Large genome assembly

Rayan Chikhi, CNRS, Univ Lille

BiG talk, Lund University
5th December 2018

Personal background

CNRS junior researcher in bioinformatics, France

Computer science background

@RayanChikhi on Twitter

http://rayan.chikhi.name

Software and methods for de novo assembly & k-mers

Minia

Kmergenie

DSK

BCALM

Collaborations with biology groups

Metagenome assembly

Condition-specific variants detection: RNA-Seq, Alzheimer's

Workshop on Genomics instructor

Large ~~(eukaryote)~~ genome assembly

Rayan Chikhi, CNRS, Univ Lille

BiG talk, Lund University
5th December 2018

Contents

1. Genome assembly short intro
2. Software methods
3. Giraffe, spruce, & axolotl genomes
4. Present & future sequencing technologies

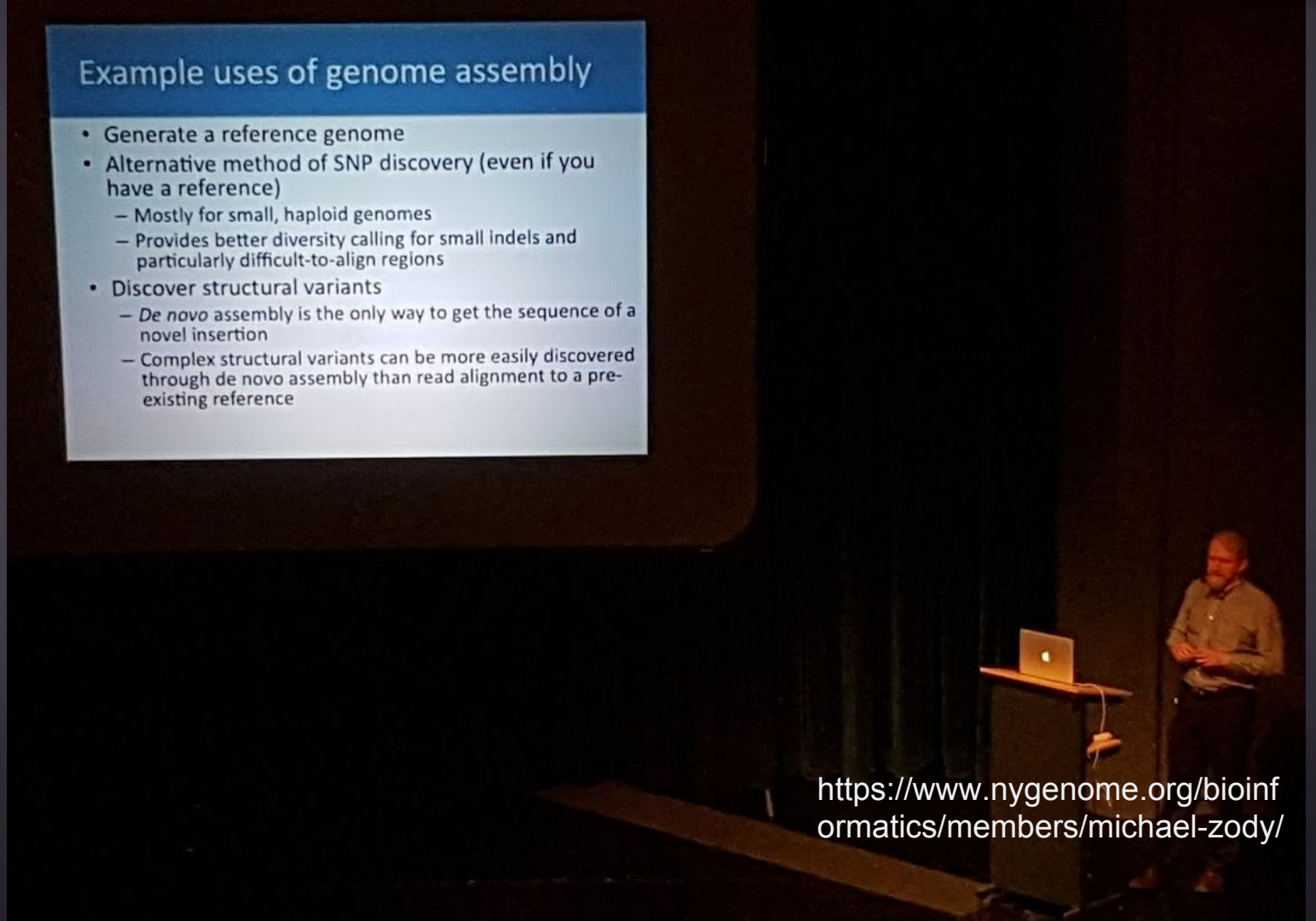
WHAT'S ASSEMBLY?

genome
not known

reads
overlapping substrings that cover the genome redundantly

assembly
what we think the genome is

"A set of sequences which best approximate the original sequenced material."


Example uses of genome assembly
• Generate a reference genome
• Alternative method of SNP discovery (even if you have a reference)
– Mostly for small, haploid genomes
– Provides better diversity calling for small indels and particularly difficult-to-align regions
• Discover structural variants
– *De novo* assembly is the only way to get the sequence of a novel insertion
– Complex structural variants can be more easily discovered through de novo assembly than read alignment to a pre-existing reference
https://www.nygenome.org/bioinformatics/members/michael-zody/

Genome assembly pre-history (from 1994 to 2001/2005)

Sanger-era sequencing

Greedy algorithms

What happened in 2001 and 2005?

An Eulerian path approach to DNA fragment assembly.
https://www.ncbi.nlm.nih.gov/pubmed/11504945
de PA Pevzner - 2001 - Cité 1211 fois

Today's paradigm for short reads assembly

The fragment assembly string graph.
https://www.ncbi.nlm.nih.gov/pubmed/16204131
de EW Myers - 2005 - Cité 318 fois

Today's paradigm for long reads assembly

Current assembly pipelines

with <u>short reads:</u>

De Bruijn graph,

Removal of errors, variants,

Repeat-resolution,

...

areas I've worked on

with <u>long reads:</u>

Error-correction,

Overlap graph,

...

area I'm working on

Assembly

Hybrid assembly

Assembly

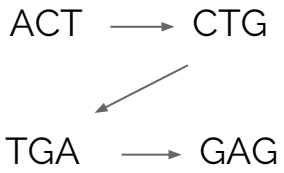
Short-read assembly

Short-read assembly outline

0) start from raw reads

ACTG
CTGA
TGAG

1) Cut reads in smaller parts, find overlaps



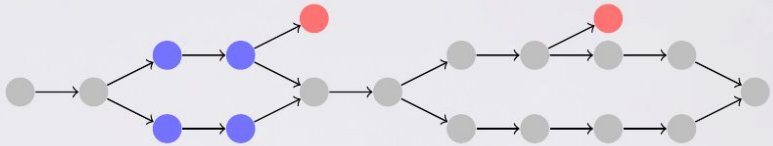
2) Refine graph
(see next slide)

3) Contigs created by graph traversal

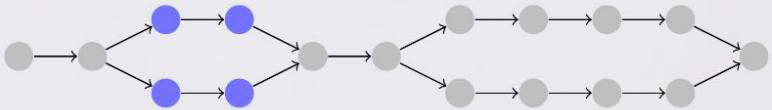
ACTGAG

Genome assembly using short reads

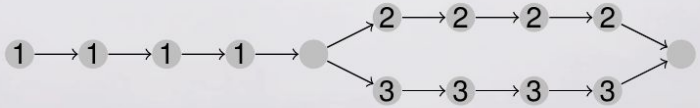
1) de Bruijn **graph** construction



2) Likely sequencing errors are removed.



