ECFG15
ROME • ITALY 2020

UPPSALA UNIVERSITET

# Chromosome-level genome assembly of fungi

## A CASE STUDY OF THE FAIRY-RING FUNGUS

Markus Hiltunen, Martin Ryberg, Hanna Johannesson

2020-02-19

# Why bother about assembly contiguity?



- Synteny between divergent genomes
- Structural variation
- Linkage of genetic variants
- Specific hard-to-assemble genetic features, e.g. transposons, gene clusters, etc.

# The fairy-ring fungus



- *Marasmius oreades,* basidiomycete
- Vegetative accumulation of genetic variation
- Long-term, natural study system

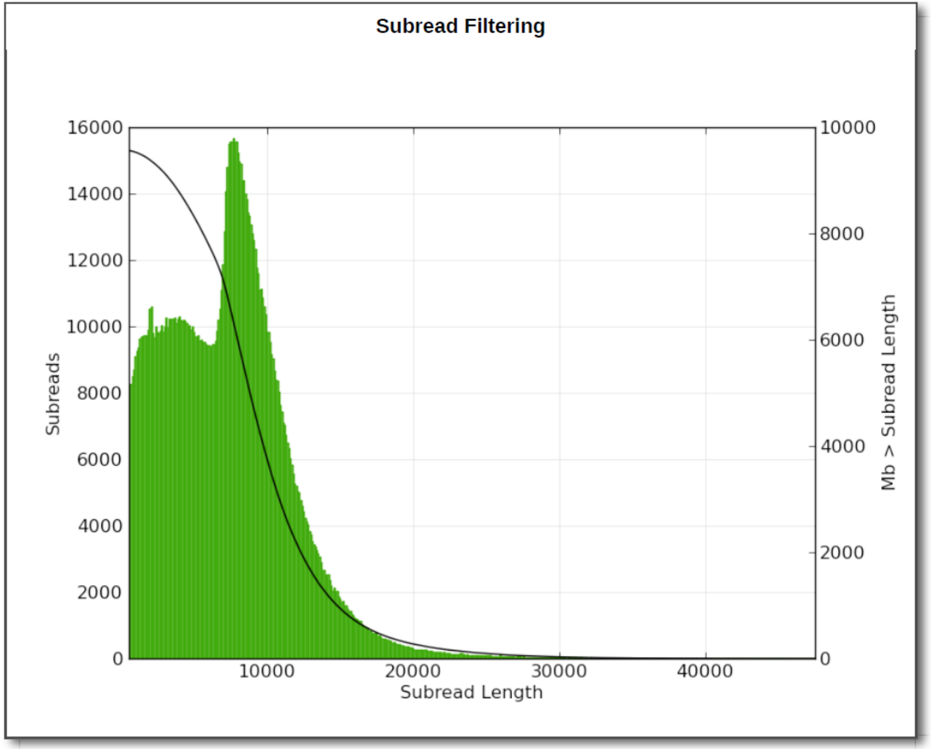
...but what about the genome?

Hiltunen *et al.* 2019. Curr. Biol.

# PacBio sequencing



- 8 SMRT cells RSII
- ~140X / bp
- Read N50: 9000 bp
- PacBio assembly:
  - 45 Mb
  - ~20-24 telomere repeats
  - ~120-160 contigs



