Lecture 8:

Genome assembly

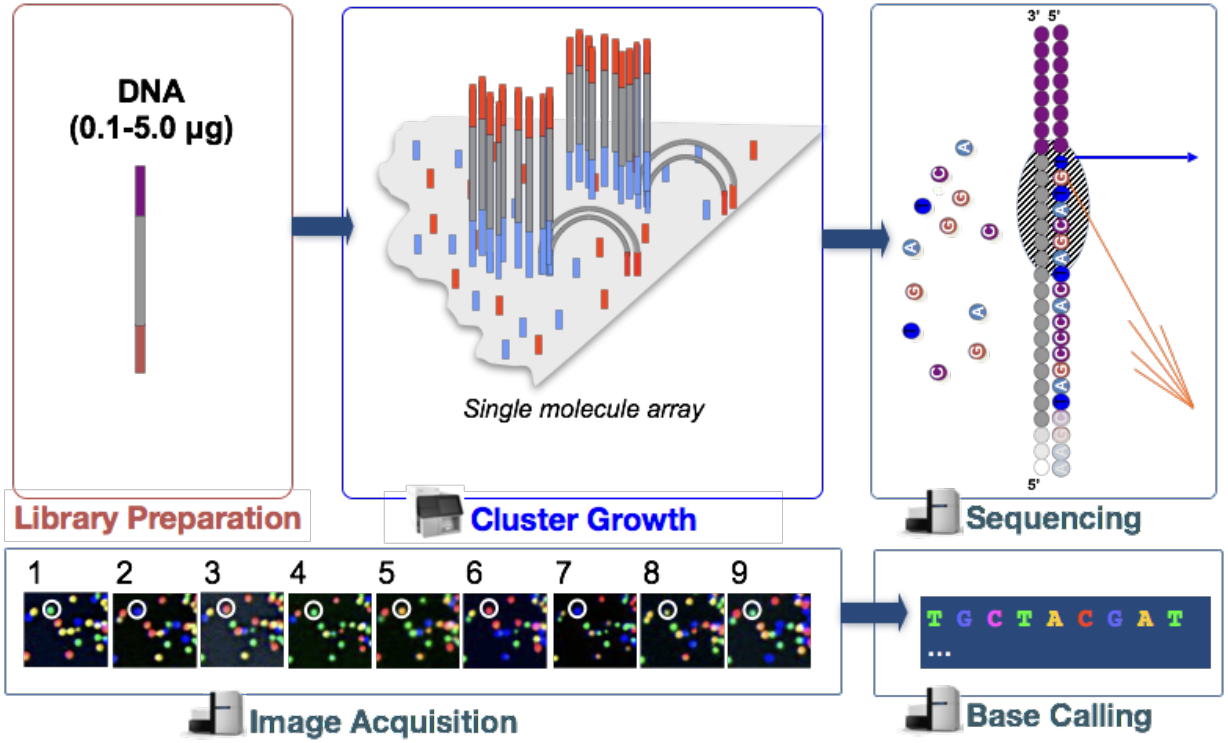
Stitching the pieces

Course 485

Introduction to Genomics

What is the outcome of sequencing?

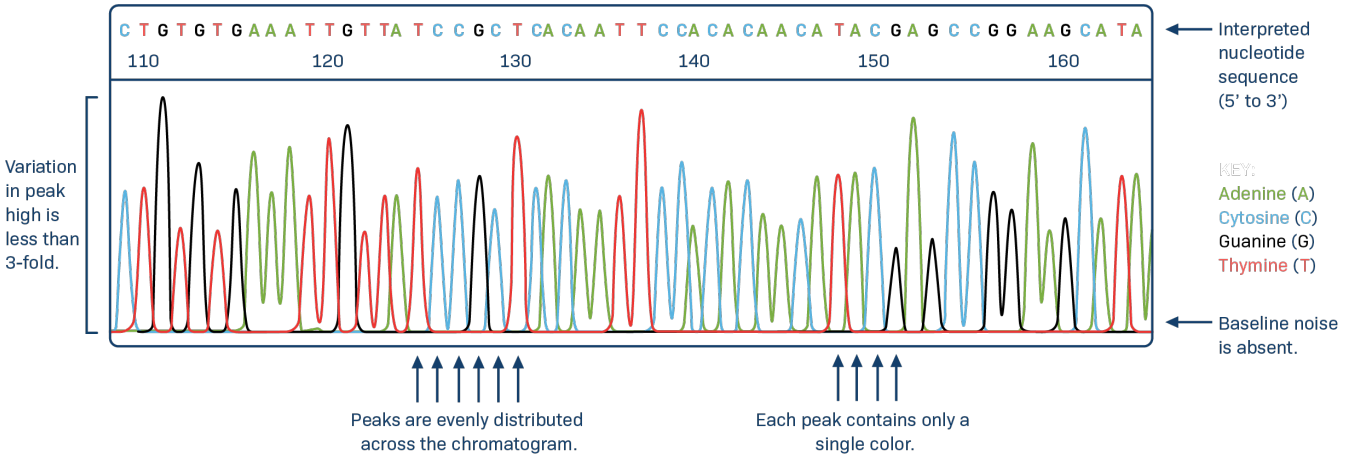
What comes out of the sequencing apparatus/machine?



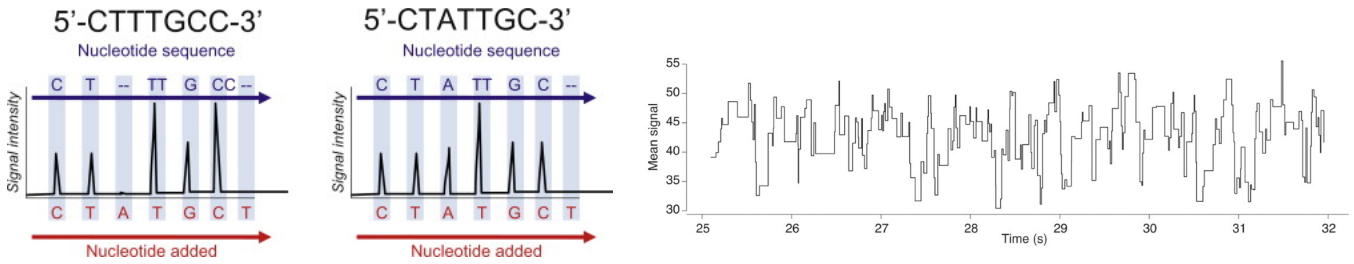
What is the outcome of sequencing?

What comes out of the sequencing apparatus/machine?

Sequence **reads** are the raw data and the output of the sequencing machine



Do all sequencing methods have identical sequence reads?



fasta files

```
>sequence1
ACCCATGATTTGCGA
```

qual files

```
>sequence1
40 40 39 39 40 39 40 40 40 40 20 20 36 39 39
```

fastq files

```
@sequence1
ACCCATGATTTGCGA
+
IIHHIHIIII55EHH
```

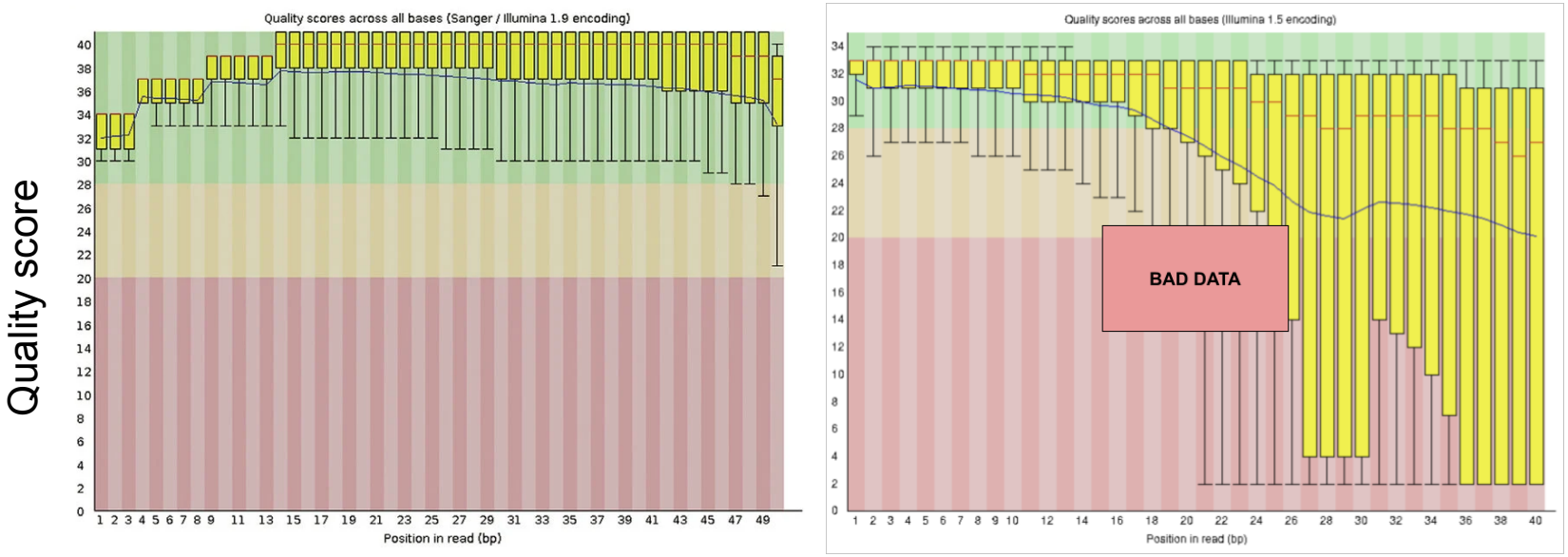
Regardless to the sequencing method used, the most important outputs are:

1) The identity of a nucleotide

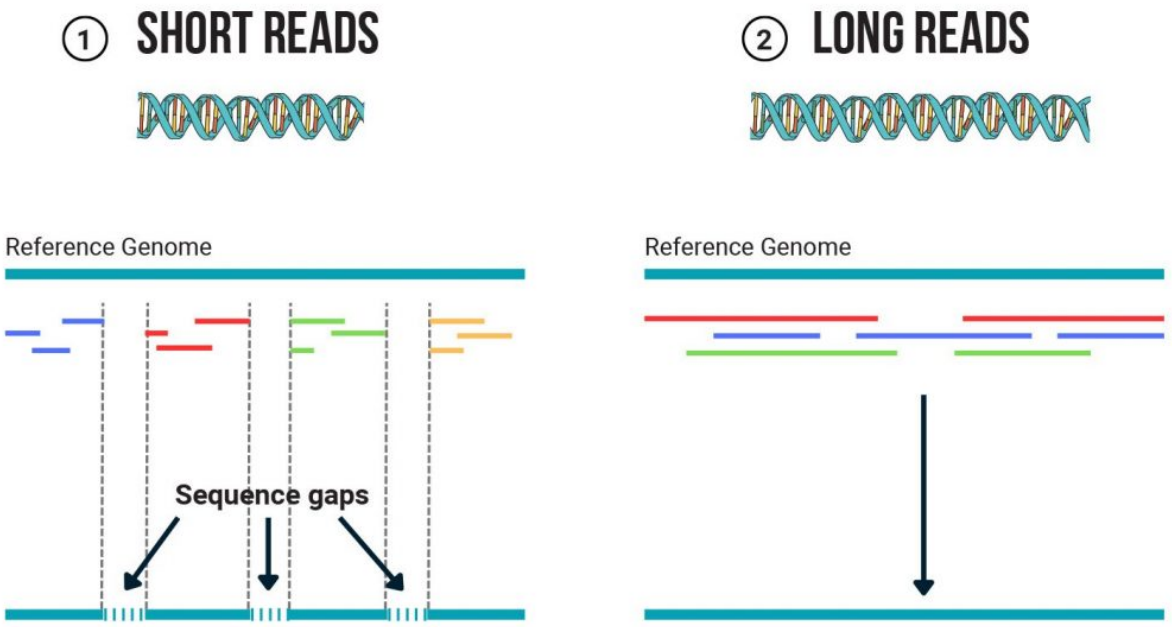
2) Confidence of the call (quality)

Are all nucleotides <u>calls</u> in a sequence read correct and trustworthy?

```
@HWI-ST193:397:D16B3ACXX:2:1101:1091:2467 1:N:0:CGATGT
ATCACAGACAGAAGAGGATTGTACAGAGGAGCTCTTTGACTTCCTGCATC
+
:=:ABBDAFFDDFHHIGCEEB:CEBF<+A??F3?D*?D*?B*:?B<?)?#
```



Are all sequence reads equal in length?



Do you remember the average sequence read's length across sequencing methods?

B Whole-genome short read data

C Whole-genome long read data

RESEARCH ARTICLE

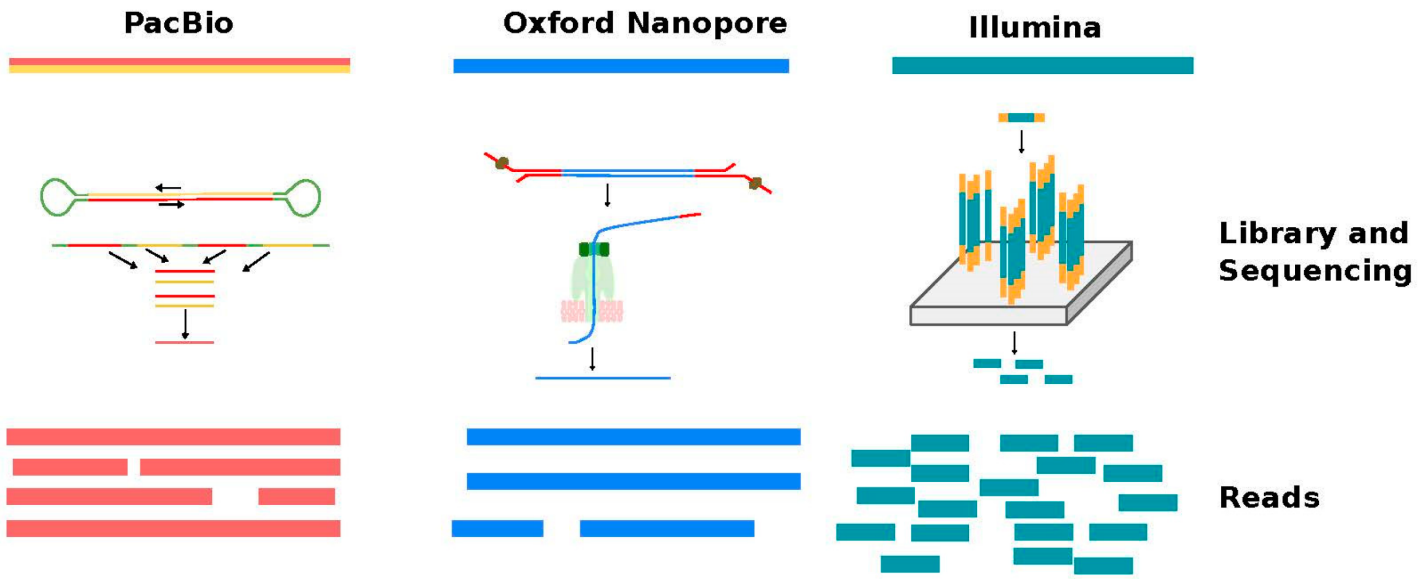
An Error Correction and DeNovo Assembly Approach for Nanopore Reads Using Short Reads

Mehdi Kchouk[1,2] and Mourad Elloumi[2,*]

Table 1. Characteristics of sequencing technologies.

Sequencing Technology	Average Read Length (pb)	Run Time	Advantages	Drawbacks	Error Types	Year
First Generation						
Sanger	~1000	~2 Hours	Long individual read lenght	First sequencing technology and expensive cost.	Insertion - Deletion - Substitution	1977
Second Generation						
Roche 454	700	24 Hours	Long read length	High error rate and expansive runs	Insertion - Deletion	2005
Illumina/Solexa	2*(36 to 100) paired-end	6 Days or 2 Days in rapid mode	Low cost and low error rate	Long run time	Substitution	2006
AB SOLID	85	8 Days	Low cost per base and low error rate	Very long run time	Substitution	2007
Helicos Biosciences	55	Depending on read length	Low error rate	Long Run time	Insertion - Deletion	2009
Ion Torrent	400	2 Hours	Short run time and Less expansive equipment	High error rate	Insertion - Deletion	2010
Third Generation						
Pacific Biosciences	3000 (and up to 15000)	20 Min to 4 Hours	Longest read length and fast run time	High error rate	Insertion - Deletion	2010
Oxford Nanopore (MinIon)	9545	< 6 Hours	Long read; small cost USB device	High error rate	Insertion – Deletion - Substitution	2014

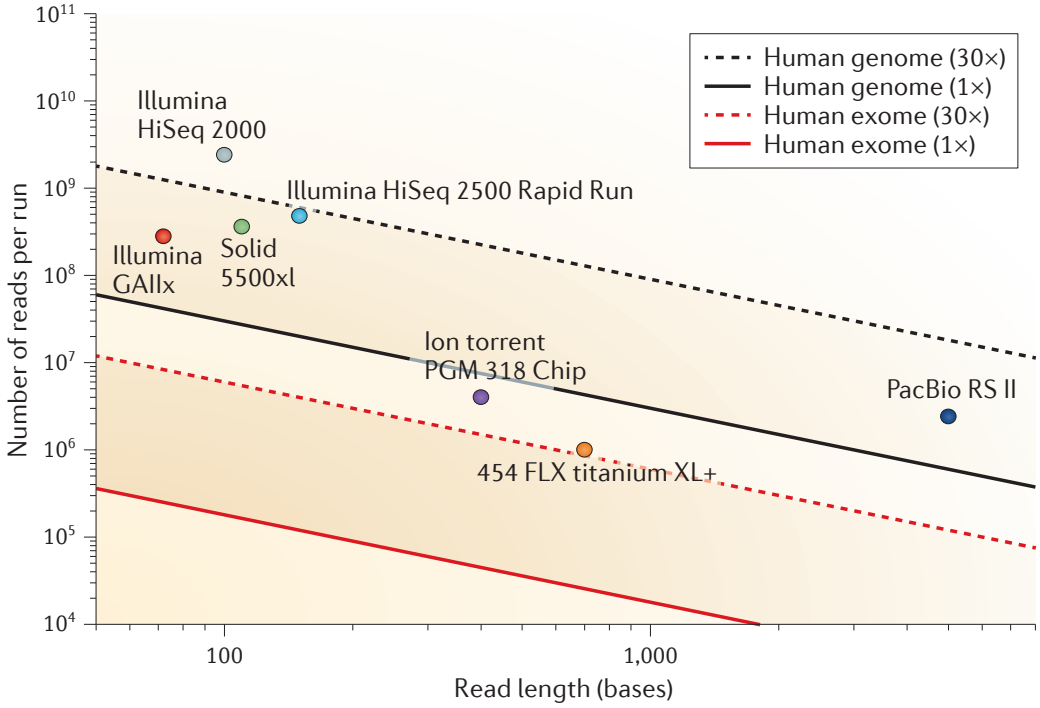
Does the number of reads resulting from sequencing a genome differs between sequencing technologies?



