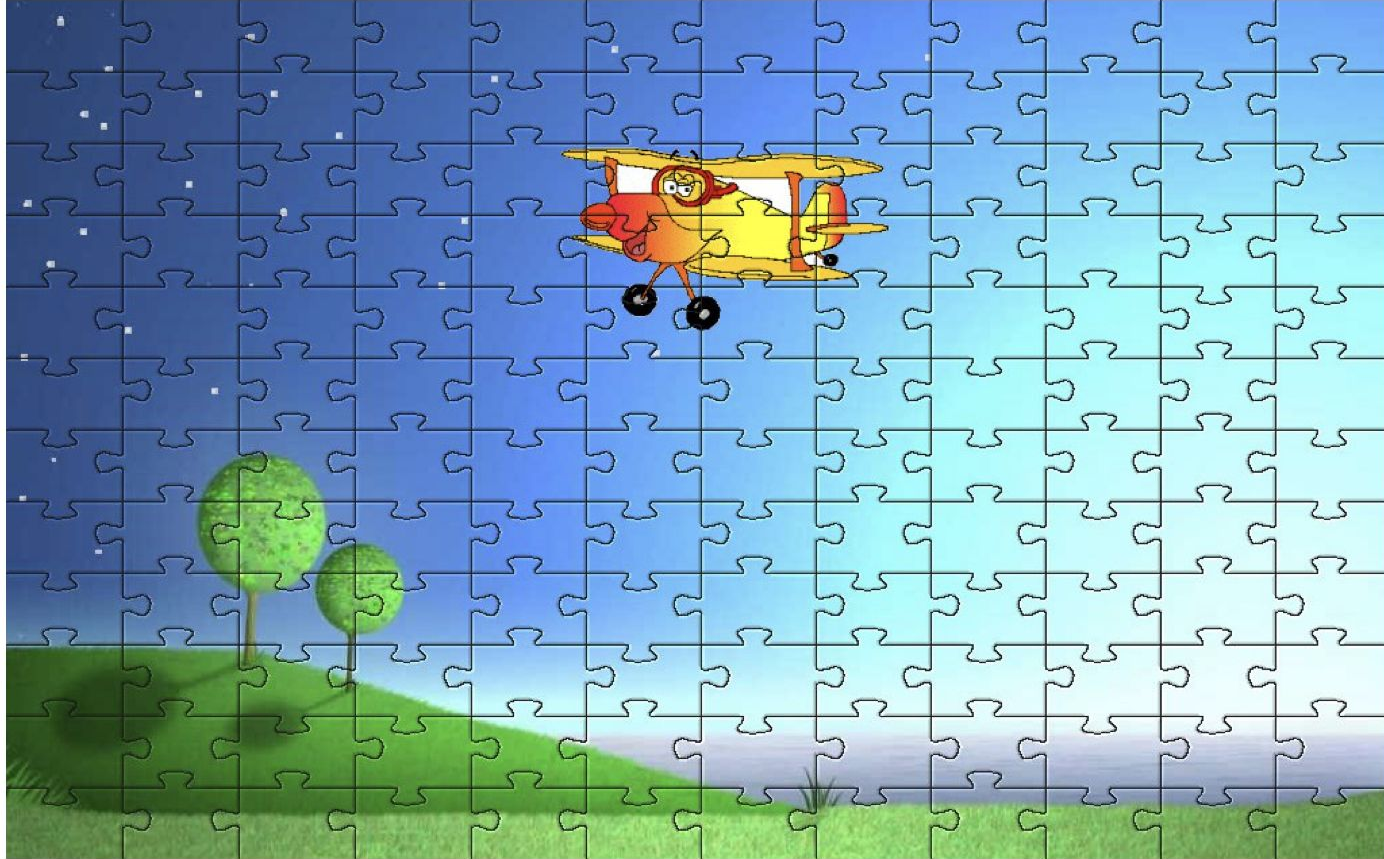


What makes a jigsaw puzzle hard?

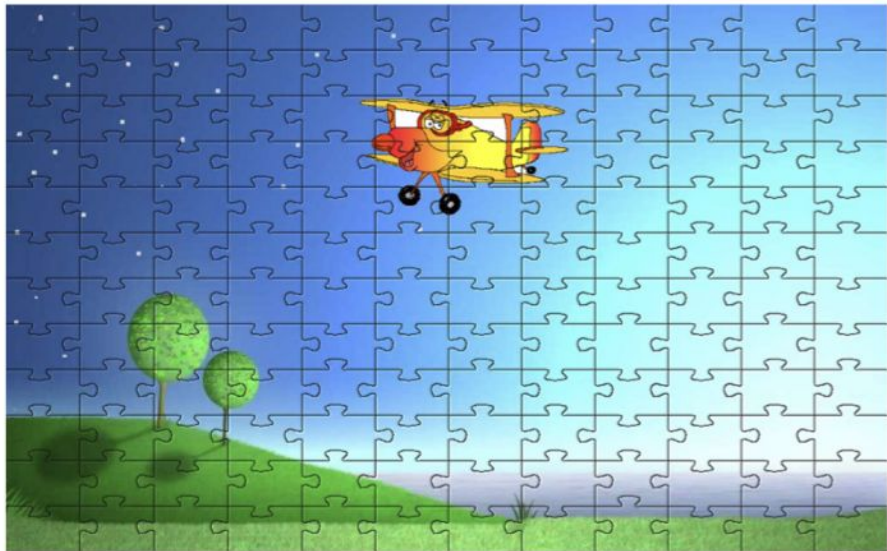


- What helps? Larger pieces (read length); fewer dirty or frayed pieces (errors in reads). fewer repeats and copies...

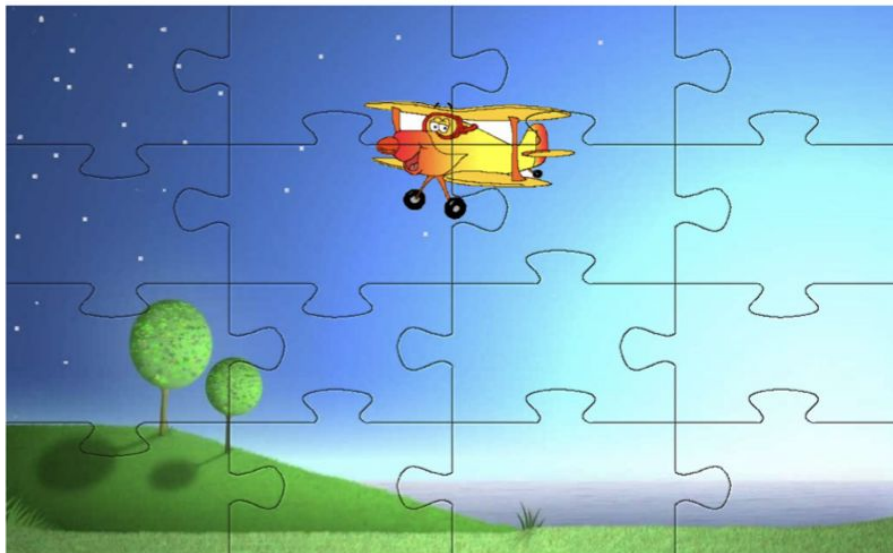
I'M A EUKARYOTIC GENOME - THE BLUE AND GREEN ARE MY REPEATS



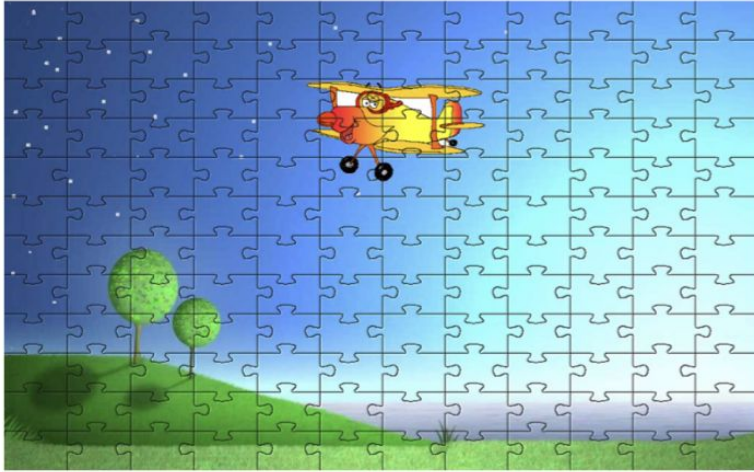
Genome assembly with short reads



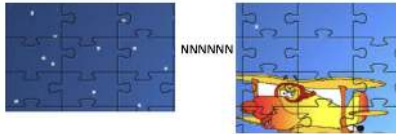
Genome assembly with long reads



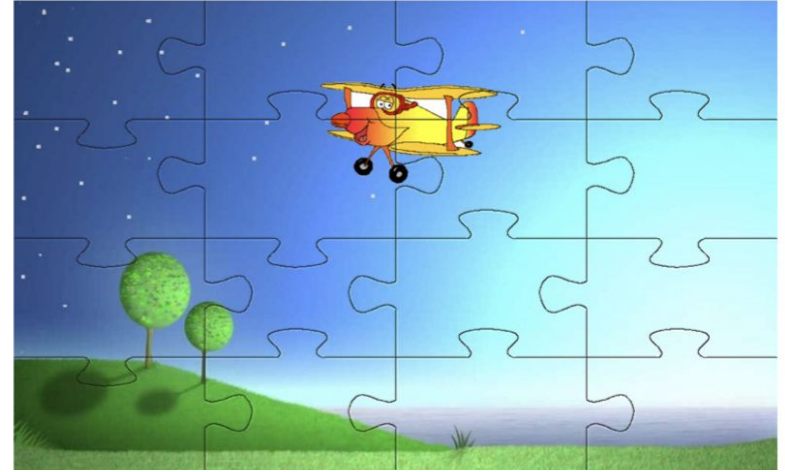
Genome assembly with short reads



- Incomplete assemblies
- Chimeras
- Many errors



Genome assembly with long reads



- Complete assemblies
- High accuracy for biological inferences

**I WANT TO TALK TO YOU ABOUT
LONG READ SEQUENCING**

Genome sequencing and assembly project: long reads

Genome assembly in the telomere-to-telomere era

Heng Li^{1,2}✉ & Richard Durbin³✉

Table 1 | Common data types for high-quality assembly

Data type	Technologies	Description	Roles
Accurate long reads	PacBio HiFi, ONT duplex	>10 kb in length; error rate <0.5%	Initial assembly graph construction; phasing over heterozygous variants that are less than 10 kb apart
Ultra-long reads	ONT ultra-long	>100 kb in length; error rate <10%	Resolving tangles; phasing through homozygous regions over 100 kb in length
Trio data	Short-read	Standard whole-genome shotgun sequencing of parents	Whole-genome phasing
Long-range data	Hi-C, Pore-C, Strand-seq	Information over 1 kb to over 10 Mb in length	Chromosomal phasing; chromosome-scale scaffolding

PACBIO HIFI READS



Start with high-quality double stranded DNA



Ligate SMRTbell adapters and size select



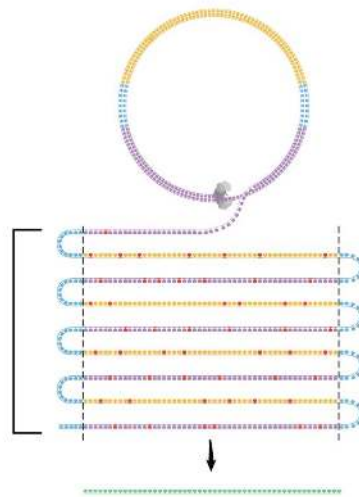
Anneal primers and bind DNA polymerase



Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus is called from subreads



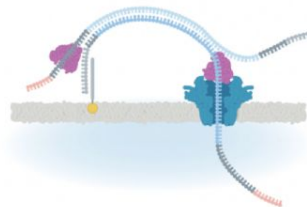
HiFi READ
(>99% accuracy)

• Nanopore Duplex

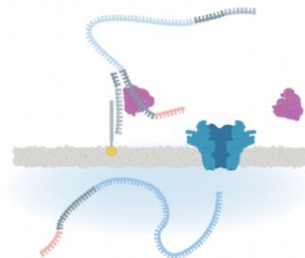
- >10 kb reads
- 99.9% (Q30) read quality
- 99.999% (Q50+) assembly quality



Linear dsDNA molecule adapted on both ends and first strand sequenced



Second strand captured and sequenced subsequently



YOUR GENOME ASSEMBLY PROJECT STARTS IN THE LAB

High Molecular Weight DNA extraction is key

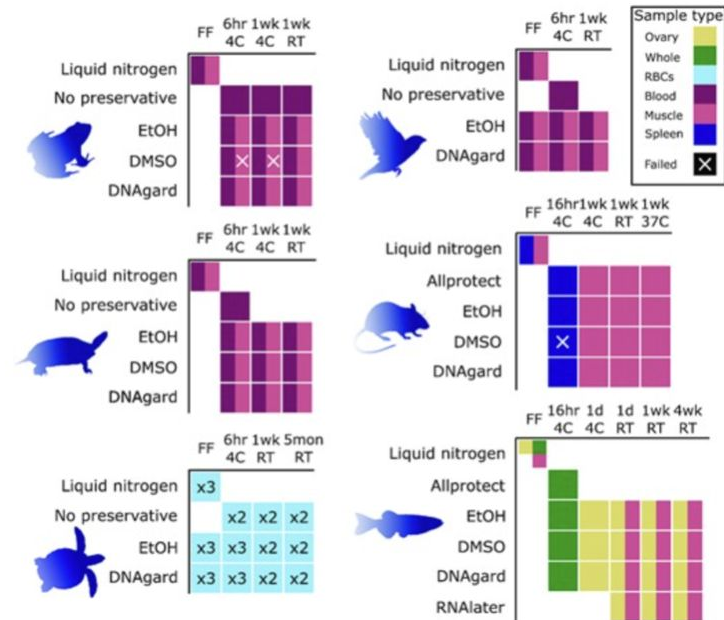
Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing

Hollis A. Dahn^{1,†}, Jacquelyn Mountcastle^{2,†}, Jennifer Balacco², Sylke Winkler³, Iliana Bista^{4,5}, Anthony D. Schmitt⁶, Olga Vinnere Pettersson⁷, Giulio Formenti², Karen Oliver⁴, Michelle Smith⁴, Wenhua Tan³, Anne Kraus³, Stephen Mac⁶, Lisa M. Komoroske⁸, Tanya Lama⁸, Andrew J. Crawford⁹, Robert W. Murphy¹, Samara Brown², Alan F. Scott¹⁰, Phillip A. Morin¹¹, Erich D. Jarvis^{2,12} and Olivier Fedrigo^{2,*}

No one-size-fits-all protocol!

