Genome assembly

BIOL 7210: Computational Genomics - 2019

Hanying Pan, Jinkinson Smith, Linglin Zhang, Patrick Howard, Shrinkhla Sharma, Tianci Li and Zainab Arshad

What is genome assembly?

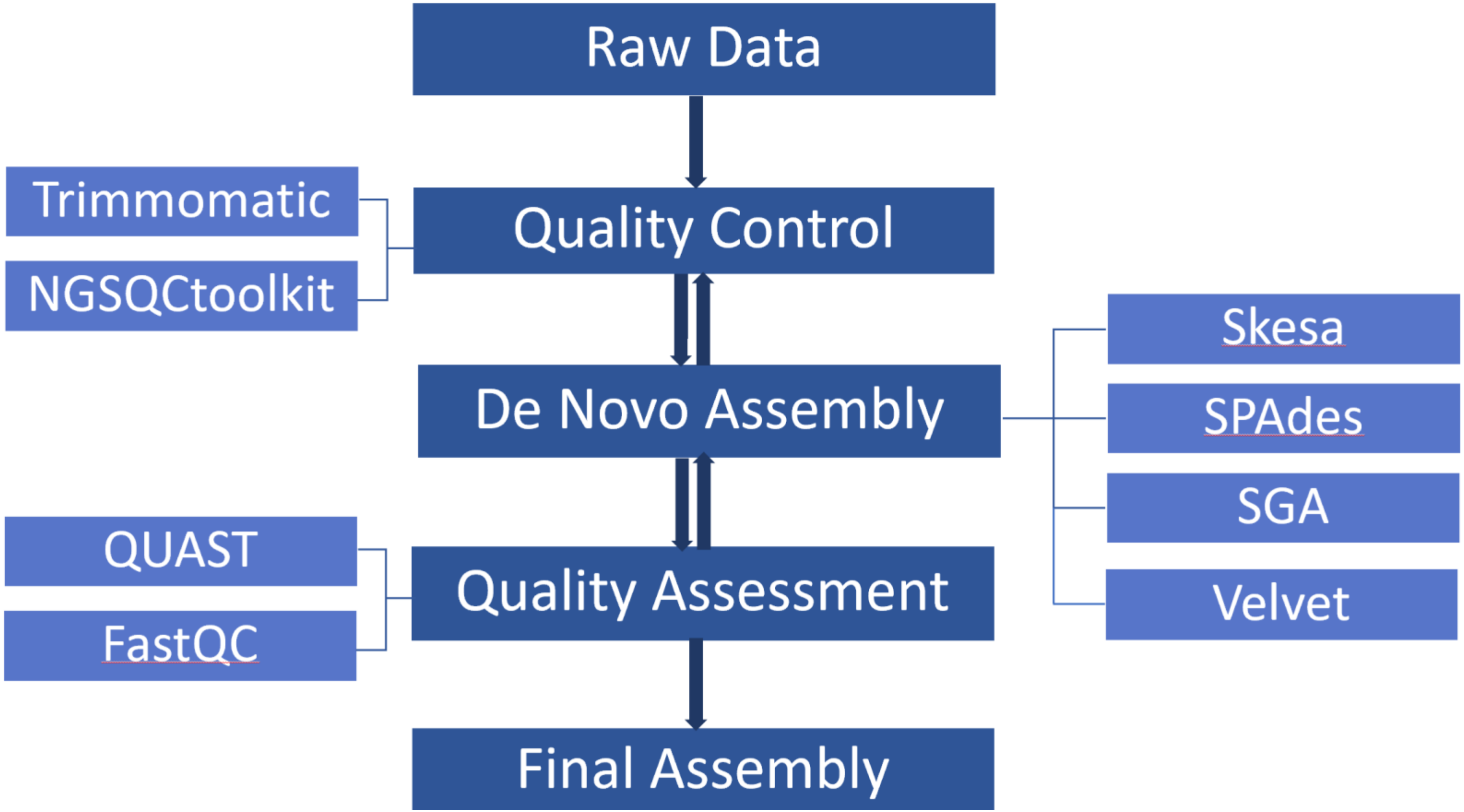
- Genome assembly is the process of taking individual, small DNA sequences, or reads, and reconstructing an organism’s genome
- Like many related genomics processes, genome assembly has become considerably faster and cheaper to perform in recent years
- De novo assembly, which we will focus on here, involves assembling a new genome for which there is no existing reference genome, i.e. “from scratch”

