Genome Sequencing & Assembly

Deb Triant
University of Virginia, USA
Dept. Biochemistry & Molecular Genetics

Programming for Biology
Cold Spring Harbor Labs
23 October 2023

1

Lecture outline

1. General background & theory of genome assembly
2. Comparison of sequencing technologies
3. Assembly quality and annotation
5. Assembly workshop with Python programming

2

2

Sequencing DNA

Objective: determine sequence of nucleotides in organism or DNA molecule

- Need many copies of your sequence
- Requires shearing your DNA into fragments
- Sequence fragments - "reads"
- Assemble original sequence from reads
  - How many possibilities exist?

3

3

History of Genome Assembly

1977. Sanger *et al.* 1st Complete Organism bacteriophage 5375 bp

1995. Fleischmann *et al.* 1st Free Living bacteria; *Haemophilus influenzae;* TIGR Assembler. 1.8Mb

1998. *C.elegans* SC 1st Multicellular Organism BAC-by-BAC Phrap. 97Mbp

2000. *Drosophila genome*; Myers *et al.* 1st Large WGS Assembly Celera Assembler. 116 Mbp

Human Genome

Public: 13-year project began 1990, Dept Energy & NIH,
$3 billion; millions of small fragments
2003 – announced as complete

Private: Craig Venter, Celera Genomics; 1998, $300 million
Could not be patented.

Human genome "finished" ~2003

4

4

History of Genome Assembly

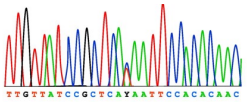
- Human genome 99% finished
  - heterochromatic regions
  - centromeres, telomeres
  - tandem duplications
  - Repeats!!

5

5

History of Genome Assembly

- First software to analyze Sanger sequencing, Roger Staden, 1979



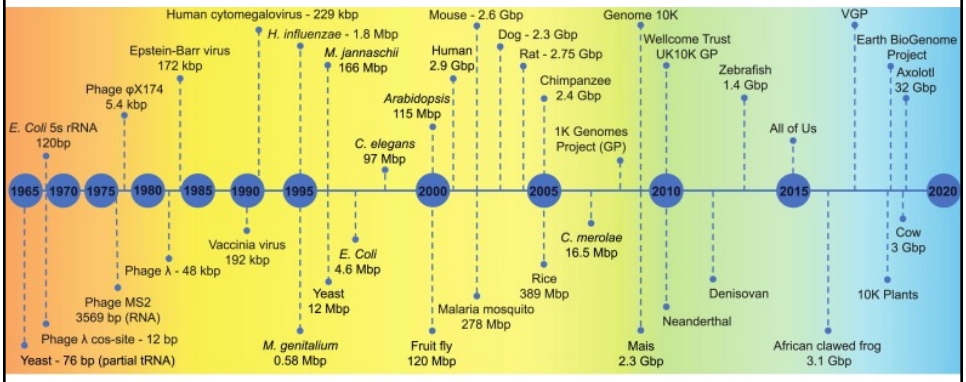
~800 bp reads, low error rate, costly

"If the 5′ end of the sequence from one gel reading is the same as the 3′ end of the sequence from another the data is said to overlap. If the overlap is of sufficient length to distinguish it from being a repeat in the sequence the two sequences must be contiguous. The data from the two gel readings can then be joined to form one longer continuous sequence."

6

6

History of Genome Assembly



Sanger, NGS, TGS

Giani et al. 2020

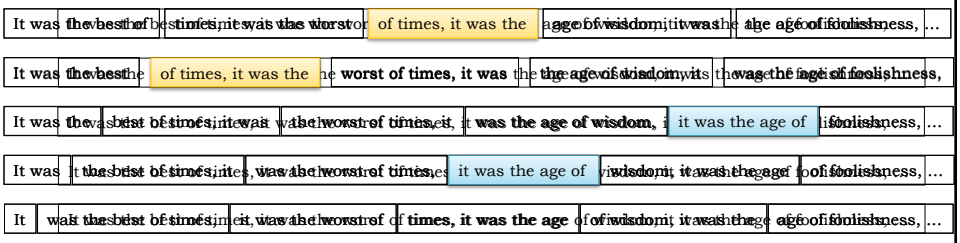
7

7

Shredded Book Reconstruction

**Based on example by Michael Schatz

- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



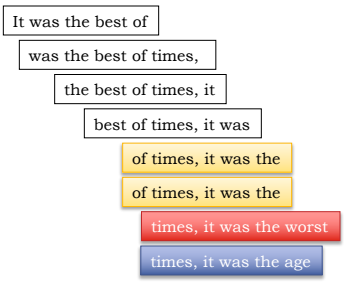
- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

8

8

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

Shredded Book Reconstruction



The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

9

9

Graph Construction

- Graph representing overlaps between subfragments
- Best match between end of one read and beginning of another
- Often not an exact match - sequencing errors

Original Fragment

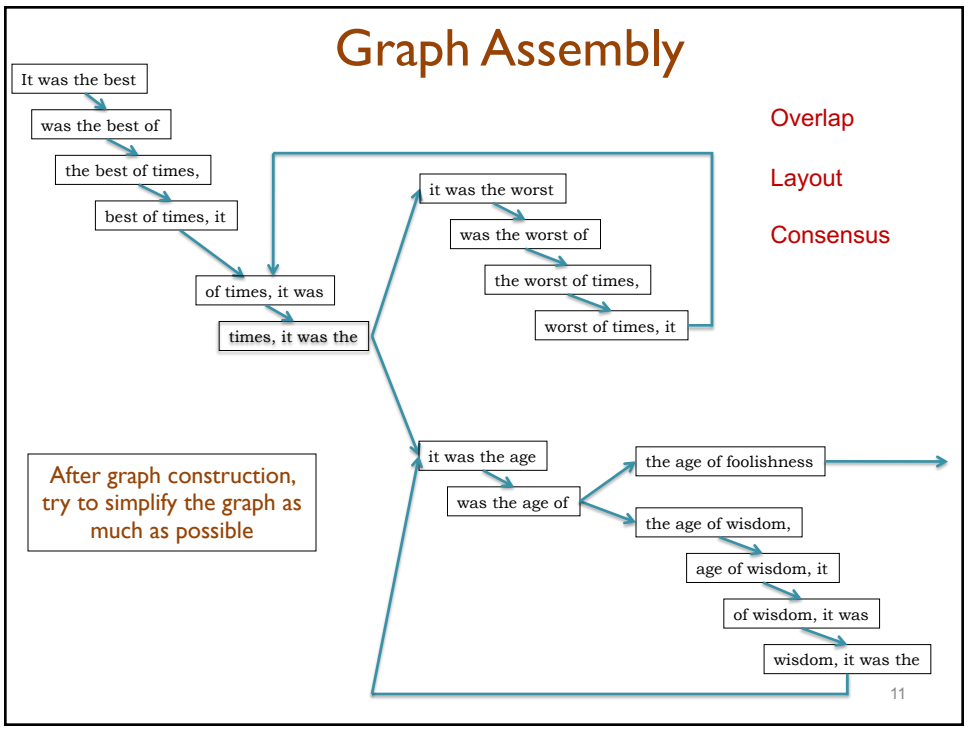
It was the best of

Directed graph with overlaps between subfragments

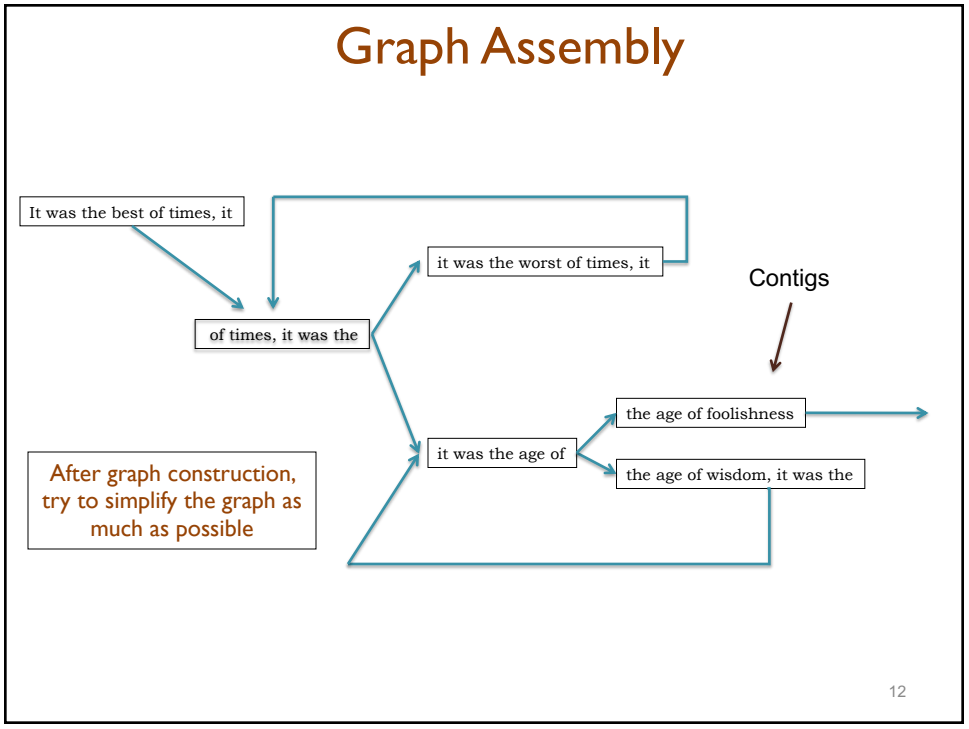
- Construct optimal alignment
  - pairs of reads that share common subfragments (kmers)
  - identify shortest path between overlaps
  - repeats - shortest might not be correct path