# PacBio and Nanopore sequencing

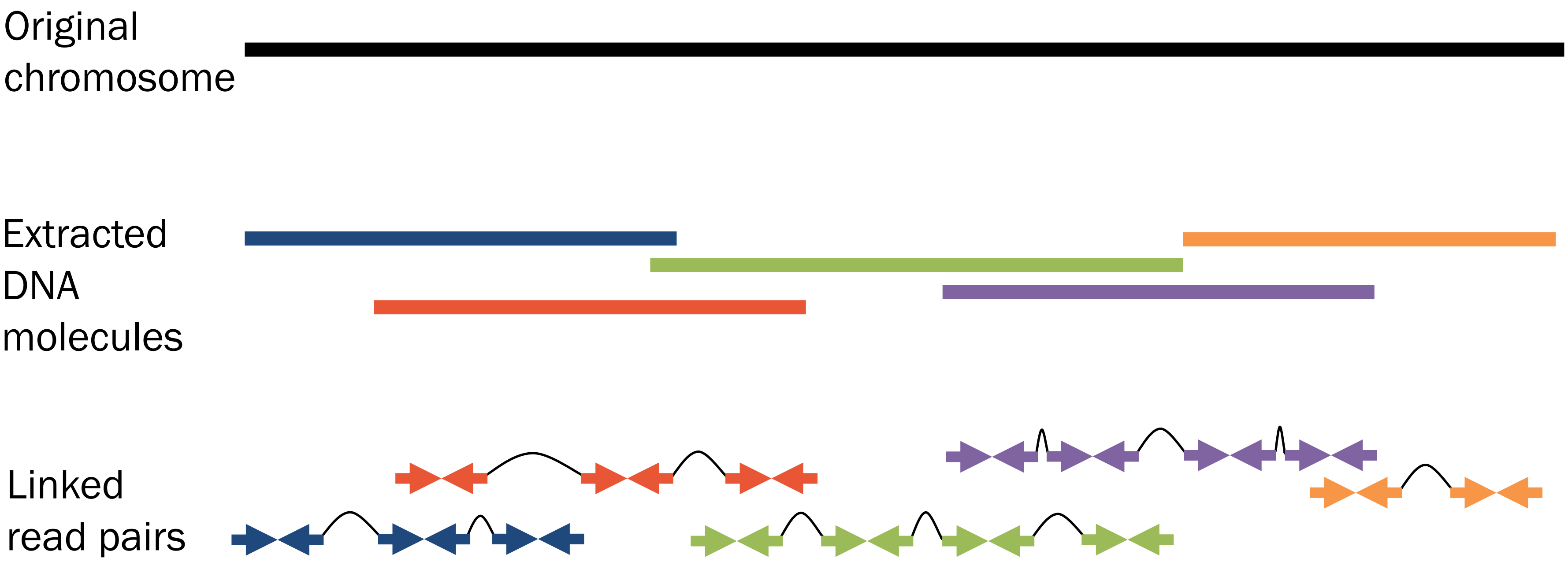


- 8 SMRT cells RSII
- ~140X / bp
- Read N50: 9000 bp
- PacBio assembly:
  - 45 Mb
  - ~20-24 telomere repeats
  - ~120-160 contigs

- MinION
- ~26X / bp
- Read N50: 11000 bp
- PacBio & Nanopore assembly:
  - 45 Mb
  - ~20-24 telomere repeats
  - ~100-140 contigs

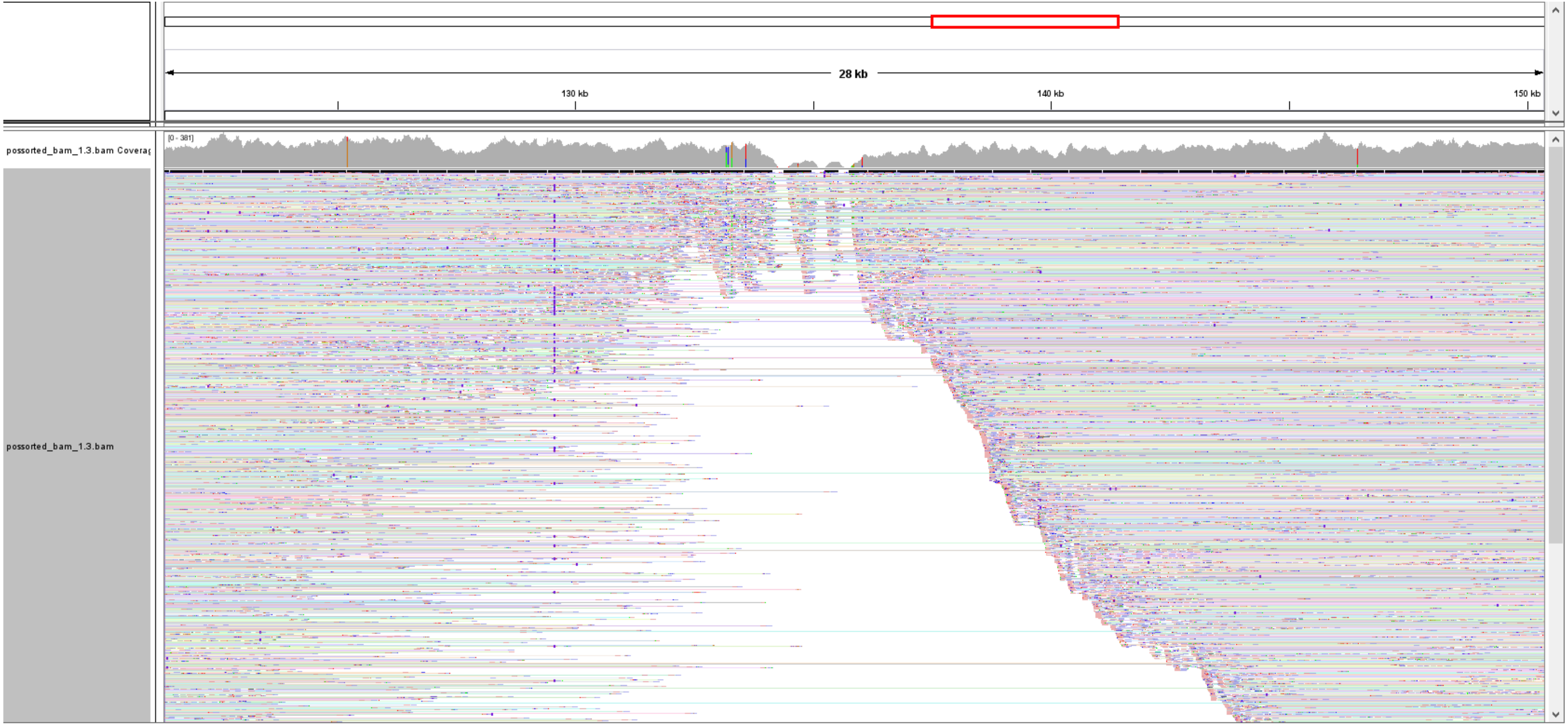
# What about 10X Chromium data?