De novo assembly

- De novo assembly is the most common type of genome assembly for short read sequences
- It involves reconstructing an entire genome solely from overlapping sequence reads
- The quality of such an assembly depends on the size of the reads and the number of gaps between them
- The programs that perform de novo assemblies use either de Bruijn graphs or Overlap graphs
- This method can generate new, accurate reference sequences, even for complex genomes
- It takes more time when used to assemble longer genomes (e.g. those from eukaryotes)

Our objective

- We aim to perform de novo assembly based on 50 isolates
- To identify species from which the sequences were obtained
- Evaluate the performance of several tools relating to specific steps in the assembly pipeline
- Each tool was tested one isolate for which we had preliminary results


Raw Data
Quality Control
Trimmomatic
NGSQCtoolkit
De Novo Assembly
Skesa
SPAdes
SGA
Velvet
Quality Assessment
QUAST
FastQC
Final Assembly

NGSQCtoolkit

NGSQCtoolkit is a set of perl package that can do quality control assignment, convert file format, trimming and statistics. We only focus on the trimming package for this time.

- Trimming packages:
- TrimmingReads.pl: Tool for trimming reads from 5’ and/or 3’ end of the read(FASTQ or FASTA format)
- HomoPolymerTrimming.pl: Tool for trimming 3’ end of the reads from the first base of homopolymer of given length
- AmbiguityFiltering.pl: Tool for filtering reads containing ambiguous bases or trimming flanking ambiguous bases

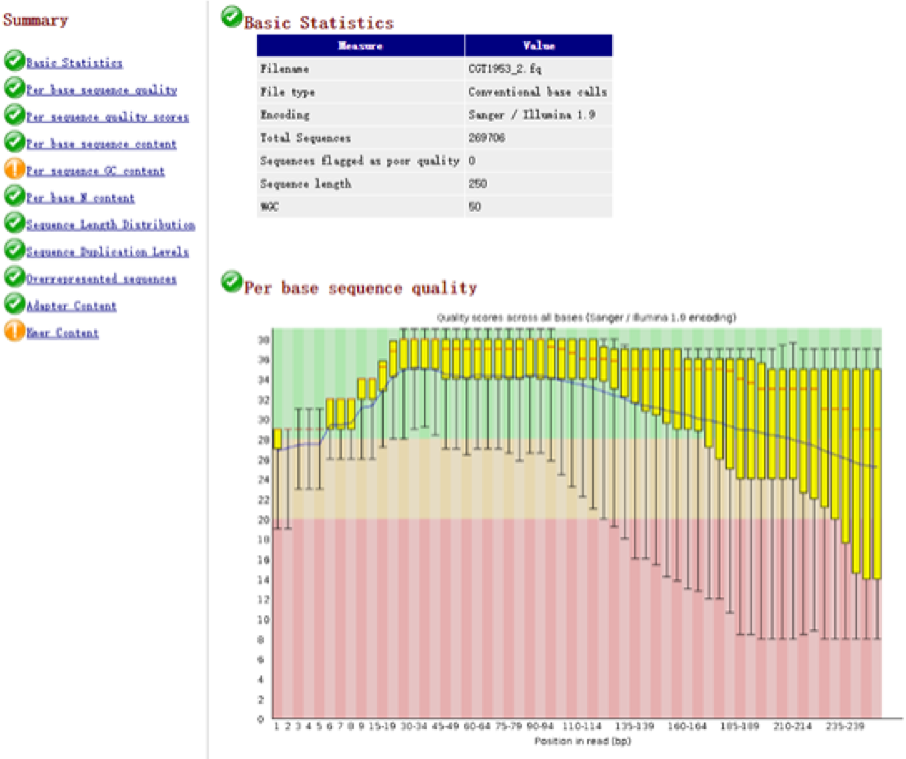
NGSQCtoolkit - Key value for trimming

- -l | -leftTrimBases <Integer> Number of bases to be trimmed from left end (5' end) default: 0
- -r | -rightTrimBases <Integer> Number of bases to be trimmed from right end (3' end)default:0
- -q | -qualCutOff <Integer> (Only for FASTQ files) Cut-off PHRED quality score for trimming reads from right end (3' end) For eg.: -q 20, will trim bases having PHRED quality score less than 20 at 3' end of the read Note: Quality trimming can be performed only if -l and -r are not used default: 0 (i.e. quality trimming is OFF)
- -n | -lenCutOff <Integer> Read length cut-off Reads shorter than given length will be discarded default: -1 (i.e. length filtering is OFF)