10

10


Graph Assembly
It was the best
was the best of
the best of times,
best of times, it
of times, it was
times, it was the
it was the worst
was the worst of
the worst of times,
worst of times, it
it was the age
was the age of
the age of foolishness
the age of wisdom,
age of wisdom, it
of wisdom, it was
wisdom, it was the
Overlap
Layout
Consensus
After graph construction, try to simplify the graph as much as possible
11

11


Graph Assembly
It was the best of times, it
of times, it was the
it was the worst of times, it
Contigs
it was the age of
the age of foolishness
the age of wisdom, it was the
After graph construction, try to simplify the graph as much as possible
12

12

13

13

Assembly outline

Overlap → Layout → Consensus

1. Overlap: identify overlapping reads
2. Layout: order reads
3. Consensus: merge reads into a sequence while correcting errors

14

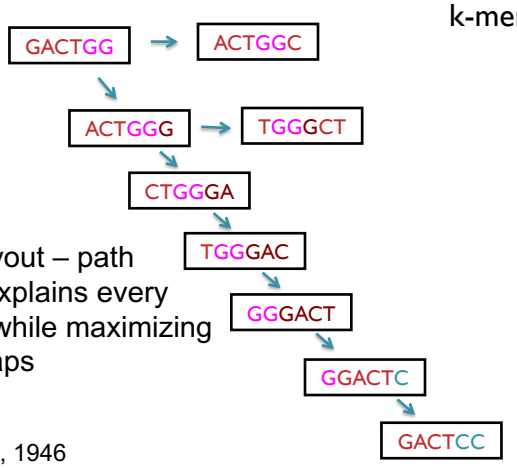
14

Graph Assembly

1. Overlap
   - Shredded words → k-mer

GACTGGGACTCC

k-mer - substring of length *k*
k = 6



2. Layout – path that explains every read while maximizing overlaps

de Bruijn, 1946

15

Graph Assembly

3. Consensus
   - sequence alignments to represent the same position

G TGGGACTCCGGCATTAGC TA
GACTCAGACTCCGGCATTAGCCTA
GACTGGGACTCCGGCATTAGCCTA
GACTGGG TCCGGCATTCGCCTA
GACTGGGACTCCGGCATTAGCCTA
GACTGGGACTCCGGCATTA CCTA

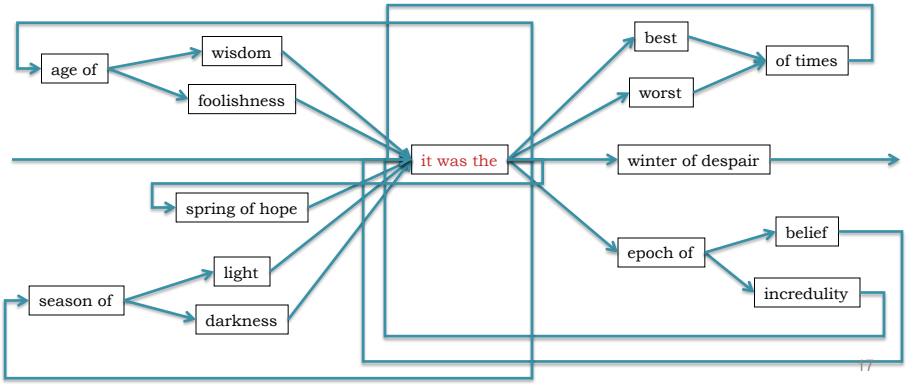
GACTGGGACTCCGGCATTAGCCTA

16

16

The full tale

… 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 epoch of belief it was the epoch of incredulity …
… it was the season of light it was the season of darkness …
… it was the spring of hope it was the winter of despair …

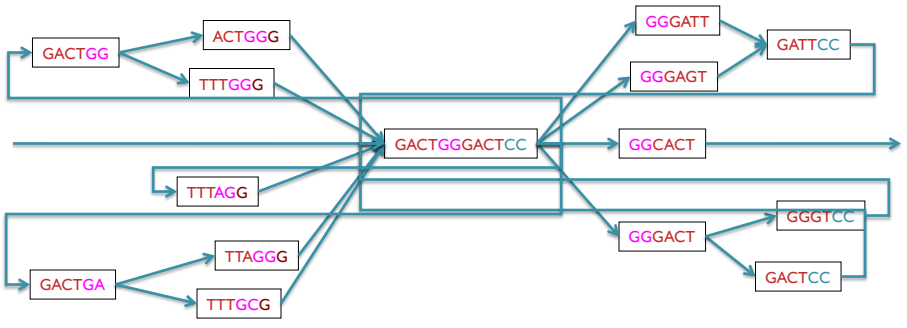


17

17

Graph Assembly

TAGACTGGGACTCCAG
AAGACTGGGACTCCGG
GGACTGGGAGTCCCTG
CGTTTGGGGGTCCTTA

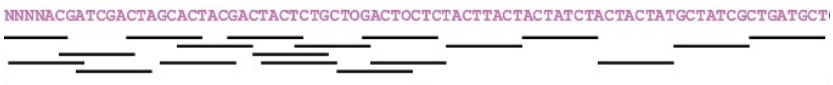


18

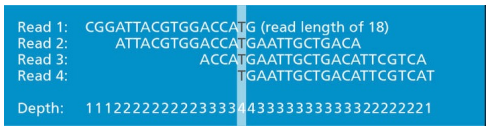
18

Considerations for assembly projects

Coverage



- How many times has the genome been sequenced?
  - Number of reads that cover sequence in your assembly
  - Too little? Few reads lead to sequencing errors



- Too much? Aim for oversampling ~40-60x.

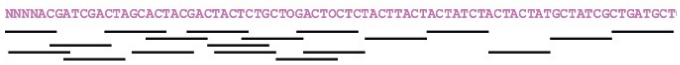
- Raw read depth vs Mapped read depth
  - -driven by efficiency of alignment process

19

19

Considerations for assembly projects

Coverage



Much is driven by funding and application
- SNPs and genome rearrangements
- structural variants
- particular coding regions
- 'complete' genome

Depth vs Coverage

Coverage - how deeply genome is sequenced, how many times each nucleotide represented

Depth - number of reads aligned to a particular location

20

20

Calculating coverage

- Lander-Waterman Model (1988)
  - Assumes reads randomly positioned in genome & same probability of covering each region of a genome

Coverage c = L * N/G

- L = Read length
- G = Genome size
- N = Number of reads

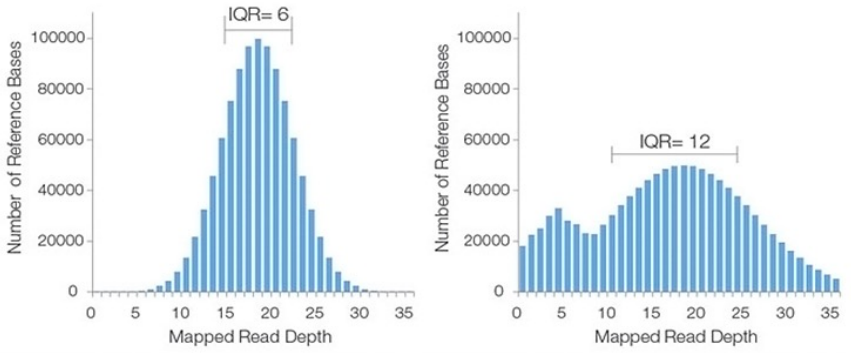
Can be modeled by a Poisson distribution

Many coverage calculators available online

21

21

Coverage histograms



Assumes reads randomly distributed across the genome

https://www.illumina.com/science/education/sequencing-coverage.html

22

22

Correction by coverage.
Figure S16, evaluates the ability of high-coverage to correct for sequencing errors. With sufficient coverage, even high error rates can be compensated for.