Does the number of reads resulting from sequencing a genome differs between sequencing technologies?

Sequencing depth and coverage: key considerations in genomic analyses

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger and Chris P. Ponting



Plant Communications
Research article

CellPress
Partner Journal

Pistachio genomes provide insights into nut tree domestication and ZW sex chromosome evolution

illumina®

495,729

100bp

50 Mb

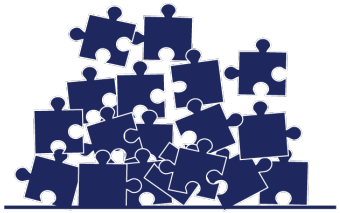
PacBio

9,536

5000bp

50 Mb

What is the effect of sequence length on genome assembly?



>99.9% accuracy sequencing data for large batches and low budget per sample

VS

Continuous long sequence data with improve accesibility to AT/GC-rich regions

Short-read sequencing

The genome is randomly fragmented in fragments of 100-600bp

Libraries are sequenced to produce reads up to 300bp

NovaSeq X | MiSeq | NextSeq | NovaSeq 6000

Long-read sequencing

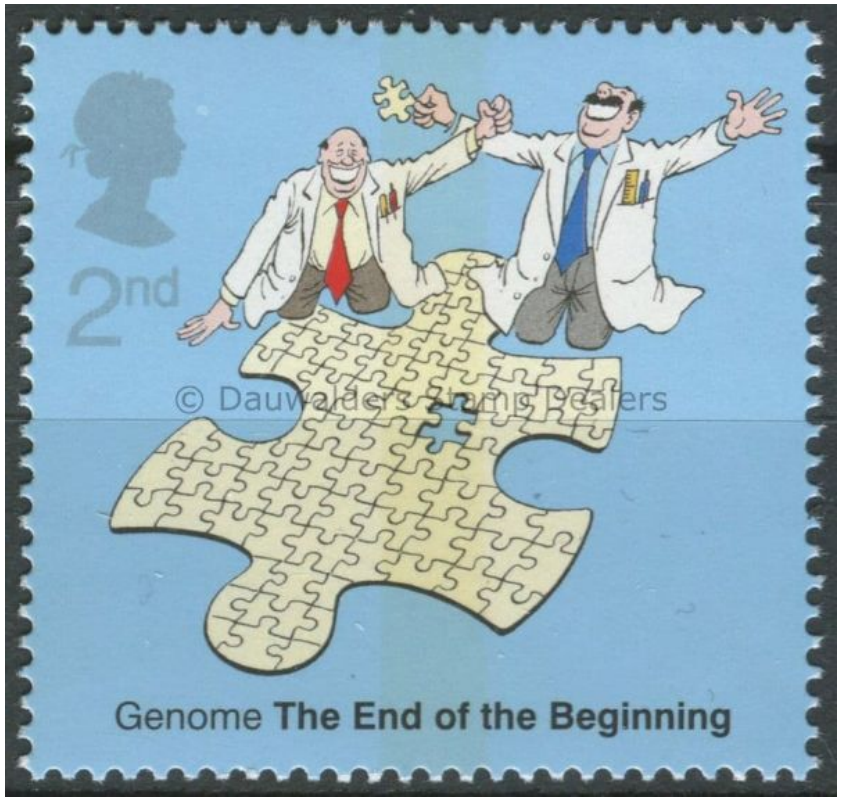
The genome is randomly fragmented in fragments of thousand of base pairs

Libraries are sequenced to produce reads up to 300Kbp

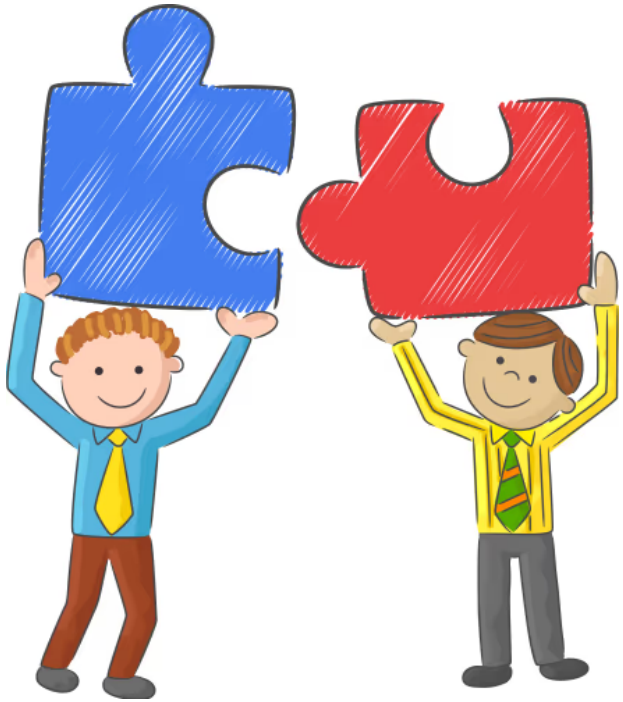
Sequel IIe | Revio | GridION

What is the total length of combined sequence reads in relation to the genome size?

Genome assembly: An analogy




Edge Pieces First
Grouping by Color or Pattern
Working on Small Sections



What do we need to assemble a genome under a hierarchical strategy?

1) Ordered clones

2) Paired-end sequences of segments

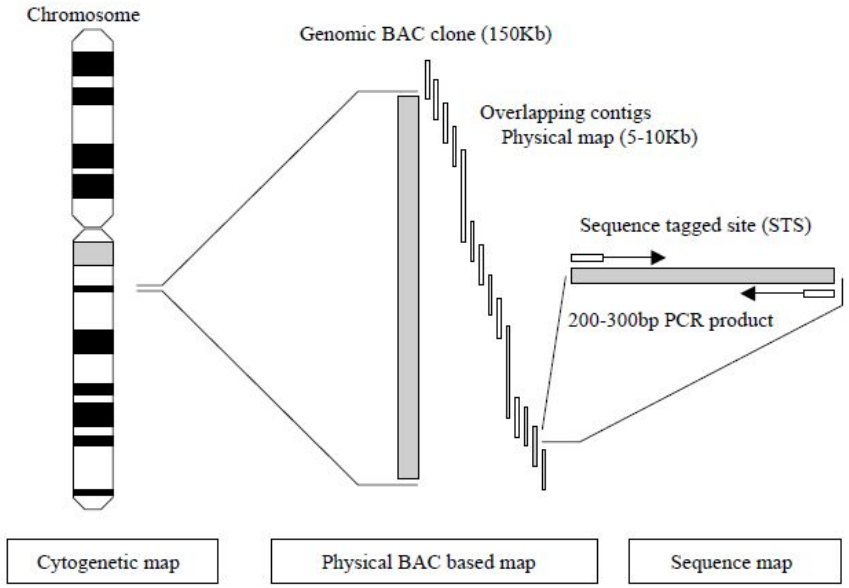
3) Computational assembly

What do we need to assemble a genome under a whole-genome shotgun strategy?

1) Single-end or paired-end sequences

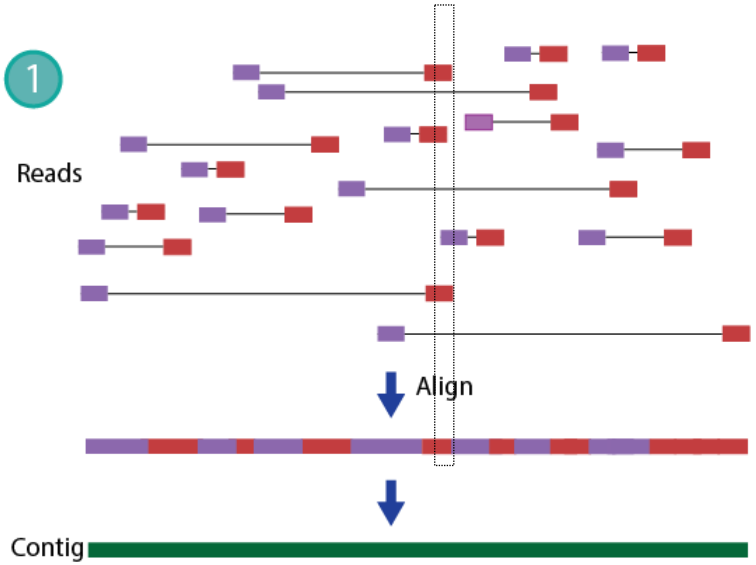
2) Perform de-novo assembly

What does ordered clones and paired-end sequence provide us?



