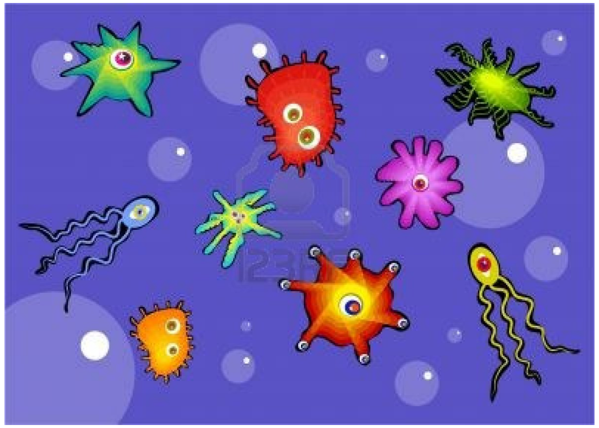
Comparison of microbial eukaryotic single cell genome assembly tools

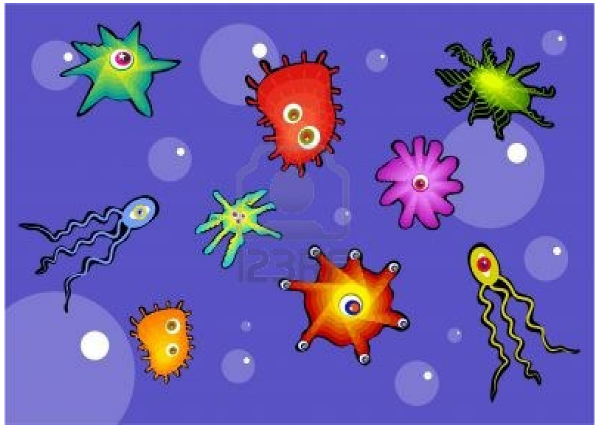
Rajat S. Roy[1,3], Alexander Schliep[1,2], Debashish Bhattacharya[3]

[1]Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA.
[2]BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854, USA.
[3]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA.

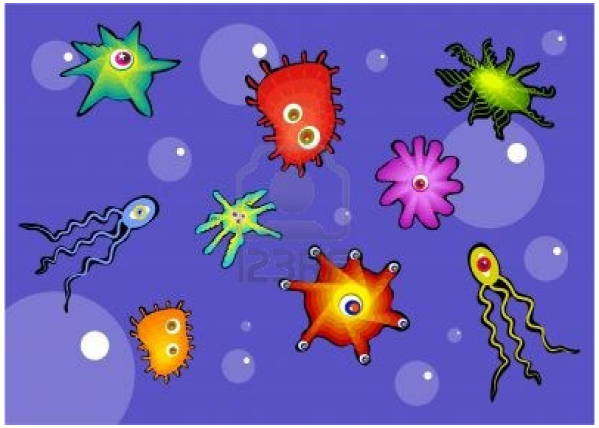
Why single cell *de novo* genome assembly?



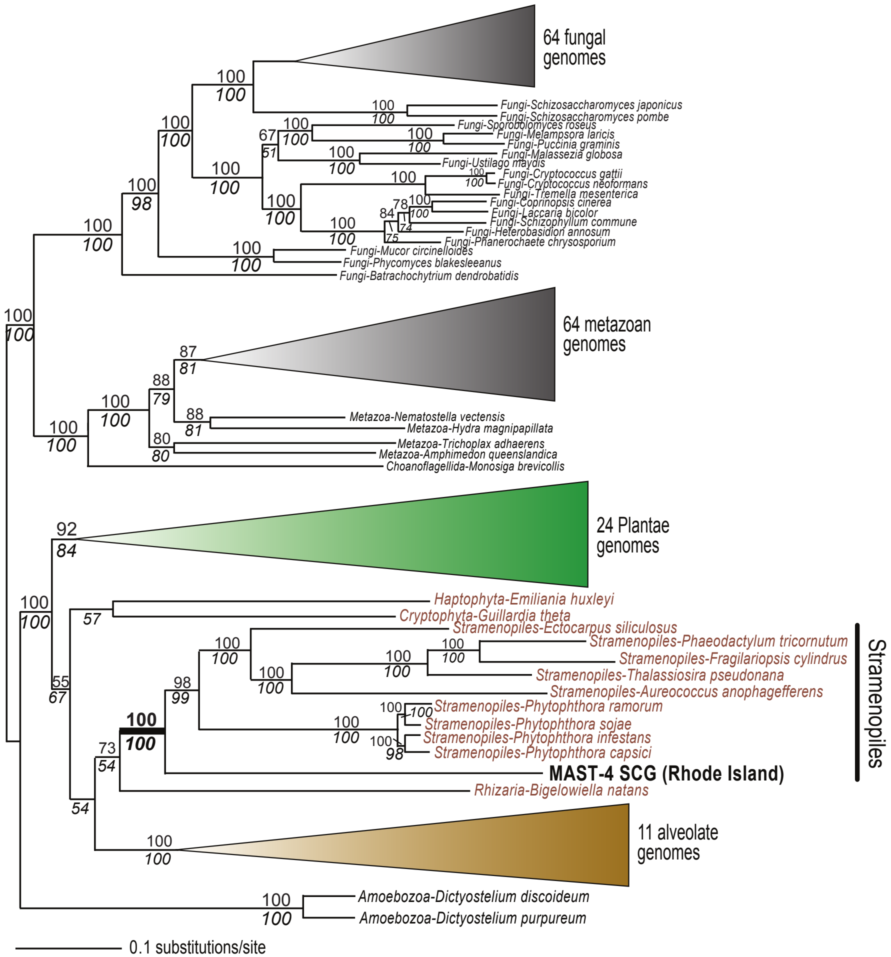
Why single cell *de novo* genome assembly?