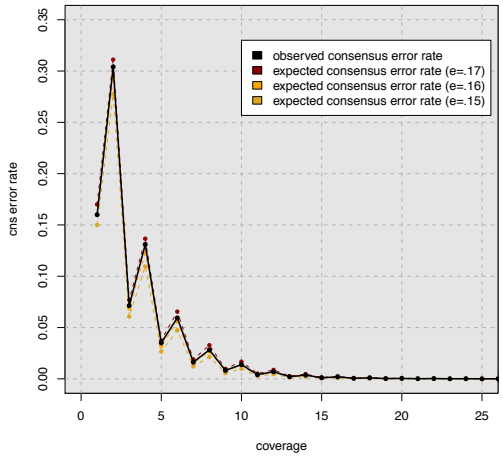


Figure S16. Coverage can overcome most random errors. 1,000 bp reads for *E. coli* K12 were simulated with random errors and the resulting consensus accuracy was measured. Even with high errors, coverage over 10X is sufficient to generate an accurate consensus. The periodic fluctuation in consensus error rate is an artifact of the tie-breaking scheme used in the consensus simulation (even numbers of reads can have ties and odd cannot).

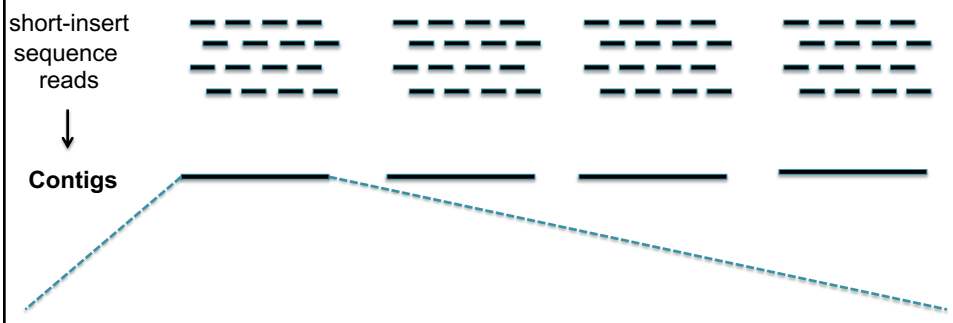
Hybrid error correction and de novo assembly of single-molecule sequencing reads.

Koren et al (2012) *Nature Biotechnology. doi:10.1038/nbt.2280*

23

23

Assembly construction - Hierarchical process

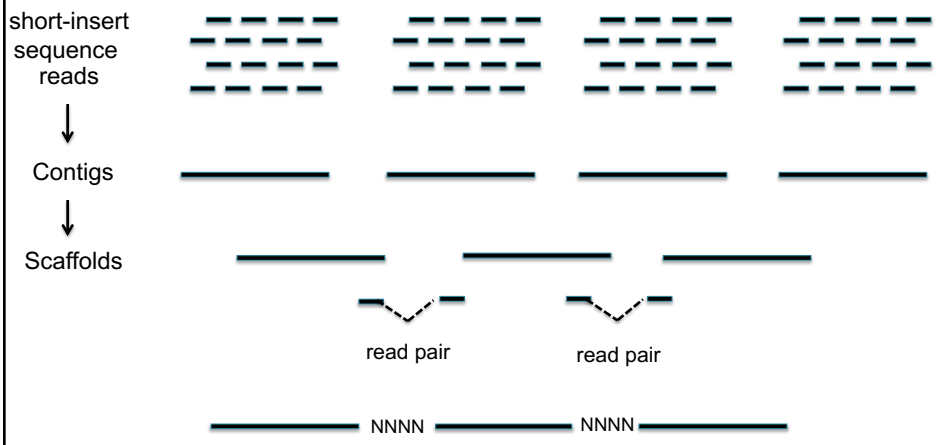


CCAAGTAGGGATCAAGGCCAGCCGGTGCGCTGAAAACCTGGAGTGCAGTTTCTGGGGTGGGACGTGTT

Contig - contiguous sequence of nucleotides

24

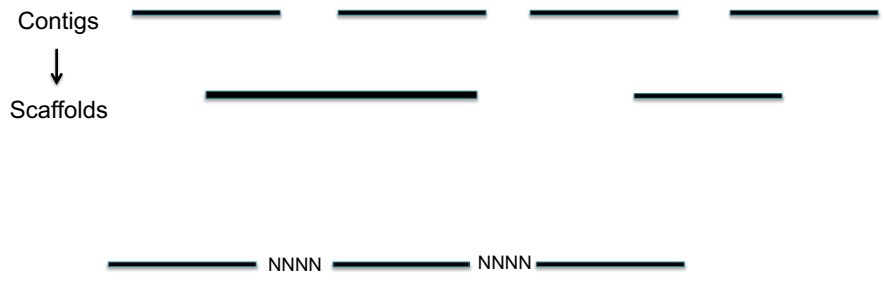
Assembly construction



Scaffold - sequence of contigs, separated by gaps - Ns are predicted gap size

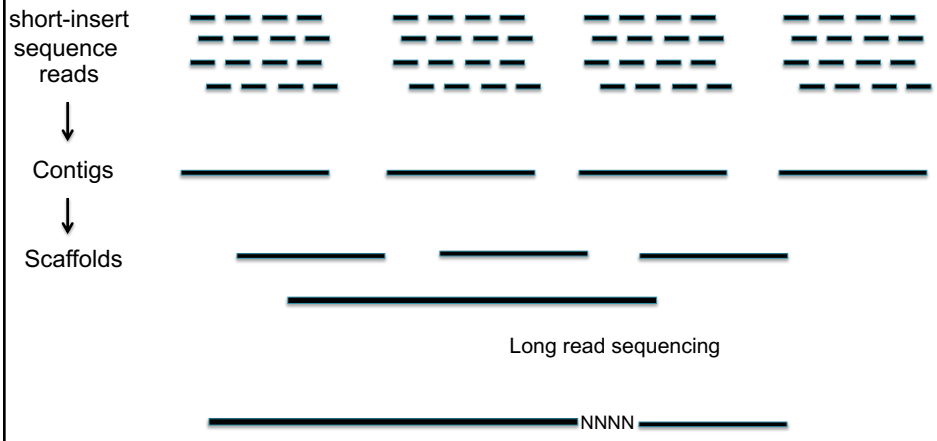
>GTAGTATTCTAGAAAATGTTTTAAATATAGATAGTTGTTTTANATCTGTTAGTTGTCAGATGCTTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN
NNNNNNTGTGAGGTTTTTAGCTCATGAAAGTTTATGATTATTGCACCCCTACTCACAACGAATCCCTATTCTTATCTTTTNNNNNNNNNNNNNNNCATGT
CACTGGTTTATTTATTTTTTGTGGCTGCAGAAGTCCTTTGTCGCTGTTAATTTTTTGGGAGTTCTCCTGTCGTATATTAAAGCTTTCTTTCTTTTCAGTTT
TAAATTTATTTGAACCTTACTATCTTTCTAACAATAAATTGTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTTATACGAAAGCTATATAG
TGTTAGGCTTTTCTTTTTTTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCCTTTTCTGGTAATCTTACTAAATCAGTTTGCTTGTTACTTGTATAGTTGT

Assembly construction



How to resolve long repeats?!

Assembly construction

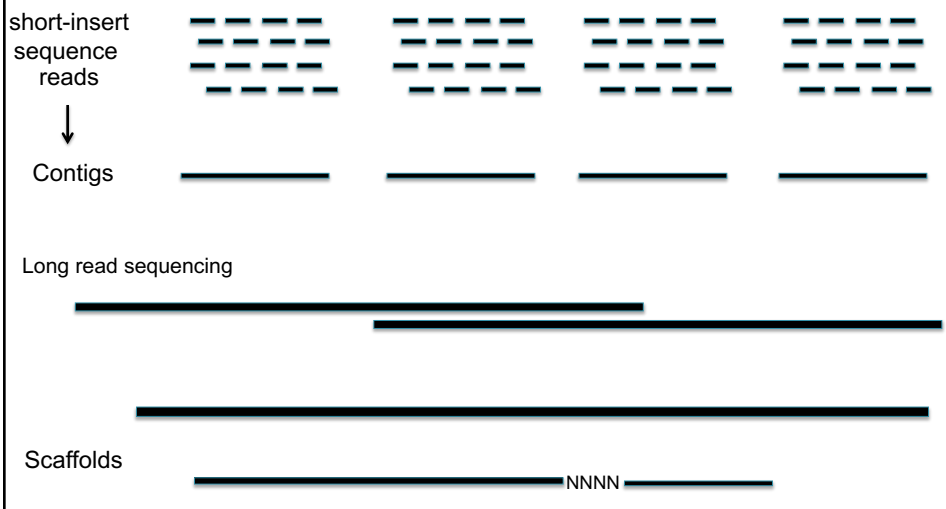


Scaffold - sequence of contigs, separated by gaps - Ns are predicted gap size

>GTAGTATTCTAGAAAATGTTTTAAATATAGATAGTTGTTTTANATCTGTTAGTTGTCAGATGCTTACTGAATAGTTGGAANNNNNNNNNNNNNNNNNN
NNNNNNTGTGAGGTTTTTAGCTCATGAAAGTTTATGATTATTGCACCCCTACTCACAACGAATCCCTATTCTTATCTTTTNNNNNNNNNNNNNNNCATGT
CACTGGTTTATTTATTTTTTGTGGCTGCAGAAGTCCTTTGTCGCTGTTAATTTTTTGGGAGTTCTCCTGTCGTATATTAAAGCTTTCTTTCTTTTCAGTTT
TAAATTTATTTGAACCTTACTATCTTTCTAACAATAAATTGTGGAATTATCAACGAAAACATAGGNNNNNNNNNNNGTCCTTTATACGAAAGCTATATAG
TGTTAGGCTTTTCTTTTTTTTTNNNNNNNNGGTGATGTTGTTAATGGTGCCCTTTTCTGGTAATCTTACTAAATCAGTTTGCTTGTTACTTGTATAGTTGT

