Genome assembly quality control

Bioplatforms Fungi Genomics Workshop 2024 Australian National University

Rita Tam



ritahltam



ritatam

"I have my first genome assembly (.fasta)! Now what?"

Evaluate assembly quality!

Never assume your genome assembly is "error-free".

This is a "draft" assembly.

TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC CCTAACCCAAACCCTAACCCTAAACCCTAAACCCTAACCCTAAACCCTAAA CCTTTATCATACTTCACTTATTGAATTCTAATACTTAACATACTCTGC TCATCTTGTTTGATGCTCATACTCATATCAGCTGGCCGTTATCGCTCTGTC TCCATTATACTTGACATGCTCCGGCCCCTCAATATACTTGACATACTCTTG TACTCTTGTCATATATAGTATCTTGTTTGATGGTAATAATCATACTGCTAG TCAGTCGGATTCTGAATATCTCATCTACGTCTCATACCATTAACTTGAGCT TCGACACAGACGTAGACTTTTTAAAATTTCCTGAAAAAAGTTTGACAATTA CACAGCTTTCCAAATCTGTTCTCAGAATGTTCGTAGCTATCATAGTTTTTG GGGATTTCTTTCCAGATTCGCGCAAATCTGGGATTTAGAAAAATCCGACTT CTGAGGAATGAGGGCTCATTCCAACCCGCCCGAGGGTTAGAAAAAATAGTT CAATTATTATTTGATATTTCAGGTCTCAGGACTCTGATTCCTCCAGAACTT TTTTCGAGTTAGAAATCCCAAATTCGCGCGAATTTGGGATTGCAGTCCAAA TTTTCAAATAAAAAAGTCAATTTTTTTTGGAAAAATCAGGCCAAAAATTC/ ATTTGACAGGTTTGACTTTGGATGTTCGATCTTGAAGTACTCTAGTGCTAG TTCAGCTAAGGGCAACTCTCTCTCGGCAGTTTCGATCTCGTCTTCAGGATC TCTCCCTTGGGCATGAGTTTCGAATGTGATTTTTCTGGCAGA AGAATCTTCTATTGAGGCCTGAATCTTCAAAAATTGTTCGGAGTGATTCTA CGGATAATACCCAAACCTCTTTGCTTCAAGGTTAATGAGGTGAGACAAGAT CAAGAGGCAGGTGGATGCTTAGTAGAGAAGGTTGCCCGAGTGATTTTCGAT GACAGGGTAAGTTCCCTTCATATCGACCATTCATTCGGAGTTCTTCGTGTT GACAATCCTTGATCGTTAACCGGTCGAAGGTGAGGAAACATTTTCCAACAC GTTGAGCATCTCATTGAATACACGAGCTCAATTGGACCCCTCGACAAATGC GGCAAGGTCAGGGCATCGTTGAGTTCTTGGAGATCCAGCATGTTTTCGACC GAGTCGGTTAGCGAGTTCGACTGTTAATCGTGAGAGAAAAACAGATTTATG ATAGAAGAACACAAGAGAGTCAAAAATATACTTAATGATACTAGGATCAAA/ ATTATGGAATAATCGAAAGCGATTAGTGAATAATCGATGGAGGATTAGCGA TTCAATGAGGGTGAAGGATGTTGTTGAGAGTATGGGATTAAGTATTGTTCC GAGCGGAGAGCAAGGAGCAGAGAGCGCGGAGCGACGAGCGGAAGACAGAAT CGACATTGATGCCTACCTTGAAGCTCTGAAGACAGAGGAGGCTGAGAAGAA TTACACTGTAGAGACTGACTGCTTGGACAACCTGGTCCAAGCAGTTGTCGG ACCAATGCTAGTAATGAACCCTGGTACTTGATAATTGCCGAGCTCATTGCT TGACAGCATTGGAGTTCCATGTCAAATATCCGAAGGTGAAGAGGACAAAGA TTTTAGAAATTGGGTAGCGGTAATCATCAAGTCTCAGCAATCGTT TGACCTCCTAGAAACTCTGGTTTGGGCGGATTTTTGAAAATGCGTTGAGCT CCTCCTGTCTGATGTGCGCATTTCCACGAGGGACGAATCAAATGGACTGGA TAACACATGTTGTCAAGGAAACAGGTTTGGCCCGAGTTGCATCGACGCACCG GGCCAACGCCGAGCTGTTTCTGCTTCGCGAAGCCATCAGCATGAACCGCTT

How assembly can fail perfection

Sequencing data

- Insufficient read depth
- DNA sequencing errors (biases)
 - · Lower read quality towards read 3' end
 - Prefers low-GC
 - Homopolymers (e.g. AAAAAAA or CCCCCC)
- Contaminant DNA
 - Bacteria, unexpected fungal species, host plant, etc.