3) Variations (e.g. SNPs, similar repetitions) are removed.
4) **Simple paths** (i.e. contigs) are returned.



Following steps:

Scaffolding

Long-range improvements

Annotation

Custom analysis

Scaffolding

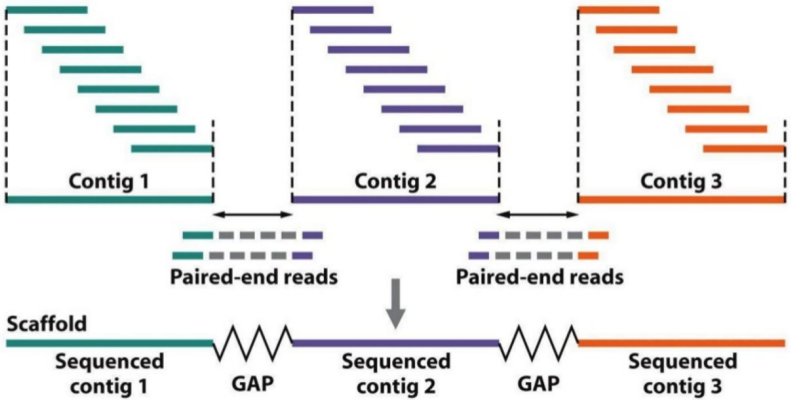


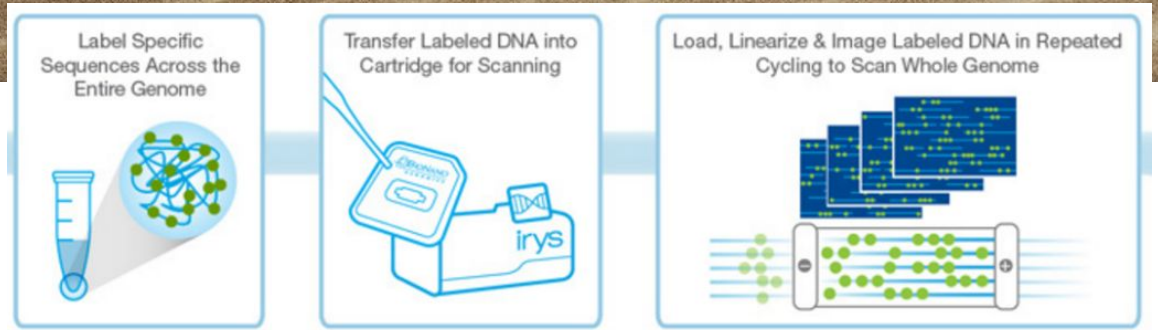
Fig: Griffiths et al 2012

Creates more contiguous assemblies using mate-pair sequencing

Source of **misassemblies**

Slowly being made obsolete by long reads

Long-range improvements



Bionano optical mapping

Hi-C / Dovetail

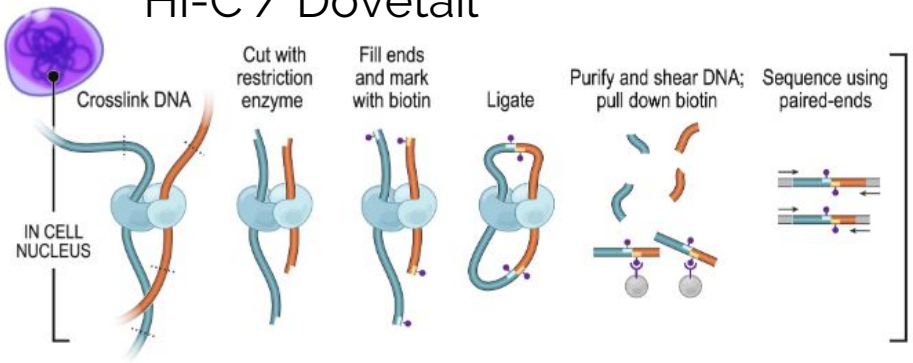
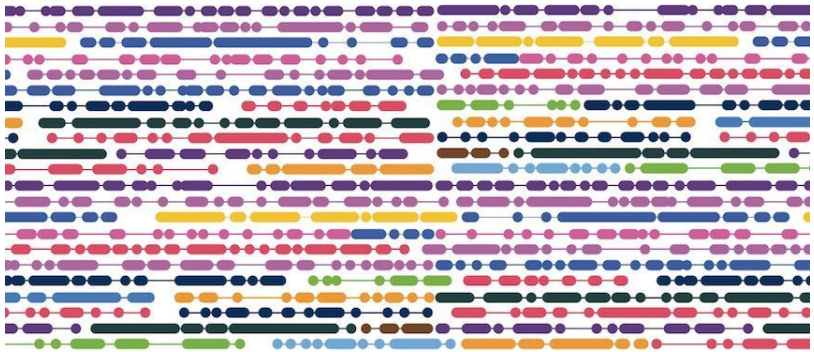


Fig: Rao et al 2014

Linked-reads (10XGenomics)