Trimmomatic

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length

NGSQCtoolkit



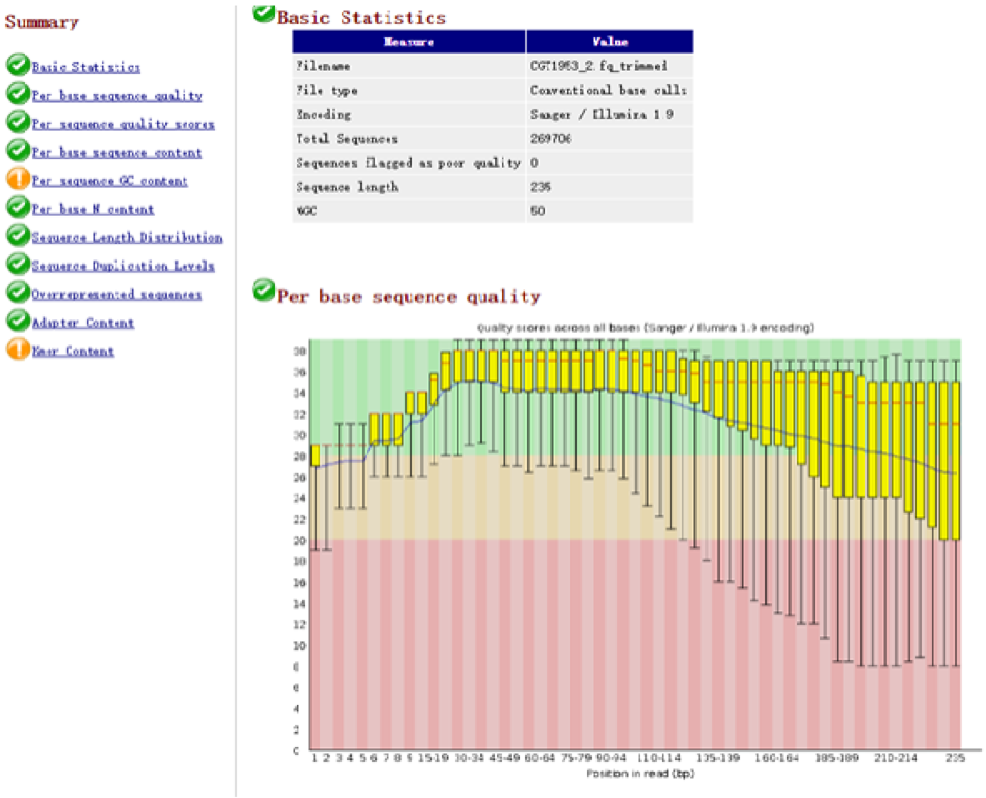
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

Measure	Value
Filename	CGT1953_2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	269706
Sequences flagged as poor quality	0
Sequence length	250
%GC	50

✓ Per base sequence quality



Summary

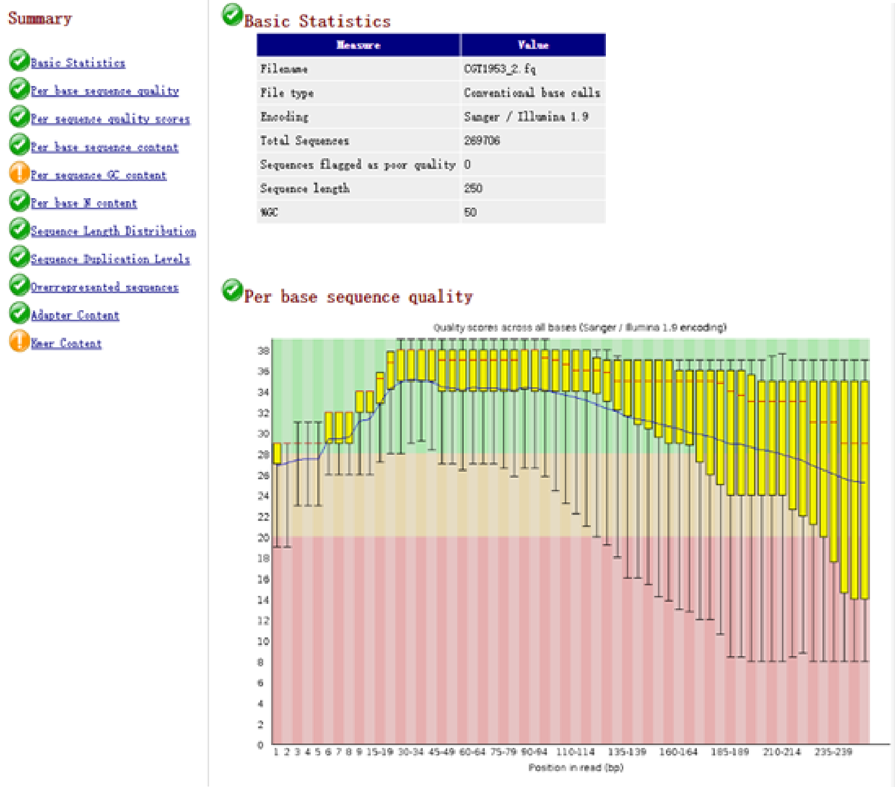
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