Linked reads = barcode-tagged short reads



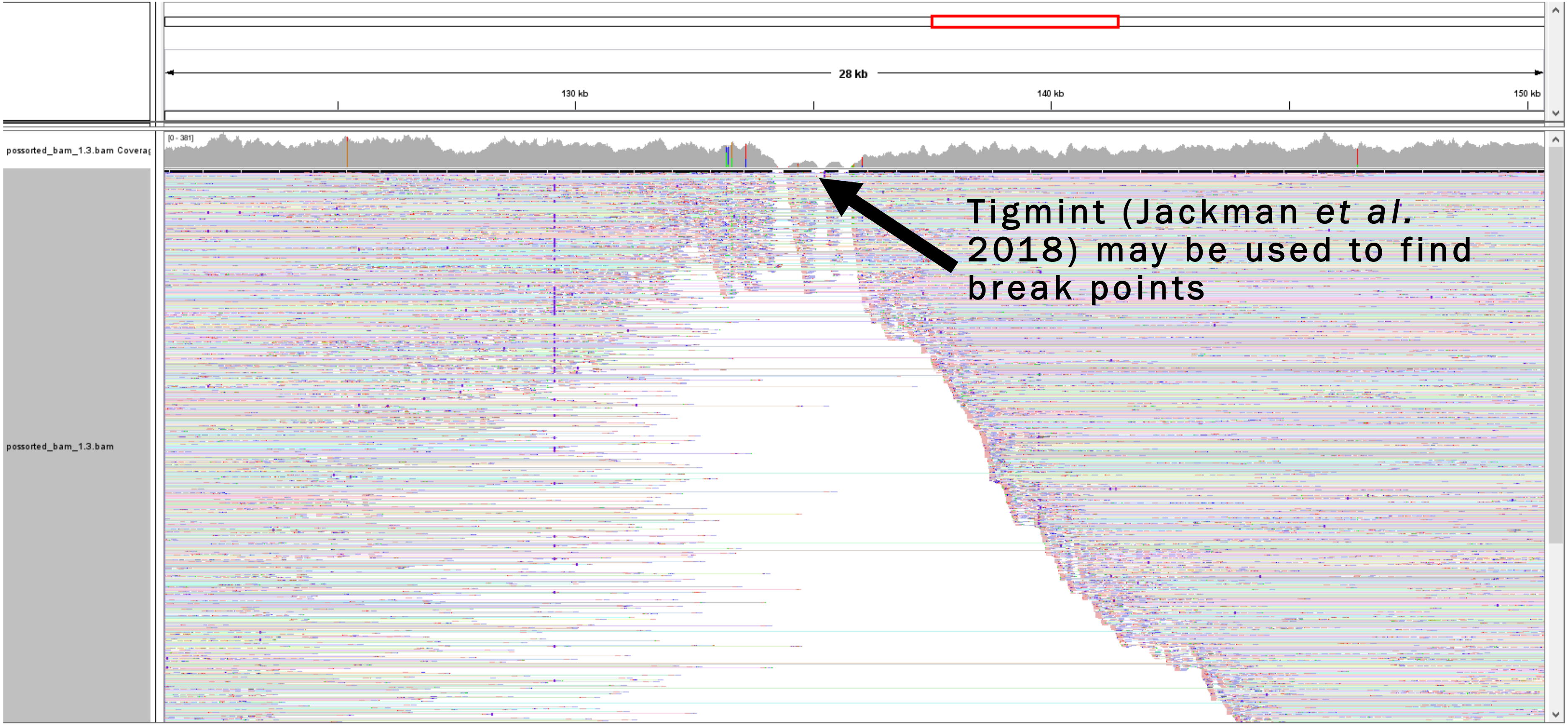
# Linked reads for misassembly detection

ECFG15 ROME • ITALY 2020



# Linked reads for misassembly detection






Tigmint (Jackman *et al.* 2018) may be used to find break points

# Linked reads can be used for scaffolding



- Barcodes can be used to infer linkage of pre-assembled contigs
- Map reads → look at barcodes at contig edges → contigs share many barcodes? They belong together!
- ARCS (Yeo *et al.* 2018) + LINKS (Warren *et al*. 2015)
  - Scaffolding by gap insertion
  - Designed mainly for large genomes
  - Many steps from mapped reads → new scaffolds

# Linked reads can be used for scaffolding



- Barcodes can be used to infer linkage of pre-assembled contigs
- Map reads → look at barcodes at contig edges → contigs share many barcodes? They belong together!
- ARCS (Yeo *et al.* 2018) + LINKS (Warren *et al*. 2015)
  - Scaffolding by gap insertion
  - Designed mainly for large genomes
  - Many steps from mapped reads → new scaffolds
- In small genomes, e.g. fungi, linked contigs often overlap
- Gaps can be avoided in those cases

# AnVIL: Assembly Validation and Improvement using Linked reads

ECFG15
ROME • ITALY 2020

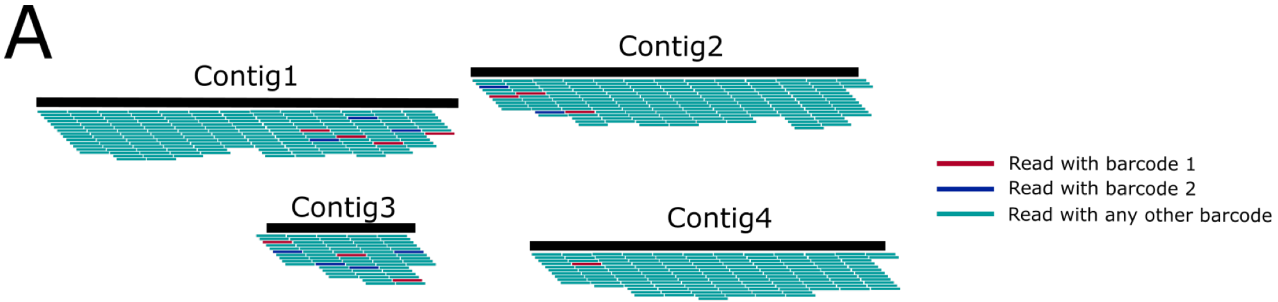
- Links resolved by Overlap-Layout-Consensus
  - Number of gaps is thereby minimized
- ”Push-button” pipeline
- Written and implemented in Python 3
  - Only dependent on Python packages
- https://github.com/markhilt/AnVIL

Hiltunen *et al.* In prep.