27

Assembly construction

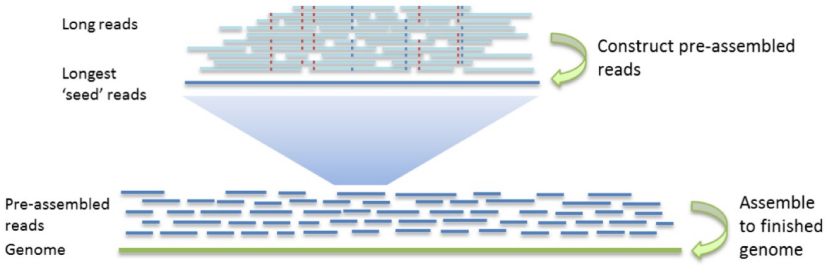


28

Assembly construction - long reads

Still hierarchical but proceeds in two rounds

1. Seed read selection – longest reads in dataset
2. Shorter reads aligned to seed reads for consensus



29

29

Repetitive regions

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - *Arabidopsis* - 10% composition
  - SINEs - Short Interspersed Nuclear Elements
  - LINES - Long Interspersed Nuclear Elements
  - LTR - Long Terminal Repeats, retrotransposons
  - Segmental duplications
  - Low-complexity - Microsatellites or homopolymers

30

30

Sequencing challenges

- Resolving repeats

- Repeats longer than reads

- Merge reads up to repeat boundaries

- Tandem repeats often not resolved

RepeatMasker - http://repeatmasker.org

screens DNA sequences for interspersed repeats and low complexity DNA sequences

- Error correction

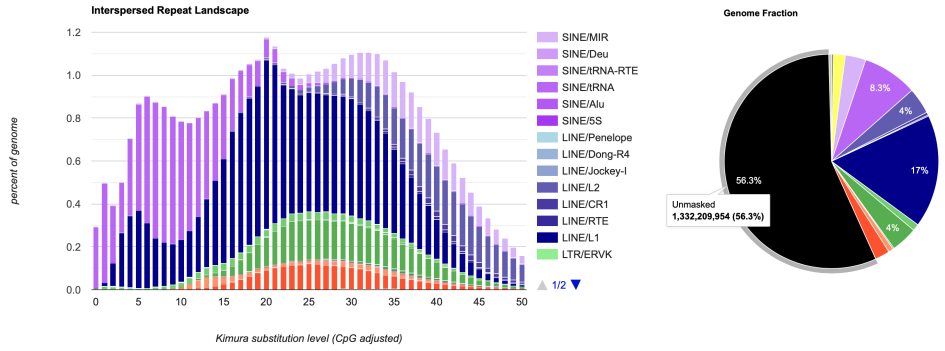
- incorporation of long reads

- correction algorithms

31

31

F. catus RepeatMasker output



NNNNNNNNNNNNcaaacaaatatacaaagacAAAAATTGCCACAGCAAAGACAAAGAGATAAATAAAAGGCACAAAATT

32

32

Lecture outline

1. General background & theory of genome assembly
2. Comparison of sequencing technologies
3. Assembly quality and annotation
5. Assembly workshop with Python programming

33

33

Sequencing Technologies

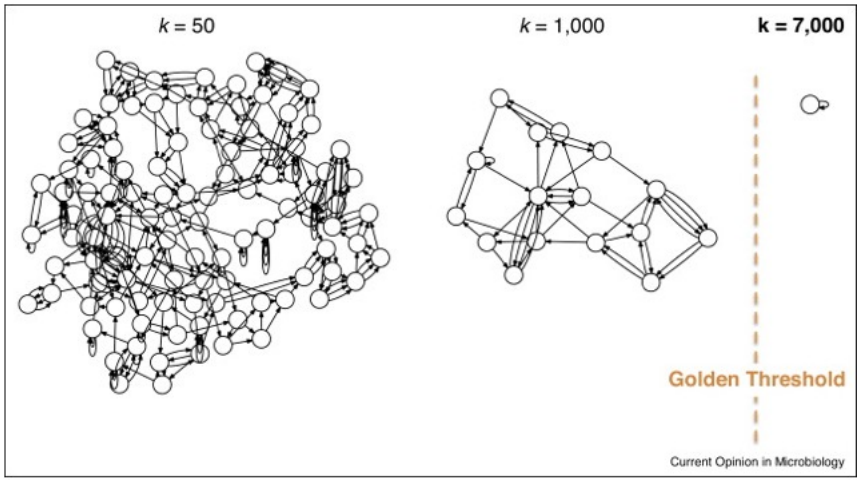
- Illumina
- PacBio
- Nanopore

Thank you Anoja!

34

34

Read lengths

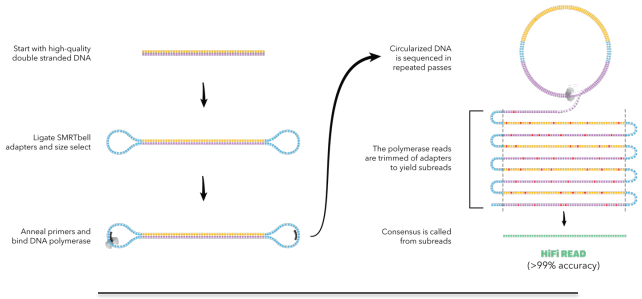


Koren & Phillippy 2015 Current Opinion in Microbiology

35

35

PacBio Hi Fi reads



- Long reads - 20kb
- 99.9% read quality (Q30) - better than some short-read contigs
- Distinguish repeats rather than spanning full repeats

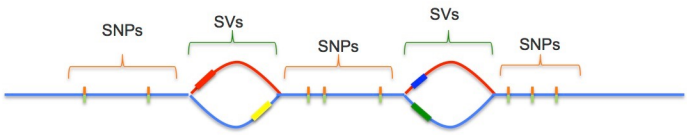
- Limited length
- Lower throughput than Nanopore

Wenger et al. 2019 *Nature Biotechnology*

36

36

Phased genome assemblies



"Pseudo-haplotypes" - mixture of both types

How do we get phased genomes?

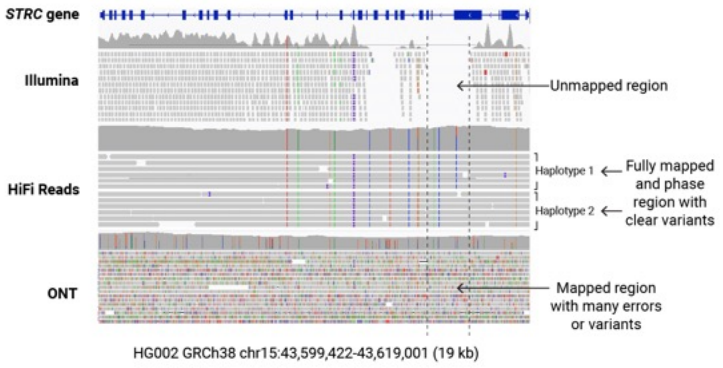
Higher heterozygosity - easier to phase.

Chin et al. *Nat Meth* 2016

37

37

PacBio Hi Fi reads



HiFi reads provide the accuracy needed to call single nucleotide variants, while improving mappability and enabling phasing with no systematic bias. STRC gene alignments from *Genome in a Bottle* (GIAB), HG002_NA24385_son. (IGV settings)

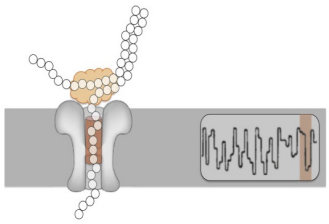
https://www.pacb.com

38

38

Nanopore sequencing

- "Ultra-long" reads
- Up to 1Mb read length
- 95% read quality (Q13)



- Long lengths
- Able to span repeat regions

- Limited quality
- Took a while for technology to catch up

Shafin et al. *Nature Biotechnology*, 2020

39

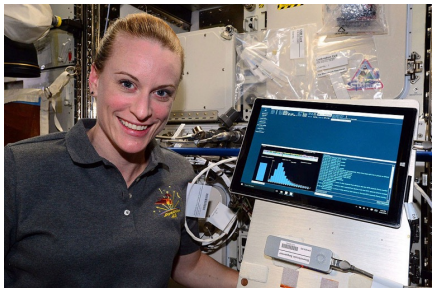
39

Nanopore sequencing

COMMENT Open Access

Mobile real-time surveillance of Zika virus in Brazil

Nuno Rodrigues Faria[1], Ester C. Sabino[2], Marcio R. T. Nunes[3,4], Luiz Carlos Junior Alcantara[5], Nicholas J. Loman[6*] and Oliver G. Pybus[1]