Genome assembly metrics

Total size

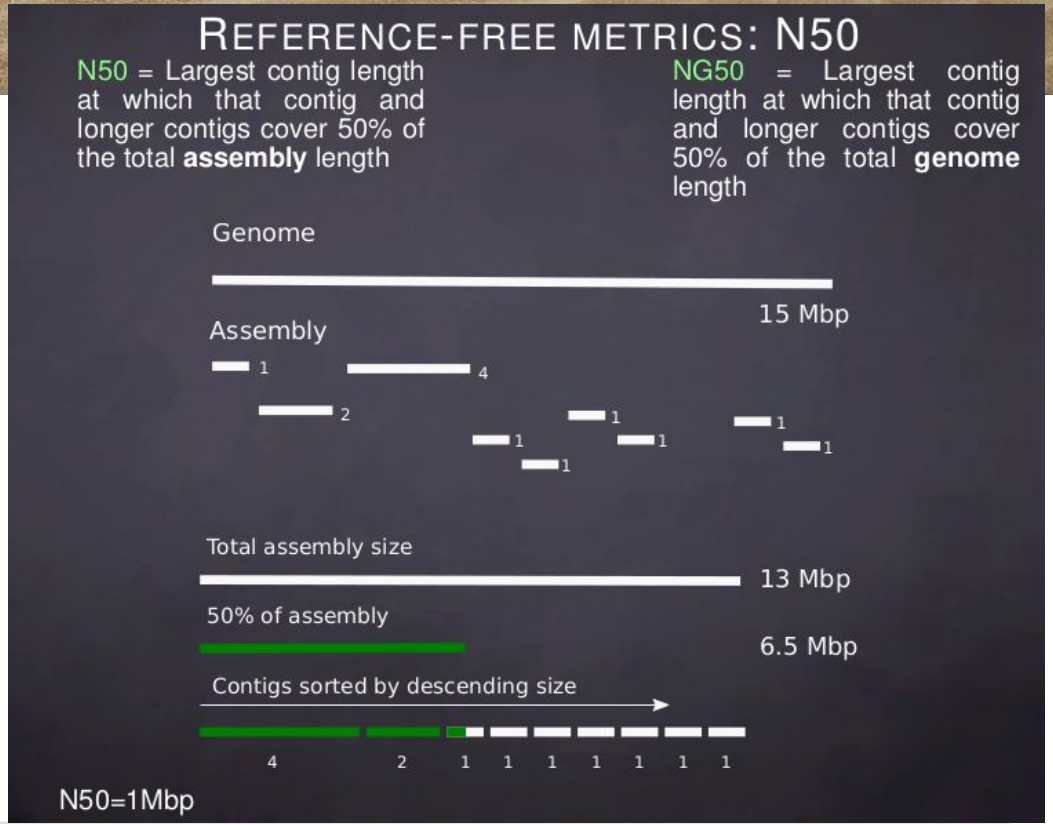
N50

% of reads mapping back to assembly

core genes found

...

Tools: **QUAST, Bandage**



Software recommendations (non exhaustive)

Small genomes:

SPAdes

Metagenomes:

SPAdes

MEGAHIT

Minia

Large genomes: unclear

ABySS

Soapdenovo2

Discovar_denovo + BESST

10x:

Supernova

Giraffe genome assembly

Data

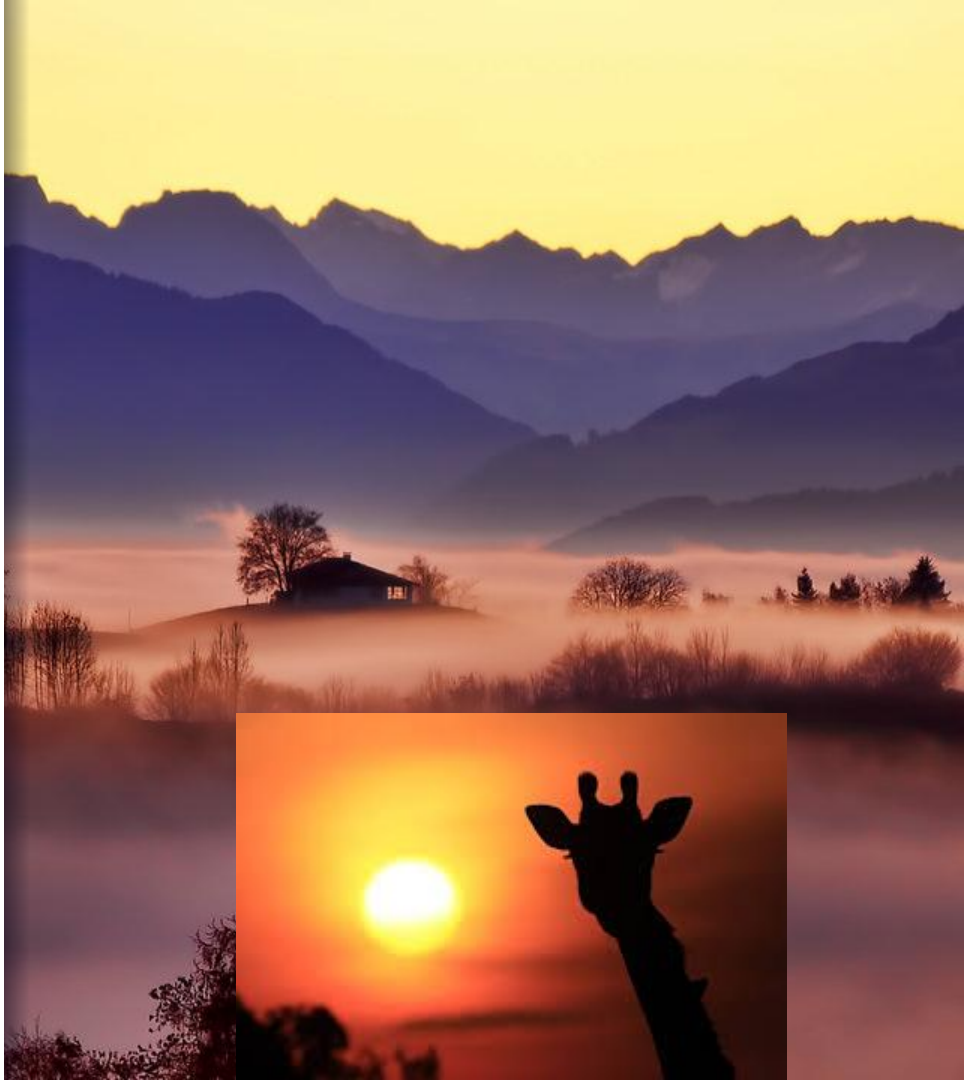
2012-style sequencing

Illumina only

30x paired-end

8x Nextera mate-pairs (4-8 kbp)

(on the very low end of sufficient)