Why single cell *de novo* genome assembly?



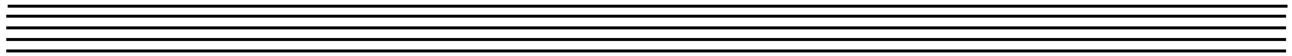
Evolutionary analysis of a wild-caught stramenopile



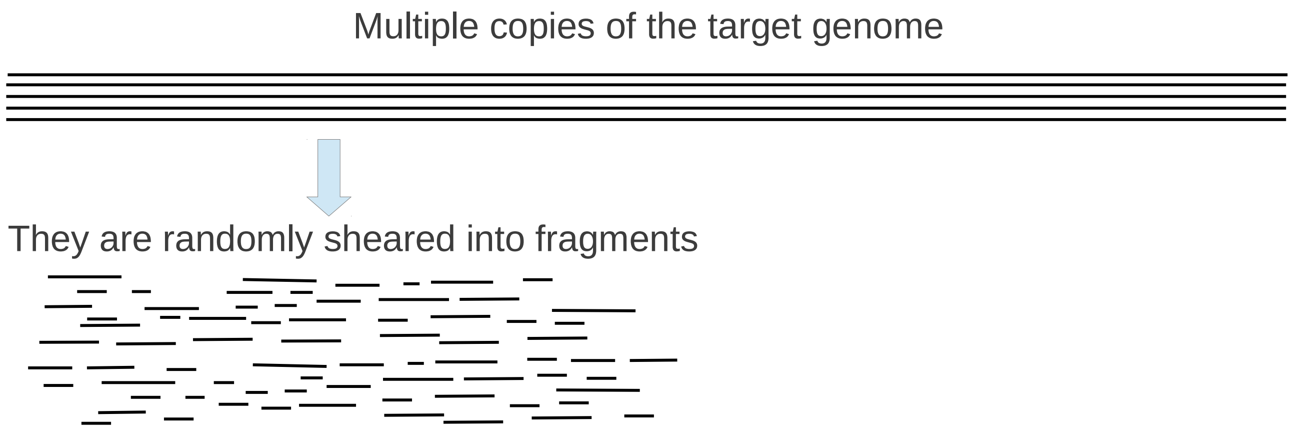
Evolutionary analysis of single cell genome data from a wild-caught marine stramenopile. Rajat S. Roy, Dana C. Price, Alexander Schliep, Guohong Cai, Anton Korobeynikov, Hwan Su Yoon, Eun Chan Yang, and D Bhattacharya. (In preparation), 2013.