Order clones and paired-end sequences provide map to assemble large genomic fragments

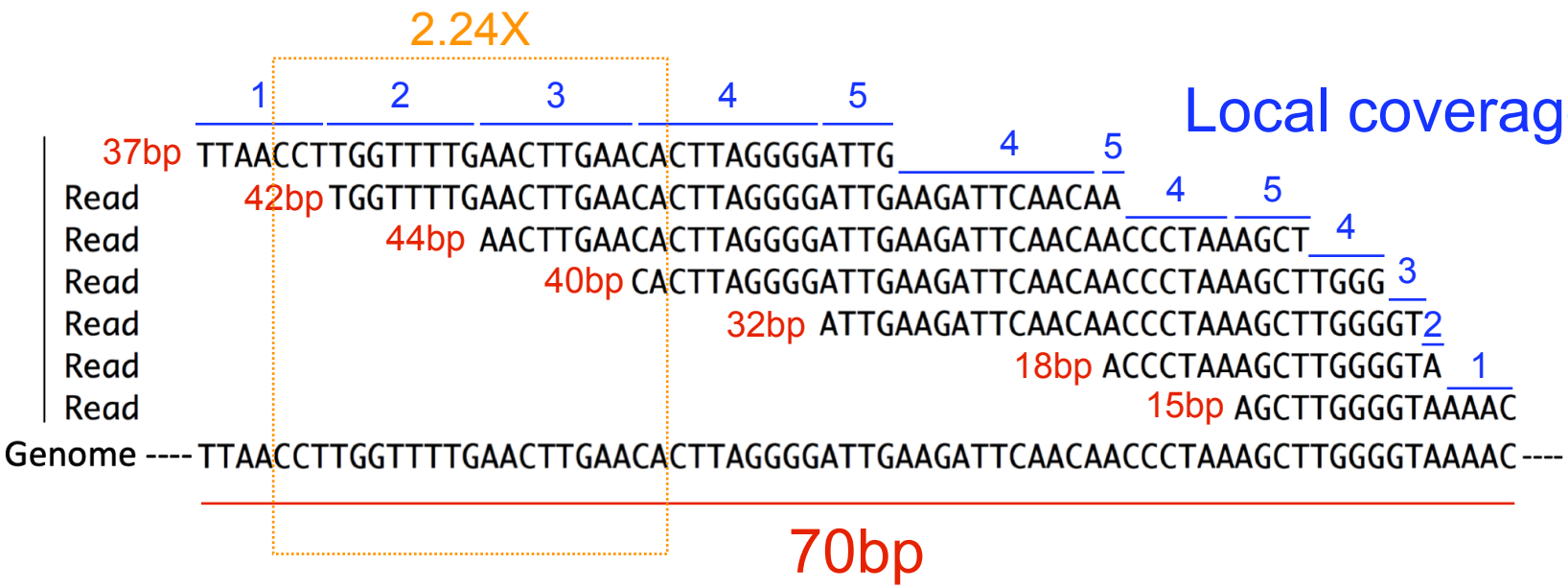
How exactly can we assemble sequences or stitch sequence reads together?



Computational assembly: looking at overlapping aligned sequence reads to get longer sequences (stitched pieces)

Sequence coverage/depth

21bp 1+1+1+2+2+2+2+2+2+3+3+3+3+3+3+3+4+4=47



$$\text{Sequence coverage} = \frac{37+42+44+40+32+18+15\text{bp}}{70\text{bp}} = \frac{228\text{bp}}{70\text{bp}} = 3.26\text{X}$$

Sequence coverage/depth

The size of the genome of X species is 96Mb. Its genome was sequenced using Illumina sequencing generated a total of 480 Mb. What is the sequence coverage of X genome?

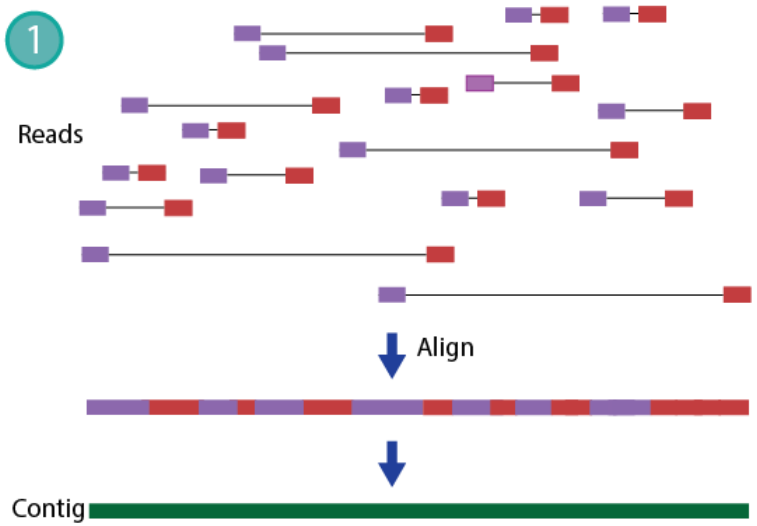
$$\text{Sequence coverage} = \frac{\text{Total sequences generated}}{\text{size of the genome}} = \frac{480\text{Mb}}{96\text{Mb}} = 5\text{X}$$

What do we need to align sequence reads and stitch them together?

1) A method of alignment

2) a test of the significance of the alignment

Information theory!



Using a method of alignment, we order and connect overlapping sequence reads.

Contigs: the result of ordering, aligning, and connecting sequence read.

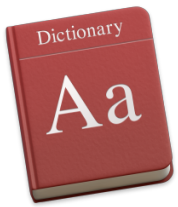
What is a contig?

con·tig·u·ous | kən'tigyəwəs |

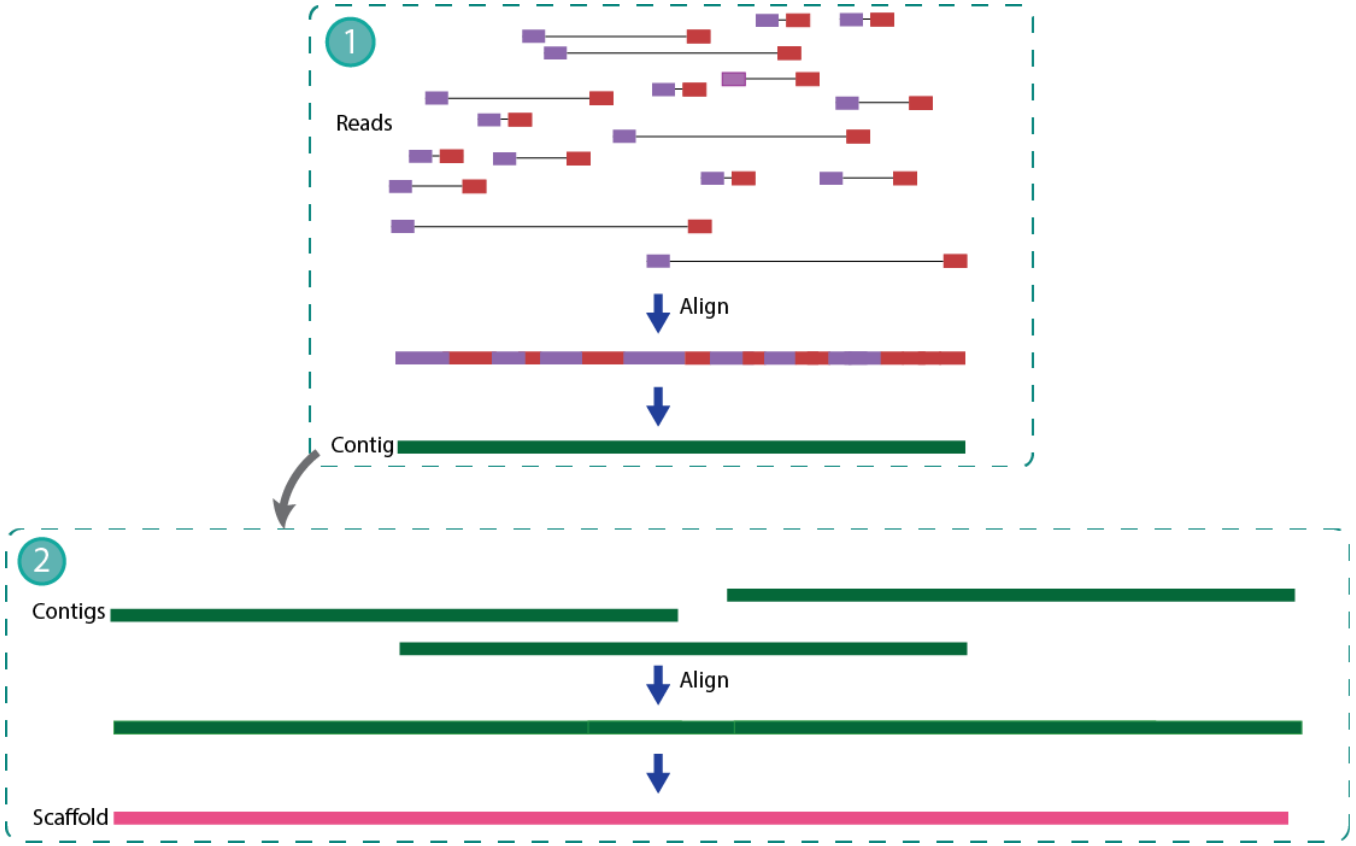
adjective

sharing a common border; touching: *the 48 contiguous states.*

- next or together in sequence: *five hundred contiguous dictionary entries.*



Scaffolds: the result of ordering, aligning, and connecting contigs



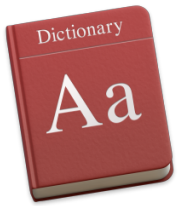
What is a scaffold?