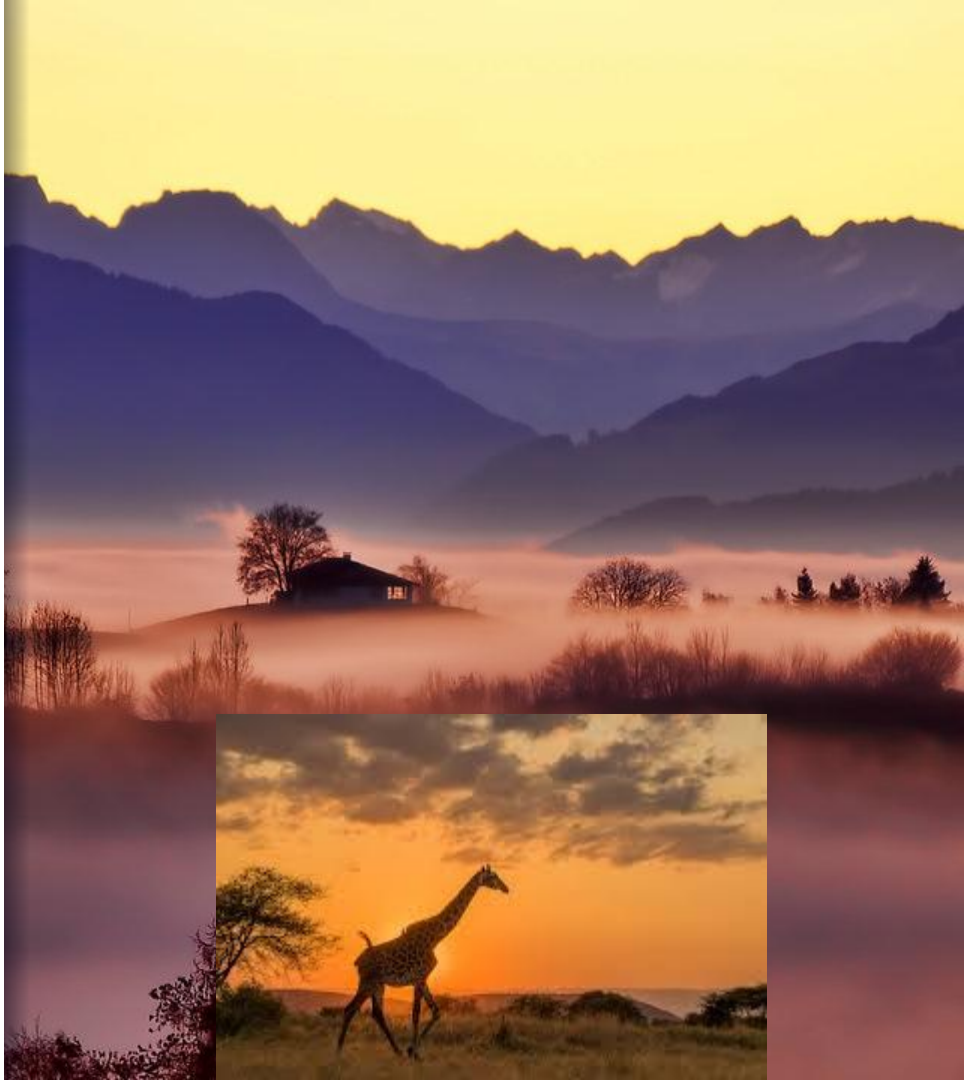
Assembly strategy

SOAPdenovo2

Scaffold N50=330 kbp

Contig N50=47 kbp

Mediocre contiguity



Insights [Agaba et al, Nature Comm 2016]

Gene analysis, comparison with assembled okapi genome

Found 70 genes that exhibit multiple signs of adaptation

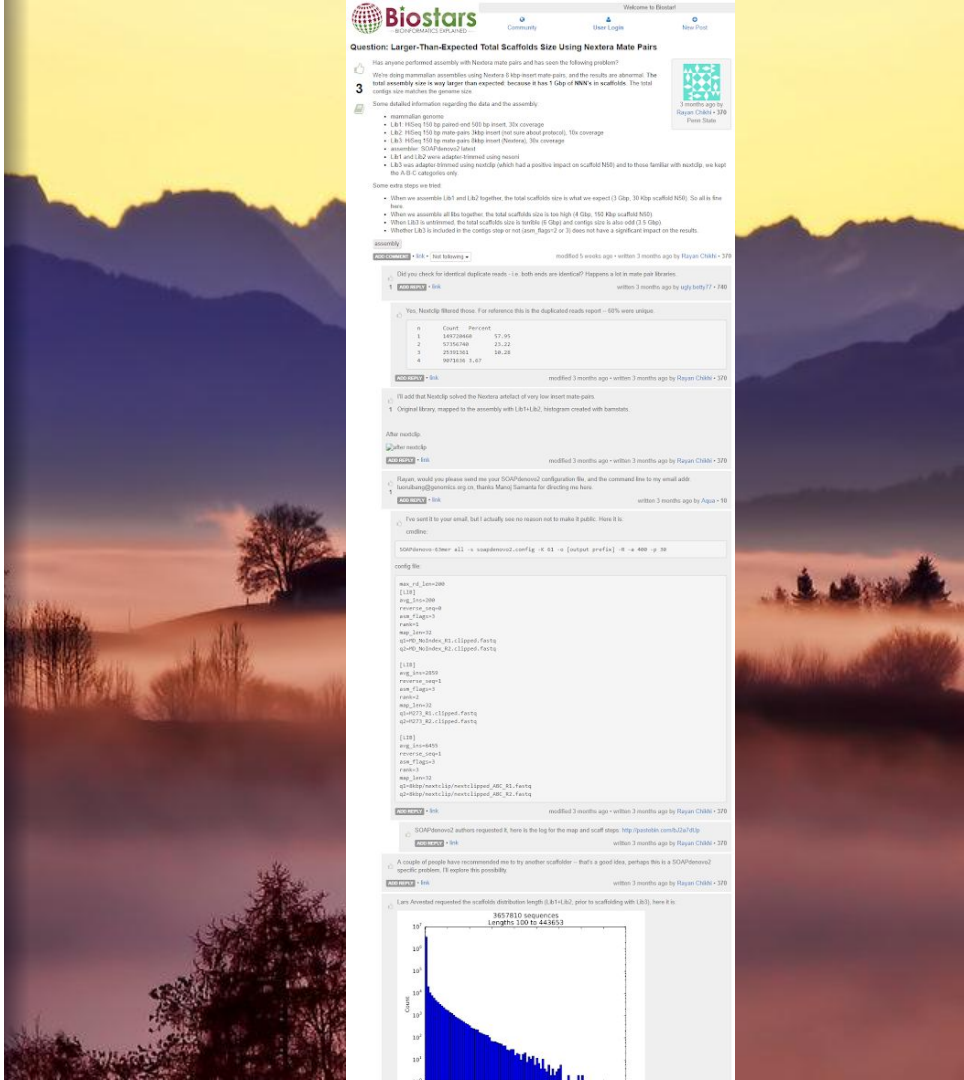
"*Giraffe's stature and cardiovascular adaptations evolved in parallel through changes in a small number of genes*"

Behind the scenes

Skewed sequencing data.

Standard pipeline ran, but results were off.