Genome Assembly

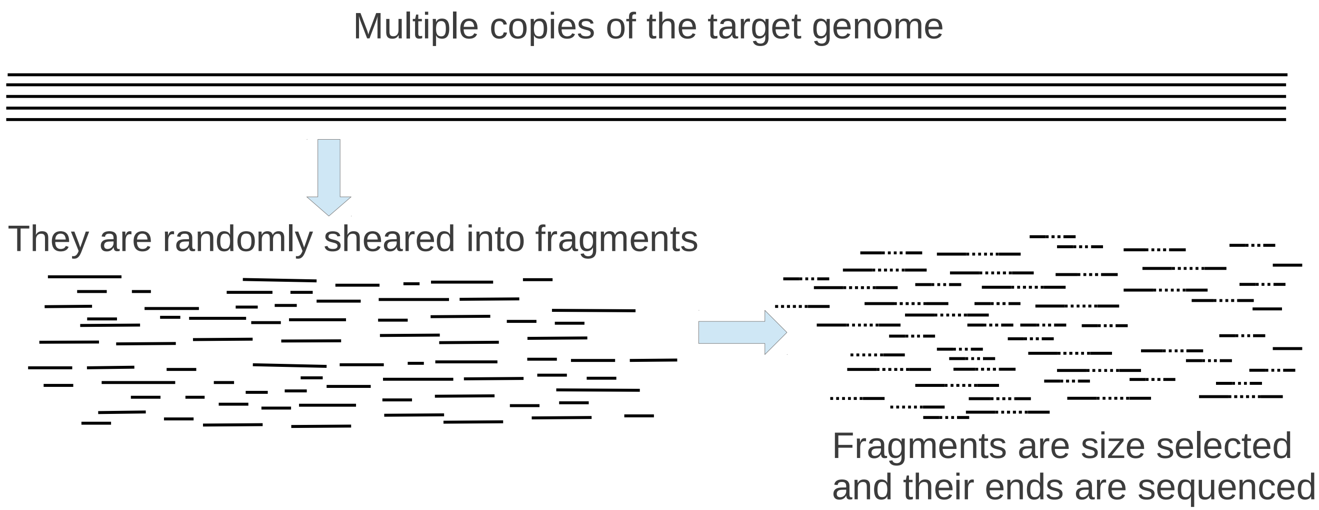
Multiple copies of the target genome



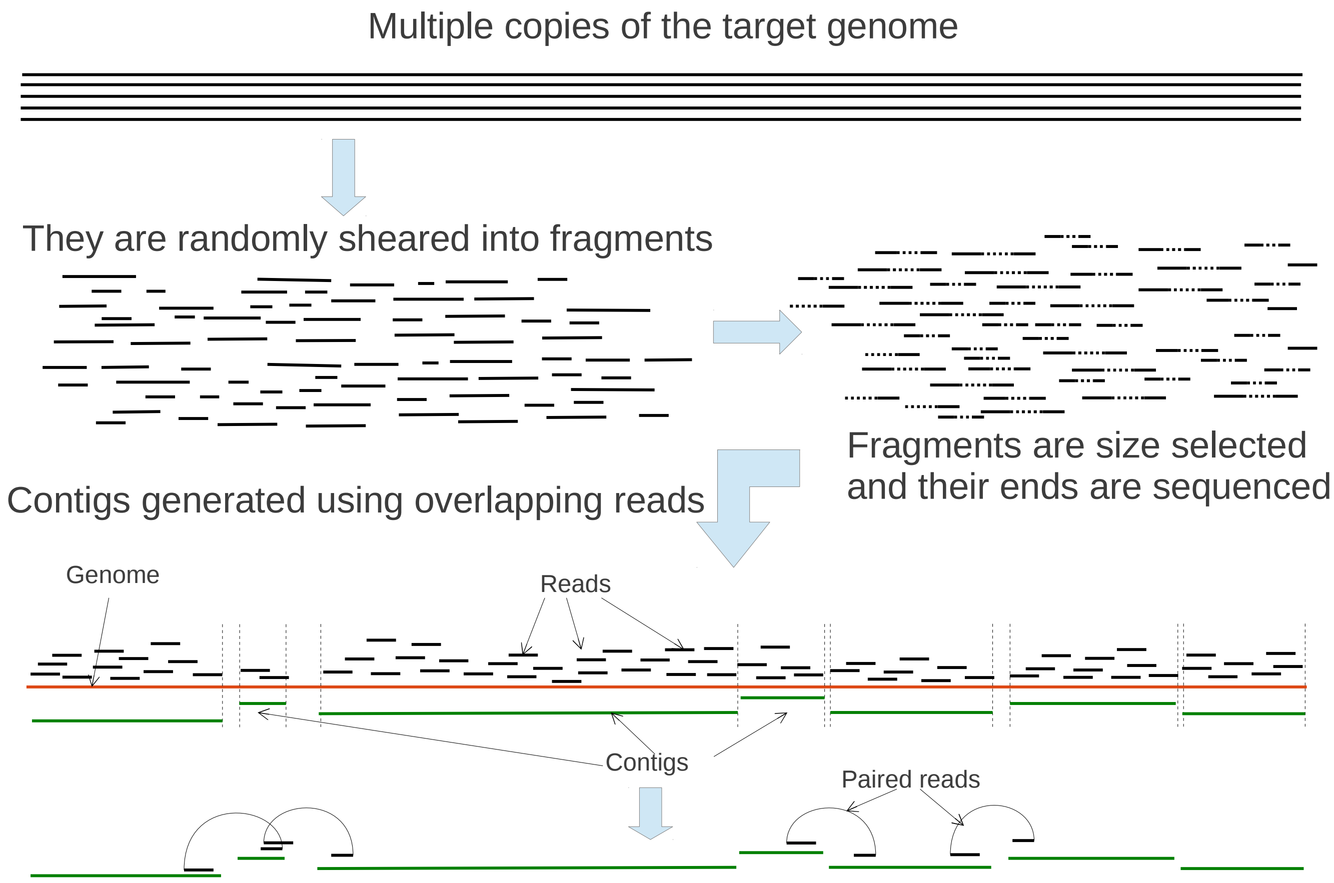
Genome Assembly



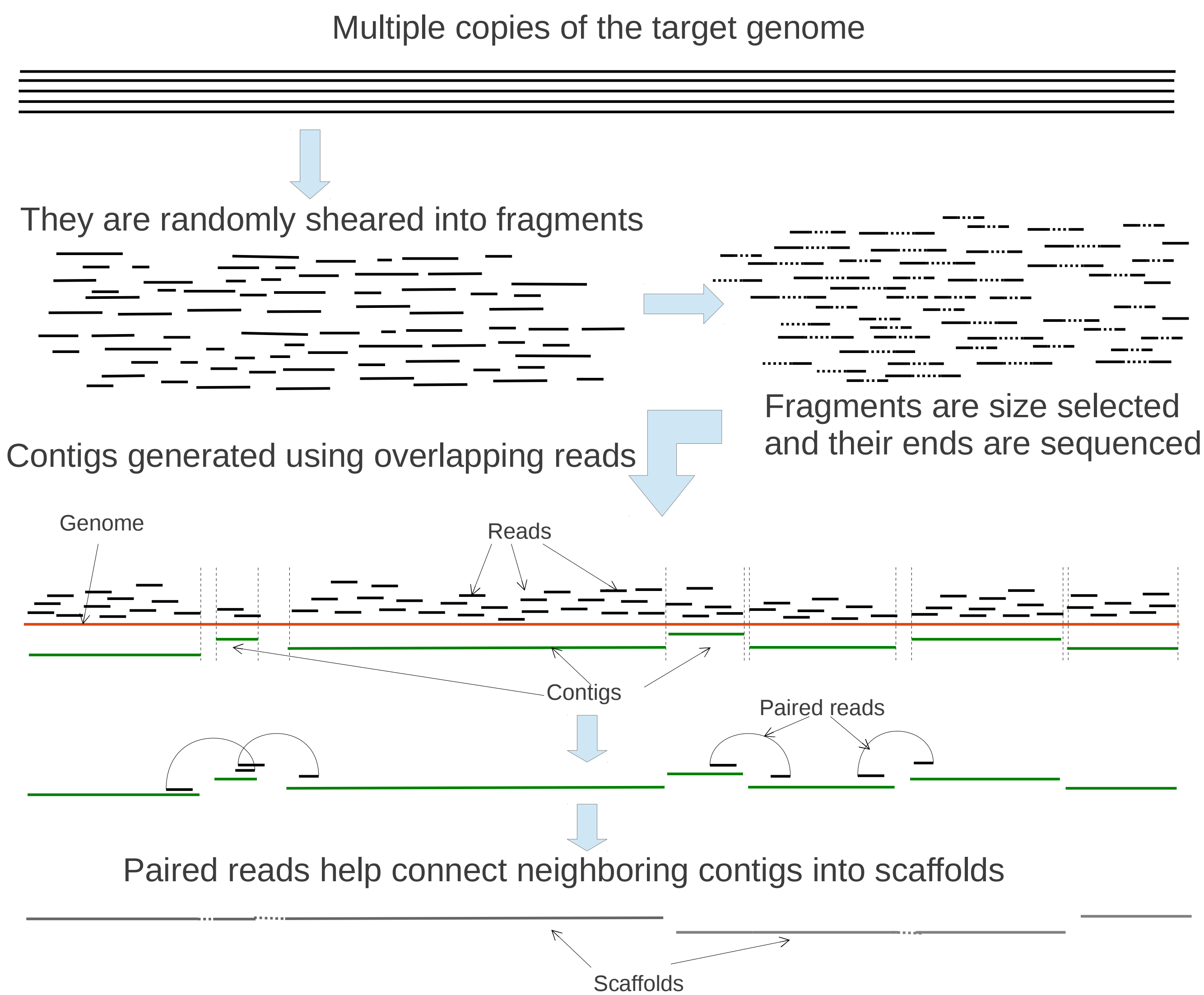