# AnVIL pipeline



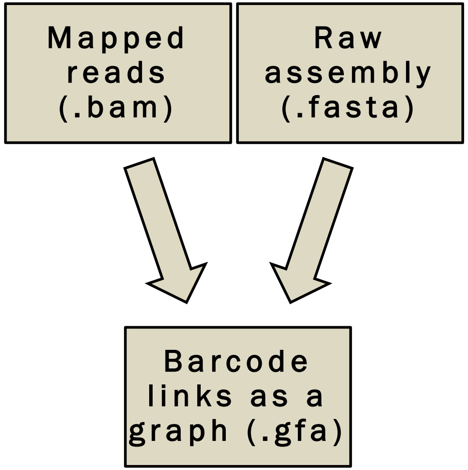
| | |
|---|---|
| Mapped reads (.bam) | Raw assembly (.fasta) |

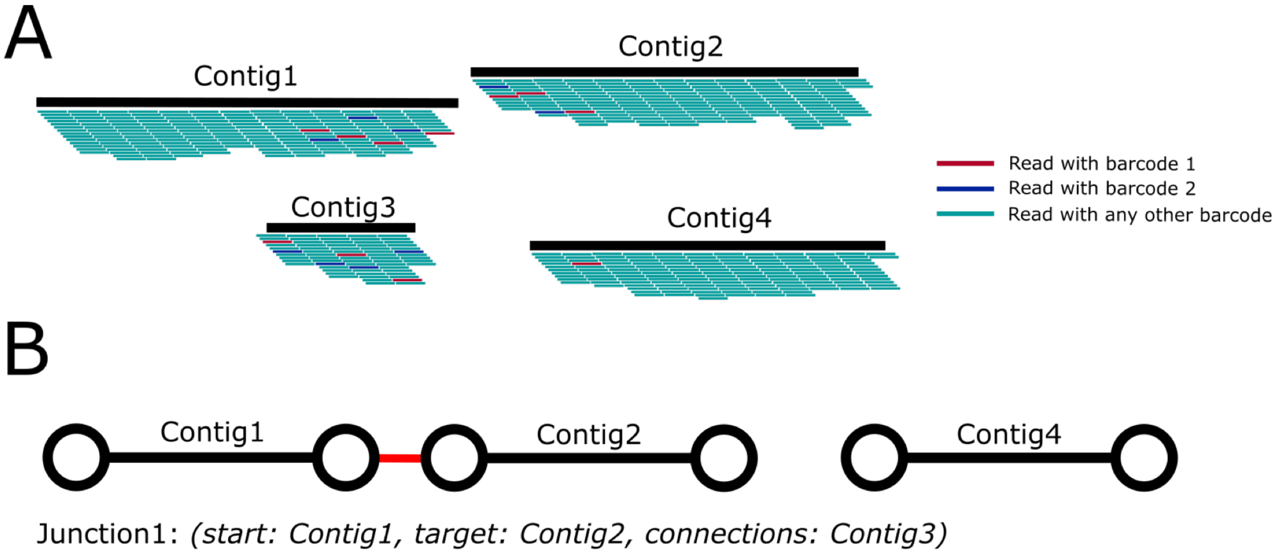



Hiltunen *et al.* In prep.