4
-2-
-2-
-2- -2-
5
E
-2-
-2-
3
-2-
<u>5</u> 7
<u>19</u>
-2-
4
5
-2-
_2
T T T T G T G G T G G T G G G G G G G G

How assembly can fail perfection



Genome complexity

- Reads not long enough to span repeats
- Heterozygosity (if diploid)
 - e.g. confusion at highly homozygous regions

Assembly algorithms

- Tolerance for sequencing errors
- Ploidy-awareness



Why is it important to evaluate assembly quality?

Know how trustworthy your genome assembly is!

If errors go unchecked, they can be

propagated downstream

next steps?

scaffolding

annotation

curation

misinterpreted as true biological events (e.g. a deletion)!

Users can make careful decisions Short/long reads revisit library prep? Contig assembly Scaffold/chromosome assembly

First – BLAST check your assembly!

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.



Web BLAST



First – BLAST check your assembly!

Puccinia striiformis f. sp. tritici (wheat stripe rust fungi)





BLAST to detect and remove contaminant (e.g. wheat, bacteria) & mtDNA contigs.



Herbaspirillum genome fully assembled along with the fungal genome!

The "3C" rules for assmembly QC

Contiguity

Completeness

Correctness





What we want: Fewer and longer contigs, chromosome-scale

Fragmented

Chromosome-scale



Contiguity











N50

Sequence <u>length</u> of the shortest contig at 50% of the assembly length. Generally, the higher the better.

L50

Smallest <u>count</u> of contigs whose length sum make up 50% of the assembly. Generally, the lower the better.





50% of total length is contained within sequences of at least 8 (25+10+10+8 = 53, >= 50)





Nx curve

QUAST

Quality Assessment Tool for Genome Assemblies by CAB



Contiguity measured at repeats

- LTR retrotransposons are abundant in fungi (Muszewska et al. 2017)
- Assembly contiguity can be measured at repeats, e.g. LTR transposons

	Category	LAI
LTR Assembly Index (LAI)	Draft	$0 \leq LAI$
(Ou et al. 2018)	Reference	10 ≤ L/

< 10AI < 20 Gold $20 \leq LAI$

Limitations

- Less suitable for species with low LTR abundance
- Intrinsic degradation rate of LTRs varies between species
- General guide only, low LAI score =/= bad!

LTR retrotransposon-dominated region near P. striiformis f.sp. tritici mating type locus



Genome graph

Quick visualisation for a "feeling" of contiguity



Good sign:

Chromosome-scale contigs

Not so good sign:

- Chromosomes broken into small contigs
- Ambiguous contig connections



. . .







Assembly size Gene space completeness • BUSCO • Reference-based

Telomere counting

Assembly gap

Assembly size vs Expected

Assembly size

Expected genome size

Compare the assembly size to expected genome size

Very rough estimation

Limitations

- Empirical evidence can have inaccuracies
- Genome size can vary between species, even genotypes
- Contaminant DNA can inflate it

BUSCO: Benchmarking Universal Single-Copy Orthologs

Evaluate "core" gene content using a predefined set of highly conserved orthologs





lineages

- agaricales_odb10
- agaricomycetes_odb10
- archaea_odb10
- ascomycota_odb10
- bacillales_odb10
- bacteria_odb10
- endopterygota_odb10
- eudicots_odb10
- eukaryota_odb10
- fungi_odb10
- hemiptera_odb10
- insecta_odb10
- mammalia_odb10
- metazoa_odb10
- methanococcales_odb10
- natrialbales_odb10
- primates_odb10
- saccharomycetes_odb10
- solanales_odb10

Fungi datasets Odb10

Phylum level: Ascomycota Basidiomycota Microsporidia Mucoromycota

Kingdom: <u>Fungi Odb10</u>

Class level: <u>Agaricomycetes</u> <u>Dothideomycetes</u> <u>Eurotiomycetes</u> <u>Leotiomycetes</u> <u>Saccharomycetes</u> <u>Sordariomycetes</u> <u>Tremellomycetes</u>

Order level:

Agaricales Boletales Capnodiales Chaetothyriales Eurotiales Glomerellales Helotiales Hypocreales Mucorales Onygenales Pleosporales Polyporales