Genome Assembly



Genome Assembly



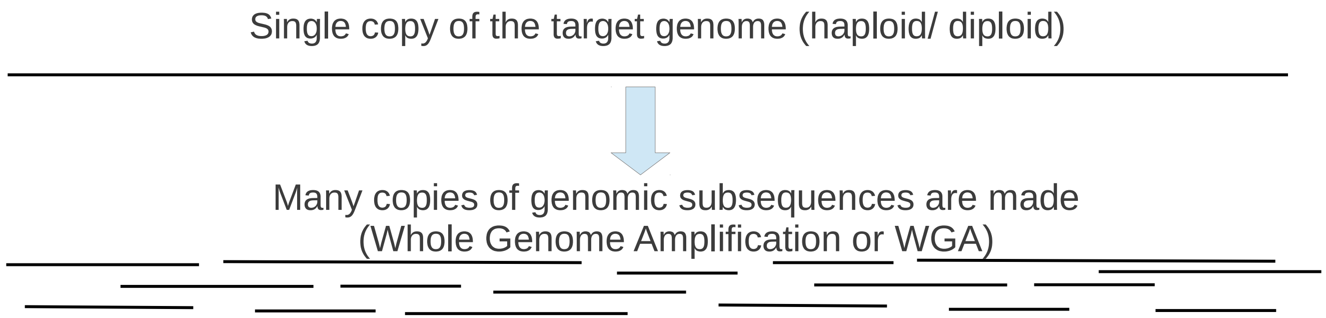
Genome Assembly



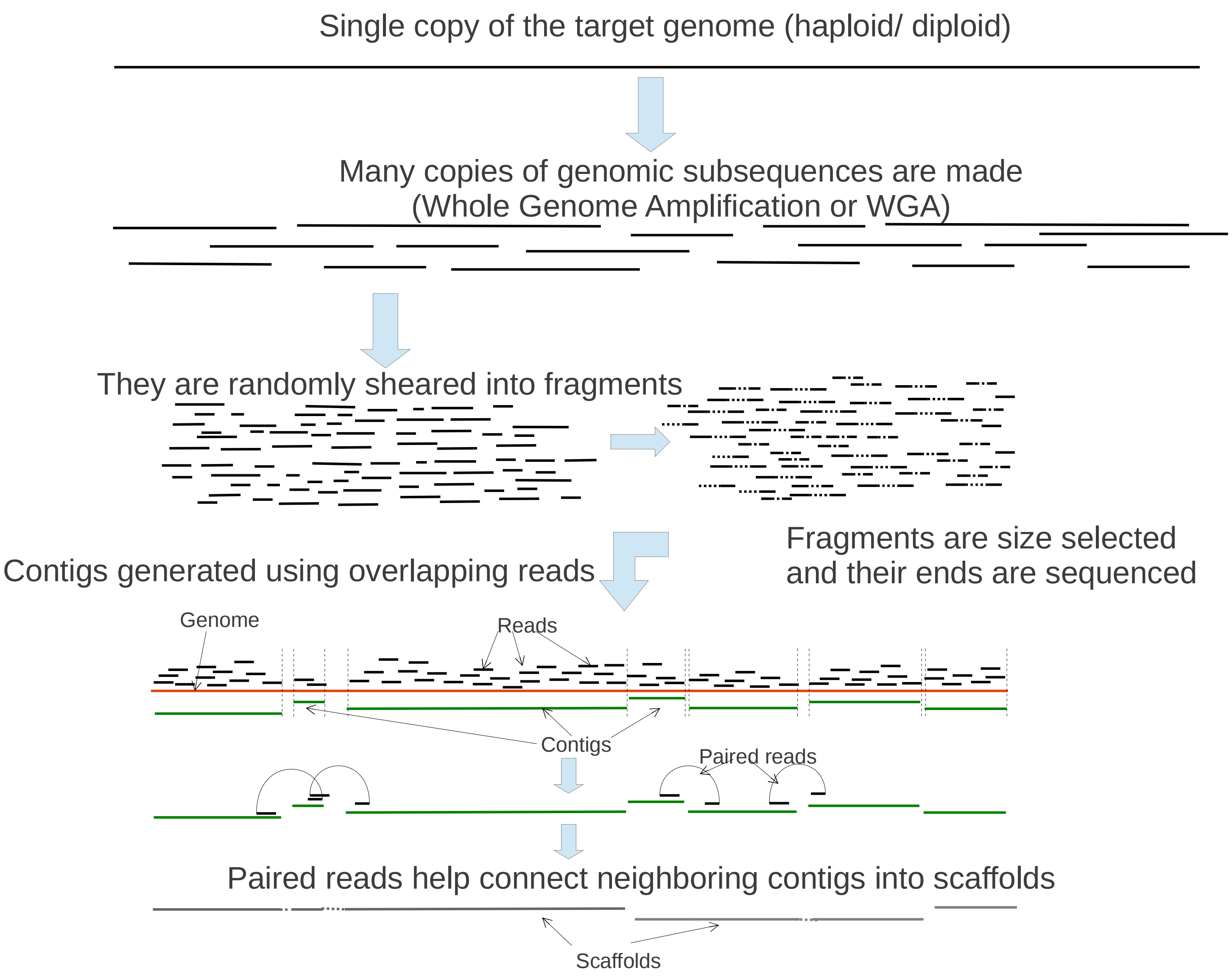
Single Cell Genomics

Single copy of the target genome (haploid/ diploid)