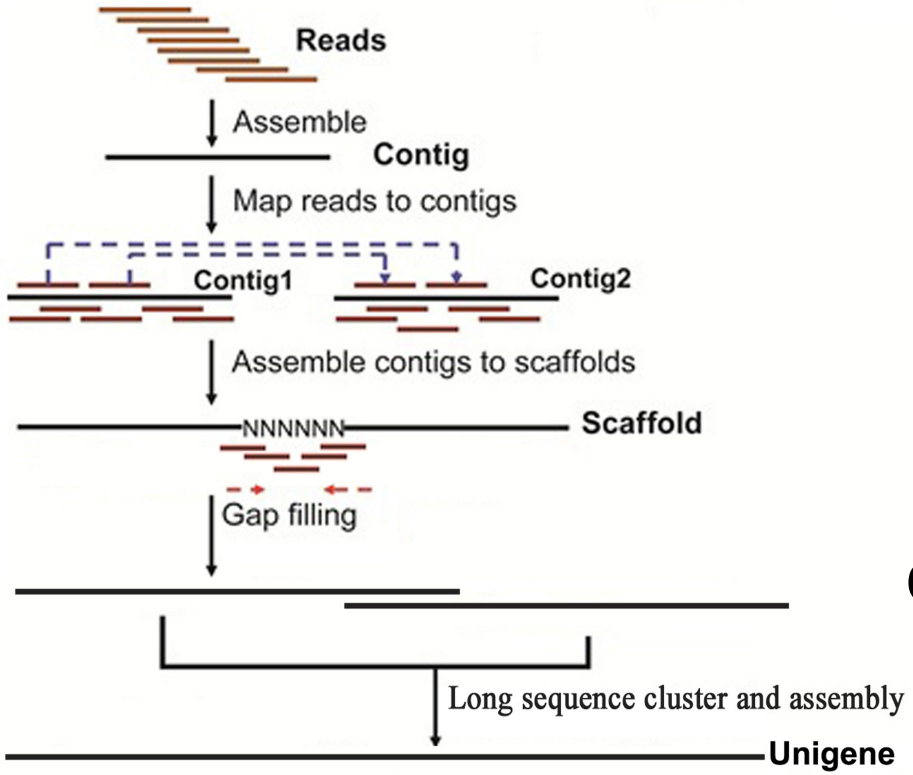
scaf·fold | 'skafəld |

noun

1 a raised wooden platform used formerly for the public execution of criminals.

2 a structure made using scaffolding: *[as modifier] : scaffold boards.*





How do we connect disconnected contigs?

Use molecular markers that order the contigs or identify BAC clone the fills the gap

How many contigs are needed to assemble a genome?

It depends on,

1) Type of the genome

2) Type of sequencing technology

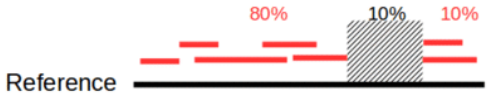
How can we assess the quality of a genome assembly?

1) Completeness

2) Correctness

3) Contiguity

Genome sequence completeness



$$C = \frac{\text{\# area covered by reads}}{\text{\# reference area}}$$

RESEARCH ARTICLE

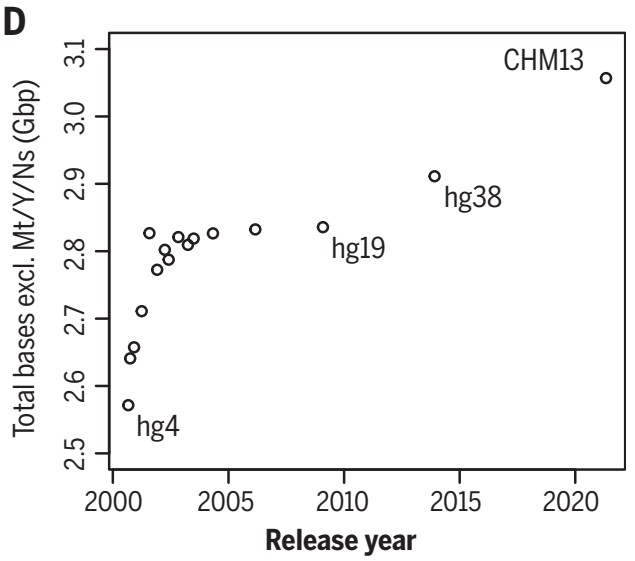
HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

STATISTICS	GRCH38	T2T-CHM13	DIFFERENCE (±%)
	Summary		
Assembled bases (Gbp)	2.92	3.05	+4.5
Unplaced bases (Mbp)	11.42	0	−100.0
Gap bases (Mbp)	120.31	0	−100.0
Number of contigs	949	24	−97.5
Contig NG50 (Mbp)	56.41	154.26	+173.5
Number of issues	230	46	−80.0
Issues (Mbp)	230.43	8.18	−96.5

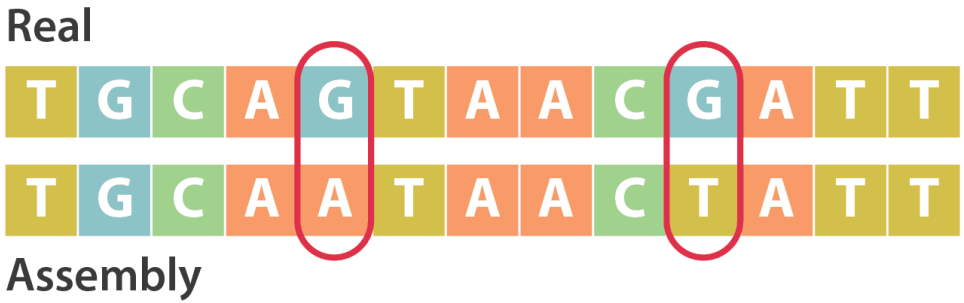


Genome sequence completeness

The genome of X species is 2.65Gb and was sequenced using pyrosequencing resulting in an assembled genome of 2.3Gb. How much of the genome was covered with sequencing?

$$\% \text{ genome sequenced} = \frac{2.30\text{Gb}}{2.65\text{Gb}} \text{x}100 = 86.8\%$$

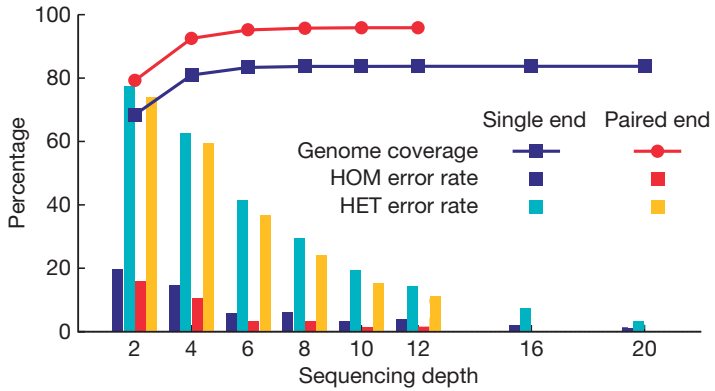
Genome sequence correctness



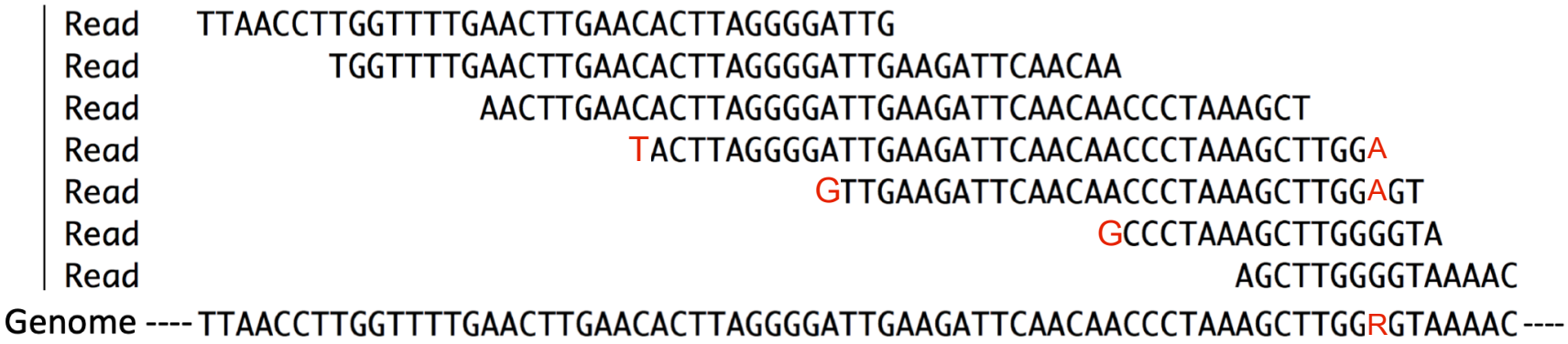
nature Vol 456|6 November 2008|doi:10.1038/nature07484