 **slack** Channel: [all.things.up.to.assembly](https://allthingsuptoassembly.slack.com)

e 1:



No one-size-fits-all protocol!



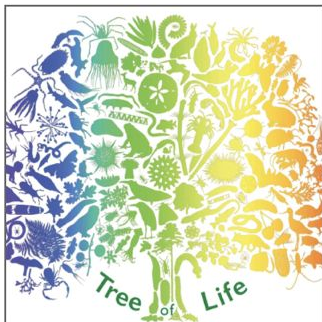
protocols.io

FEATURES

PLANS

BLOG

CASE STUDY



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 Denton¹, Halyna Yatsenko¹, Jessie Jay¹, kh¹,
Caroline Howard¹

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

Tree of Life at the Wellcome Sanger Institute



Tree of Life Genome Note Editor



Scan me!



COPY / FORK

MORE



I EXTRACTED HMW DNA: WHAT DO I DO NOW?



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)

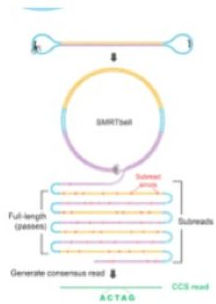


DToL Current Pipeline

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

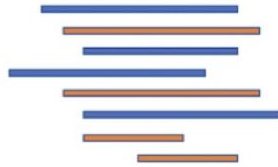


*For mitochondria genome
assembly*



Assembly

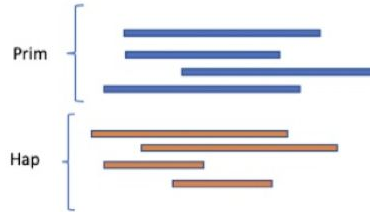
Hicanu
or Hifiasm



2 - asmstats,
BUSCO, merquy

Haplotype separation

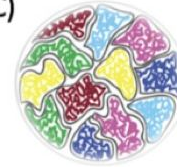
Purge dups



3 - asmstats,
BUSCO, merquy

Scaffolding

Yahs scaffolding
(Arima or Qiagen
HiC)



4 - asmstats,
BUSCO,
merquy, HiC
heatmap

Curated assembly

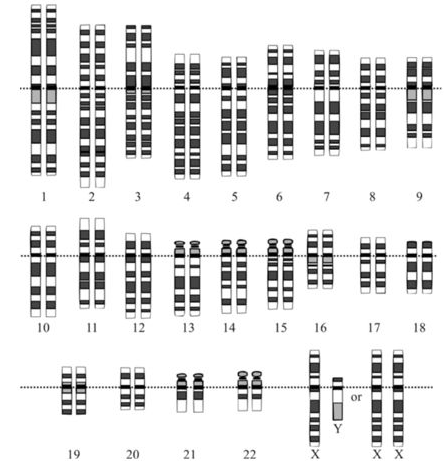
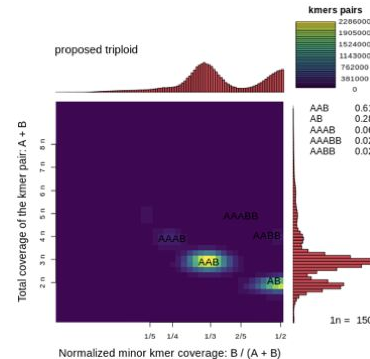
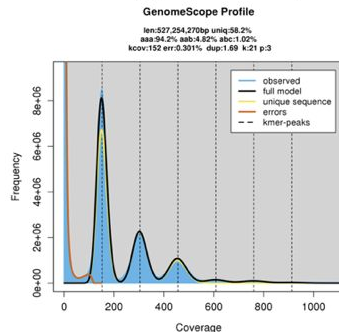
5 - asmstats,
BUSCO,
merquy, HiC
heatmap

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

Key considerations to start your genome assembly project

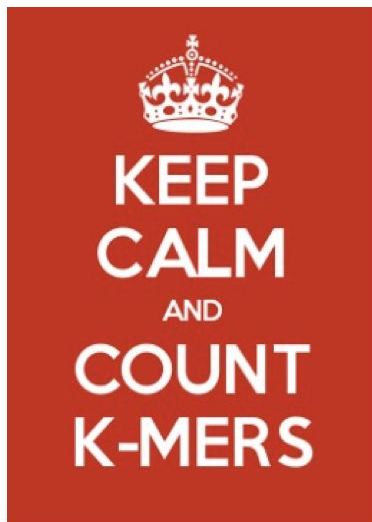


- Genome size (flow cytometry, Kmer analysis, GoaT) <https://goat.genomehubs.org/>
- Heterozygosity (kmer analyses: jellyfish, genomescope)
- Repetitive content (kmer analyses: jellyfish, genomescope)
- Ploidy (kmer analyses: smudgeplots)

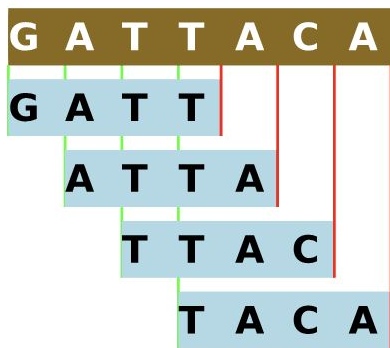


**I HAVE HIGH-QUALITY DATA
(ILLUMINA, PACBIO HIFI, DUPLEX
NANOPORE)**

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 Bruijn 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/>

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
...
```

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
```