Single Cell Genomics

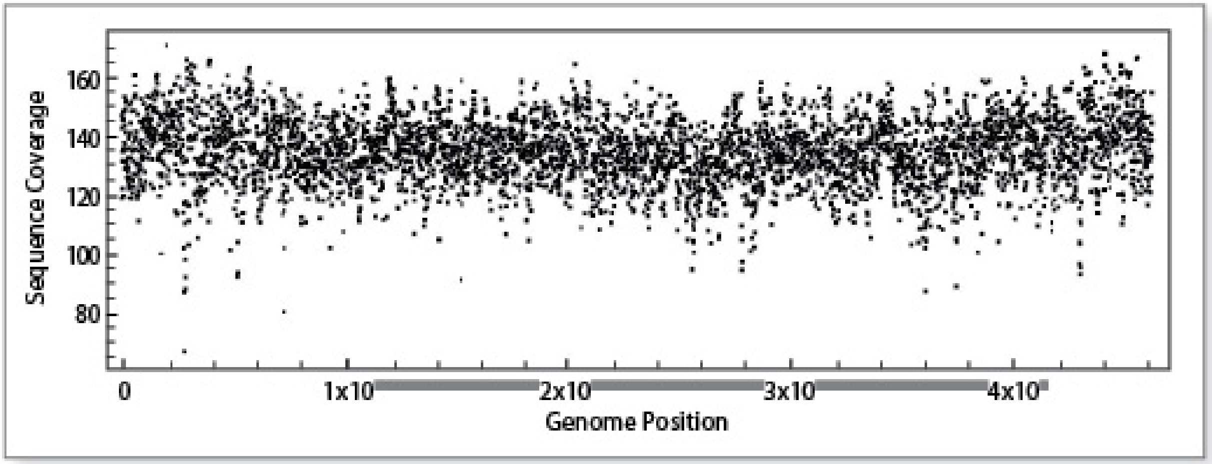


Single Cell Genomics



Cultured vs. amplified library coverage

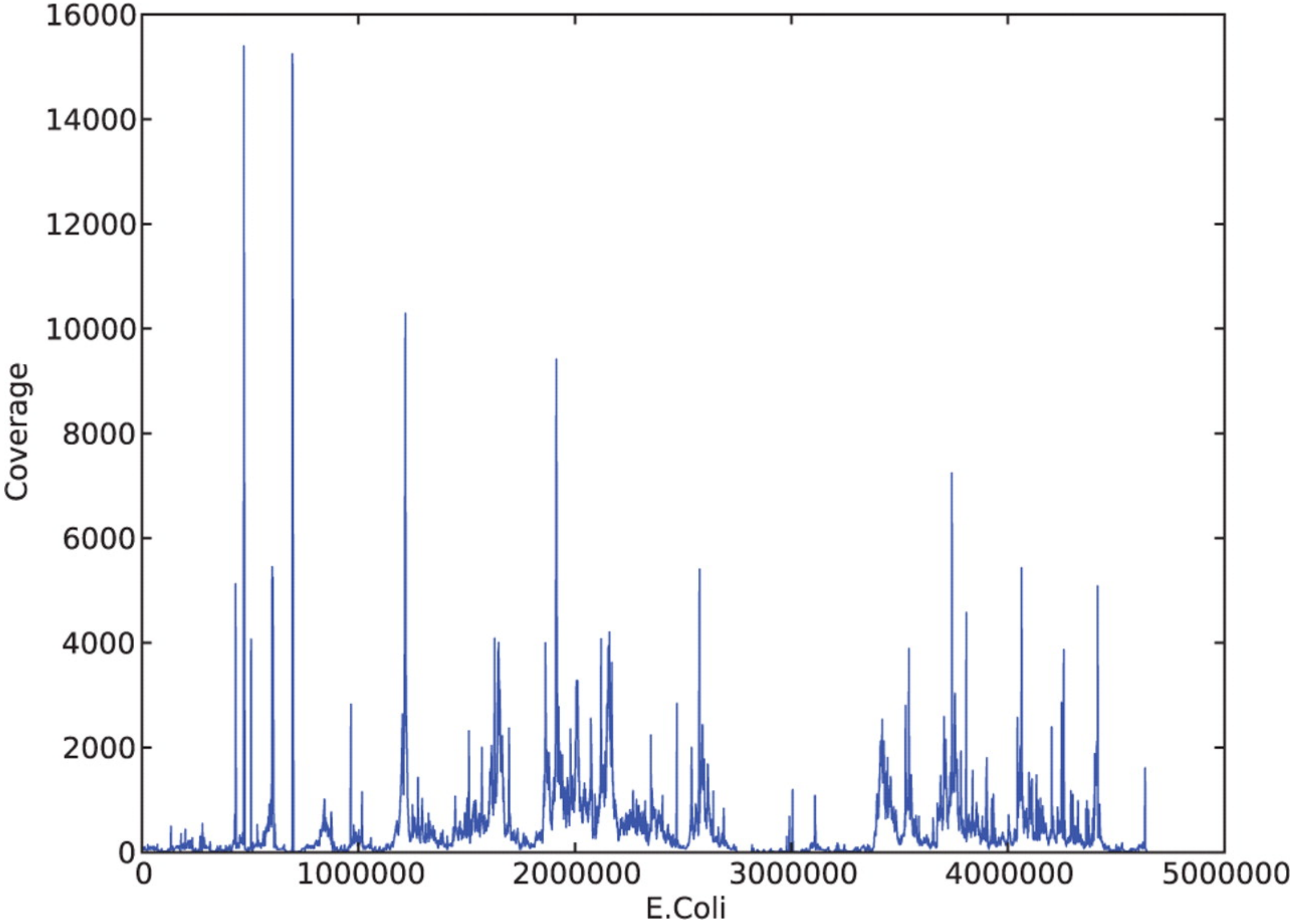
Sequencing libraries from culture show an almost uniform coverage.