Measure	Value
Filename	CGT1953_2.fq_trimmed
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	269706
Sequences flagged as poor quality	0
Sequence length	235
%GC	50

✓ Per base sequence quality

Trimmomatic



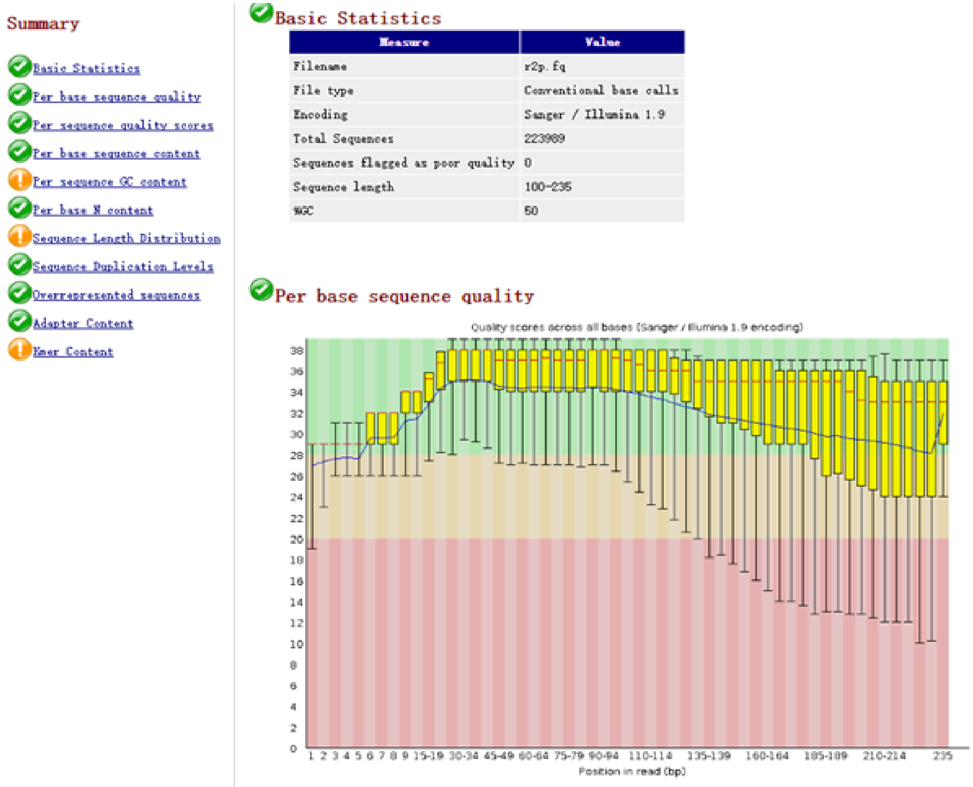
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

Measure	Value
Filename	CGT1953_2.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	269706
Sequences flagged as poor quality	0
Sequence length	250
%GC	50