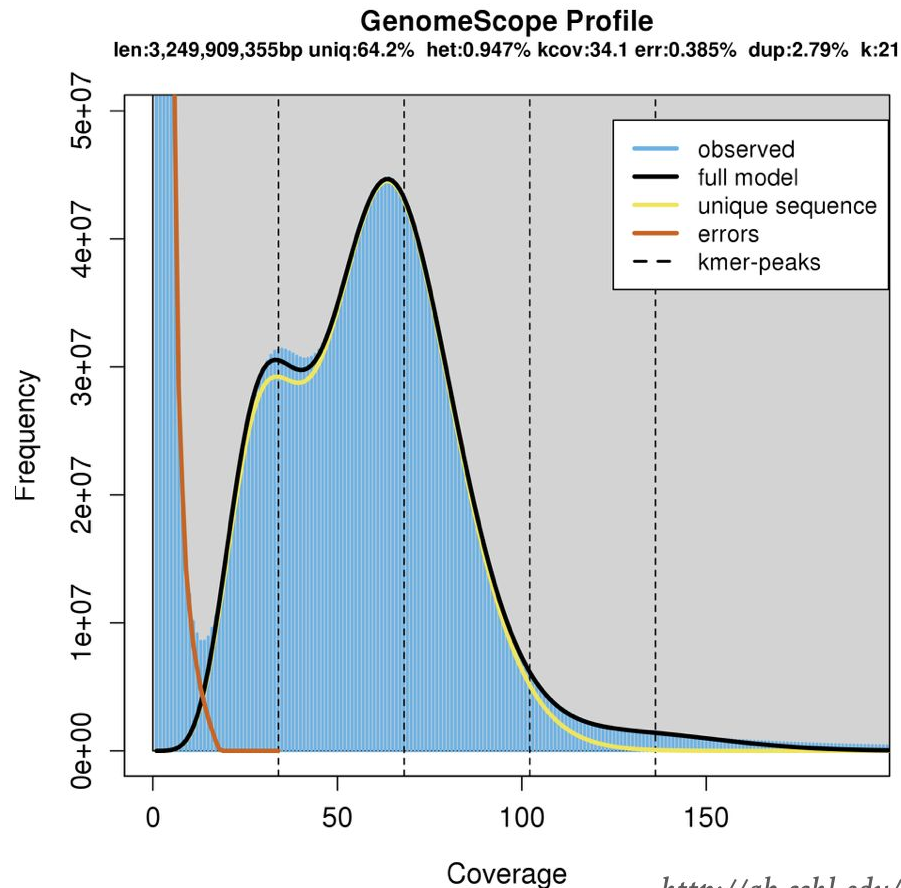
Now the histogram file looks like this:

```
# 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



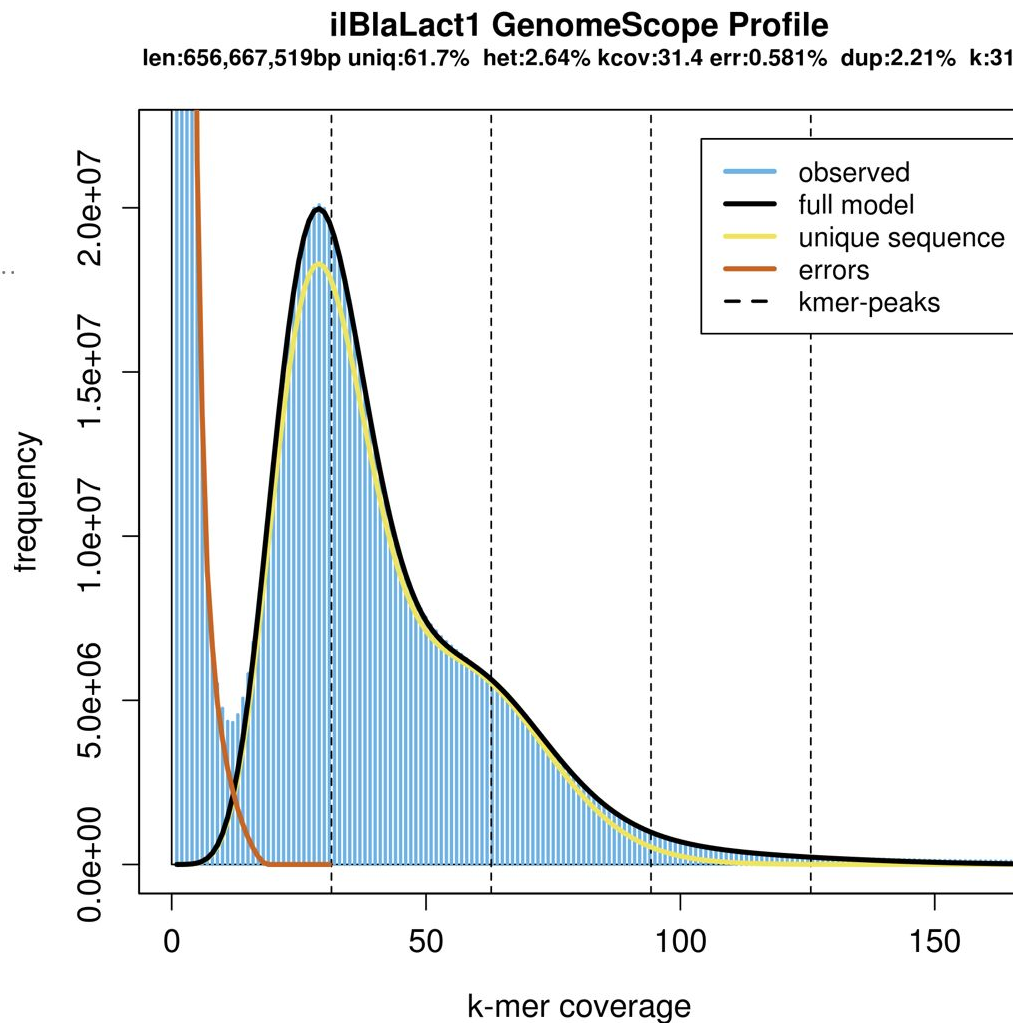
Choloepus didactylus (VGP)



A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH HIGH HETEROZYGOSITY

Blastobasis lacticolella (DToL)

Wakely's dowd



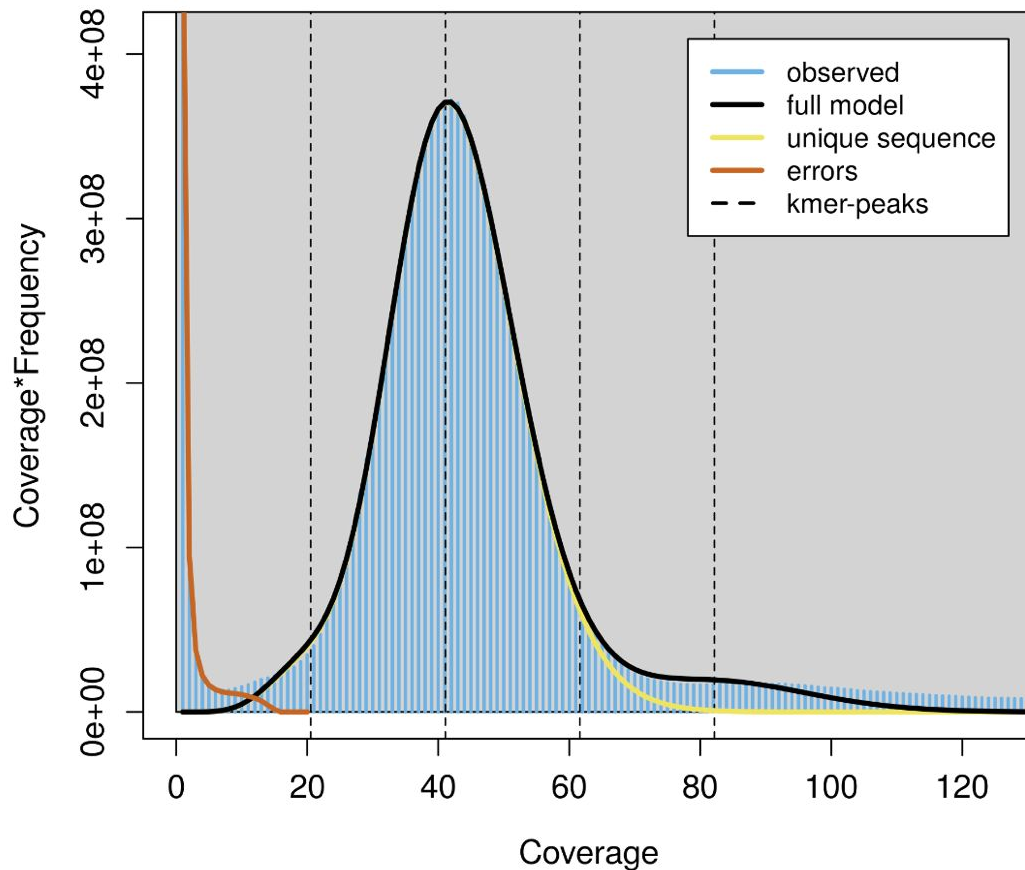
A TYPICAL KMER PLOT FOR A DIPLOID SPECIES WITH LOW HETEROZYGOSITY

Urtica urens



GenomeScope Profile

len:438,762,965bp uniq:51.8%
aa:99.8% ab:0.183%
kcov:20.5 err:0.135% dup:1.2 k:31 p:2



KMER SIZE

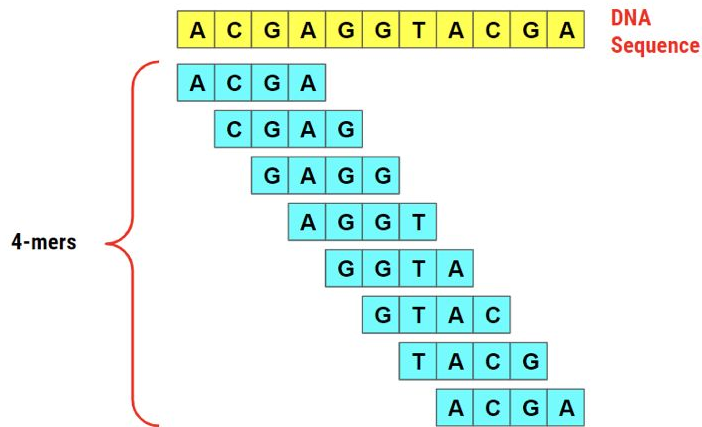
Choosing k : specificity vs. Sensitivity

- 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)

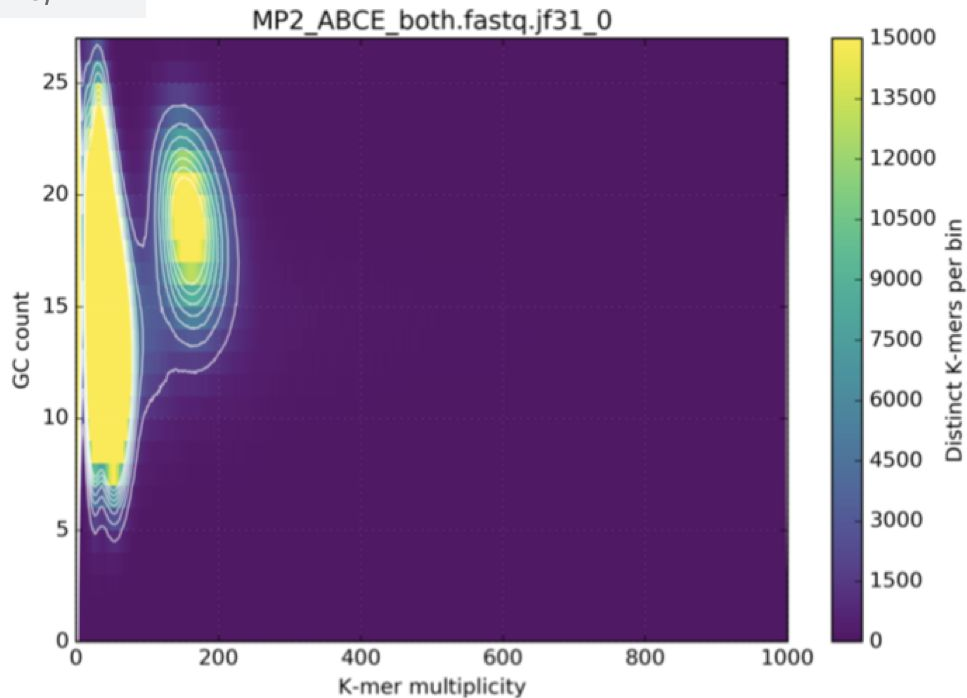
Why do we chose $k=31$ so often?

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.



SPOTTING BACTERIAL CONTAMINATION: KMER AND ITS GC CONTENT

github.com/TGAC/KAT



You can use KAT to plot this!

README.md



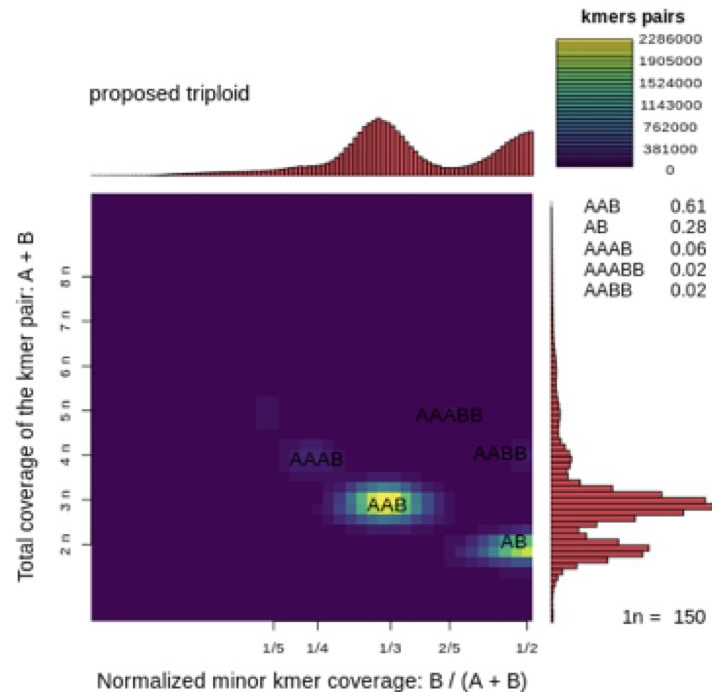
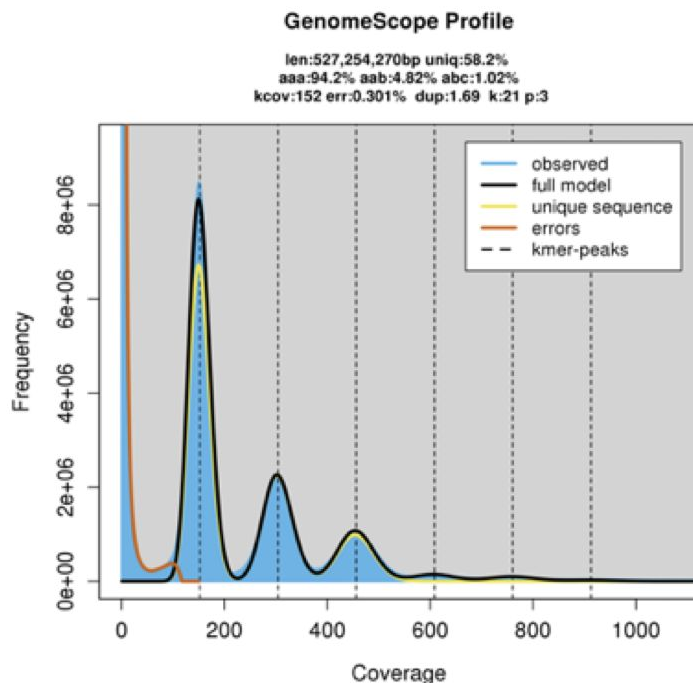
KAT - The K-mer Analysis Toolkit

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**: SSequence Coverage estimator Tool. Estimates the coverage of each sequence in a file using K-mers from another sequence file.

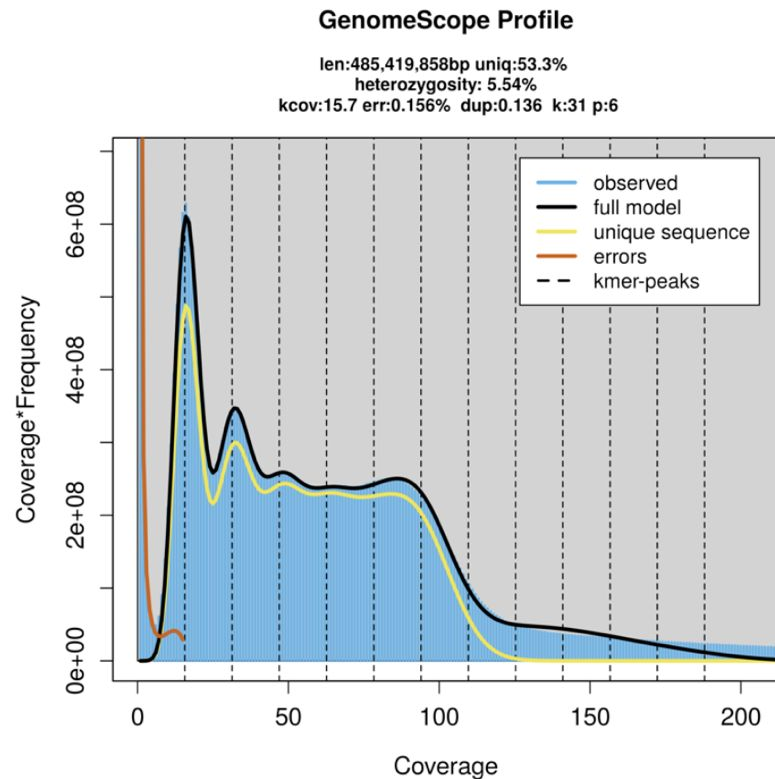
Tubastraea tagusensis

KMER PROFILE FOR A TRIPLOID SPECIES

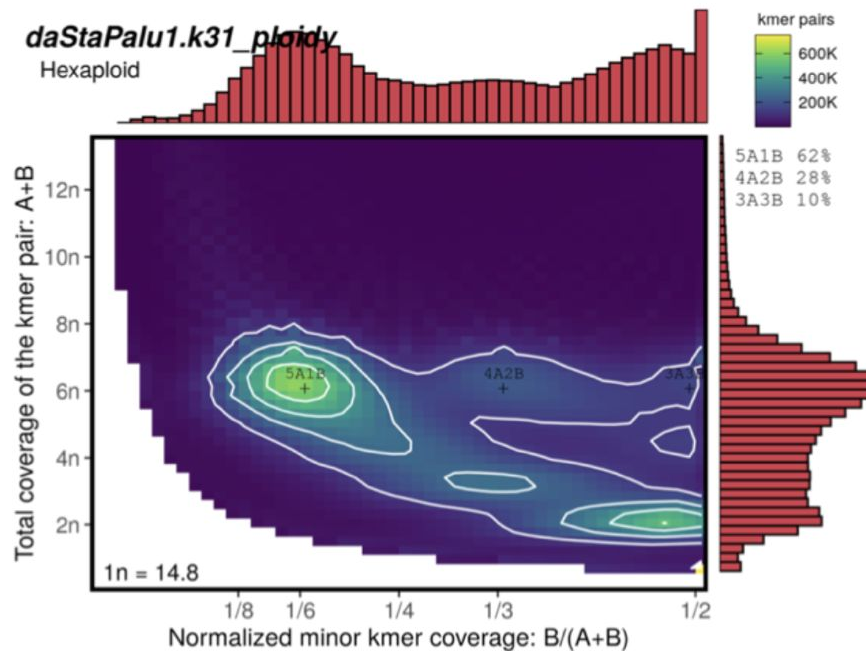


KHMER PROFILE FOR A POLYPLOID SPECIES

pacbio daStaPalu1 GenomeScope 2.0 linear plot



Ploidy stack plot daStaPalu1



MORE ON SMUDGEPLOTS



<https://www.youtube.com/watch?v=8vuNSvrAloA>

Mudanças climáticas

United Nations • As mudanças climáticas são transformações a longo prazo nos padrões de temperatura e clima. As atividades humanas têm sido o principal impulsionador das mudanças climáticas, principalmente devido à queima de combustíveis fósseis como carvão, petróleo e gás.

BGA24: Smudgeplot



The Biodiversity Genomics Academy

207 inscritos



Compartilhar



Download



Clipe



[Submitted on 1 Apr 2024]

Guide to k-mer approaches for genomics across the tree of life

<https://arxiv.org/abs/2404.01519>

Katharine M. Jenike, Lucía Campos-Domínguez, Marilou Boddé, José Cerca, Christina N. Hodson, Michael C. Schatz, Kamil S. Jaron

The wide array of currently available genomes display a wonderful diversity in size, composition and structure with many more to come thanks to several global biodiversity genomics initiatives starting in recent years. However, sequencing of genomes, even with all the recent advances, can still be challenging for both technical (e.g. small physical size, contaminated samples, or access to appropriate sequencing platforms) and biological reasons (e.g. germline restricted DNA, variable ploidy levels, sex chromosomes, or very large genomes). In recent years, k-mer-based techniques have become popular to overcome some of these challenges. They are based on the simple process of dividing the analysed sequences (e.g. raw reads or genomes) into a set of sub-sequences of length k , called k-mers. Despite this apparent simplicity, k-mer-based analysis allows for a rapid and intuitive assessment of complex sequencing datasets. Here, we provide the first comprehensive review to the theoretical properties and practical applications of k-mers in biodiversity genomics, serving as a reference manual for this powerful approach.