ARTICLES

The diploid genome sequence of an Asian individual



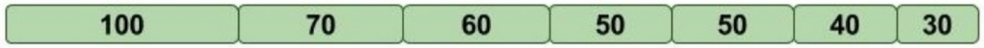
Genome sequence correctness



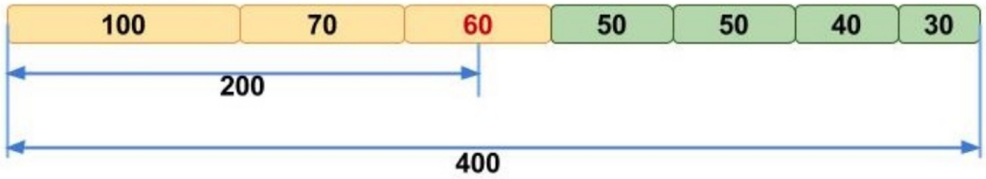
How can we assess the contiguity of a genome assembly?

1) Contig N50 size

2) Scaffold N50 size



1a. Contigs, sorted according to their lengths.



1b. Calculation of N50 using sorted contigs.

These measures are arbitrary measures and statistical parameters such as the mean. The measures are used only to compare between different assemblies.

Contig/Scaffold N50: the size of the contig/scaffold at which 50% of the genome is assembled into contigs/scaffold

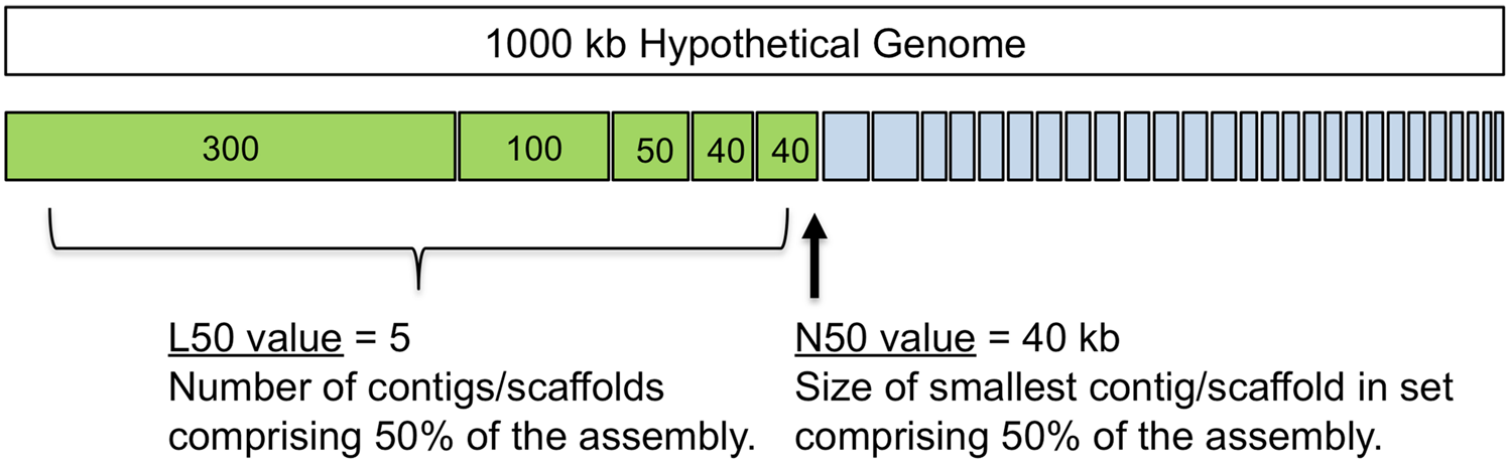
Conservation Genetics
https://doi.org/10.1007/s10592-019-01224-x

REVIEW ARTICLE

Genome projects in invasion biology

Michael A. McCartney[1] · Sophie Mallez[1] · Daryl M. Gohl[2,3]

Contiguity Measures



Contig/Scaffold L50: the number of the contig/ scaffold at which 50% of the genome is assembled into contigs/scaffold

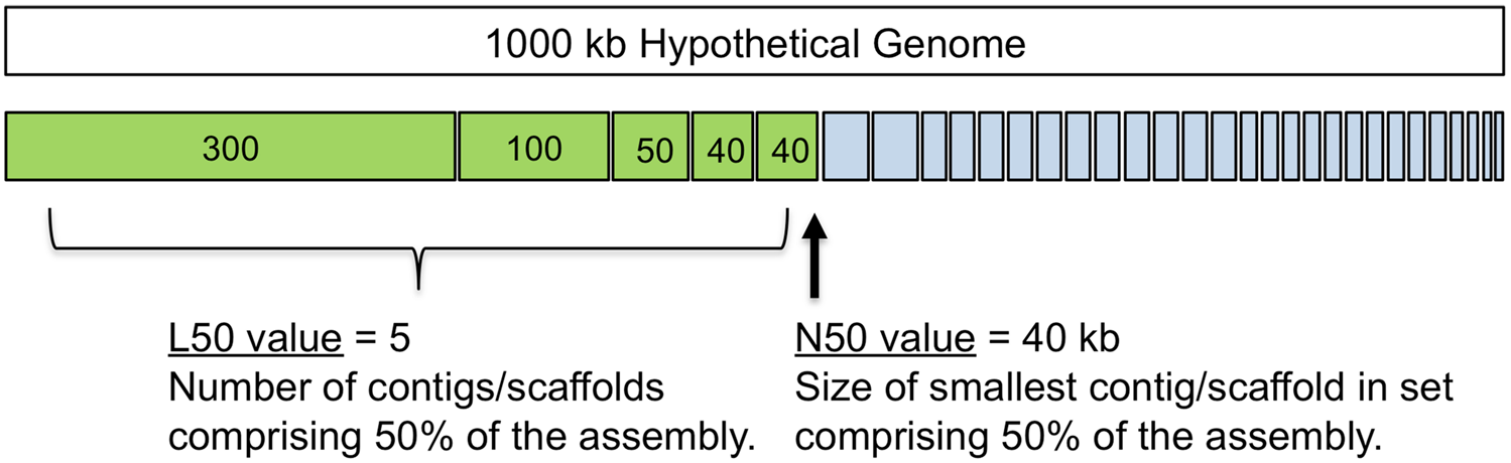
Conservation Genetics
https://doi.org/10.1007/s10592-019-01224-x

REVIEW ARTICLE

Genome projects in invasion biology

Michael A. McCartney[1] · Sophie Mallez[1] · Daryl M. Gohl[2,3]

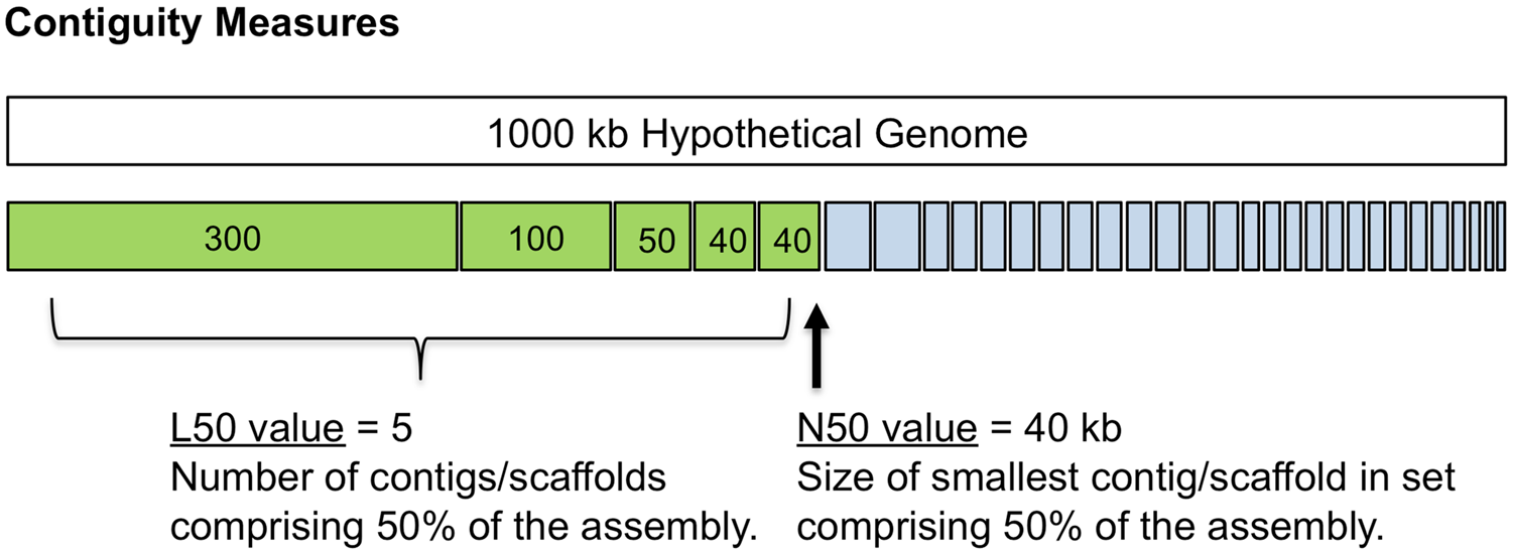
Contiguity Measures



How to calculate/find the contig N50 or L50?