✓ Per base sequence quality



Summary

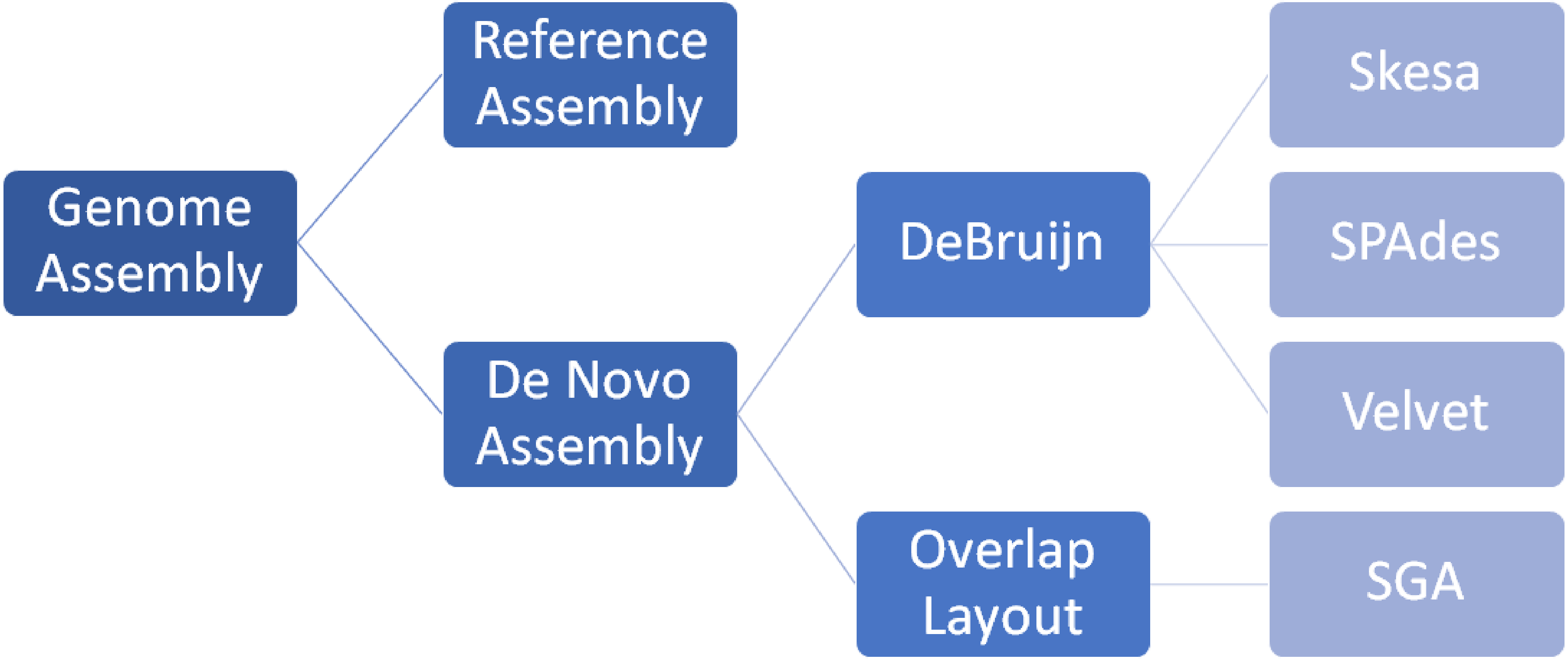
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Basic Statistics

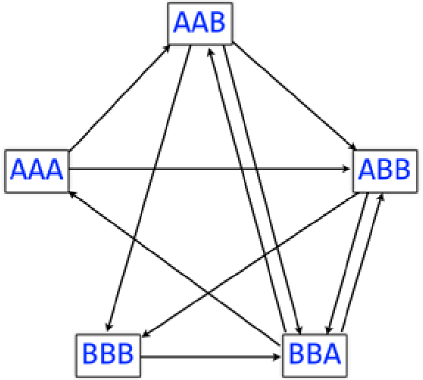
Measure	Value
Filename	r2p.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	223989
Sequences flagged as poor quality	0
Sequence length	100-235
%GC	50

✓ Per base sequence quality

Genome Assembly

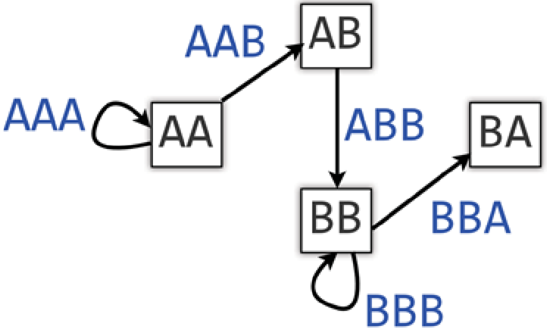


Overlap Graphs



- Nodes: reads
- Edges: Overlap > a threshold between two reads
- Graph simplification involves removing transitive edges
- Hamiltonian path gives is assembly
- Time Complexity O(N^2) because of we compare all pairs of reads for graph construction.