Comments: Main text is 25 pages, 4 figures; With supplement it is 44

 Subjects: **Genomics (q-bio.GN)**

 Cite as: [arXiv:2404.01519](https://arxiv.org/abs/2404.01519) [q-bio.GN]

 (or [arXiv:2404.01519v1](https://arxiv.org/abs/2404.01519v1) [q-bio.GN] for this version)

<https://doi.org/10.48550/arXiv.2404.01519>

Submission history

 From: Kamil Jaron [[view email](#)]

[v1] Mon, 1 Apr 2024 22:58:30 UTC (5,205 KB)

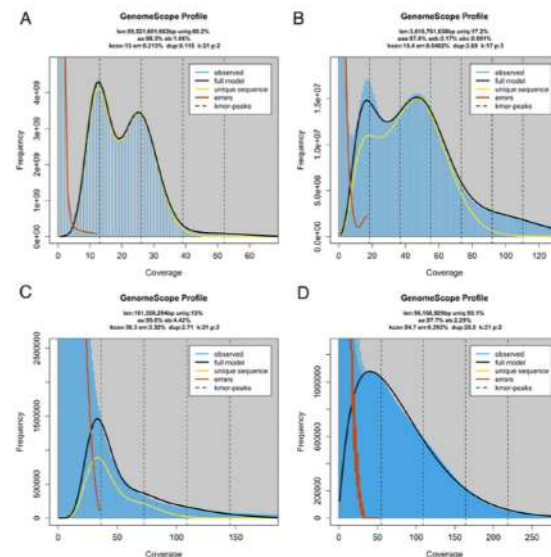
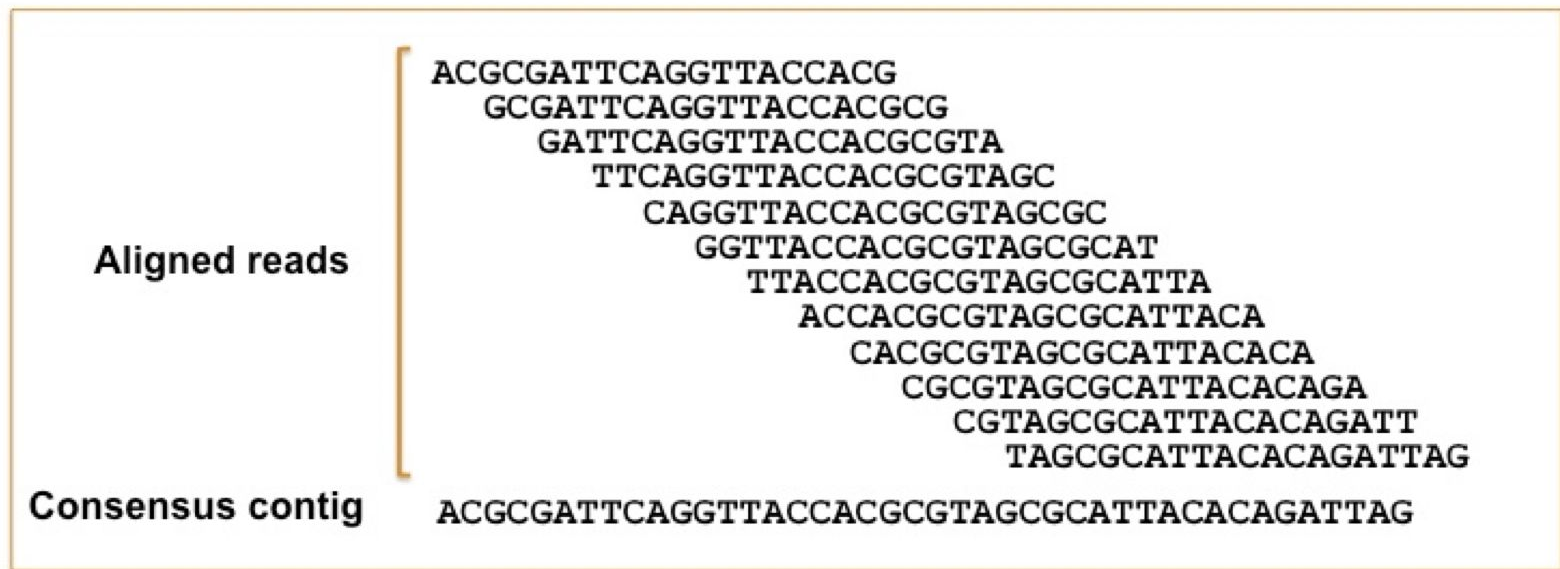


Figure 3: Examples of k-mer spectra. A. *Viscum album*: a diploid spectra with enough data to observe two distinct peaks and fit a model that accurately reflects genomic features despite the large size of the genome. B. *Procamburus virginalis*: k-mer spectra of a sample with low coverage, barely sufficient for a model fit. Notably, we used $k = 17$ to increase the k-mer coverage and make the model fit possible. C. *Allium schoenoprasum*: The sequencing coverage of this data set is approximately 1x, error k-mers and genome k-mers are

KNOWING THE CHALLENGE, YOU GO AND BUILD CONTIGS WITH ASSEMBLERS

CONTIG



Check point

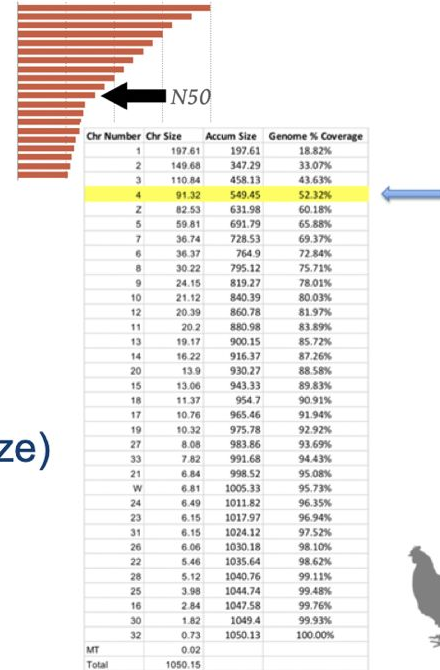
**DOES MY ASSEMBLED SIZE
CORRESPONDS WITH MY
ESTIMATED GENOME SIZE?**

Genomics is a game of going back and forth

Basic Assembly Metrics

- Total assembled sequence length
- Number of sequences (contigs and scaffolds)
- Average length (contigs and scaffolds)
- Largest/smallest (contigs and scaffolds)
- N50 = X means 50% of the genome is in sequences larger than X
- NG50 (N50 scaled by the expected genome size)
- Number of gaps

N50 = what is the smallest contig at 50% of genome?



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

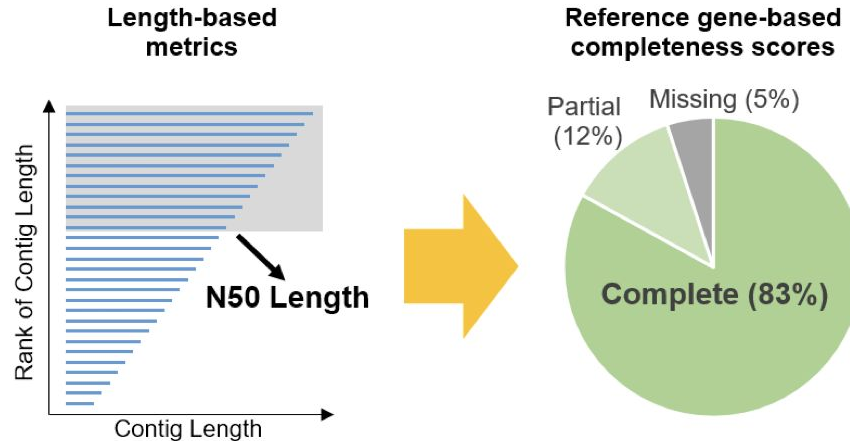




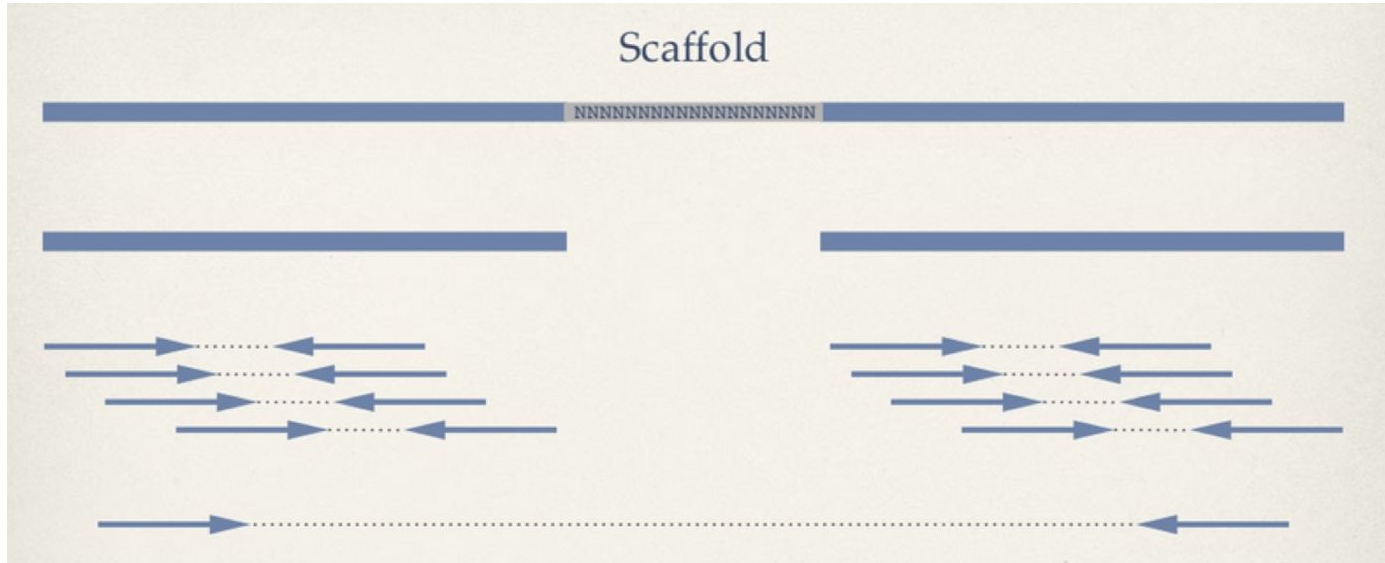
Assessing genome assembly and annotation completeness with
Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

- 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!



Scaffolding methods



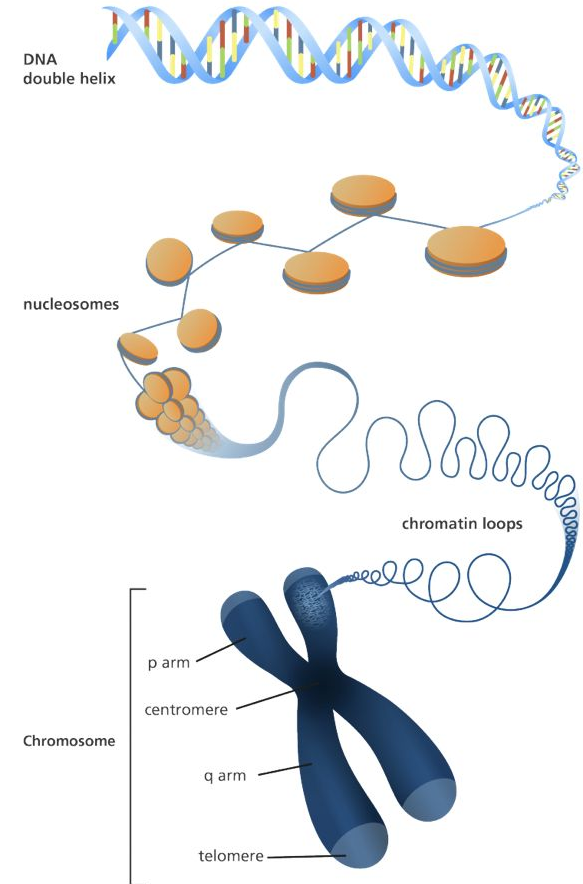
Scaffold: joining and orienting contigs

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

HOW DO I BUILD UP SCAFFOLDS AND CHROMOSOMES?

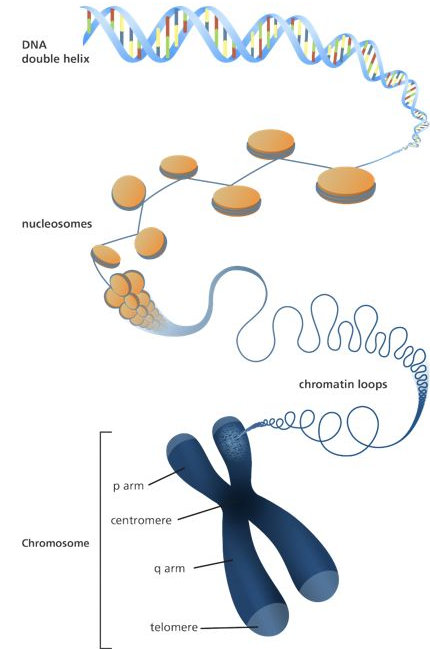
Hi-C and Ultralong Nanopore

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)



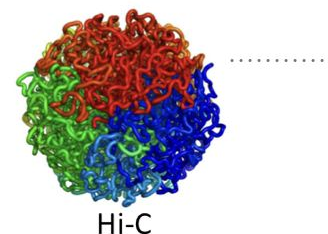
Chromosome conformation

- *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.



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 Williams¹, Maxim Imakaev², Tobias Ragoczy^{6,7}, Agnes Telling^{6,7}, Ido Amit¹, Bryan R. Lajoie⁵, Peter J. Sabo⁸, Michael O. Dorschner⁸, Richard Sandstrom⁸, Bradley Bernstein^{1,9}, M. A. Bender¹⁰, Mark Groudine^{6,7}, Andreas Gnirke¹, John Stamatoyannopoulos⁸, 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.