# AnVIL pipeline

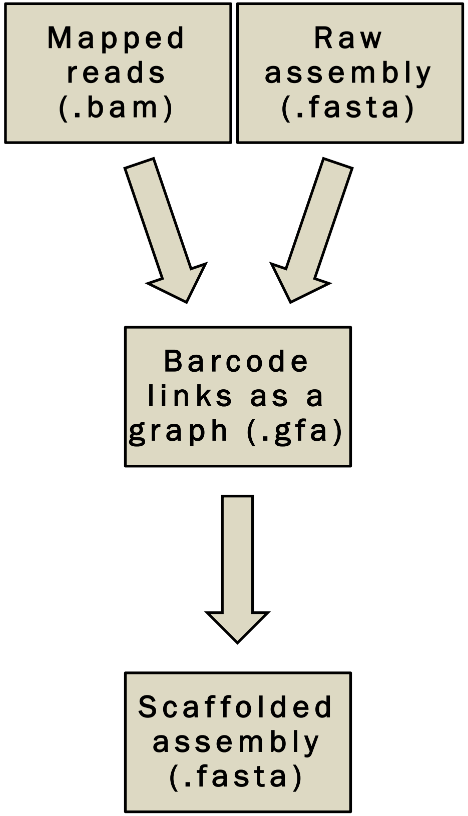


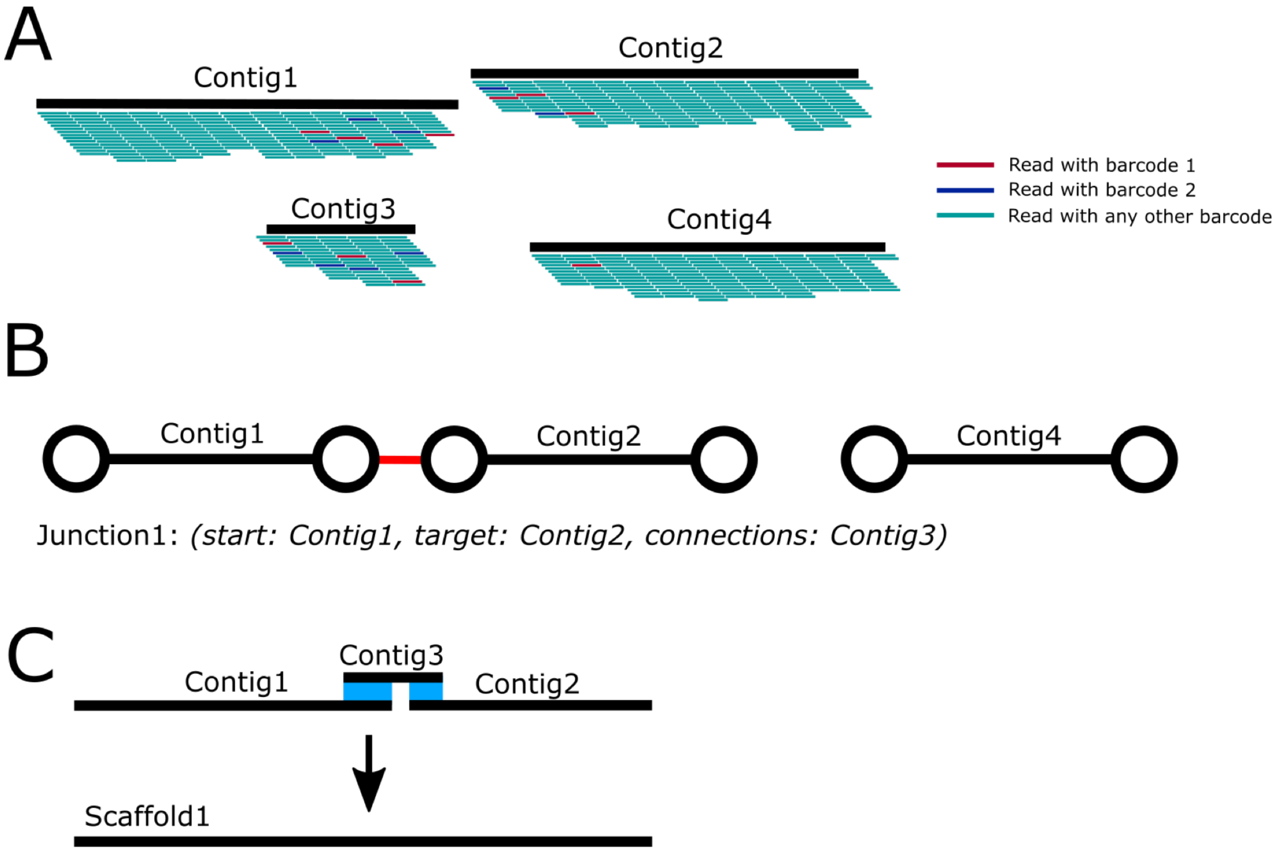



Hiltunen *et al.* In prep.