Requested specific knowledge from experts & tool authors



Message

A finished assembly is not always necessary for downstream insights, draft quality may be good enough

Beware even if tools run to completion without warnings

Spruce genome assembly

White Spruce (*Picea glauca)*

22 Gbp genome

ABySS assembly

70 kbp scaffold NG50

Annotation: MAKER-P

[Warren et al, 2015]

White spruce

Assembled with **ABySS 2**

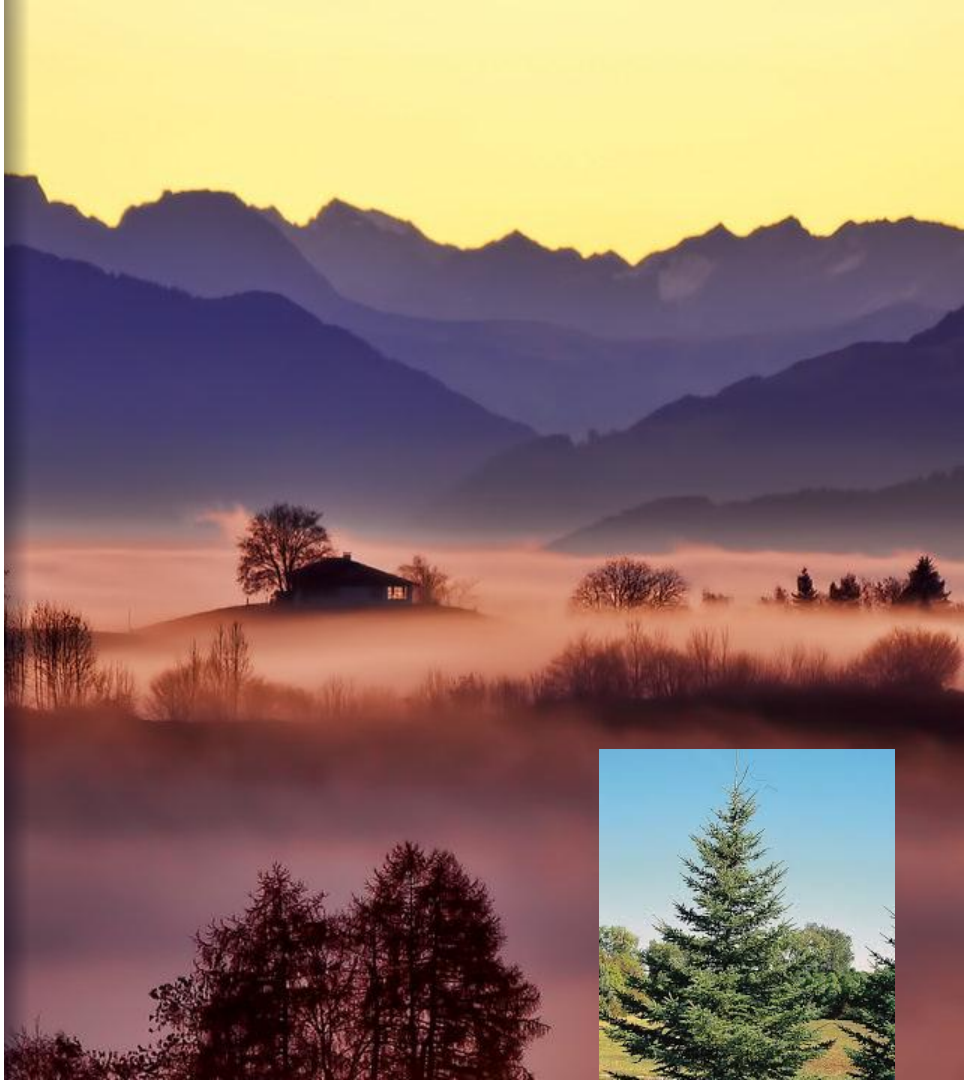
3.2 CPU-years, 500 GB RAM

Source: http://sjackman.ca/abyss2-slides/#/spruce-genome-assemblies

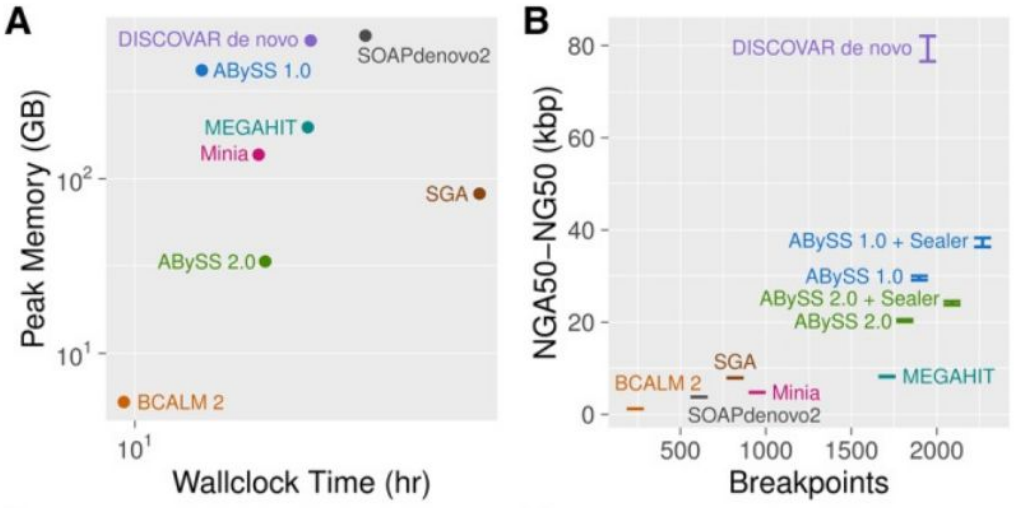
We proposed **BCALM2.1** for initial unitigs step:

4 CPU-days, 18 GB RAM

[Chikhi et al, ISMB 2016]