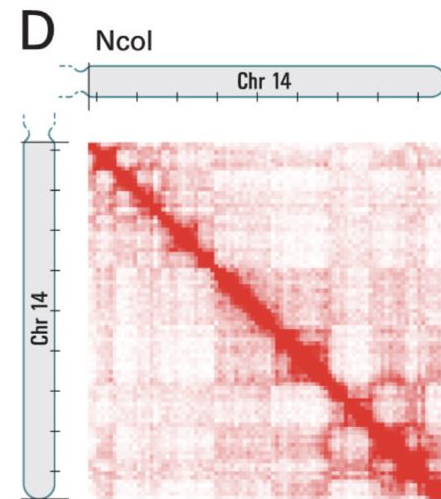
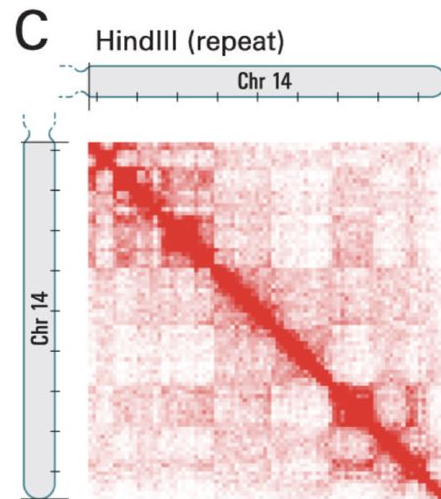
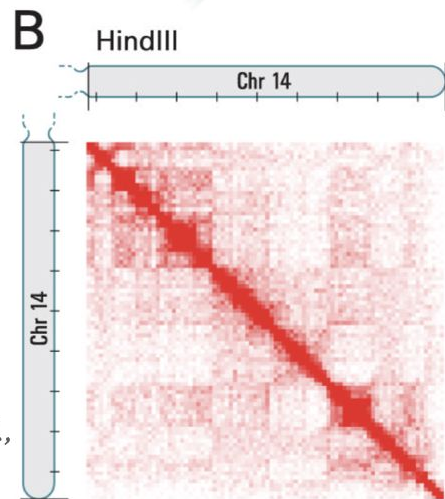
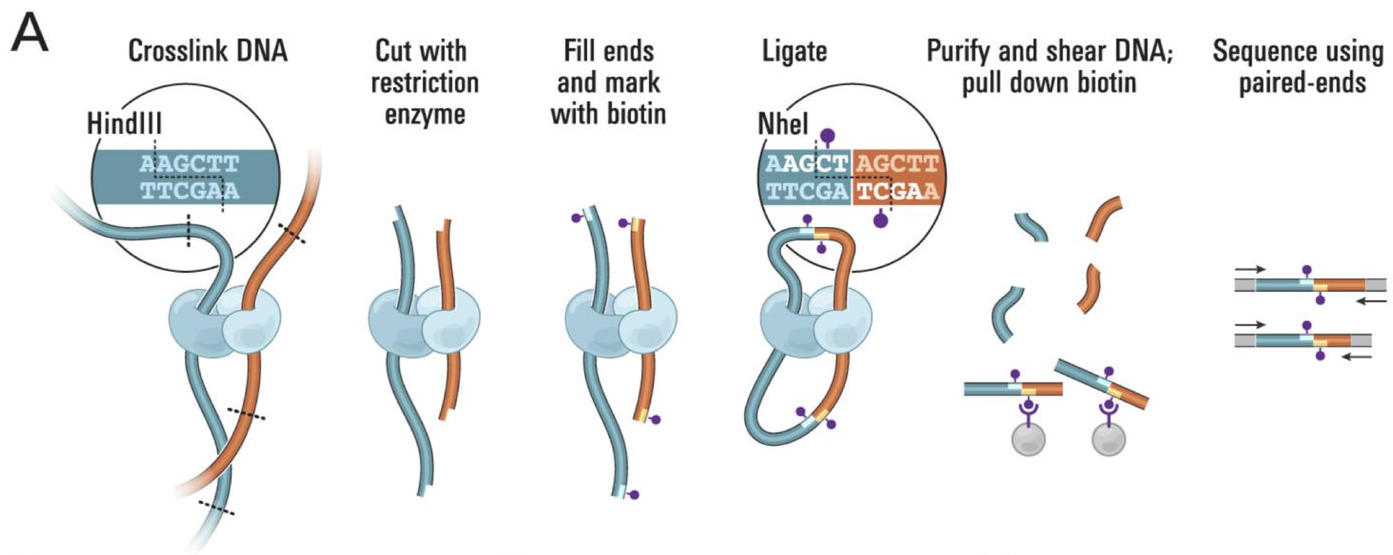
⁹ Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.

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.



Hi-C

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

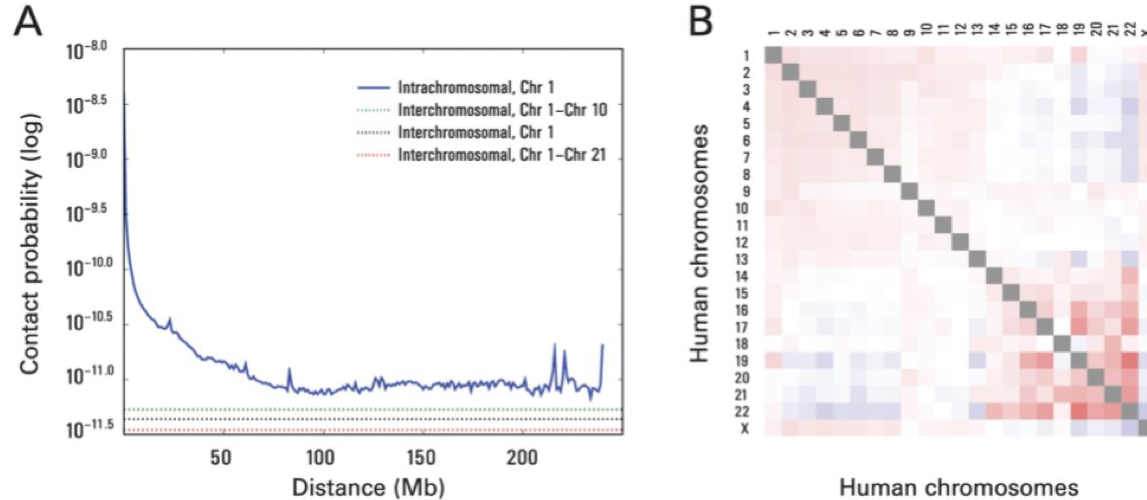
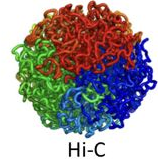
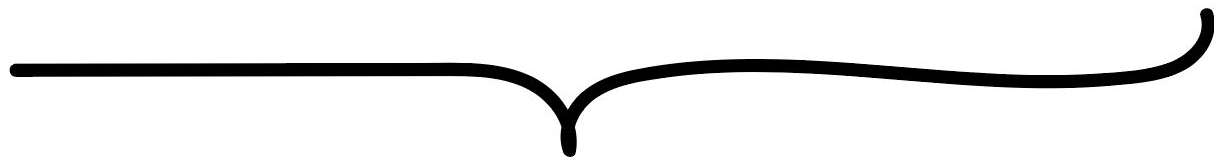
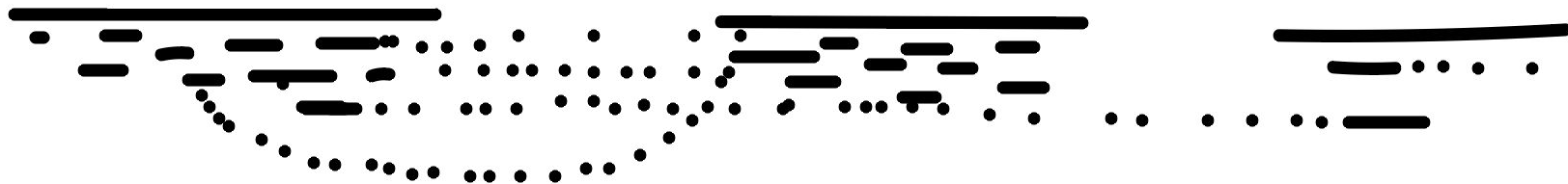


Fig. 2.

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 mem -5SP)
- Ran YaHS and/or Salsa for scaffolding
- Build and look at Hi-C HeatMaps



—NNN—



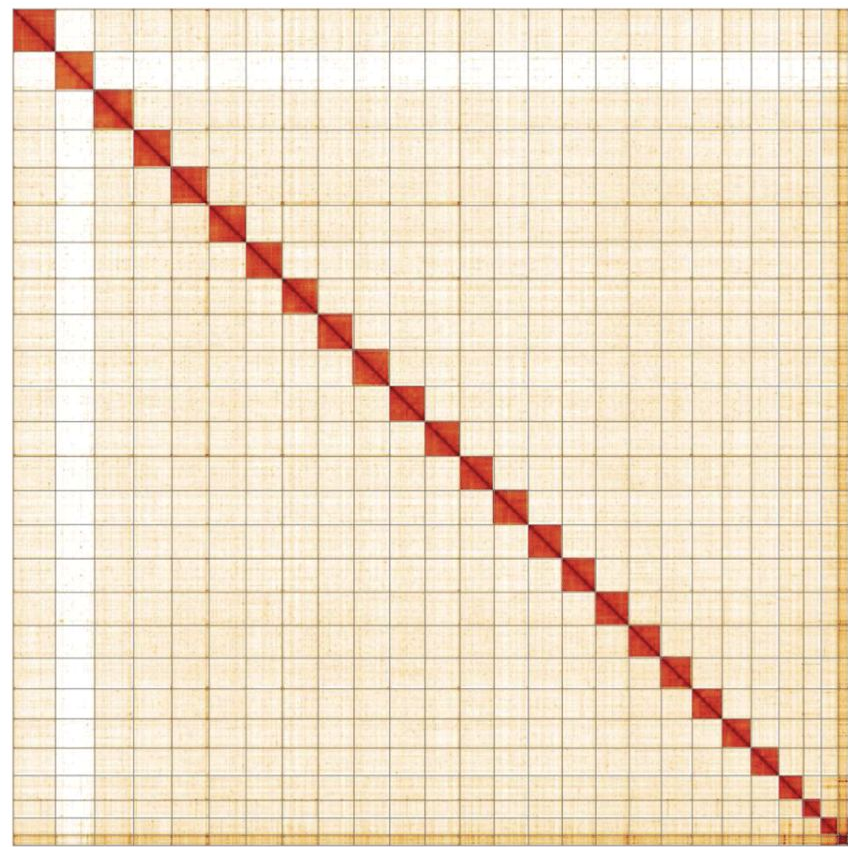
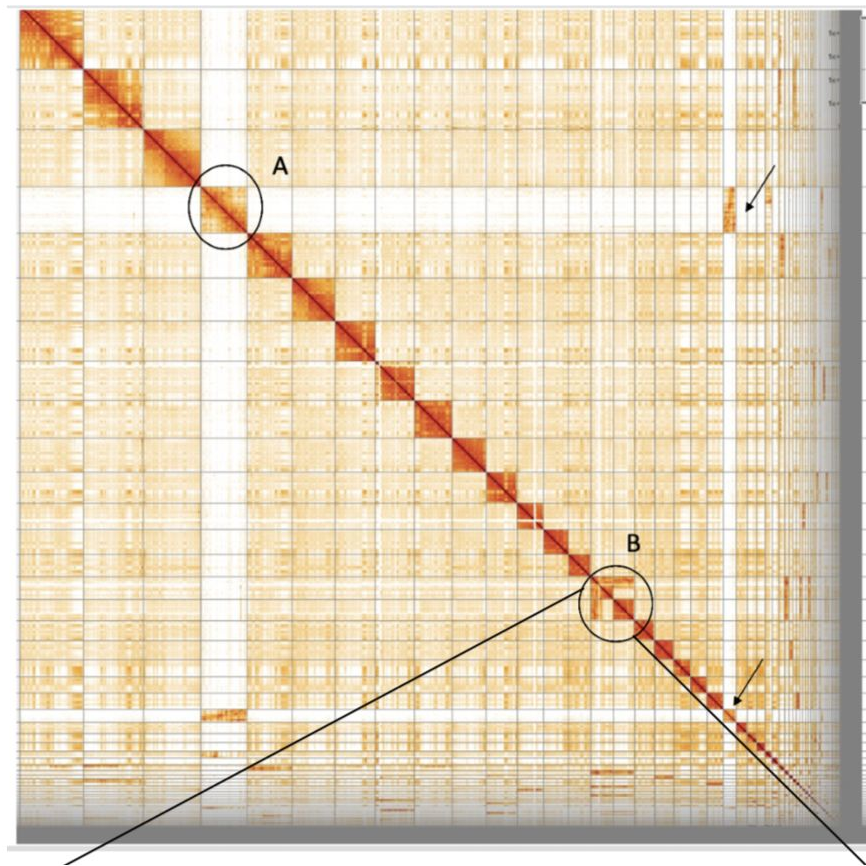


Figure 5. Genome assembly of *Pieris rapae*, ilPieRapa1.1: Hi-C contact map.

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

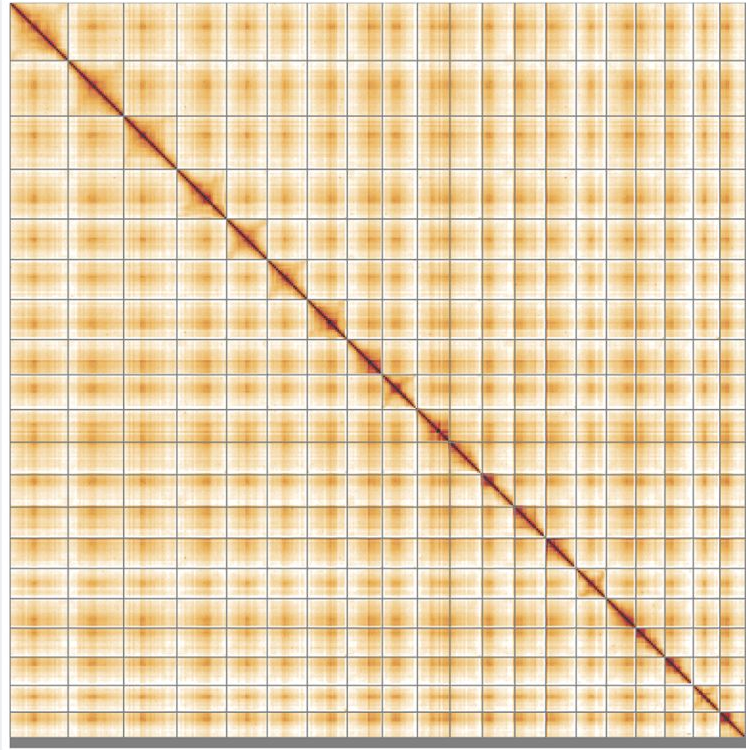


Figure 5. Genome assembly of *Ilex aquifolium*, drlleAqui2.1: Hi-C contact map of the drlleAqui2.1 assembly, visualised using HiGlass.

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

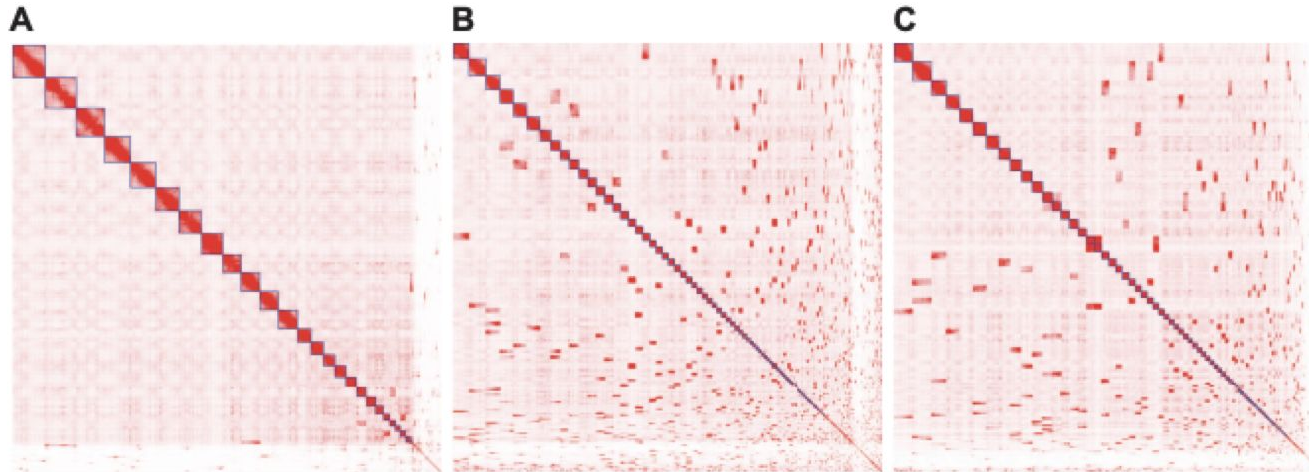


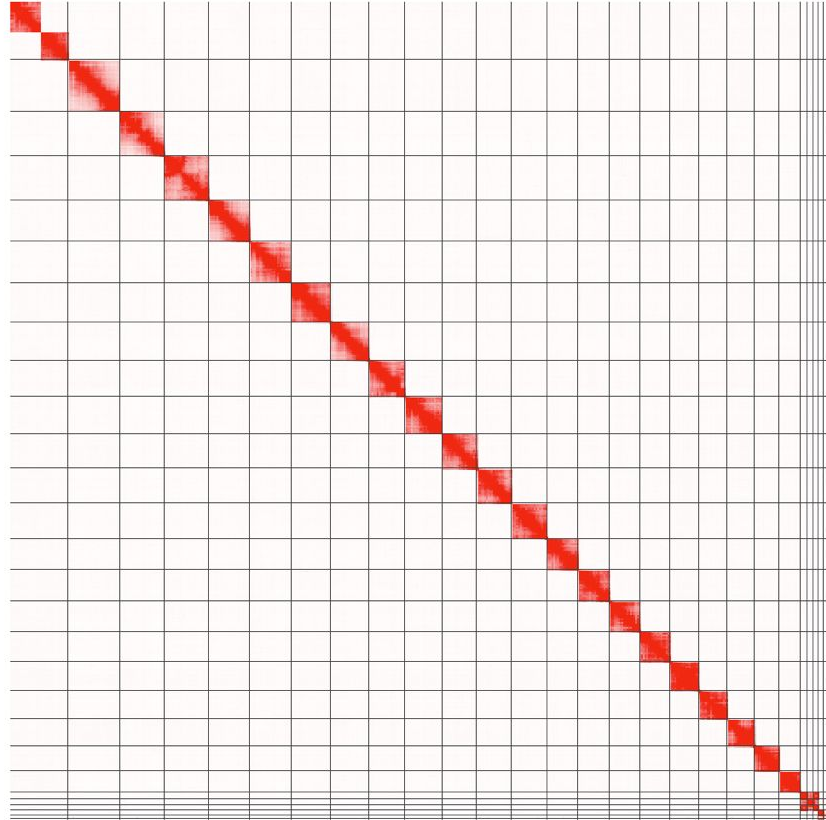
Figure 1. Hi-C contact maps of genome assemblies constructed with YaHS (A), SALSA2 (B) and pin-hic (C) for the simulated T2T data without contig errors. The blocks highlighted with blue squares in diagonal line are scaffolds. The contact maps were plotted with Juicebox (Durand *et al.*, 2016).

HI-C: DETECTING MISASSEMBLES

Look at me!!!!



Lycaena phlaeas - iLycPhla1

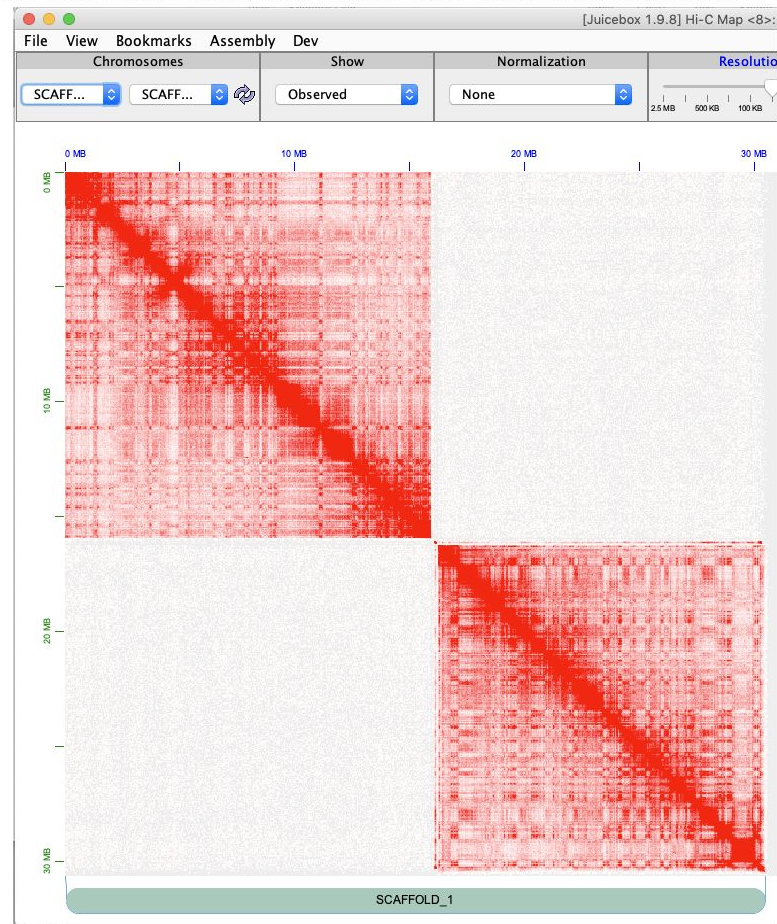


Darwin
TREE
of
LIFE

HI-C: DETECTING MISASSEMBLES



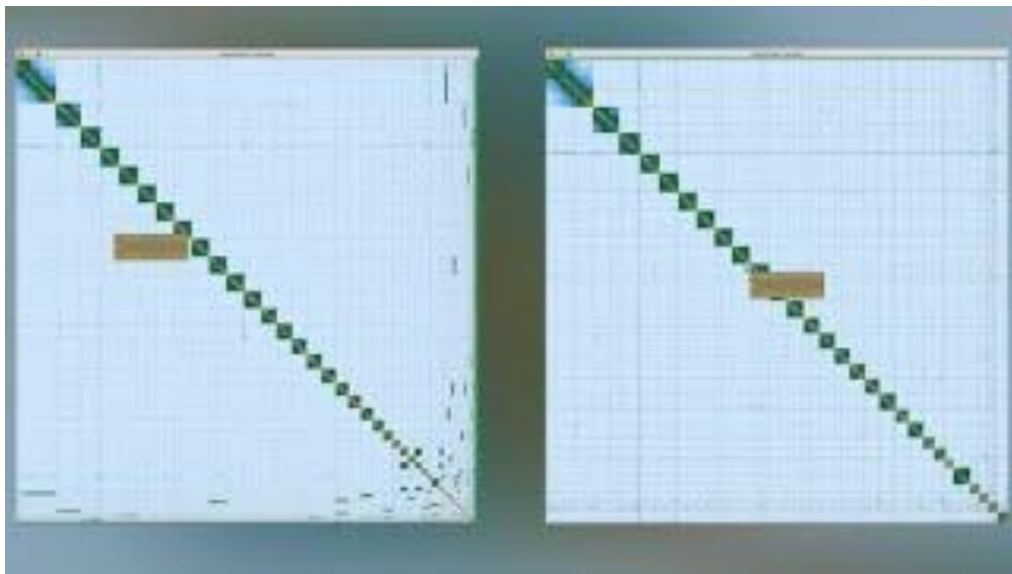
Lycaena phlaeas - ilLycPhla1



Resources for the Sanger Grit curation team

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

<https://gitlab.com/wtsi-grit/rapid-curation/-/blob/main/PretextView%20-%20Tutorial.pdf>



<https://www.youtube.com/watch?v=3lL2Q4f3k3I>

Phase 1 VGP Genomes: 1st data release of 15 genomes, 14 species

Mammals
(4 species)



GREATER HORSESHOE BAT



SPEAR-NOSED BAT



CANADIAN LYNX



PLATYPUS

Birds
(3 species)
4 genomes



ANNA'S HUMMINGBIRD



ZEBRA FINCH
(male) (female)



KAKAPO



Dedicated to Jane, the
Kakapo parrot

Reptiles
(1 species)



GOODE'S DESERT TORTOISE

Amphibians
(1 species)



TWO-LINED CAECILIAN

Fishes
(5 species)



FLIER CICHLID



EASTERN HAPPY



CLIMBING PERCH



TIRE TRACK EEL



BLUNT-SNOURED
CLINGFISH



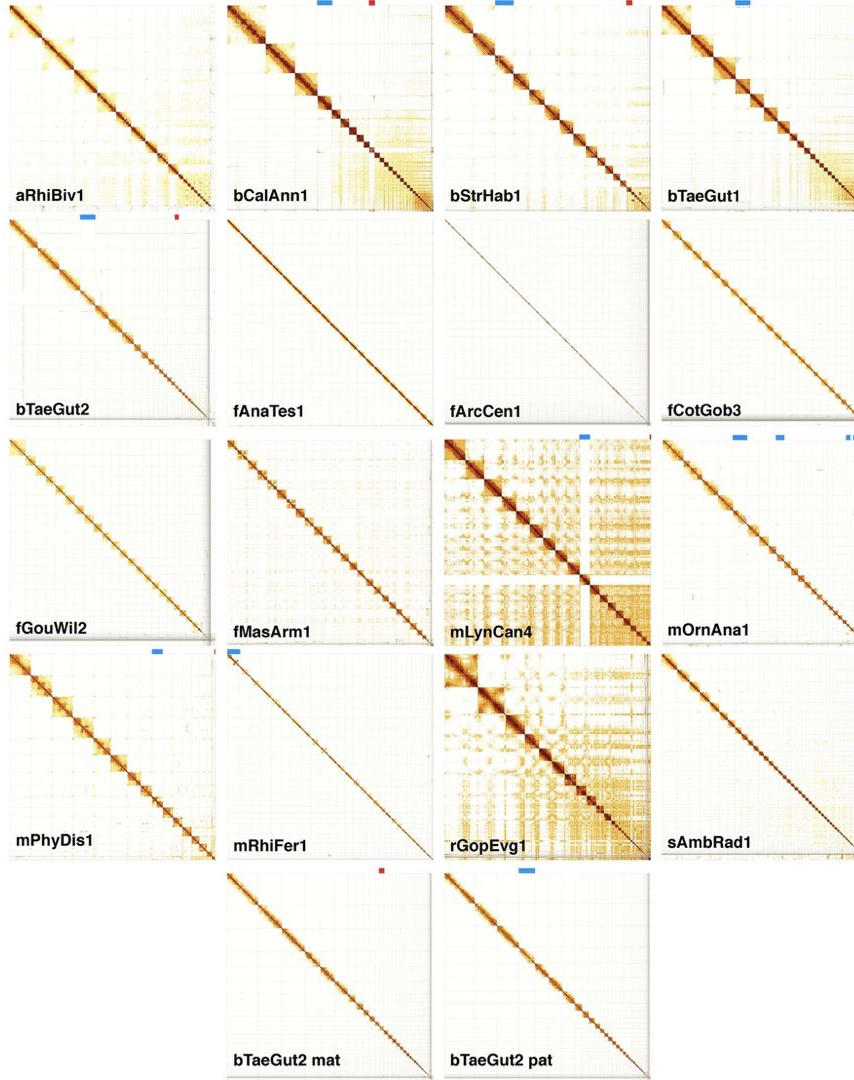
six major vertebrate classes, with a wide diversity of genomic characteristics.

Towards complete and error-free genome assemblies of all vertebrate species