# AnVIL pipeline

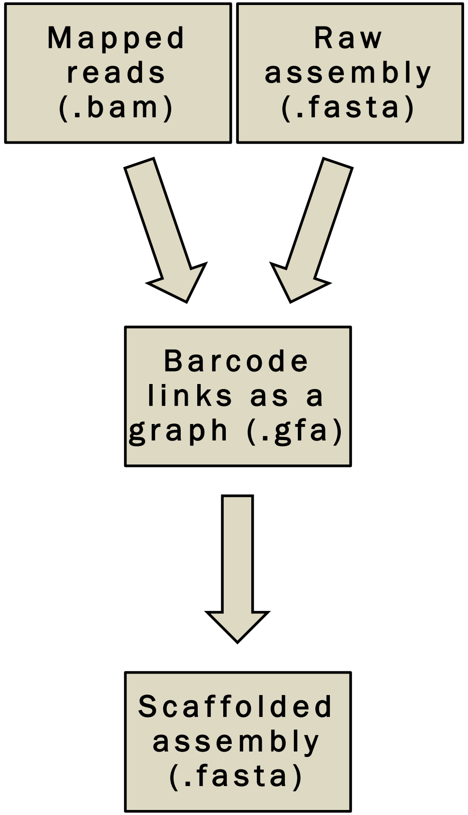


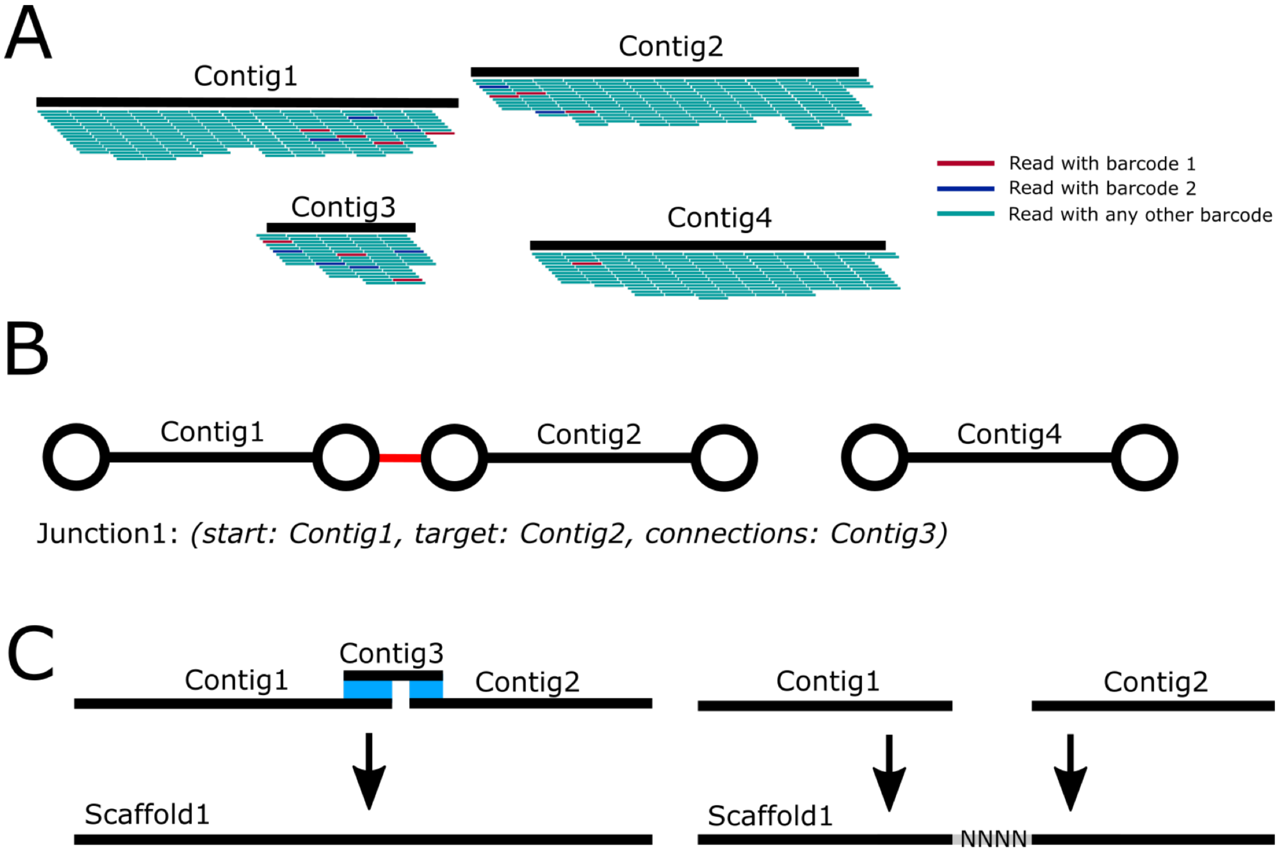




Hiltunen *et al.* In prep.

# AnVIL pipeline






Hiltunen *et al.* In prep.

# AnVIL & ARCS on *Marasmius* data



1. De novo assembly of PacBio and Nanopore long reads
2. Break assembly at regions of low barcode support
3. Scaffold using either ARCS or AnVIL

# AnVIL & ARCS on *Marasmius* data



1. De novo assembly of PacBio and Nanopore long reads
2. Break assembly at regions of low barcode support
3. Scaffold using either ARCS or AnVIL

| Genome | Scaffolds | Gaps + aligned merges | N50 (kb) | L50 | Longest scaffold (Mb) |
|---|---|---|---|---|---|
| Raw assembly | 171 | NA | 846 | 18 | 2.6 |
| AnVIL scaffolds | 67 | 54 + 47 | 2 562 | 5 | 5.1 |
| ARCS scaffolds | 138 | 33 + 0 | 1 229 | 12 | 3.7 |

# AnVIL & ARCS on *Marasmius* data



1. De novo assembly of PacBio and Nanopore long reads
2. Break assembly at regions of low barcode support
3. Scaffold using either ARCS or AnVIL

| Genome | Scaffolds | Gaps + aligned merges | N50 (kb) | L50 | Longest scaffold (Mb) |
|---|---|---|---|---|---|
| Raw assembly | 171 | NA | 846 | 18 | 2.6 |
| AnVIL scaffolds | 67 | 54 + 47 | 2 562 | 5 | 5.1 |
| ARCS scaffolds | 138 | 33 + 0 | 1 229 | 12 | 3.7 |

# AnVIL & ARCS on *Marasmius* data



1. De novo assembly of PacBio and Nanopore long reads
2. Break assembly at regions of low barcode support
3. Scaffold using either ARCS or AnVIL

| Genome | Scaffolds | Gaps + aligned merges | N50 (kb) | L50 | Longest scaffold (Mb) |
|---|---|---|---|---|---|
| Raw assembly | 171 | NA | 846 | 18 | 2.6 |
| AnVIL scaffolds | 67 | 54 + 47 | 2 562 | 5 | 5.1 |
| ARCS scaffolds | 138 | 33 + 0 | 1 229 | 12 | 3.7 |

# AnVIL & ARCS on *Marasmius* data



1. De novo assembly of PacBio and Nanopore long reads
2. Break assembly at regions of low barcode support
3. Scaffold using either ARCS or AnVIL

| Genome | Scaffolds | Gaps + aligned merges | N50 (kb) | L50 | Longest scaffold (Mb) |
|---|---|---|---|---|---|
| Raw assembly | 171 | NA | 846 | 18 | 2.6 |
| AnVIL scaffolds | 67 | 54 + 47 | 2 562 | 5 | 5.1 |
| ARCS scaffolds | 138 | 33 + 0 | 1 229 | 12 | 3.7 |