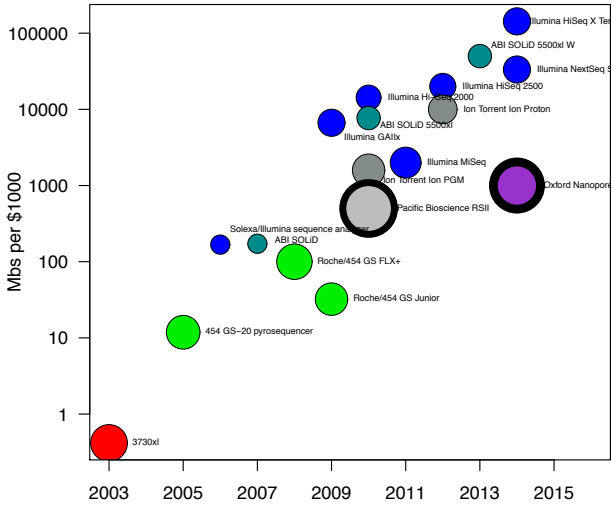


Kate Rubins, ISS

Castro-Wallace, Nature Scientific Reports, 2017

40

40



(http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-186.pdf)

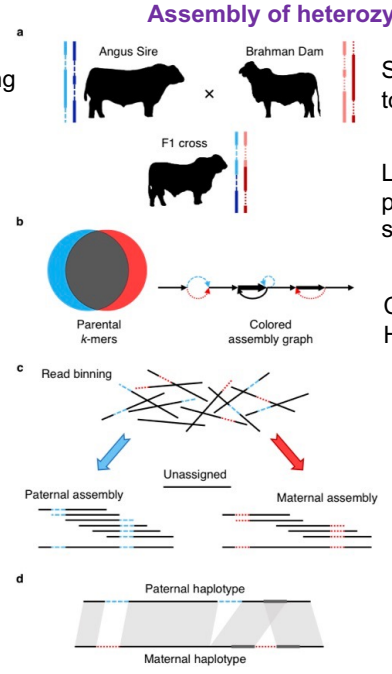
Border widths indicate error rates

41

41

Assembly of heterozygous genomes

Trio-binning



Short reads from two parental genomes to identify unique k-mers - breed specific

Long reads from offspring partitioned into into haplotype-specific sets prior to assembly

Consistent path through each haplotype Homozygotes rep. twice

Each haplotype assembled separately to represent diploid assembly

Koren et al. 2018. Nature Biotechnology 42

42

History of Genome Assembly

1977. Sanger *et al.* 1st Complete Organism bacteriophage 5375 bp

1995. Fleischmann *et al.* 1st Free Living bacteria; *Haemophilus influenzae;* TIGR Assembler. 1.8Mb

1998. *C.elegans* SC 1st Multicellular Organism BAC-by-BAC Phrap. 97Mbp

2000. *Drosophila genome*; Myers *et al.* 1st Large WGS Assembly Celera Assembler. 116 Mbp

Human Genome

Public: 13-year project began 1990, Dept Energy & NIH,
$3 billion; millions of small fragments
2003 – announced as complete

Private: Craig Venter, Celera Genomics; 1998, $300 million
Could not be patented.

Human genome "finished" ~2003

43

43

Sequencing Technologies

Combine technologies to improve assemblies.

Article

Telomere-to-telomere assembly of a complete human X chromosome

https://doi.org/10.1038/s41586-020-2547-7
Received: 30 July 2019
Accepted: 29 May 2020
Published online: 14 July 2020
Open access
Check for updates

Karen H. Miga[1,24✉], Sergey Koren[2,24], Arang Rhie[2], Mitchell R. Vollger[3], Ariel Gershman[4], Andrey Bzikadze[5], Shelise Brooks[6], Edmund Howe[7], David Porubsky[3], Glennis A. Logsdon[3], Valerie A. Schneider[8], Tamara Potapova[7], Jonathan Wood[9], William Chow[9], Joel Armstrong[1], Jeanne Fredrickson[10], Evgenia Pak[11], Kristof Tigyi[1], Milinn Kremitzki[12], Christopher Markovic[12], Valerie Maduro[13], Amalia Dutra[11], Gerard G. Bouffard[6], Alexander M. Chang[2], Nancy F. Hansen[14], Amy B. Wilfert[3], Françoise Thibaud-Nissen[8], Anthony D. Schmitt[15], Jon-Matthew Belton[15], Siddarth Selvaraj[15], Megan Y. Dennis[16], Daniela C. Soto[16], Ruta Sahasrabudhe[17], Gulhan Kaya[16], Josh Quick[18], Nicholas J. Loman[18], Nadine Holmes[19], Matthew Loose[19], Urvashi Surti[20], Rosa ana Risques[10], Tina A. Graves Lindsay[12], Robert Fulton[12], Ira Hall[12], Benedict Paten[1], Kerstin Howe[9], Winston Timp[4], Alice Young[6], James C. Mullikin[6], Pavel A. Pevzner[21], Jennifer L. Gerton[7], Beth A. Sullivan[22], Evan E. Eichler[3,23] & Adam M. Phillippy[2✉]

44

44

Human genome completion

- 120x Nanopore
- 70x PacBio
- 30x PacBio HiFi
- 50x 10X Genomics
- 100x Illumina
- 35x Arima Hi-C
- BioNano optical map
- PacBio Iso-Seq

- Telomere-to-Telomere
- Centromeres resolved
- Previous gaps filled

Article

The complete sequence of a human Y chromosome

Nature, September 2023

45

45

Genome Assembly Projects

Bhattacharya et al. 2018.
Genome Research

Wheat - hexaploid, 15.3 billion bases
Axolotl salamander - 32 billion

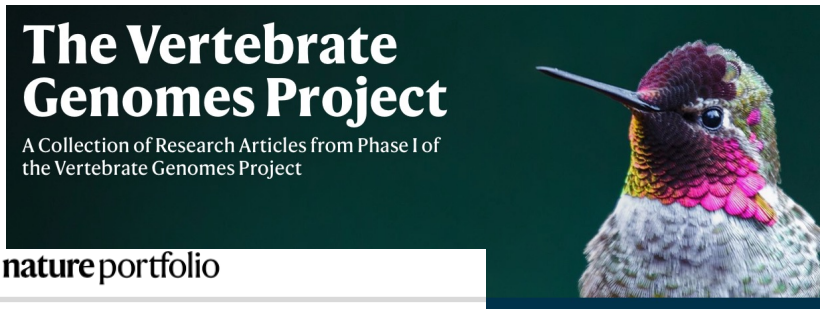
46

46

Assembly Projects

GENOME 10K

i5K



nature > special

Special | 28 April 2021

Vertebrate Genomes Project

47

47

Pan genomes

> Cell. 2020 Jul 9;182(1):162-176.e13. doi: 10.1016/j.cell.2020.05.023. Epub 2020 Jun 17.

Pan-Genome of Wild and Cultivated Soybeans

Yucheng Liu [1], Huilong Du [2], Pengcheng Li [3], Yanting Shen [4], Hua Peng [2], Shulin Liu [5], Guo-An Zhou [5], Haikuan Zhang [3], Zhi Liu [1], Miao Shi [3], Xuehui Huang [6], Yan Li [7], Min Zhang [5], Zheng Wang [5], Baoge Zhu [5], Bin Han [7], Chengzhi Liang [8], Zhixi Tian [9]

Affiliations + expand

PMID: 32553274 DOI: 10.1016/j.cell.2020.05.023

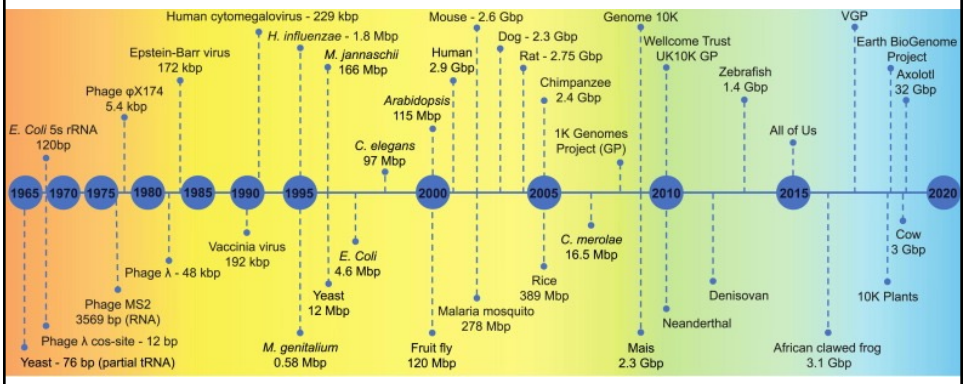
26 denovo assemblies
3 previously reported assemblies

Searching for variants that were undetectable in single reference genome

48

48

History of Genome Assembly



Sanger, NGS, TGS

Giani et al. 2020

49

History of Human Genome Assembly

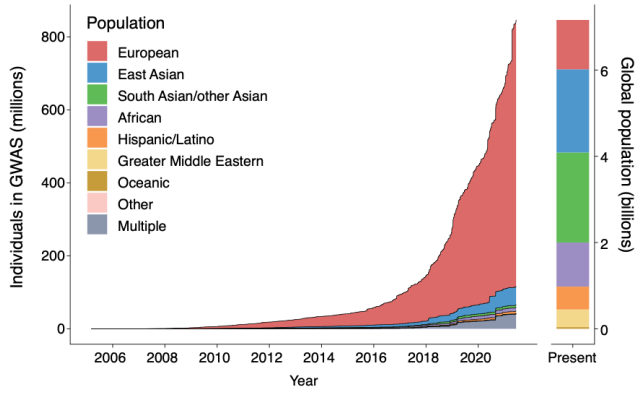


Fig. 1 | The proportion of samples from individuals cumulatively reported by the GWAS Catalog[1] as of 8 July 2021.