Arang Rhie, [Shane A. McCarthy](#), [Olivier Fedrigo](#), [Joana Damas](#), [Giulio Formenti](#), [Sergey Koren](#), [Marcela Uliano-Silva](#), [William Chow](#), [Arkarachai Fungtammasan](#), [Juwan Kim](#), [Chul Lee](#), [Byung June Ko](#), [Mark Chaisson](#), [Gregory L. Gedman](#), [Lindsey J. Cantin](#), [Francoise Thibaud-Nissen](#), [Leanne Haggerty](#), [Iliana Bista](#), [Michelle Smith](#), [Bettina Haase](#), [Jacquelyn Mountcastle](#), [Syke Winkler](#), [Sadye Paez](#), [Jason Howard](#), [Sonja C. Vernes](#), [Tanya M. Lama](#), [Frank Grutzner](#), [Wesley C. Warren](#), [Christopher N. Balakrishnan](#), [Dave Burt](#), [Julia M. George](#), [Matthew T. Biegler](#), [David Iorns](#), [Andrew Digby](#), [Daryl Eason](#), [Bruce Robertson](#), [Taylor Edwards](#), [Mark Wilkinson](#), [George Turner](#), [Axel Meyer](#), [Andreas F. Kautt](#), [Paolo Franchini](#), [H. William Detrich III](#), [Hannes Svardal](#), [Maximilian Wagner](#), [Gavin J. P. Naylor](#), [Martin Pippel](#), [Milan Malinsky](#), [Mark Mooney](#), [Maria Simbirsky](#), [Brett T. Hannigan](#), [Trevor Pesout](#), [Marlys Houck](#), [Ann Misuraca](#), [Sarah B. Kingan](#), [Richard Hall](#), [Zev Kronenberg](#), [Ivan Sović](#), [Christopher Dunn](#), [Zemin Ning](#), [Alex Hastie](#), [Joyce Lee](#), [Siddarth Selvaraj](#), [Richard E. Green](#), [Nicholas H. Putnam](#), [Ivo Gut](#), [Jay Ghurye](#), [Erik Garrison](#), [Ying Sims](#), [Joanna Collins](#), [Sarah Pelan](#), [James Torrance](#), [Alan Tracey](#), [Jonathan Wood](#), [Robel E. Dagnew](#), [Dengfeng Guan](#), [Sarah E. London](#), [David F. Clayton](#), [Claudio V. Mello](#), [Samantha R. Friedrich](#), [Peter V. Lovell](#), [Ekaterina Osipova](#), [Farooq O. Al-Ajli](#), [Simona Secomandi](#), [Heeбал Kim](#), [Constantina Theofanopoulou](#), [Michael Hiller](#), [Yang Zhou](#), [Robert S. Harris](#), [Kateryna D. Makova](#), [Paul Medvedev](#), [Jinna Hoffman](#), [Patrick Masterson](#), [Karen Clark](#), [Fergal Martin](#), [Kevin Howe](#), [Paul Flicek](#), [Brian P. Walenz](#), [Woori Kwak](#), [Hiram Clawson](#), [Mark Diekhans](#), [Luis Nassar](#), [Benedict Paten](#), [Robert H. S. Kraus](#), [Andrew J. Crawford](#), [M. Thomas P. Gilbert](#), [Guojie Zhang](#), [Byrappa Venkatesh](#), [Robert W. Murphy](#), [Klaus-Peter Koepfli](#), [Beth Shapiro](#), [Warren E. Johnson](#), [Federica Di Palma](#), [Tomas Marques-Bonet](#), [Emma C. Teeling](#), [Tandy Warnow](#), [Jennifer Marshall Graves](#), [Oliver A. Ryder](#), [David Haussler](#), [Stephen J. O'Brien](#), [Jonas Korlach](#), [Harris A. Lewin](#), [Kerstin Howe](#) ✉, [Eugene W. Myers](#) ✉, [Richard Durbin](#) ✉, [Adam M. Phillippy](#) ✉ & [Erich D. Jarvis](#) ✉ [-Show fewer authors](#)

Nature **592**, 737–746 (2021) | [Cite this article](#)

72k Accesses | **52** Citations | **546** Altmetric | [Metrics](#)



I have my assembly, how do I know its correct?

Final checks:

- Does final assembled size corresponds to predicted genome size?
- How are my general metrics? How is my BUSCO?
- How does my Hi-C heatmap looks like? Clean? Correct karyotype?
Any scaffolding mistakes?
- Reads coverage and **Merqury (important!)**

Minimal supplementary materials for your genome paper: (i) kmer plot of your data (genomescope plot), (ii) general assembly stats, (iii) merqury plots, (iv) busco and (v) HiC heatmap.

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. McCarthy², Jonathan Wood³, Kerstin Howe³,
Yadong Wang^{1,*} and Richard Durbin^{2,3,*}