1) Order contigs/scaffolds according to size in a descending order

2) Identify the number/size of the contig where 50% of the genome is assembled.



A group of three great students (Sara, Nouf, and Hayat) have sequenced the genome of a local ant. The sequencing machine generated **4,105,387** sequence reads, which accounted to a total of **480 Mb**. Below are the details of the initial assembled contigs.

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

The three students decided to assemble the genome independently.

Sara used ALL sequences.

Nouf used sequences > 150 bp.

Hayat used sequences > 200 bp.

A group of three great students (Sara, Nouf, and Hayat) have sequenced the genome of a local ant. The sequencing machine generated **4,105,387** sequence reads, which accounted to a total of **480 Mb**. Below are the details of the initial assembled contigs.

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

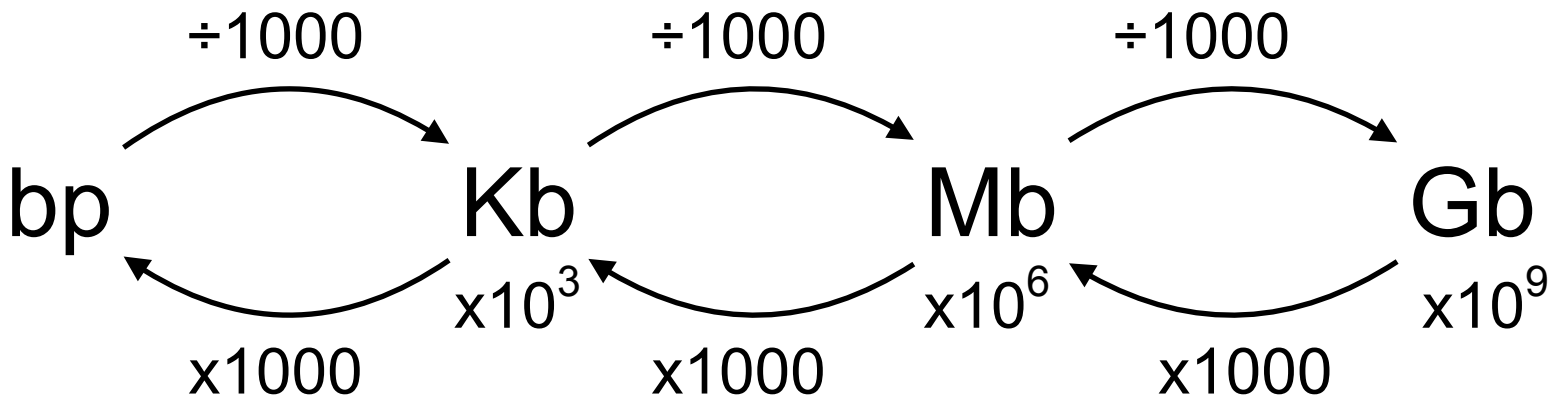
The three students decided to assemble the genome independently.

Sara used ALL sequences.

Nouf used sequences > 150 bp.

Hayat used sequences > 200 bp.

Conversion of genome length units



What does the contigs' size catagories and numbers mean?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

What does 150bp and 200bp contigs mean?

What the most likely sequencing technology used and why?

What is the contig number for each student's genome assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431

13,181

7,181

1,431

What is the assembled genome size for each student's assembly?

Contig size	# of Contig	Combined length
150 bp	6,000	150bpx6,000 = 900,000bp = 900Kb = 0.9Mb
200 bp	5,750	200bpx5,750 = 1,150,000bp = 1,150Kb = 1.15Mb
1 Kb	1,200	1Kbx1,200 = 1,200Kb = 1.2Mb
5 Kb	50	5Kbx50 = 250Kb = 0.25Mb
10 Kb	110	10Kbx110 = 1,100Kb = 1.1Mb
100 Kb	59	100Kbx59 = 5,900Kb = 5.9Mb
500 Kb	5	500Kbx5 = 2,500Kb = 2.5Mb
1 Mb	4	1Mbx4 = 4Mb
5 Mb	3	5Mbx3 = 15Mb

29.95Mb **31.1Mb** **32Mb**

What is the assembled genome size for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb

What is 50% the size of the assembled genome for each student's assembly?

Contig size	# of Contig	Combined length
150 bp	6,000	150bpx6,000 = 900,000bp = 900Kb = 0.9Mb
200 bp	5,750	200bpx5,750 = 1,150,000bp = 1,150Kb = 1.15Mb
1 Kb	1,200	1Kbx1,200 = 1,200Kb = 1.2Mb
5 Kb	50	5Kbx50 = 250Kb = 0.25Mb
10 Kb	110	10Kbx110 = 1,100Kb = 1.1Mb
100 Kb	59	100Kbx59 = 5,900Kb = 5.9Mb
500 Kb	5	500Kbx5 = 2,500Kb = 2.5Mb
1 Mb	4	1Mbx4 = 4Mb
5 Mb	3	5Mbx3 = 15Mb

	29.95Mb	**31.1Mb**	**32Mb**
50% genome = genome size x 0.5 =	**14.975Mb**	**15.55Mb**	**16Mb**

What is 50% the size of the assembled genome for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb

What is the mean contig length for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb

$$\text{Mean contain length} = \frac{\text{Genome size}}{\text{\# contigs}} = \frac{32\text{Mb}}{13{,}181} = 0.00242\text{Mb} = 2.42\text{Kb}$$

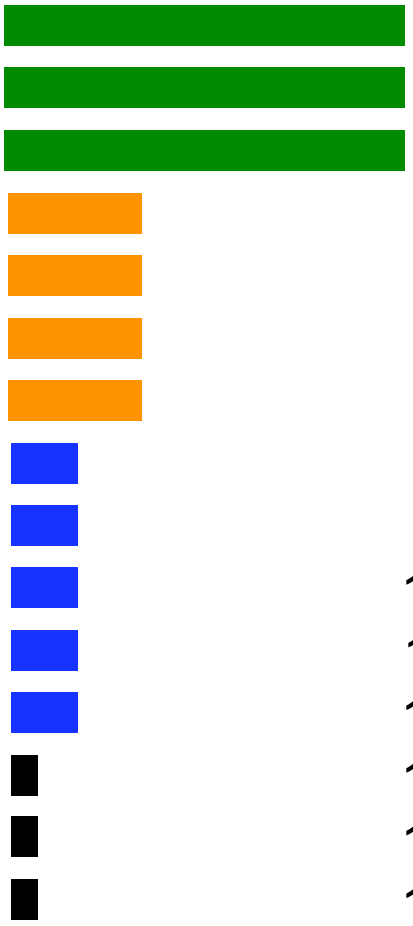
What is the mean contig length for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb

What is the contig N50 for each student's assembly?

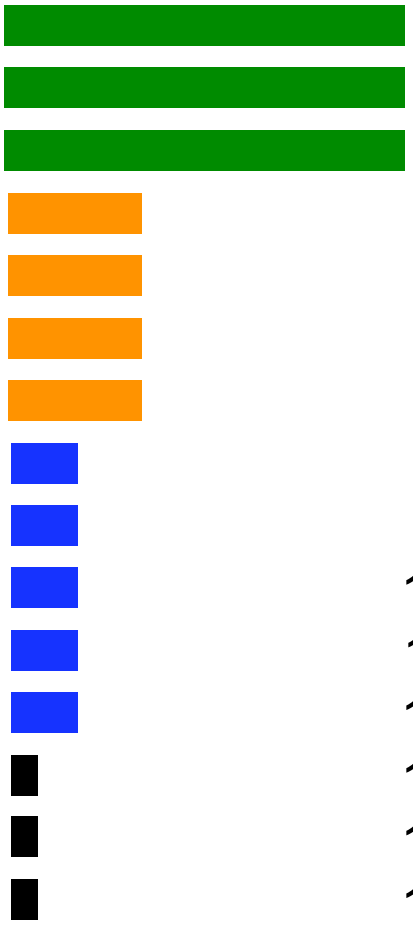
Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size
1	5Mb
2	5Mb
3	5Mb
4	1Mb
5	1Mb
6	1Mb
7	1Mb
8	500Kb
9	500Kb
10	500Kb
11	500Kb
12	500Kb
13	100Kb
14	100Kb
15	100Kb