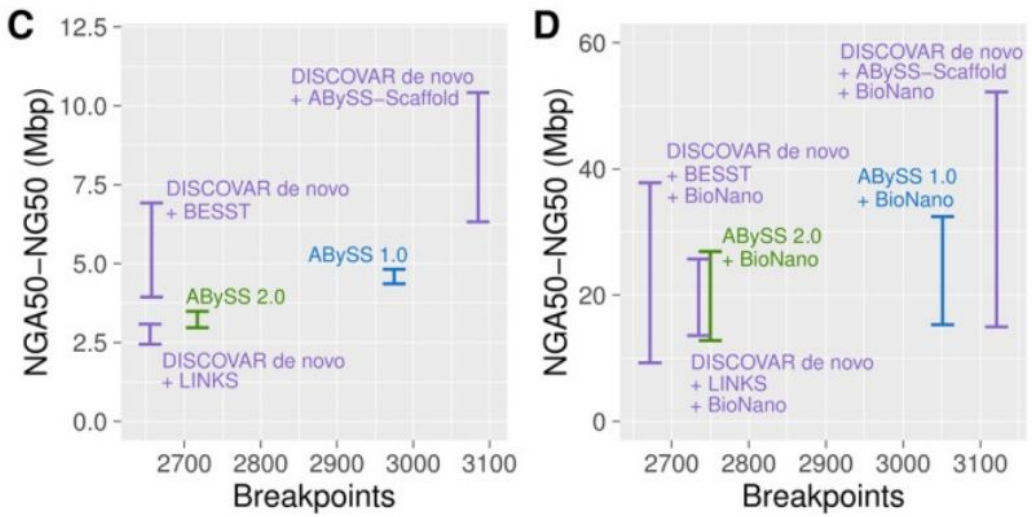


Performance of ABySS and BCALM2 assemblers



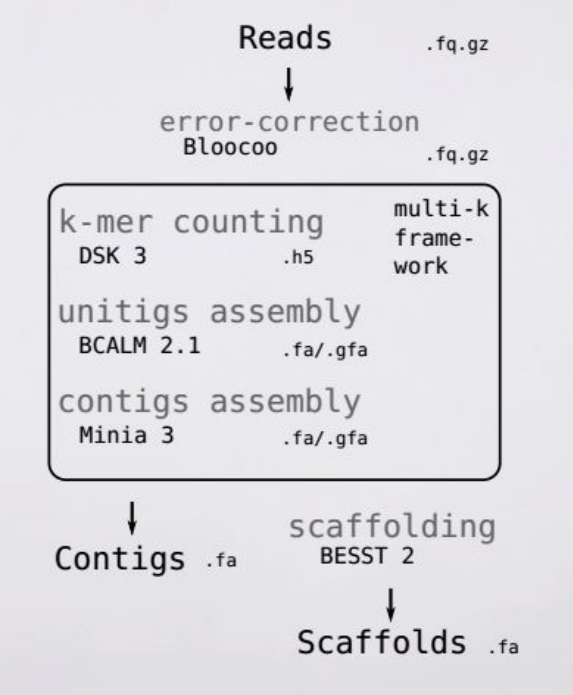
Human genome assembly, **contigs** Fig: [Jackman at al 2017]

ABySS scaffolding



Human genome assembly, **scaffolds** Fig: [Jackman at al 2017]

Minia-pipeline



[Chikhi, Rizk 2012] [Rizk, Chikhi, Lavenier 2013]
[Chikhi, Limasset, Medvedev 2016]
[Sahlin et al 2015] [in prep (2019?)]

Message

Larger genomes => challenging-er assembly

Unsupported by 10X

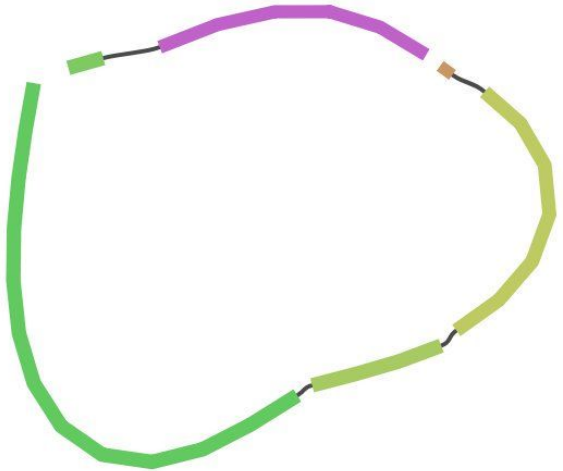
Long reads assembly

Long-reads assembly methods

Error-correction (or not)

Construction of a large read overlap graph

Consensus refinement



Previously known as *overlap-layout-consensus*

Software: Canu, FALCON, miniasm, smartdenovo, ...

Axolotl genome assembly

Axolotl facts

Mexican salamander

Species nearly extinct

Can regenerate limbs

32 Gbp genome,

65% repetitive

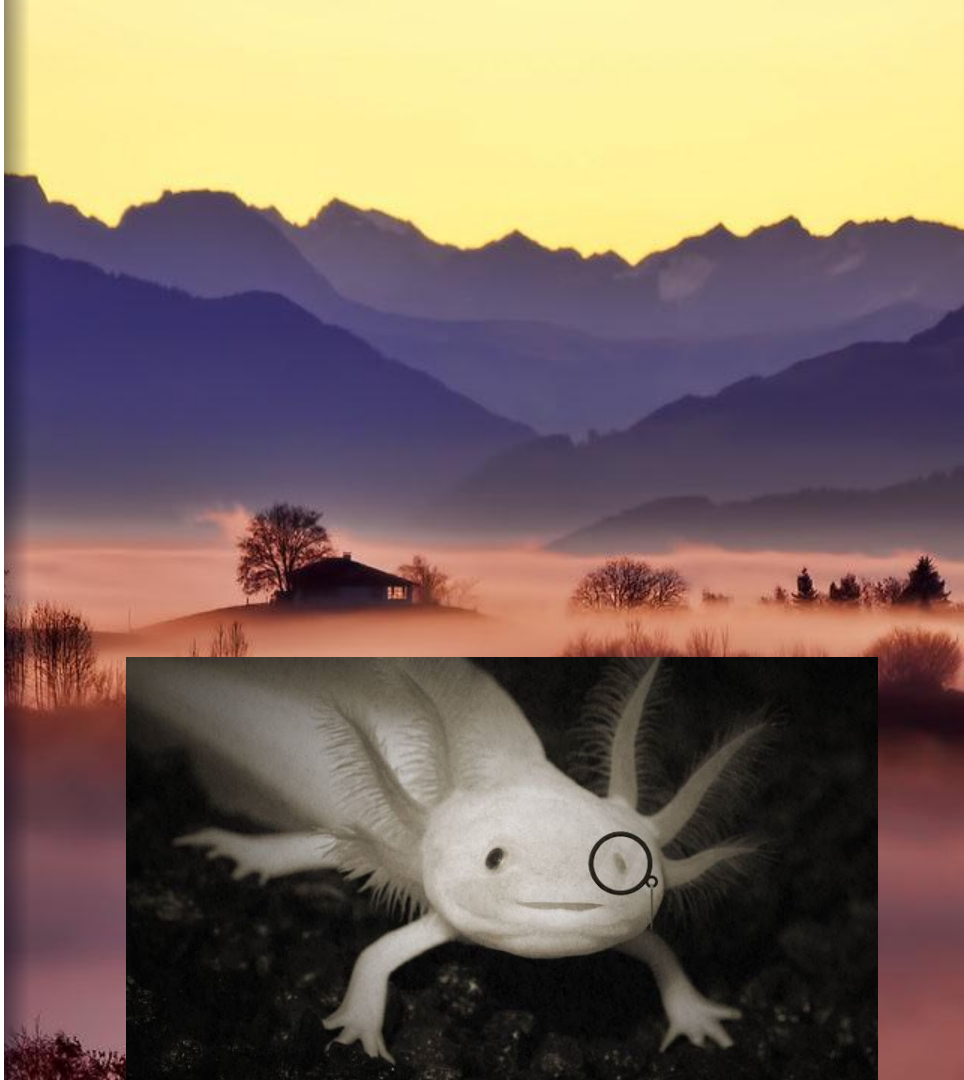
Axolotl genome, 2014 pre-history

First sequencing attempt

19x coverage Illumina

"Attempts to directly assemble [..] fail due to memory limitations (beyond 1 terabyte of RAM)."