Fatumo, Nature Medicine, 2021

50

History of Human Genome Assembly

nature medicine
SERIES | PERSPECTIVE
https://doi.org/10.1038/s41591-021-01672-4
Check for updates

A roadmap to increase diversity in genomic studies

Segun Fatumo[1,2], Tinashe Chikowore[3,4], Ananyo Choudhury[3], Muhammad Ayub[5], Alicia R. Martin[6,7] and Karoline Kuchenbaecker[5,8]

Article

The GenomeAsia 100K Project enables genetic discoveries across Asia

https://doi.org/10.1038/s41586-019-1793-z
Received: 29 January 2018
Accepted: 11 October 2019
Published online: 4 December 2019
Open access

GenomeAsia100K Consortium*

The underrepresentation of non-Europeans in human genetic studies so far has limited the diversity of individuals in genomic datasets and led to reduced medical relevance for a large proportion of the world's population. Population-specific reference genome datasets as well as genome-wide association studies in diverse populations are needed to address this issue. Here we describe the pilot phase of the GenomeAsia 100K Project. This includes a whole-genome sequencing reference dataset from 1,739 individuals of 219 population groups and 64 countries across Asia. We catalogue genetic variation, population structure, disease associations and founder effects. We also explore the use of this dataset in imputation, to facilitate genetic studies in populations across Asia and worldwide.

Million-person U.S. study of genes and health stumbles over including Native American groups

By Jocelyn Kaiser | May. 29, 2019 , 1:40 PM

History of Human Genome Assembly

Christian Happi at Redeemer's University in Ede, Nigeria, plans to sequence human genomes.

Sequence three million genomes across Africa

Wonkam, Nature 2021

Sequenced 426 genomes with African ancestry and discovered 3 million unknown variants

Choudhury et al. Nature 2020

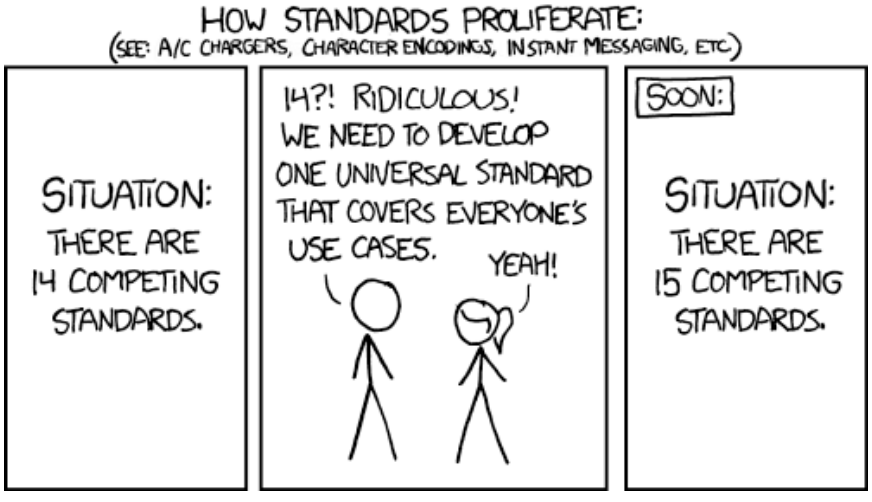
Lecture outline

1. General background & theory of genome assembly
2. Comparison of sequencing technologies
3. Assembly quality and annotation
5. Assembly workshop with Python programming

53

53

Data Formats



http://xkcd.com/927/

54

Assembly analysis

Base calling, quality control, trimming

- Most data returned in FASTQ format with quality scores included

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATA
+
!''*((((***+))%%%++)(%%%%).1***-
```

← id
← sequence
← description line
← base qualities

55

55

FASTQ

```
@M00747:32:000000000-A16RG:1:1112:15153:29246 1:N:0:1
TCGATCGAGTAACTCGCTGCTGTCAGACTGGTTTTTGGTCGATCGACTATTGTTTCAGTCGCAAGAAT
ATTGTGTCCAGTCGATCGACTGAATTCTGCTGTACGGCCACGGCGGATGCACGGTACAGCAGGCTCAG
ACGGATTAAACTGTT
+
5=9=9<=9,-5@<<55>,6+8AC>EE.88AE9CDD7>+7.CC9CD+++5@=-FCCA@EF@+***+*--
55--AA---AA-5A<9C+3+<9)4++=E=+===<D94)00=9)))2@624(/(/2/-
(.(6;9(((((.(.'((6-66<6(///
@M00747:32:000000000-A16RG:1:1112:15536:29246 1:N:0:1
GTAAAATTGAGGTAAATTGTGCGGAATTTAGCAATACCGTTTTTTTTATTATCACCGGATATCTATTC
TGCTGTACGGCCAAGGAGGATGTACGGTACAGCAGGTGCGAACTCACTCCGACGCTCAAGTCAGTGAC
TTAATGATAAGCGTG
+
?????<BBBBBB5<?BFFFFFFECHEFFECCFF?9AAC>7@FHHHHHHFG?EAFGF@EEDEHHDGHHC
BDFFGDFHF)<CCD@F,+3=CFBDFHBD++??DBDEEEDE:):CBEEEBCE68>?))5?**0?:AE*A
*0//:/*:*:**.0)
@M00747:32:000000000-A16RG:1:1112:15513:29246 1:N:0:1
GCTAGTCTTGTGTTTAGTTTTATGTTTTGCATGTTGTAACGGATTCATAAACATAGGTGTTTGTTTCT
TTTTATGGTTGTACAATTTGGCCCTAAGGCCCTACACTTACTTGTTTGTTTCTTTTATGGTACGACAT
TTGAGTGGTGGTTGA
+
```

56

56

FASTQ