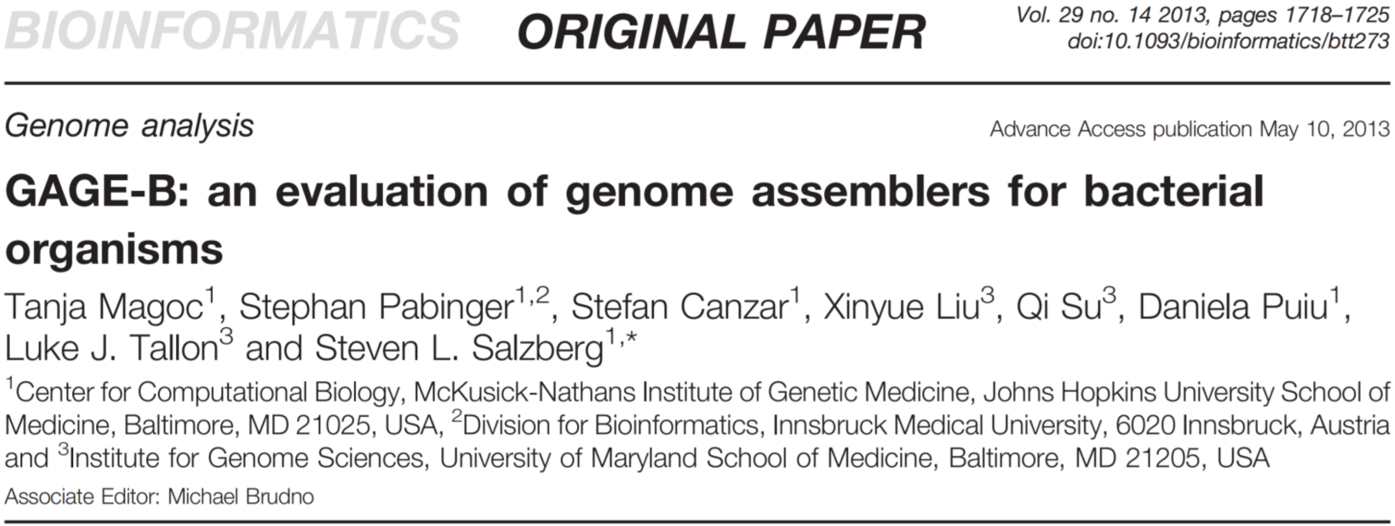
De Bruijn Graphs



- Nodes: (k-1)mers
- Edges: k-mers
- All dead-end, bubbles and cross edges removed for graph simplification
- Eulerian Walk is assembly
- Time Complexity O(N)
- Information is lost when reads are broken down into k-mers

Courtesy of Ben Langmead. http://www.langmead-lab.org/teaching-materials/

Reference Paper for Genome Assemblers



BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1718–1725
doi:10.1093/bioinformatics/btt273

Genome analysis

Advance Access publication May 10, 2013

GAGE-B: an evaluation of genome assemblers for bacterial organisms

Tanja Magoc[1], Stephan Pabinger[1,2], Stefan Canzar[1], Xinyue Liu[3], Qi Su[3], Daniela Puiu[1], Luke J. Tallon[3] and Steven L. Salzberg[1,*]

[1]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21025, USA, [2]Division for Bioinformatics, Innsbruck Medical University, 6020 Innsbruck, Austria and [3]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21205, USA