¹Department of Computer Science and Technology, Center for Bioinformatics, Harbin Institute of Technology, Harbin 150001, China,

²Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK and ³Wellcome 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

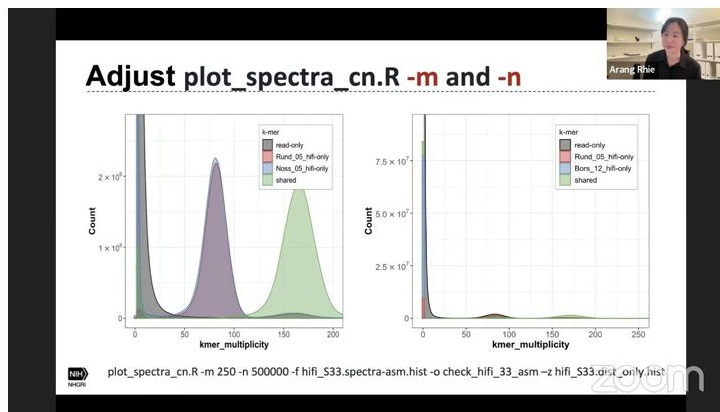


* Correspondence: arang.rhie@nih.gov
Genome Informatics Section,
Computational and Statistical
Genomics Branch, National Human
Genome Research Institute, National
Institutes of Health, Bethesda, MD,
USA

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



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](#)

4164 Accesses | **8** Citations | **14** Altmetric | [Metrics](#)

“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.

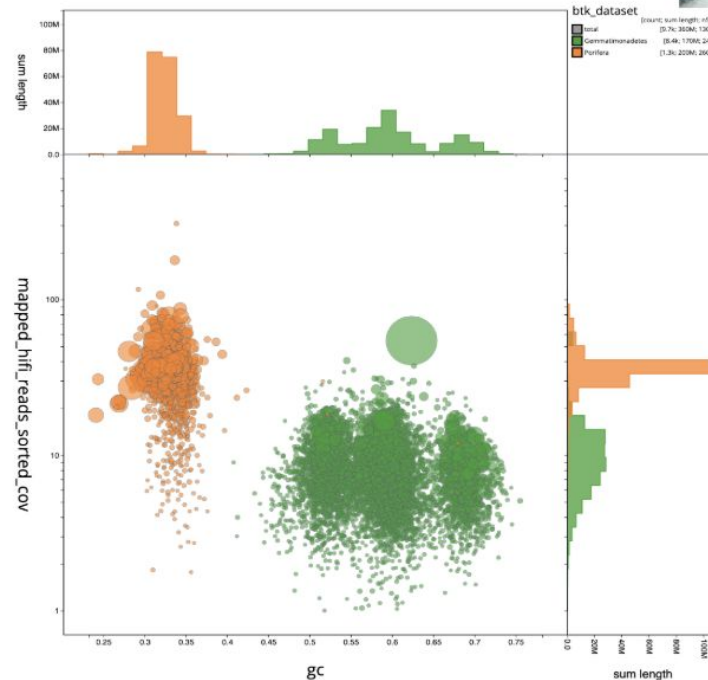
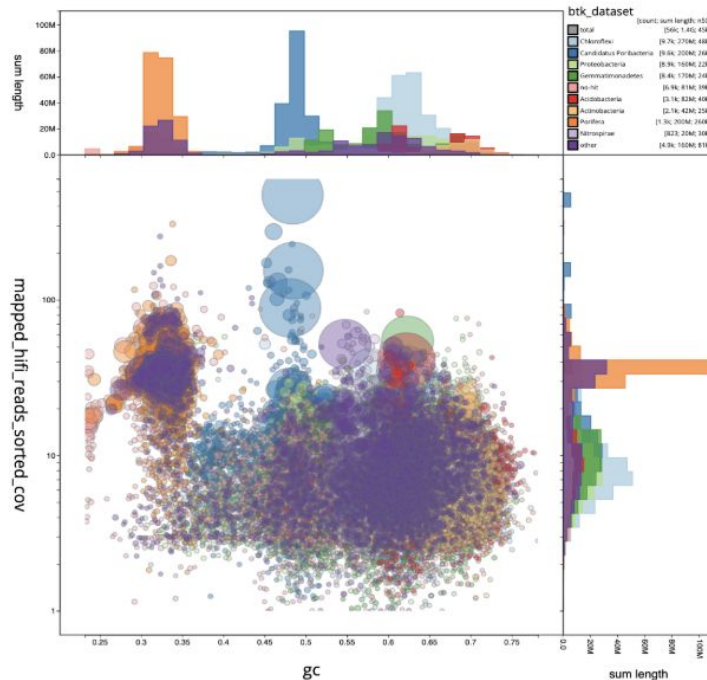
The main source of the false duplications is heterotype duplications, where the haplotype sequences were relatively more divergent than other parts of the genome leading the assembly algorithms to classify them as separate genes or genomic regions.” Kim et al, 2022

More than the target species...



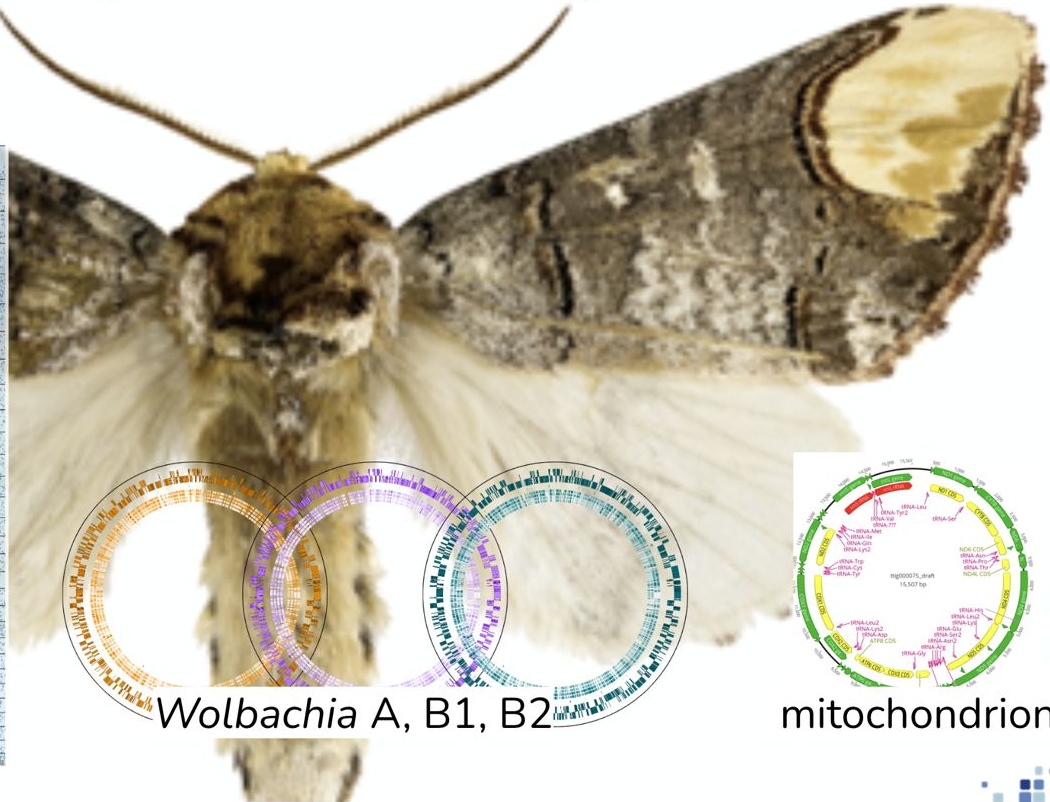
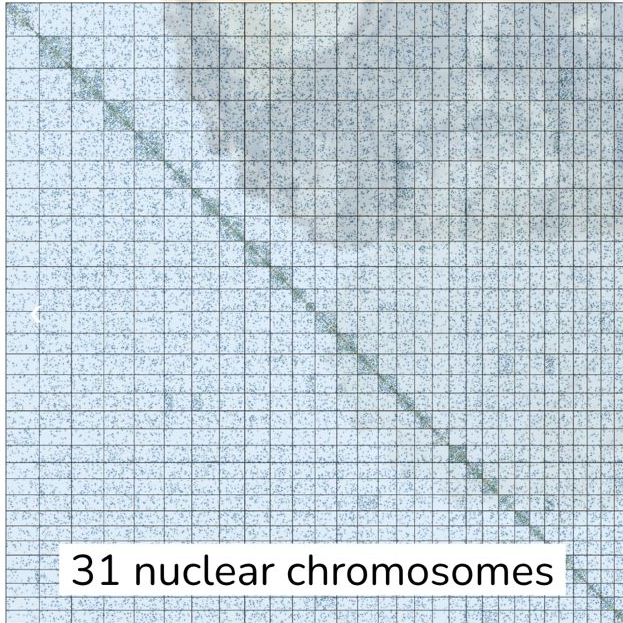
BlobToolKit analysis of *Xestospongia muta* (giant barrel sponge) odXesMuta1

github.com/blobtoolkit



Assembling genomes of target *and* microbiome

Phalera bucephala: one moth, five genomes



Open access to all data



protocols.io	Wet lab protocols	www.protocols.io/workspaces/wellcome-sanger-institute13
Darwin Tree of Life Data Portal Aquatic Symbiosis Project Data Portal	Project portals	portal.darwintreeoflife.org portal.aquaticsymbiosisgenomics.org
Tree of Life QC	Raw data & assembly progress	tolqc.cog.sanger.ac.uk
Ensembl	Ensembl genome annotation & browser	projects.ensembl.org/darwin-tree-of-life
BlobToolKit	Interactive genome viewer	blobtoolkit.genomehubs.org
Wellcome Open Research	Genome Notes	wellcomeopenresearch.org/treeoflife
Goat	Global coordination (GoaT)	goat.genomehubs.org
Genome After Party	Standard post-genome analyses	gap.cog.sanger.ac.uk
sanger-tol Sanger-TOL	Informatics pipelines and tools	pipelines.tol.sanger.ac.uk

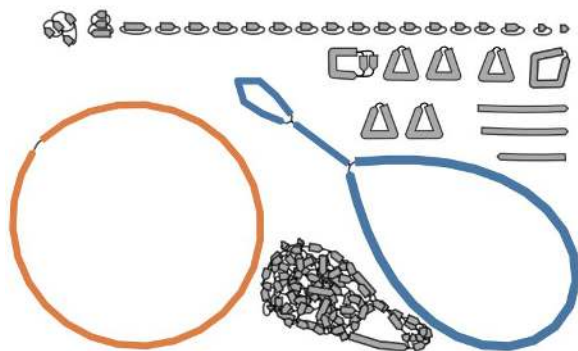
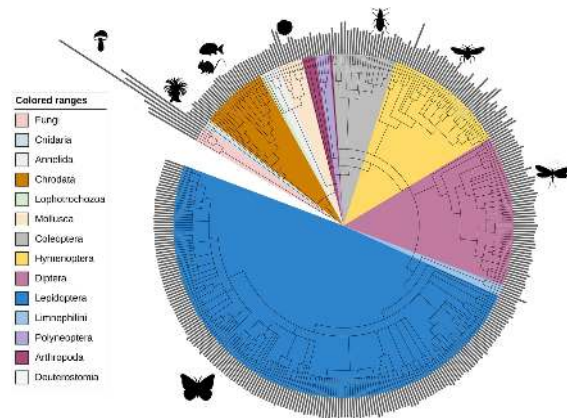
Mito and chloroplast assembly with Long Reads

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](#)



Oatk: a de novo assembly tool for complex plant organelle genomes

Chenxi Zhou^{1,2}, Max Brown^{2,3}, Mark Blaxter², The Darwin Tree of Life Project Consortium², Shane A. McCarthy^{1,2}, and Richard Durbin^{1,2,*}

¹Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK

²Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

³Faculty of Science and Engineering, Anglia Ruskin University, East Road, Cambridge, CB1 1PT, UK

*Correspondence: rd109@cam.ac.uk

Studying further: Biodiversity Genomics Academy



bga24



+ Criar



October 1-23
SAVE THE DATE

Welcome to BGA24's session on:

De novo assembly with Colora
Lia Obinu

Lia Obinu: Hello, everybody, and welcome to the workshop about Colorado. And I'm going to share my screen in a minute. But