BUSCO lineage dataset for Basidiomycota

Complete and single-copy BUSCOs (S)

Complete and duplicated BUSCOs (D)

Fragmented BUSCOs (F)

Total BUSCO groups searched

Missing BUSCOs (M)

71

1563

16

114

1764

BUSCO version is: 5.5.0
The lineage dataset is: basidiomycota_odb10 (Creation date: 2024-01-08,
number of genomes: 133, number of BUSCOs: 1764)
Summarized benchmarking in BUSCO notation for file /media/ssd/rita/proj
ect/104e/assembly_versions/v3.9_gapfill/v3.9.chr.haplotype-paired.fasta
BUSCO was run in mode: euk_genome_met
Gene predictor used: metaeuk
***** Results: *****
C:92.6%[S:4.0%,D:88.6%],F:0.9%,M:6.5%,n:1764
1634 Complete BUSCOS (C)

BUSCO completeness results

BUSCO limitations

- "Completeness" inferred from a small gene subset
- Some species can lose conserved genes and yield lower BUSCO score, even if well-assembled
- Genes with long introns are hard to detect
- Better option: annotate genes with RNA-seq evidence, then assess BUSCO

Gene space completeness – reference-based

Reference-based

If previously annotated genes are available, can count full and partial genes in your draft assembly

QUAST gene counting output

	assembly	genes	partial genes
draft assembly V1 draft assembly V2	verkko_duplex_assembly verkko_herro_assembly	31032 31058	6 2

Yunshen Liu

Telomeres

- "Endpoints" of a chromosome assembly
- Telomere-to-telomere assembly now highly achievable, thanks to long reads



Article Open access Published: 30 May 2024

Complete telomere-to-telomere genomes uncover virulence evolution conferred by chromosome fusion in oomycete plant pathogens

Zhichao Zhang, Xiaoyi Zhang, Yuan Tian, Liyuan Wang, Jingting Cao, Hui Feng, Kainan Li, Yan Wang, Suomeng Dong, Wenwu Ye 🖾 & Yuanchao Wang 🎦

RESOURCE ANNOUNCEMENT

A Telomere-to-Telomere Genome Assembly Resource of *Bipolaris sorokiniana*, the Fungal Pathogen Causing Spot Blotch and Common Root Rot Diseases in Barley and Wheat

Yueqiang Leng, Yang Du, Jason Fiedler, Sajeet Haridas, Igor V. Grigoriev, and Shaobin Zhong 🖂

Affiliations \lor

Published Online: 17 Jan 2024 https://doi.org/10.1094/PHYTOFR-08-23-0108-A

Article Open access Published: 14 September 2022

Telomere-to-telomere genome sequence of the model mould pathogen *Aspergillus fumigatus*

Paul Bowyer ⁽²⁾, Andrew Currin, Daniela Delneri ⁽²⁾ & Marcin G. Fraczek ⁽²⁾

Telomere counting

Assembled contigs



##########

356 sequences to analyze for telomeric repeats (TTAGGG/CCCTAA) in file assembly.raw.fasta ############

haplotype1-0000001	Forward	(start of sequence)	CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype1-0000001	Reverse	(end of sequence)	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
haplotype1-0000002	Forward	(start of sequence)	CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
haplotype1-0000002	Reverse	(end of sequence)	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000003	Forward	(start of sequence)	ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000003	Reverse	(end of sequence)	AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
haplotype1-0000004	Reverse	(end of sequence)	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000005	Forward	(start of sequence)	CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
haplotype1-0000006	Forward	(start of sequence)	ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000006	Reverse	(end of sequence)	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000007	Forward	(start of sequence)	AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
haplotype1-0000007	Reverse	(end of sequence)	TTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
haplotype1-0000008	Forward	(start of sequence)	CTAAACCCTAAACCCTAACCCTAAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000008	Reverse	(end of sequence)	TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA
haplotype1-0000009	Forward	(start of sequence)	CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype1-0000009	Reverse	(end of sequence)	GTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000010	Forward	(start of sequence)	CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
haplotype1-0000010	Reverse	(end of sequence)	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
haplotype1-0000011	Forward	(start of sequence)	ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
haplotype1-0000012	Forward	(start of sequence)	ACCCTAACCCTAACCCTAACCCTAACCCTAAACCCTAAACCCTA
haplotype1-0000014	Forward	(start of sequence)	CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACC
haplotype1-0000014	Reverse	(end of sequence)	GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
haplotype2-0000022	Forward	(start of sequence)	CCTAACCCTAAACCCTAACCCTAACCCTAAACCCTAACCCTAACCCTAAC
haplotype2-0000022	Reverse	(end of sequence)	GTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
haplotype2-0000023	Forward	(start of sequence)	CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
haplotype2-0000023	Reverse	(end of sequence)	GGGTTAGGGTTAGGGTTAGGGTTAGGGTATTAGGGTTAGGGACTTGAATT
haplotype2-0000025	Forward	(start of sequence)	CCCTAACCCTAACCCTAACCCTAACCCTCTAAACCCTAACCCTAACCCTA
haplotype2-0000025	Reverse	(end of sequence)	GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGG
haplotvpe2-0000026	Forward	(start of sequence)	ΤΑΑCCCTAACCCTAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA

Telomeres found: 64 (32 forward, 32 reverse)

Trap: a problematic read causing extension beyond telomere

				3 112 hn			
	5,601,000 bp		5,602,000 bp	5,112.00	5,603,000 bp		5.
			1			I	
plex to upphased barn Cover.	(p- 13)						
	-69-111 I 11						
		1 1					
plex_to_unphased.bam							
		I					
	D-13	I					
uplex_to_unphased.2ndno.bam rage	6. ul	and the second second					
		1 11					
		1					
		I					
plex_to_unphased.2ndno.bam		I					
		I ++					
		I I+++					
		I					
_							
equence 🍑							
				1991 AURUN 1991 AURUN 1994 AURUN 1			
							$ \rightarrow $
		Actual telomere			Where telomere coun	iter wants to find telom	ere, but f



Assembly gap "unknown" nucleotides

- Assembly gap (Ns) should be introduced after scaffolding (not covered in this workshop)
- But there are assemblers that might do some for you (e.g. verkko)



Gaps between contigs \rightarrow filled with NNNNN...

Assembly gap at ribosomal DNA array

Highly repetitive tandem repeats, identical copies

-

rDNA

chr17



Correctness

SNP/indels Misjoins Mapping statistics k-mer based consensus quality



SNP/indels

- Map reads back against the assembly
- Ideally should be completely identical
- Spot errors by calling SNP/indels

417.000 bp 418.000 bp 418.000 bp 410.000 bp 420.000 bp	
	421,000 bp
dpier.x3.9.ham Coverage P-AI Image: Aim of the second sec	

Structural errors

- Increase in coverage unresolved repeats/duplications
- Drop in coverage misjoins, misassemby

When ultralong reads come to rescue



WGSCoveragePlotter

Whole genome coverage plotter

Ultralong ONT reads filled it up!

homopolymer

k-mer based assembly evaluation

Merqury

Find concordance between your read dataset and assembly

Reports k-merquality values (QV)

v3.9.chr.haplotype-paired 8597 152306908 57.3972 1.82086e-06 duplex_merqury.qv (END)







- Never assume your assembly is "error-free"
- Know how trustworthy your assembly is for any analysis
- The "3C" rules
 - Contiguity
 - Completeness
 - Correctness

"Is my assembly ready for scaffolding?"

"Is my assembly ready for gene annotation?"

Wanna compare multiple assemblies and choose the best one?

Try QUAST circo plot!

Different read data types + assembler tool

- 1: ONT reads + Verkko
- 2: error-corrected ONT reads + Verkko
- 3: ONT reads + Hifiasm (failed run)
- 4: error-corrected ONT reads + Hifiasm (failed run)



Self alignment dotplot to find repetitive regions Gepard, Chromeister



Last important tip

Keep track of your assembly versions as you curate it!

assembly_version_history.txt

raw: assembly.raw.fasta
v3.1: assembly.extend_telomere.mtDNA_lowcov_rm.fasta
v3.2: telomeres extended. telomeric motif copy number corrected for all chromosomes.
v3.2.1: further telomere correction of v3.2 after coverage check. extension stops when last supported by at least two reads (2 simplex, or simplex 1 duplex.)
v3.3: after HiC scaffold. (v3.2.1.FINAL.fasta output from 3d-dna) orientation swapped to make sure p arm starts first.
v3.4: replace chr5A with v2.6 manually scaffolded chr5A. telomeres extended.
v3.5: fixed chr17B gap associated with a ~500bp GAAAA tandem repeat.
v3.5.1: fixed misassembly before the 5kbp rDNA-associated gap Ns. (Flye t assemble HiC_scaffold_44,45,46 into a high quality local assembly. mapped