Associate Editor: Michael Brudno

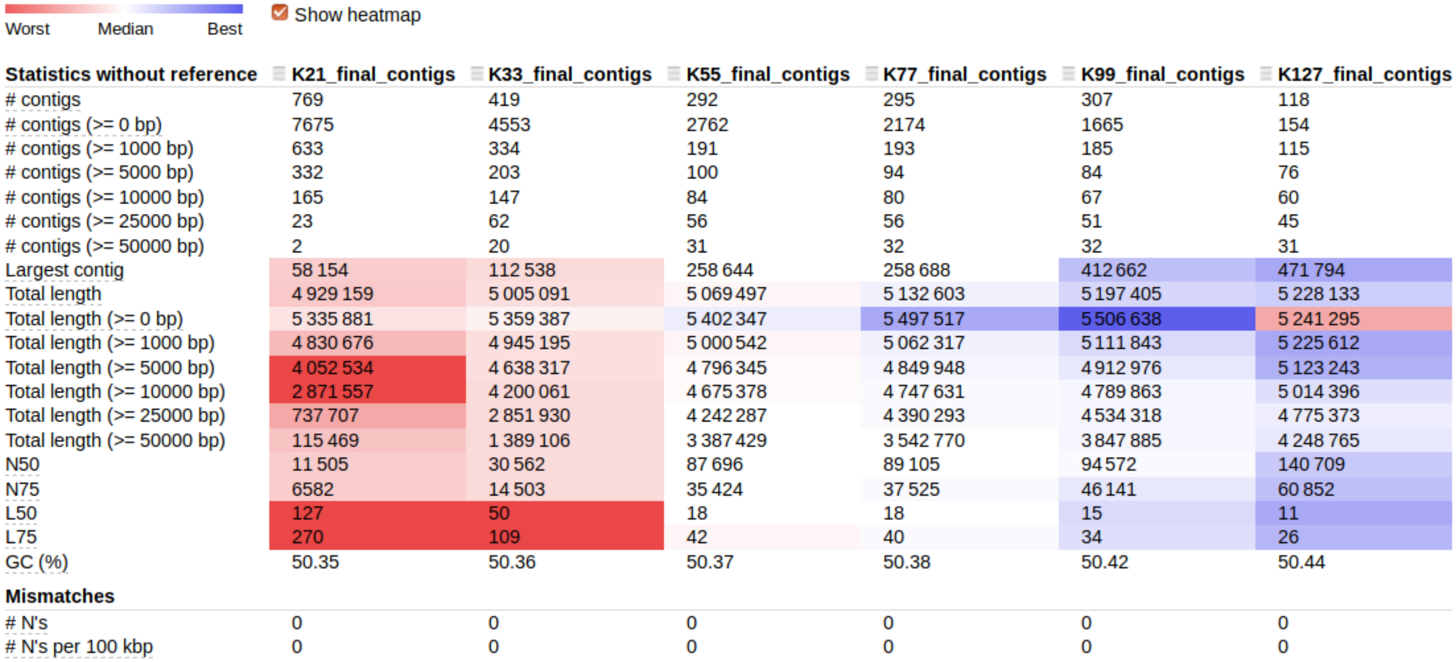
SPAdes

Pros:

- High quality assemblies with high N50 value and small number of contigs

Cons:

- Time consuming compared to Skeasa and Velvet
- Results are not perfectly reproducible