# Can we trust linked reads for scaffolding?



- We can verify the overall structure of the linked assembly by generating a genetic map

# Can we trust linked reads for scaffolding?



- We can verify the overall structure of the linked assembly by generating a genetic map
- Map construction:
  - 95 single spore isolates (haploid, meiotic progeny)
  - Sequenced by multiplexed Illumina WGS
  - Lep-Map3 + Allmaps for map-based assembly scaffolding

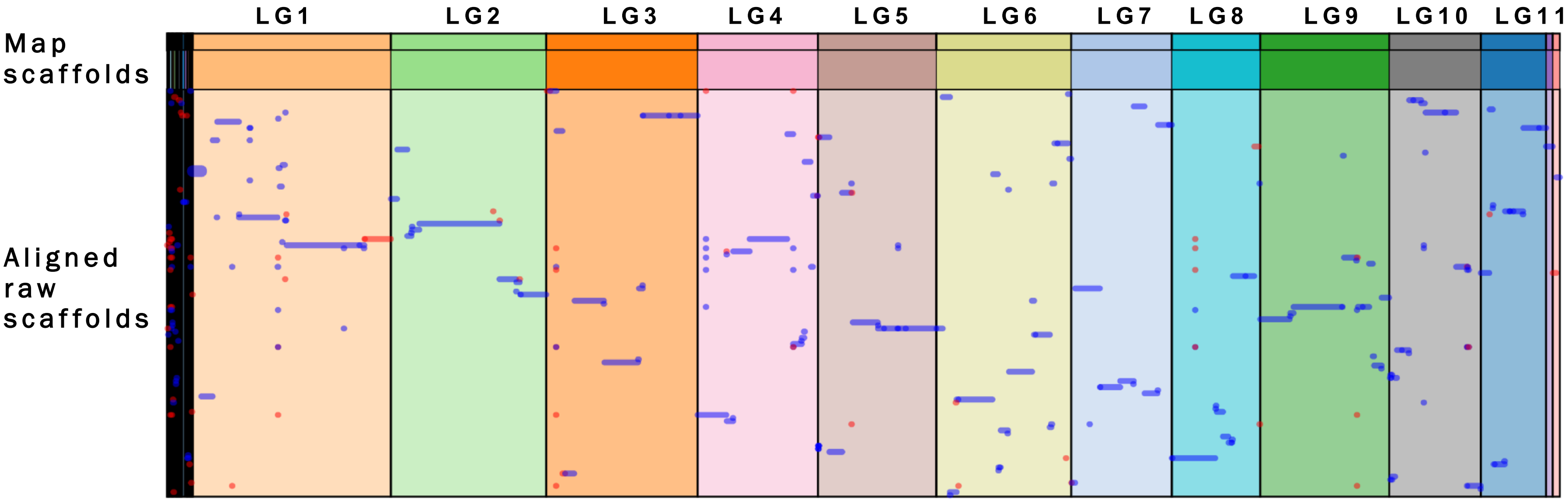
Tang *et al.* 2015, Rastas 2017

# Synteny of assemblies

ECFG15
ROME • ITALY 2020

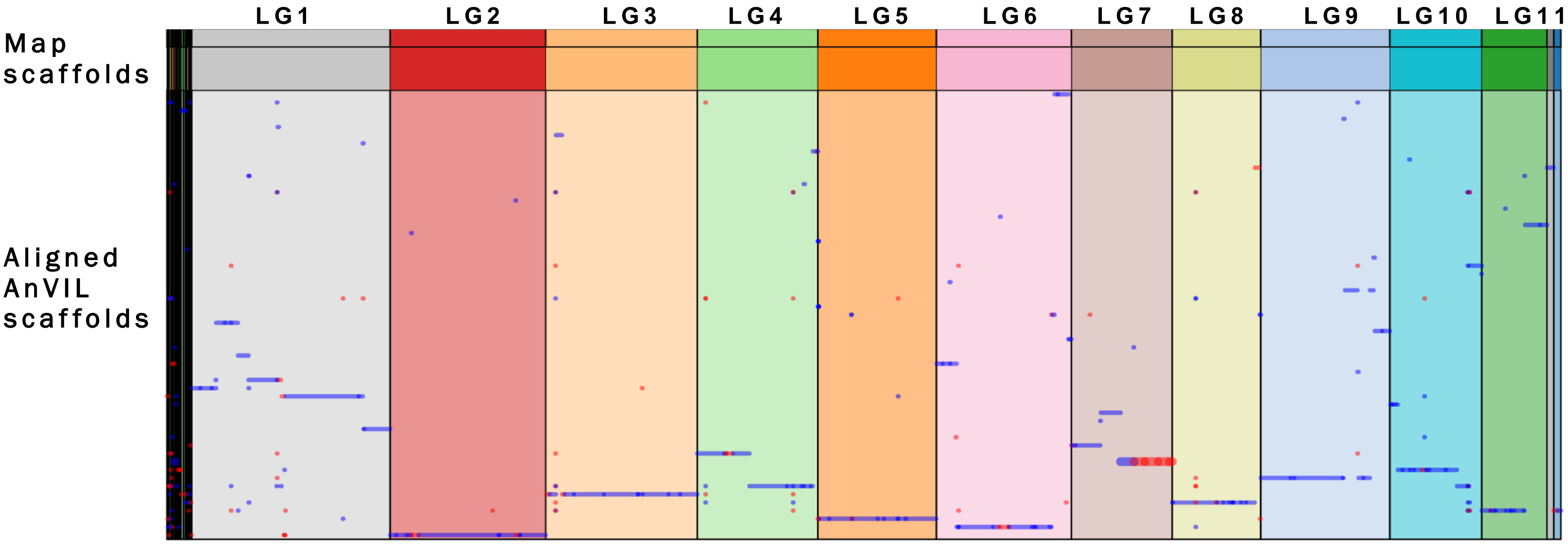
# Synteny of assemblies





# Synteny of assemblies





# Conclusions



- Long-read and linked-read data in combination can generate chromosome-level genome assemblies of fungi
- AnVIL can be used to scaffold genome assemblies with 10X Chromium data to minimize the number of gaps
- Linkage maps, while useful, are not necessary for chromosome assembly of fungi

https://github.com/markhilt/AnVIL

# Acknowledgements

Johannesson lab:

Martin Ryberg

Magdalena Grudzinska-Sterno

Ola Wallerman

Reneh Kostines

Max Käller

Mattias Ormestad

UPPSALA UNIVERSITET

UPPSALA MYCOLOGY

NATIONAL GENOMICS INFRASTRUCTURE

Next-Generation Sequencing and Genotyping for Swedish Research

SciLifeLab

https://github.com/markhilt/AnVIL

# Supernova assembly



- Native 10X Chromium assembler
- Kmer-based short read assembler
- Utilizes barcode information for scaffolding

| Assembler | Total size (bp) | Scaffolds | N50 | L50 | Longest scaffold (bp) |
|---|---|---|---|---|---|
| Spades | 44,168,907 | 52,408 | 30,382 | 336 | 463,765 |
| Supernova | 39,642,918 | 1,288 | 119,258 | 83 | 1,364,318 |

# AnVIL utility on other data

ECFG15
ROME • ITALY 2020

| Species | Genome | Scaffolds | Gaps + aligned merges | N50 (Mb) | L50 | Longest scaffold (Mb) |
|---|---|---|---|---|---|---|
| *C. elegans* | Reference | 7 | NA | 17.2 | 3 | 20.2 |