What is the contig N50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size	Sum
1	5Mb	5Mb
2	5Mb	10Mb
3	5Mb	15Mb
4	1Mb	16Mb
5	1Mb	17Mb
6	1Mb	18Mb
7	1Mb	19Mb
8	500Kb	19.5Mb
9	500Kb	20Mb
10	500Kb	20.5Mb
11	500Kb	21Mb
12	500Kb	21.5Mb
13	100Kb	21.6Mb
14	100Kb	
15	100Kb	

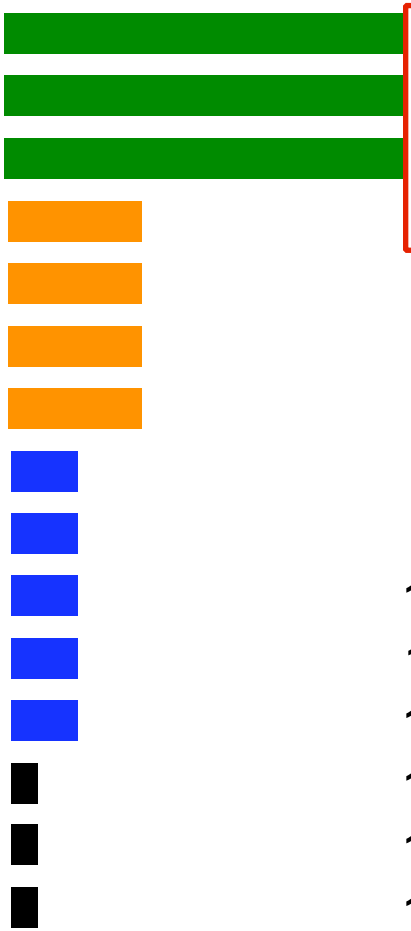
What is the contig N50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb

What is the contig N50 for each student's assembly?

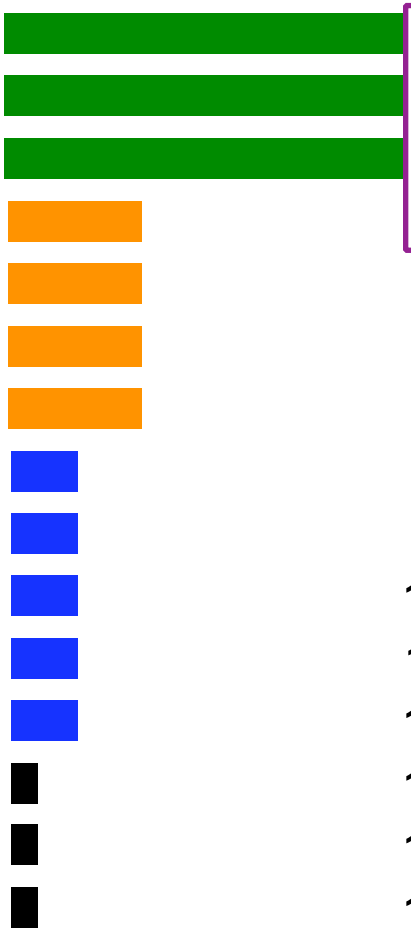
Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size	Sum
1	5Mb	5Mb
2	5Mb	10Mb
3	5Mb	15Mb
4	**1Mb**	**16Mb**
5	1Mb	17Mb
6	1Mb	18Mb
7	1Mb	19Mb
8	500Kb	19.5Mb
9	500Kb	20Mb
10	500Kb	20.5Mb
11	500Kb	21Mb
12	500Kb	21.5Mb
13	100Kb	21.6Mb
14	100Kb	
15	100Kb	

What is the contig N50 for each student's assembly?

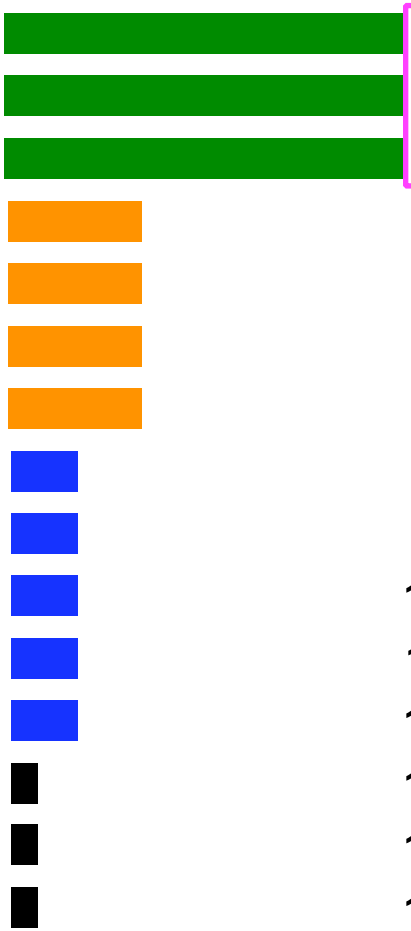
Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size	Sum
1	5Mb	5Mb
2	5Mb	10Mb
3	5Mb	15Mb
4	**1Mb**	**16Mb**
5	1Mb	17Mb
6	1Mb	18Mb
7	1Mb	19Mb
8	500Kb	19.5Mb
9	500Kb	20Mb
10	500Kb	20.5Mb
11	500Kb	21Mb
12	500Kb	21.5Mb
13	100Kb	21.6Mb
14	100Kb	
15	100Kb	

What is the contig N50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size	Sum
1	5Mb	5Mb
2	5Mb	10Mb
3	5Mb	15Mb
4	**1Mb**	**16Mb**
5	1Mb	17Mb
6	1Mb	18Mb
7	1Mb	19Mb
8	500Kb	19.5Mb
9	500Kb	20Mb
10	500Kb	20.5Mb
11	500Kb	21Mb
12	500Kb	21.5Mb
13	100Kb	21.6Mb
14	100Kb	
15	100Kb	

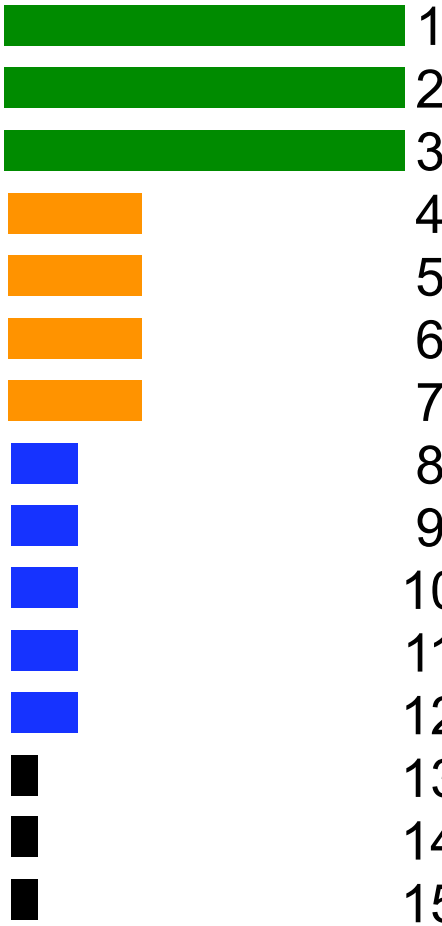
What is the contig N50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb

What is the contig L50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3



Contig #	Contig Size	Sum
1	5Mb	5Mb
2	5Mb	10Mb
3	5Mb	15Mb
4	1Mb	16Mb
5	1Mb	17Mb
6	1Mb	18Mb
7	1Mb	19Mb
8	500Kb	19.5Mb
9	500Kb	20Mb
10	500Kb	20.5Mb
11	500Kb	21Mb
12	500Kb	21.5Mb
13	100Kb	21.6Mb
14	100Kb	
15	100Kb	

What is the contig L50 for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb
L50 contig	4	4	3

What is the sequence coverage for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

$$\frac{480\text{Mb}}{32\text{Mb}} = 15\text{x}$$

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb
L50 contig	4	4	3
~Coverage	15x		

What is the sequence coverage for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

$$\frac{480\text{Mb}-0.9\text{Mb}}{31.1\text{Mb}} = 15.4\text{x}$$

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb
L50 contig	4	4	3
~Coverage	15x	15.4x	

What is the sequence coverage for each student's assembly?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

$$\frac{480\text{Mb}-2.05\text{Mb}}{29.95\text{Mb}} = \sim 12\text{x}$$

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb
L50 contig	4	4	3
~Coverage	15x	15.4x	12x

Who has the best assembly and why?

Contig size	# of Contig
150 bp	6,000
200 bp	5,750
1 Kb	1,200
5 Kb	50
10 Kb	110
100 Kb	59
500 Kb	5
1 Mb	4
5 Mb	3

	Sara	Nouf	Hayat
# Contigs	13,181	7,181	1,431
Genome size	32Mb	31.1Mb	29.95Mb
50% Genome	16Mb	15.55Mb	14.975Mb
Mean contig length	2.42KB	4.33Kb	20.92Kb
N50 contig	1Mb	1Mb	5Mb
L50 contig	4	4	3
~Coverage	15x	15.4x	12x

Disclaimer

Figures, photos, and graphs in my lectures are collected using google searches. I do not claim to have personally produced the material (except for some). I do cite only articles or books used. I thank all owners of the visual aid that I use and apologize for not citing each individual item. If anybody finds the inclusion of their material in my lectures a violation of their copy rights, please contact me via email.

hhalhaddad@gmail.com