Source: `http://www.neb.uk.com/whats_new/documents/NEBNext_Sample_Prep_Reagents.asp`

Cultured vs. amplified library coverage

Single cell libraries show a very uneven coverage[2].

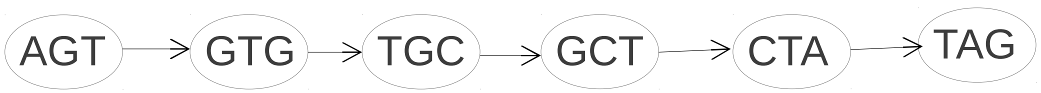


de novo Assembly with *de Bruijn* graphs

Genomic sequence: AGTGCTAG

Reads: AGTGCT, GTGCT, TGCTAG

3-mers: AGT, GTG, TGC, GCT, CTA, TAG



A walk from AGT to TAG spells out the genome AGTGCTAG

Assembling low coverage regions

Genomic sequence: AGTGCTAG

Reads: AGTGC, GCTAG

4-mers: AGTG, GTGC, GCTA, CTAG

AGTG → GTGC   GCTA → CTAG

3-mers: AGT, GTG, TGC, GCT, CTA, TAG

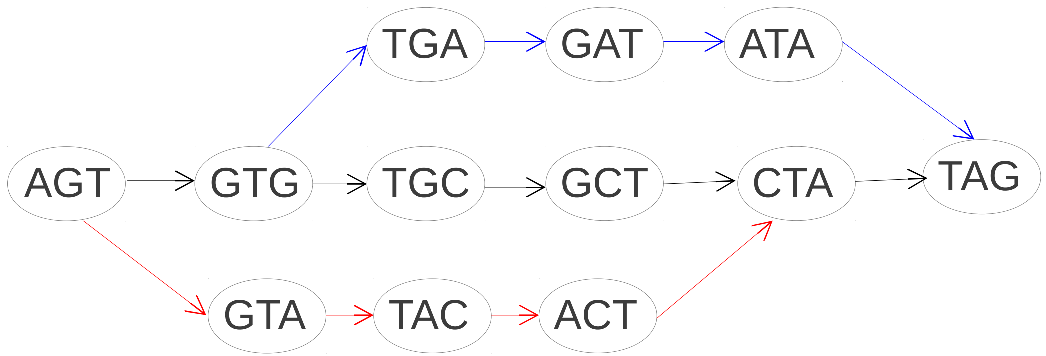
AGT → GTG → TGC → GCT → CTA → TAG

de Bruijn graph with erroneous *k*-mers

Genomic sequence: AGTGCTAG

Reads: AGTGCT, GTACT, TGATAG

3-mers: AGT, GTG, TGC, GCT, GTA, TAC, ACT, TGA, GAT, ATA, TAG



A walk from AGT to TAG spells out the genome AGTGCTAG

Common strategies for handling uneven coverage

1. Assembling low coverage regions using iterative assembly.
2. Correct reads from high coverage regions.

Single cell assembly on bacterial dataset

Assembly performance comparison of some Single Cell Assemblers. The dataset was a single cell MDA amplified read library with the following statistics: 6.3 Gbp in total, 29M reads, 2x100bp, insert size≈270bp. The reference genome is 4.64 Mbp in size.

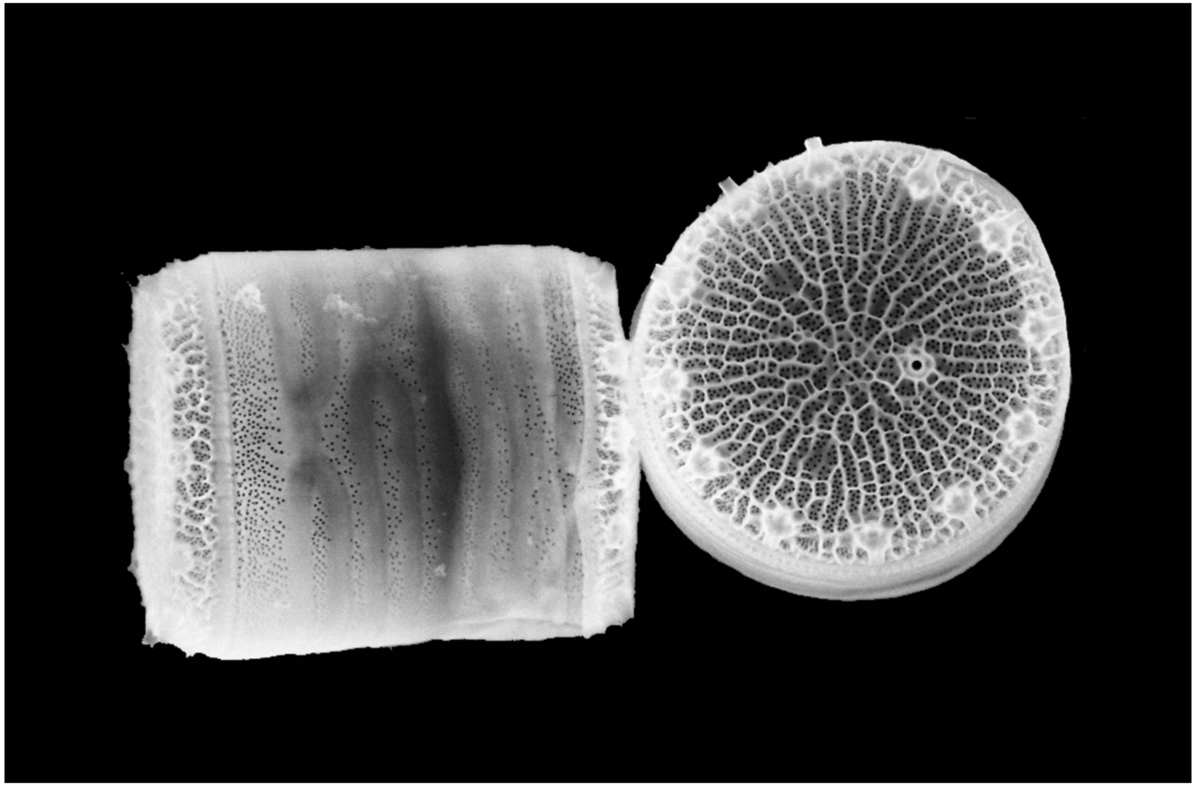
Tool	Assembly size (Mbp)	No of contigs	N50	Max Scaffold length	Genome coverage (%)
ABySS	4.35	**179**	68534	178720	88.25
IDBA1.1	4.81	244	98306	**284464**	94.89
SPAdes2.4	4.88	277	**110539**	269177	**95.62**