| | Raw assembly | 66 | NA | 5.6 | 7 | 12.2 |
| | AnVIL scaffolds | 44 | 12 + 8 | 17.1 | 3 | 20.1 |
| | ARCS scaffolds | 54 | 12 + 0 | 8.5 | 5 | 13.0 |
| *A. thaliana* | Reference | 7 | NA | 23.5 | 3 | 30.4 |
| | Raw assembly | 771 | NA | 0.5 | 79 | 2.4 |
| | AnVIL scaffolds | 472 | 243 + 59 | 4.6 | 10 | 10.7 |
| | ARCS scaffolds | 433 | 338 + 0 | 4.4 | 9 | 12.5 |
| *H. sapiens* | Reference | 639 | NA | 145.1 | 9 | 249.0 |
| | Raw assembly | 65 907 | NA | 5.6 | 147 | 26.7 |
| | AnVIL scaffolds | 65 321 | 514 + 79 | 13.6 | 61 | 47.5 |
| | ARCS scaffolds | 65 505 | 402 + 0 | 11.0 | 73 | 60.6 |

# AnVIL utility on other data

ECFG15
ROME • ITALY 2020

| Species | Genome | Total size (bp) | Scaffolds | N50 | L50 | Longest scaffold | # misassembled contigs | mismatches/100 kb | indels/100kb |
|---|---|---|---|---|---|---|---|---|---|
| *C. elegans* | Reference | 98,291,416 | 7 | 17,183,857 | 3 | 20,182,852 | NA | NA | NA |
| | Canu contigs | 98,659,475 | 66 | 5,629,213 | 7 | 12,209,780 | 3 | 1.74 | 3.30 |
| | AnVIL scaffolds | 98,608,266 | 44 | 17,140,346 | 3 | 20,110,319 | 5 | 1.82 | 3.12 |
| | ARCS scaffolds | 98,660,675 | 54 | 8,544,338 | 5 | 13,042,304 | 7 | 1.97 | 3.29 |
| *A. thaliana* | Reference | 119,667,750 | 7 | 23,459,830 | 3 | 30,427,671 | NA | NA | NA |
| | Canu contigs | 122,579,182 | 771 | 454,218 | 79 | 2,359,306 | 487 | 905.48 | 298.23 |
| | AnVIL scaffolds | 122,420,859 | 472 | 4,635,812 | 10 | 10,687,703 | 265 | 900.75 | 295.56 |
| | ARCS scaffolds | 122,612,982 | 433 | 4,354,541 | 9 | 12,501,495 | 223 | 902.43 | 298.42 |
| *H. sapiens* | Reference | 3,272,089,205 | 639 | 145,138,636 | 9 | 248,956,422 | NA | NA | NA |
| | Raw assembly scaffolds | 3,066,280,821 | 65,907 | 5,640,076 | 147 | 26,692,832 | 1,749 | 106.60 | 26.13 |
| | AnVIL scaffolds | 3,065,891,481 | 65,321 | 13,586,348 | 61 | 47,493,337 | 1,503 | 106.55 | 26.18 |
| | ARCS scaffolds | 3,066,321,021 | 65,505 | 10,981,961 | 73 | 60,607,432 | 1,599 | 106.60 | 26.13 |