[Keinath et al, Sci Rep 2015]



Axolotl genome, 2016

Unofficial & unpublished axolotl assembly made using BCALM2 & Minia assembler.

10x coverage

1.3 kbp contig N50

1 week running time

160 GB RAM

Axolotl genome, 2017

First published assembly

32x cov PacBio

7x cov Illumina

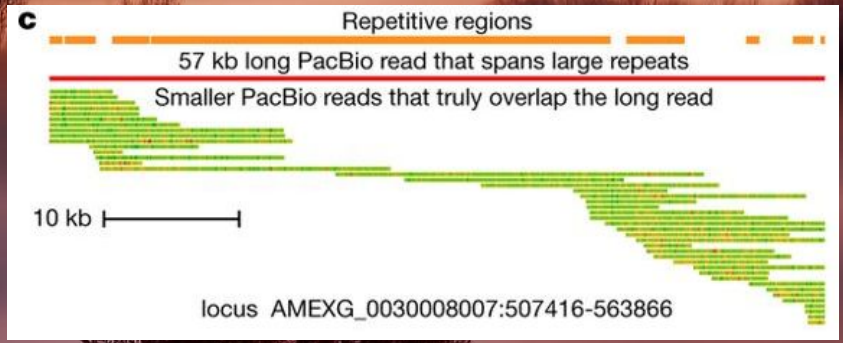
New MARVEL assembler

218 kbp contig N50

Bionano optical map

3 Mbp scaffold N50

[Nowoshilow et al, Nature 2017]



Axolotl genome, 2018 & beyond

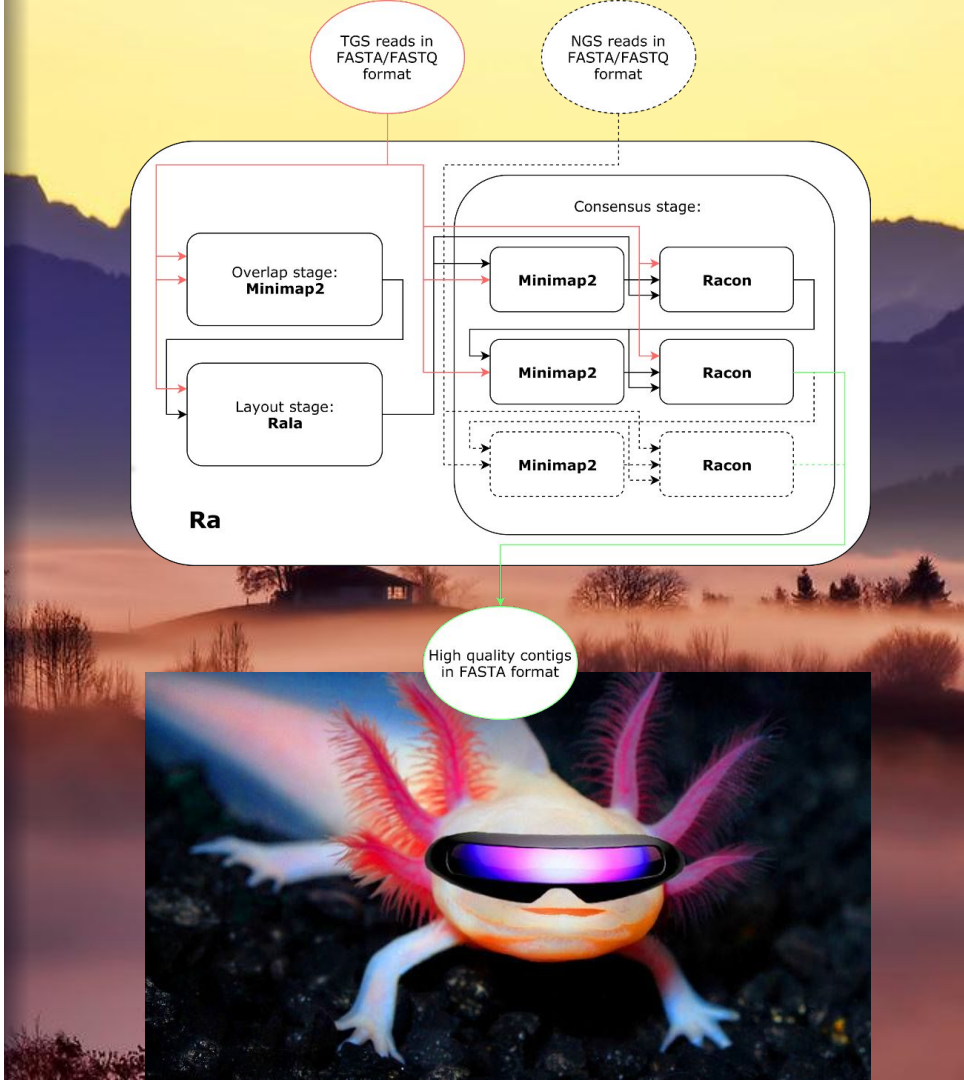
Improved techniques for long-reads

Jue Ruan's [wtdbg2]

Fuzzy de Bruijn graphs

5 CPU-months, 1.6 TB mem

Robert Vaser's [ra]



Future sequencing techniques

Sequencing advances

Sources: Twitter, & “Omics! Omics!” blog

PacBio

Illumina acquisition

MGI

Novaseq-class sequencer

Oxford Nanopore

Sub-$100 long reads for human genome?

Trio-binning [Koren et al 2018]

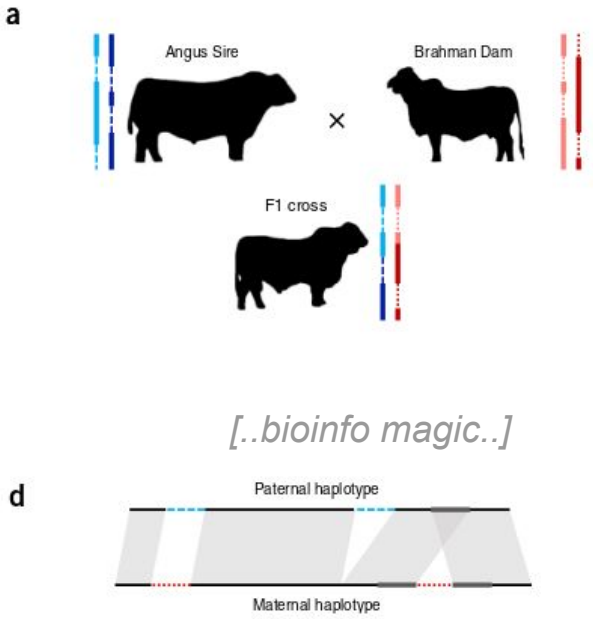
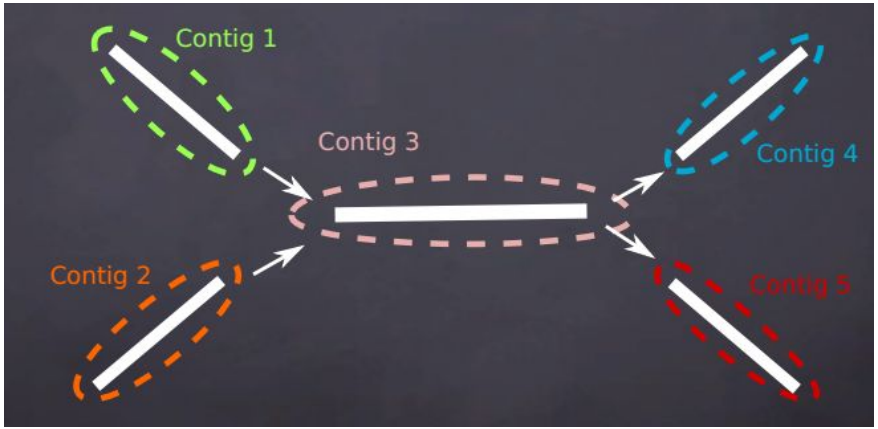