source: `http://bioinf.spbau.ru/spades`

How do they perform with eukaryotic genomes?

Our model single cell eukaryote

Thalassiosira pseudonana: Marine diatom.



Reference assembly available at `http://genome.jgi-psf.org/Thaps3/Thaps3.home.html`

source:`http://www.awi.de/fileadmin/user_upload/News/Press_Releases/2004/3._Quarter/Tpseudonana-2_p.jpg`

Test datasets.

Statistics of three (A, B, C) MDA-derived *T.pseudonana* cultured samples we used as test datasets. The reference genome length is 32.61Mbp. Bowtie 2 [1] was used for read mapping.

Sample	Read Library size (Gbp)	Mean read length	Std of read length	Mean insert length	Std of insert length	Mapped (%)
A	1.30	144.22	18.81	356.41	4189.19	92.34
B	0.98	144.28	19.26	394.33	4812.72	88.27
C	1.18	141.86	22.37	360.24	4902.67	90.16

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Total (Gbp)	Contig statistics Count	N50	Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Contig statistics: Total (Gbp)	Contig statistics: Count	Contig statistics: N50	Contig statistics: Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Contig statistics: Total (Gbp)	Count	N50	Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Contig statistics: Total (Gbp)	Count	N50	Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Contig statistics: Total (Gbp)	Count	N50	Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C. The reference assembly is 32.61Mbp.

Dataset	Assembler	Contig statistics: Total (Gbp)	Count	N50	Max (Kbp)	Time (hr:min)	Memory (GB)
A	ABySS	38.20	**10147**	11913	104	**05:54**	**3.8**
	IDBA1.1	36.22	36604	2571	47	36:48	9.0
	SPAdes2.4	35.47	12164	**39813**	**294**	14:35	19.8
B	ABySS	45.17	**9108**	17960	112	**08:50**	**4.2**
	IDBA1.1	38.83	39218	3137	87	39:50	10.9
	SPAdes2.4	41.43	18846	**26519**	**179**	20:32	23.6
C	ABySS	50.99	**9118**	22406	105	**11:14**	**4.1**
	IDBA1.1	40.70	39480	3778	87	44:12	11.6
	SPAdes2.4	43.08	16853	**28664**	**163**	28:55	22.9

Comparing contigs

Comparison of contig assembly for sample A, B, C considering only good contigs (those with $\geq$ 90% alignment to reference). Reference contigs were produced by mapping reads to the reference genome and declaring contiguous regions with a coverage of at least 3 to be contigs.

		Contig statistics			
Dataset	Assembler	Total (Gbp)	Count	N50	Max (Kbp)
A	ABySS	20.15	6276	**3005**	57
	IDBA1.1	**30.36**	27539	2214	46
	SPAdes2.4	8.34	**1240**	0	**121**
	Reference	*31.77*	*4535*	*27981*	*251*
B	ABySS	26.33	5211	**8913**	79
	IDBA1.1	**31.28**	27947	2709	87
	SPAdes2.4	19.55	**4784**	3140	**195**
	Reference	*32.20*	*1565*	*122162*	*441*
C	ABySS	30.85	5307	**13106**	104
	IDBA1.1	**32.36**	29080	2685	87
	SPAdes2.4	19.33	**4520**	3054	**155**
	Reference	*32.25*	*1018*	*250158*	*994*

Comparing contigs

Comparison of contig assembly for sample A, B, C considering only good contigs (those with $\geq$ 90% alignment to reference). Reference contigs were produced by mapping reads to the reference genome and declaring contiguous regions with a coverage of at least 3 to be contigs.

Dataset	Assembler	Contig statistics: Total (Gbp)	Count	N50	Max (Kbp)
A	ABySS	20.15	6276	**3005**	57
	IDBA1.1	**30.36**	27539	2214	46
	SPAdes2.4	8.34	**1240**	0	**121**
	Reference	*31.77*	*4535*	*27981*	*251*
B	ABySS	26.33	5211	**8913**	79
	IDBA1.1	**31.28**	27947	2709	87
	SPAdes2.4	19.55	**4784**	3140	**195**
	Reference	*32.20*	*1565*	*122162*	*441*
C	ABySS	30.85	5307	**13106**	104
	IDBA1.1	**32.36**	29080	2685	87
	SPAdes2.4	19.33	**4520**	3054	**155**
	Reference	*32.25*	*1018*	*250158*	*994*

Comparing contigs

Comparison of contig assembly for sample A, B, C considering only good contigs (those with $\geq$ 90% alignment to reference). Reference contigs were produced by mapping reads to the reference genome and declaring contiguous regions with a coverage of at least 3 to be contigs.