De novo assembly with Colora (BGA24)



Biodiversity Genomics Acade...

243 inscritos



Inscrito



11



Compartilhar



Todos

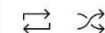
De Biodiversity Genomics Aca...

Inform



BGA24

Biodiversity Genomics Academy & Conference - 24 / 24



19

56:20

(BGA24)

Biodiversity Genomics Academy & ...

20



T2T assemblies with Verko
(BGA24)

Biodiversity Genomics Academy & ...

21



Inkscape: A crash course
(BGA24)

Biodiversity Genomics Academy & ...

22



Annotating genomes the Ensembl
way (BGA24)

Biodiversity Genomics Academy & ...

23



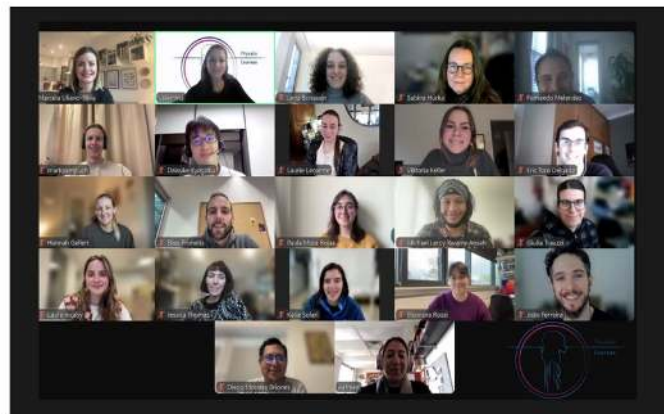
The TreeVal Pipeline (BGA24)

Biodiversity Genomics Academy & ...



De novo assembly with Colora
(BGA24)

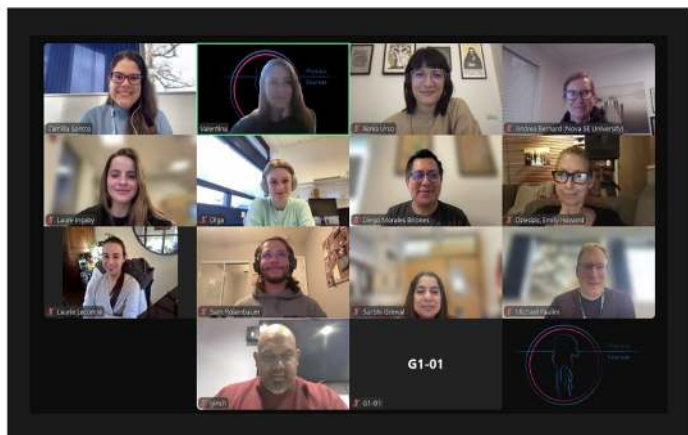
Biodiversity Genomics Academy & ...



5TH EDITION GENOME ASSEMBLY USING PACBIO AND HI-C

ONLINE, 4-8 NOVEMBER 2024

20 attendees
16 research institutes
14 different countries



GENOME MANUAL CURATION

ONLINE, 11-15 NOVEMBER 2024

11 attendees
9 research institutes
8 different countries



It takes a village...



Tree of Life teams - Mark Blaxter

ToL Sample Management, ToL Core Lab, ToL Assembly, GRIT, Delivery & Operations, ToL faculty teams

Anna Kovalevskaia
Priyanka Sethu Raman Thomas Mathers
Sarah Pelan Manuela Kieninger Julia Gries
Dominic Absolon Zeynep Goktan Barnaby Dingemans
Witold Morek Cibeles Sotero-Caio Jessie Jay Meyer Marco
Maneno Baravuga Ksenia Krasheninnikova Abitha Thomas
Jo Wood Downie Jim Eerik Aunin Remi Clare Eva van der Heijden
Adam Bates Alex Makunin Haoyu Niu Noah Gettle Anushka Mittal
Martin Wagah Kerstin Howe Luke Wilson Katie Woodcock Nicol Rueda
Halyna Yatsenko Rebecca O'Brien Witwicka Alicja Seri Kitada Victoria McKenna
Molly Carter Elizabeth Sinclair Guoying Qi Karen Brooks
Roz Malik Edward Moulds Lyndall Pereira da Conceicao Karin Näsval
Beth Yates Fiona Teltscher Sunil Dogga Camilla Santos Amy Denton
Cibin Sadasivan Baby Andrew Varley Ian Still Jemma Salmon Joana Meier
Mark Blaxter James Torrance Logan Howat Radka Platte
Claudia Weber Clothilde Chenal Francis Totanes Manuel Batista Erna King
Chafin Tyler Marilou Boddé Matthieu Muffato
Iszy Clayton-Lucey Nathan Riley Ying Sims Damon-Lee Pointon
Graeme Oatley Kiernan Harding Amjad Khalaf Ashish Mittal
Nancy Holroyd Mara Lawniczak Shane McCarthy
Ore Francis Paul Davis James Gilbert Petra Korlević Caroline Howard
Wiesia Johnson Jesse Rop Will Eagles Sam Ebdon
Emmellen Vancaester Joachim Nwezeobi Sinead Calnan
Yan Liang Charlie Hathaway Priyanka Surana Jessica Thomas
Michael Paulini Lora Downes Camilla Muyo Karen Houliston
Kamil Jaron Lewis Stevens Richard Challis Ben Jackson
Arif Maulana Marcela Uliano-Silva Charlotte Wright
Aleksandra Bliznina Martha Mulongo Raquel Vionette Do Amaral
Joanna Collins

Sanger Core Facilities

SciOps, especially the Long Read Team and R&D

Darwin collaborators

RBGE, Kew, NHM, MBA, Oxford, Cambridge, Edinburgh, Earlham, EBI
& hundreds of engaged naturalists

ASG, VGP, ERGA

and other collaborators and collectors



Obrigada! Thank you! Grazie!

mu2@sanger.ac.uk