```
@M00747:32:000000000-A16RG:1:1112:15153:29246 1:N:0:1
TCGATCGAGTAACTCGCTGCTGTCAGACTGGTTTTTGGTCGATCGACTATTGTTTCAGTCGCAAGAAT
ATTGTGTCCAGTCGATCGACTGAATTCTGCTGTACGGCCACGGCGGATGCACGGTACAGCAGGCTCAG
ACGGATTAAACTGTT
+
5=9=9<=9,-5@<<55>,6+8AC>EE.88AE9CDD7>+7.CC9CD+++5@=-FCCA@EF@+****--
55--AA---AA-5A<9C+3+<9)4++=E=+===<D94)00=9)))2@624(/(/2/-
(.(6;9(((((.(.'((6-66<6(///
@M00747:32:000000000-A16RG:1:1112:15536:29246 1:N:0:1
GTAAAATTGAGGTAAATTGTGCGGAATTTAGCAATACCGTTTTTTTTATTATCACCGGATATCTATTC
TGCTGTACGGCCAAGGAGGATGTACGGTACAGCAGGTGCGAACTCACTCCGACGCTCAAGTCAGTGAC
TTAATGATAAGCGTG
+
?????<BBBBBB5<?BFFFFFFECHEFFECCFF?9AAC>7@FHHHHHHFG?EAFGF@EEDEHHDGHHC
BDFFGDFHF)<CCD@F,+3=CFBDFHBD++??DBDEEEDE:):CBEEEBCE68>?))5?**0?:AE*A
*0//:/*:*:**.0)
@M00747:32:000000000-A16RG:1:1112:15513:29246 1:N:0:1
GCTAGTCTTGTGTTTAGTTTTATGTTTTGCATGTTGTAACGGATTCATAAACATAGGTGTTTGTTTCT
TTTTATGGTTGTACAATTTGGCCCTAAGGCCCTACACTTACTTGTTTGTTTCTTTTATGGTACGACAT
TTGAGTGGTGGTTGA
+
```


57

Assembly Quality Scores

Calculating Phred Quality Scores - Base calling accuracy

Q scores are defined as a property that is logarithmically related to the base calling error probabilities (P)2.

$Q = -10\log_{10}P$

Q - sequencing quality score of a given base Q

P - probability of base call being wrong

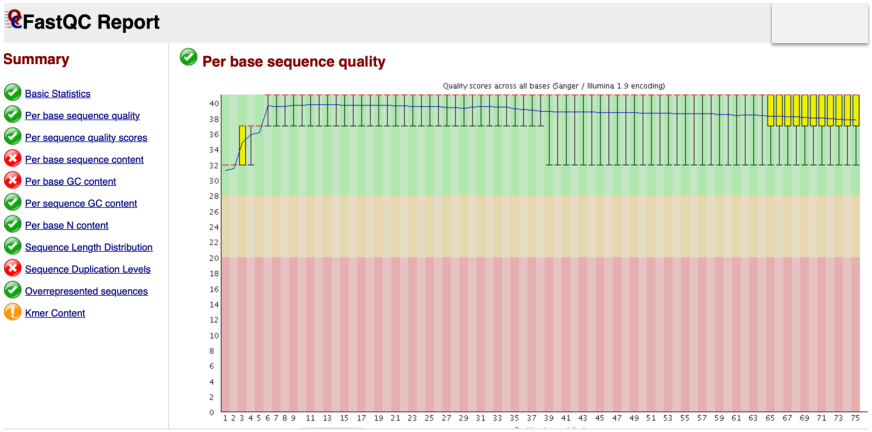
Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

https://www.illumina.com/Documents/products/technotes/technote_Q-Scores.pdf

58

Assembly analysis

Checking quality of reads - FASTQC

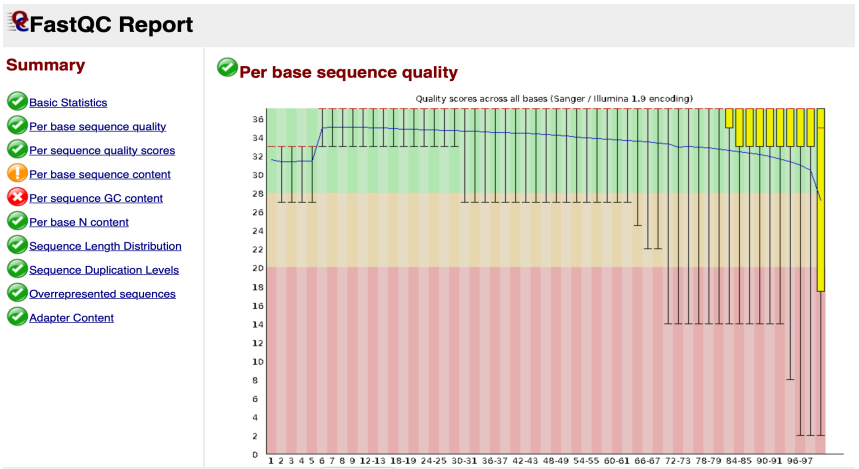


https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 59

59

Assembly analysis

Checking quality of reads - FASTQC

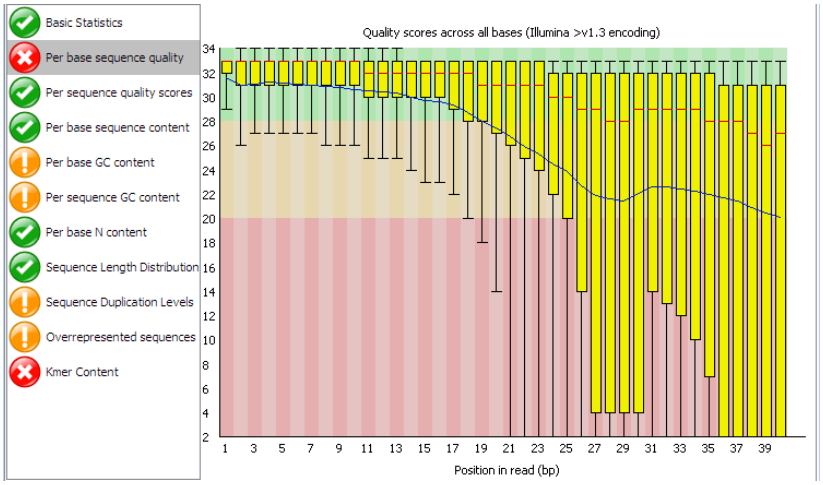


https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 60

60

Assembly analysis

Checking quality of reads - FASTQC



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ 61

61

Sequence trimming

- Trim reads for adapters and quality
- Adapter trimming depends on technology
- Quality trimming, Q20 common cut-off

Trimmomatic - Illumina data, Bolger et al. 2014
https://github.com/usadellab/Trimmomatic
DynamicTrim – SolexaQA - quality trimming
https://solexaqa.sourceforge.net
AfterQC – filter, trim, QC, Chen et al.
https://github.com/OpenGene/AfterQC

62

62

Contaminant removal

**Genome assemblies contain *only* genomic sequences from target organism

Contamination removal step often needed

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steinegger & Steven L. Salzberg

Genome Biology **21**, Article number: 115 (2020)

NCBI FCB - Foreign Contamination Screen

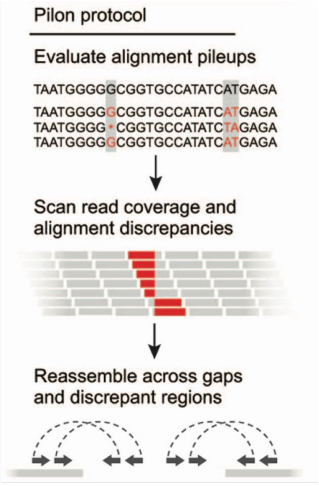
https://github.com/ncbi/fcs#readme

ContFree-NGS - https://github.com/labbces/ContFree-NGS
Kraken2 - https://github.com/DerrickWood/kraken2
HoCoRT - https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05492-w

63

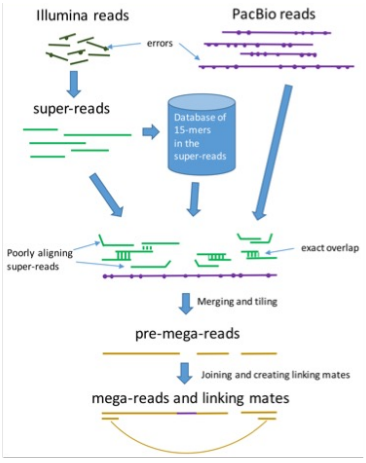
63

Genome assembly correction



Pilon: genome assembly improvement

Walker et al. PLoS ONE. 2014



MaSuRCA mega-reads algorithm

Zimin et al. Genome Res. 2017

64

64

Assessing genome quality

- Map raw reads back to assembled genome
  -mapping back uniformly
- SAM/BAM file
- BEDTools - to retrieve coverage statistics

https://github.com/arq5x/bedtools2

65

65

File formatting

- FASTA
- FASTQ
  – quality scores
- SAM/BAM
  – Developed for NGS data
  – Sequence Alignment Map
  – Stores alignment information

66

66

FASTA

```
>NP_000552.2 Human glutathione transferase M1 (GSTM1)
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGAH
KITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEFEKLKPKYLEELPEKLK
LYSEFLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPNLKDFISRFEGLEKISAYMKSSRFL
PRPVFSKMAVWGNK
```

67

67

SAM/BAM Sequence Alignment Map

- Alignment file - provides context for raw data
  - Eleven columns, tab delimited
  - One alignment record per line
- SAM is plain-text (human readable)
- BAM is a binary format
- SAMTools - suite of utilities for SAM/BAM files
- Picard - tools for sequencing data

samtools: http://samtools.sourceforge.net
Picard: https://broadinstitute.github.io/picard/

68

68

SAM/BAM Sequence Alignment Map