Fig: Koren et al 2018

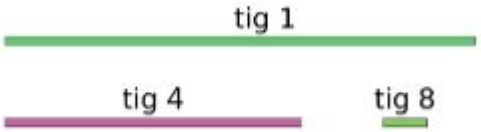
For a human genome, on par with 10x Supernova haplotypes

Assembly debugging

Assembly graphs



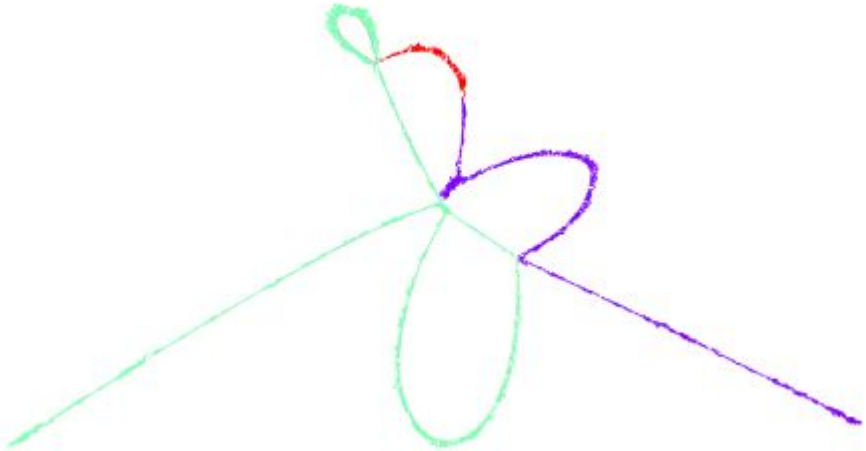
"Informative" contig graph



"Uninformative" assembly graph

PhD thesis of Pierre Marijon

Fragmented assemblies can be “debugged” using new methods



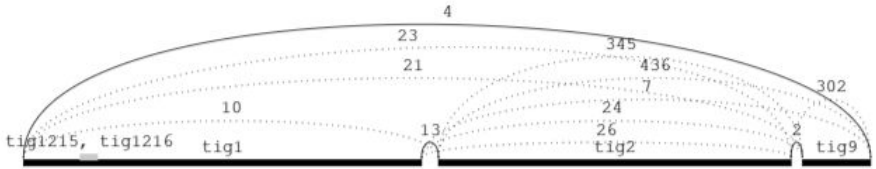
Canu contigs projected onto Minimap’s overlap graph

Analysis of 45 fragmented bacterial assemblies

NCTC3000 dataset

Recovering 35% of the missing contigs adjacencies, using raw reads.

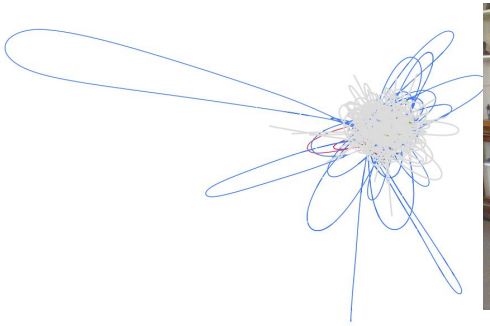
=> Finishing “for free”



We’re looking for a collaboration on PacBio-sequenced small eukaryote

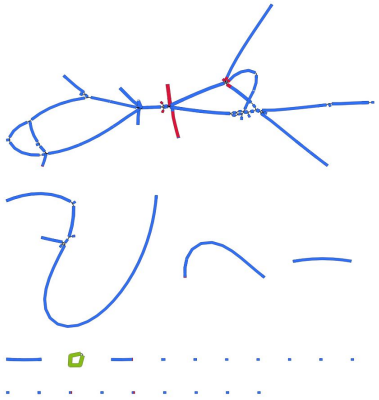
[Marijon et al, submitted]

HackSeq'18 event

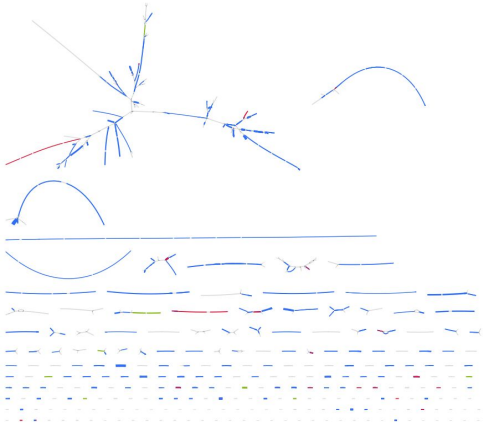


BCALM x ABySS k=99

- Many assembly tools
- Interoperable
- Mixing and matching can increase efficiency and processing power, and improve results
- GFA format



SPABySS k=99



BCALM x gfaview x ABySS k=99

Conclusion

Assembly is still unsolved

But long reads greatly improved the situation

Metagenomics: still unclear

Sequence Bioinformatics group @ Institut Pasteur, Paris

Questions?

Acknowledgements:

Univ Lille
Pierre Marijon
Jean-Stéphane Varré
Antoine Limasset

PSU
Paul Medvedev
Kristoffer Sahlin

Irisa
Charles Deltel
Dominique Lavenier
Claire Lemaitre
Pierre Peterlongo
Guillaume Rizk
(ex-Irisa:)
Patrick Durand
Erwan Drezen

Stockholm Univ
Lars Arvestad

Lund Univ
Dag Ahren

Axolotl pictures credit:
Maggie Sefton, bio data analyst for hire

Slide template:
https://www.slidescarnival.com/isabella-free-presentation-template/1989