Dataset	Assembler	Contig statistics			
		Total (Gbp)	Count	N50	Max (Kbp)
A	ABySS	20.15	6276	**3005**	57
	IDBA1.1	**30.36**	27539	2214	46
	SPAdes2.4	8.34	**1240**	0	**121**
	Reference	*31.77*	*4535*	*27981*	*251*
B	ABySS	26.33	5211	**8913**	79
	IDBA1.1	**31.28**	27947	2709	87
	SPAdes2.4	19.55	**4784**	3140	**195**
	Reference	*32.20*	*1565*	*122162*	*441*
C	ABySS	30.85	5307	**13106**	104
	IDBA1.1	**32.36**	29080	2685	87
	SPAdes2.4	19.33	**4520**	3054	**155**
	Reference	*32.25*	*1018*	*250158*	*994*

Comparing contigs

Comparison of contig assembly for sample A, B, C considering only good contigs (those with $\geq 90\%$ alignment to reference). Reference contigs were produced by mapping reads to the reference genome and declaring contiguous regions with a coverage of at least 3 to be contigs.

Dataset	Assembler	Contig statistics Total (Gbp)	Count	N50	Max (Kbp)
A	ABySS	20.15	6276	**3005**	57
	IDBA1.1	**30.36**	27539	2214	46
	SPAdes2.4	8.34	**1240**	0	**121**
	Reference	*31.77*	*4535*	*27981*	*251*
B	ABySS	26.33	5211	**8913**	79
	IDBA1.1	**31.28**	27947	2709	87
	SPAdes2.4	19.55	**4784**	3140	**195**
	Reference	*32.20*	*1565*	*122162*	*441*
C	ABySS	30.85	5307	**13106**	104
	IDBA1.1	**32.36**	29080	2685	87
	SPAdes2.4	19.33	**4520**	3054	**155**
	Reference	*32.25*	*1018*	*250158*	*994*

Observations

- SPAdes produces long but often incorrectly assembled sequences.
- IDBA-UD produces the largest amount of correctly assembled (but more fragmented) sequences.
- ABySS produces the best sequences in terms of average length.
- ABySS is the most time and space efficient assembler.

Characterizing un-culturable single-cell organisms

Protein predictions and their annotations may help to:

- Perform phylogenetic analysis.
- Metabolic pathway analysis.

No transcriptome data is available.

Measuring assembly quality via protein prediction

Comparison of protein predictions for sample A, B, C. The reference protein library has 11849 proteins. CEGMA presents 458 core eukaryotic proteins.

Dataset	Assembler	Predicted proteins				
		KOGs		Total	Correct	
		(count)	(%)	(Augustus)	($>$ 70% alignment)	(%)
A	ABySS	380	82.96	12178	**9170**	**77.39**
	IDBA1.1	340	74.23	12266	8532	72.00
	SPAdes2.4	**388**	**84.71**	10659	7450	62.87
B	ABySS	**392**	**85.59**	14039	**10469**	**88.35**
	IDBA1.1	365	79.69	13748	8671	73.18
	SPAdes2.4	385	84.06	13403	8502	71.75
C	ABySS	**402**	**87.77**	16786	**11636**	**98.20**
	IDBA1.1	370	80.78	14099	8751	73.85
	SPAdes2.4	395	86.24	15366	9314	78.60

Measuring assembly quality via protein prediction

Comparison of protein predictions for sample A, B, C. The reference protein library has 11849 proteins. CEGMA presents 458 core eukaryotic proteins.

Dataset	Assembler	Predicted proteins				
		KOGs		Total	Correct	
		(count)	(%)	(Augustus)	(> 70% alignment)	(%)
A	ABySS	380	82.96	12178	**9170**	**77.39**
	IDBA1.1	340	74.23	12266	8532	72.00
	SPAdes2.4	**388**	**84.71**	10659	7450	62.87
B	ABySS	**392**	**85.59**	14039	**10469**	**88.35**
	IDBA1.1	365	79.69	13748	8671	73.18
	SPAdes2.4	385	84.06	13403	8502	71.75
C	ABySS	**402**	**87.77**	16786	**11636**	**98.20**
	IDBA1.1	370	80.78	14099	8751	73.85
	SPAdes2.4	395	86.24	15366	9314	78.60

Major conclusion

Average sequence length seems to be more important than N50 for protein prediction.

Improving single cell assembly

- We observed[1] that contig scaffolding can become much simpler and accurate when the contigs are correct.
- Perform read correction (using a combination of Multiple sequence alignment and k-mer spectrum based approaches) and coverage normalization as a preprocessing step.
- Design a scaffolder that attempts to maximize average sequence length while maintaining accuracy and genome coverage.

1. Rajat S. Roy, Kevin C. Chen, Anirvan M. Sengupta, and Alexander Schliep. SLIQ: simple linear inequalities for efficient contig scaffolding. Journal of Computational Biology, Oct 2012.

THANK YOU FOR YOUR ATTENTION

THANK YOU FOR YOUR ATTENTION

Ben Langmead and Steven L Salzberg.
Fast gapped-read alignment with bowtie 2.
Nat Methods, 9(4):357–359, Apr 2012.

Paul Medvedev, Eric Scott, Boyko Kakaradov, and Pavel Pevzner.
Error correction of high-throughput sequencing datasets with non-uniform coverage.
Bioinformatics, 27(13):i137–i141, Jul 2011.