```
D4ZHLFP1:53:D2386ACXX:6:2115:17945:68812     0    Mle_000001     18   42   108M *    0    0
     TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTG
GCCAAGATGTACTGAGATGCAT
C@CFDFFFHHGHHHFGGBFEGGDGGGGEHGIGGGJJJJIIIGIIB9BFBBFHGGHICEAGHGEGEDHIGEEDBECCACBDDC@CCDBCDD<
?2+4>@4>>CCCCAA@@   AS:i:-5   XN:i:0    XM:i:1    XO:i:0    XG:i:0    NM:i:1    MD:Z:0A107
     YT:Z:UU
D4ZHLFP1:53:D2386ACXX:7:2110:5214:83081 0    Mle_000001     18   42   108M *    0    0
TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTGGCCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHHHGHGGEGIJIIIGJFHJJJJJIIJJIIGIJIJFHJJIIIJJIFIIIIIIJIIJJJHHFFFCEEEEDDDDDDDDDDDDD
BDCDDEEEDDDDDDDDD   AS:i:-5   XN:i:0    XM:i:1    XO:i:0    XG:i:0    NM:i:1    MD:Z:0A107
     YT:Z:UU
D4ZHLFP1:53:D2386ACXX:7:2206:9985:31556 0    Mle_000001     18   42   108M *    0    0
TCCCCCTGCATGTGTCCGTCTGGCTGGATGCCATGCTCCATGCAGTATAGCTCCCAGCATGAGTTACCGATCTGGACACCTGCTTGGCCAA
GATGTACTGAGATGCAT
CCCFFFFFHHHHHHJJJJIHJJIIIIJJIJJJJJJJJJIJJJJJJJIIJJJJJJJJJJJJJJJJJIJJJGEFFFEEEEEDDDDDDDDDDDDD
DDCD@@CDCCDDCDCDC   AS:i:-5   XN:i:0    XM:i:1    XO:i:0    XG:i:0    NM:i:1    MD:Z:0A107
     YT:Z:UU
```

Helpful site for looking up SAM flag: https://broadinstitute.github.io/picard/explain-flags.html

69

69

Genome annotation

- Describe genetic and genomic features in raw sequence.
  - genetics and genomic features, genes, repetitive elements, mobile elements, genome duplications
  - gene prediction, ORF searches, repeat region identification, homology searches
  - most genome projects use automated methods with some manual curation
  - community motivated annotation projects

70

70

Genome Assembly

- Recommended to use multiple assemblers with different parameters to assess results

- How to assess our results?

  Number of contigs

  Longest contig

  N50 - largest length for which 50% of all nucleotides are contained in the contigs of at least that length

  L50 - number of contigs that are as long or longer than the N50 value

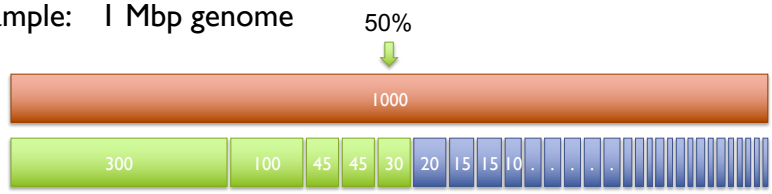
71

71

N50 size

If we place our contigs from largest to smallest on the genome, 50% of the genome in contigs as long as or larger than N50 value

Example: 1 Mbp genome



N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

A greater N50 is usually a sign of assembly improvement

- Comparable with genomes of similar size
- Genome composition can bias comparisons
- Low L50 vs High N50

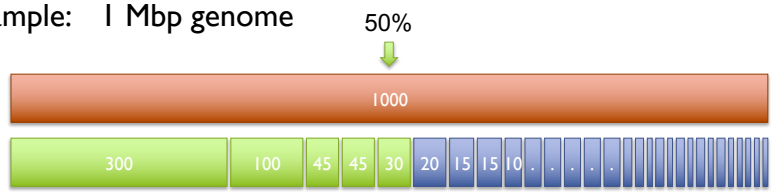
72

72

L50 size

Number of contigs that are as long or longer than the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)
L50 - number of contigs that sum to N50 length
L50 = how many?

- Low L50 vs High N50
  - longer sequences and fewer of them….in theory
  - lower stringency can inflate N50

73

73

L50/N50 size

"N50 length" (L50) first defined:

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

International Human Genome Sequencing Consortium, 2001. *Nature*, *409*(6822), p.860.

74

74

L50/N50 size

Box 1
Genome glossary

Sequence
Raw sequence Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.
Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.
Finished sequence Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.
Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).
Full shotgun coverage The coverage in random raw sequence needed from a large-insert clone to ensure that it is ready for finishing; this varies among centres but is typically 8–10-fold. Clones with full shotgun coverage can usually be assembled with only a handful of gaps per 100 kb.
Half shotgun coverage Half the amount of full shotgun coverage

Sequenced-clone contigs Contigs produced by merging overlapping sequenced clones.
Sequenced-clone-contig scaffolds Scaffolds produced by joining sequenced-clone contigs on the basis of linking information.
Draft genome sequence The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.
N50 length A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L.

Computer programs and databases
PHRED A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

75

75

Genome assembly quality

- How to assess our results?

Alignments:

- compare to a reference genome
- align reads to your assembled genome
- assess repetitive regions
- Call SNPs

Check for completeness:

- annotation, blast against reference gene set
- Busco - Simao et al. 2015. Bioinformatics

76

76

Genome assembly quality

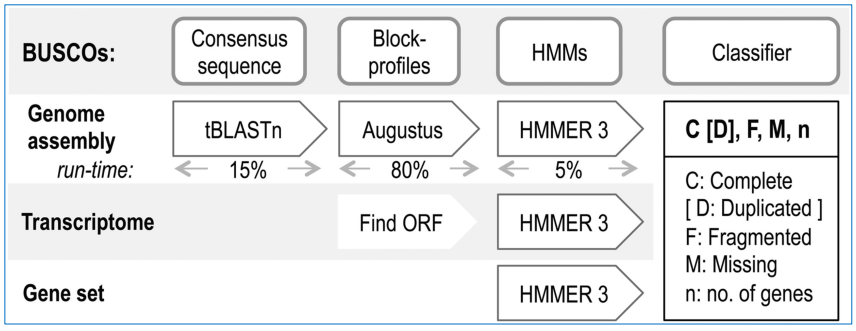
- Do not take first version. Try correcting/polishing your genome
- Always distrust your data!
  - Go back and reassess your genome
  - plotting, quality-control
  - Number of contigs/scaffolds change?
  - L50/N50 go up or down?
  - How much is fragmented?
  - Always new technologies and improvements

77

77

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Bioinformatics. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351



BUSCO assessment workflow and relative run-times

Quality of genome vs. completeness

78

Orthology vs Paralogy

- Homolog - share a common ancestor
- Ortholog - diverged after a speciation event from ancestral gene
- Paralog - diverged after a duplication event within the same lineage

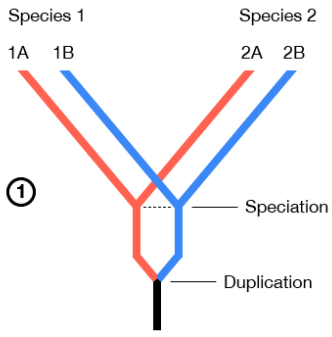
You are either homologous (share a common ancestor) or not. What varies is ability to detect homology.

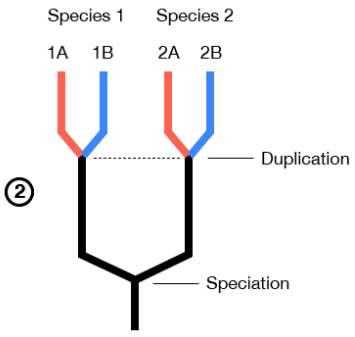
- Evolution of gene families
- Species tree constructions
- genome evolution

79

79

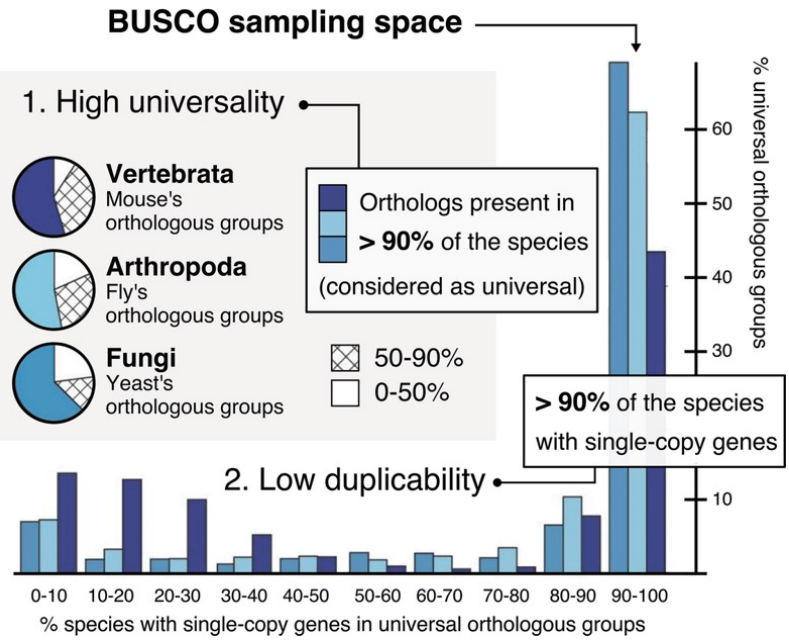
Orthology vs Paralogy





80

80



81

Meta-Analysis > Mol Ecol Resour. 2021 Jul;21(5):1416-1421. doi: 10.1111/1755-0998.13364. Epub 2021 Mar 9.

Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO

April A Jauhal [1 2], Richard D Newcomb [2]

Affiliations + expand

PMID: 33629477 DOI: 10.1111/1755-0998.13364

- High N50 reliably predicted a complete BUSCO score
- High BUSCO scores from assemblies with low N50

82

82

What should we expect from an assembly?

- Annotation of assembly
- Comparison to closely related genomes
- Gene content
- Percent repetitive
- Another estimate
  - Flow cytometry
  - kmer distribution

Genome download and stats

83

83

Genome assembly workshop

- Genome assembly with PacBio data using Canu assembler
- Python programming exercise

84

84