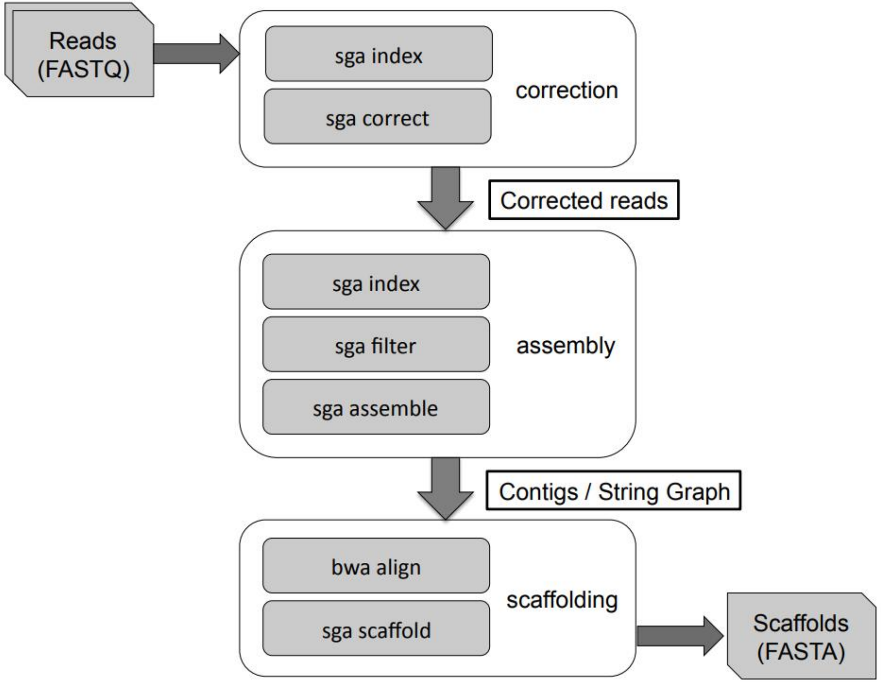


☑ Show heatmap

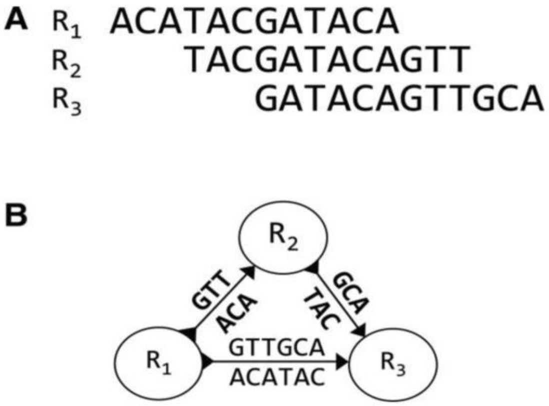
Statistics without reference	K21_final_contigs	K33_final_contigs	K55_final_contigs	K77_final_contigs	K99_final_contigs	K127_final_contigs
# contigs	769	419	292	295	307	118
# contigs (>= 0 bp)	7675	4553	2762	2174	1665	154
# contigs (>= 1000 bp)	633	334	191	193	185	115
# contigs (>= 5000 bp)	332	203	100	94	84	76
# contigs (>= 10000 bp)	165	147	84	80	67	60
# contigs (>= 25000 bp)	23	62	56	56	51	45
# contigs (>= 50000 bp)	2	20	31	32	32	31
Largest contig	58 154	112 538	258 644	258 688	412 662	471 794
Total length	4 929 159	5 005 091	5 069 497	5 132 603	5 197 405	5 228 133
Total length (>= 0 bp)	5 335 881	5 359 387	5 402 347	5 497 517	5 506 638	5 241 295
Total length (>= 1000 bp)	4 830 676	4 945 195	5 000 542	5 062 317	5 111 843	5 225 612
Total length (>= 5000 bp)	4 052 534	4 638 317	4 796 345	4 849 948	4 912 976	5 123 243
Total length (>= 10000 bp)	2 871 557	4 200 061	4 675 378	4 747 631	4 789 863	5 014 396
Total length (>= 25000 bp)	737 707	2 851 930	4 242 287	4 390 293	4 534 318	4 775 373
Total length (>= 50000 bp)	115 469	1 389 106	3 387 429	3 542 770	3 847 885	4 248 765
N50	11 505	30 562	87 696	89 105	94 572	140 709
N75	6582	14 503	35 424	37 525	46 141	60 852
L50	127	50	18	18	15	11
L75	270	109	42	40	34	26
GC (%)	50.35	50.36	50.37	50.38	50.42	50.44
Mismatches						
# N's	0	0	0	0	0	0
# N's per 100 kbp	0	0	0	0	0	0

Gurevich et al. QUAST: quality assessment tool for genome assemblies,Bioinformatics (2013) 29(8): 1072-1075

SGA - de novo sequence Assembler using String Graphs



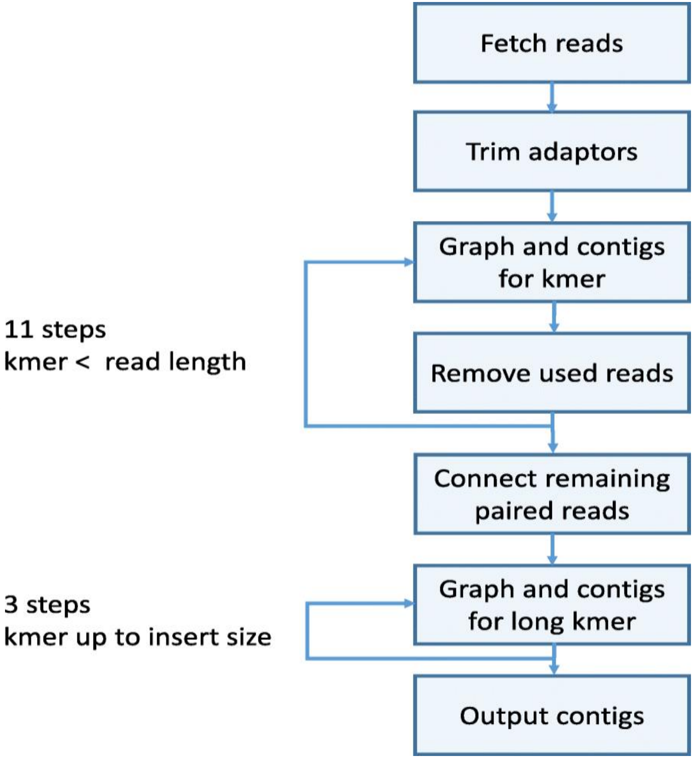
- FM-index construction
- Error correction
  - K-mer based
  - Overlap based
- Read filtering
- Read merging and assembly
- Paired end reads/Scaffolding



SKESA- Strategic k-mer extension for scrupulous assemblies

Algorithm design for SKESA

- Trimming of reads
- Assembly using different k-mer sizes
- Different k-mer sizes are used so that the shorter k-mer's can assemble the low coverage areas of the genome and longer k-mer's can resolve longer repeats
- k-mer size :
  - Varies from k-minimum(default 21 or can be entered by the user) average read length in a default of 11 iterations
  - Increases upto to insert size in 3 iterations
- At every iteration for a k-mer size De Brujin graph and contigs for that k-mer are produced and reads which are completely used up are removed as they cannot contribute any new information.
- After k-mer size has been varied upto the average read length, all the remaining paired reads are connected.



SKESA

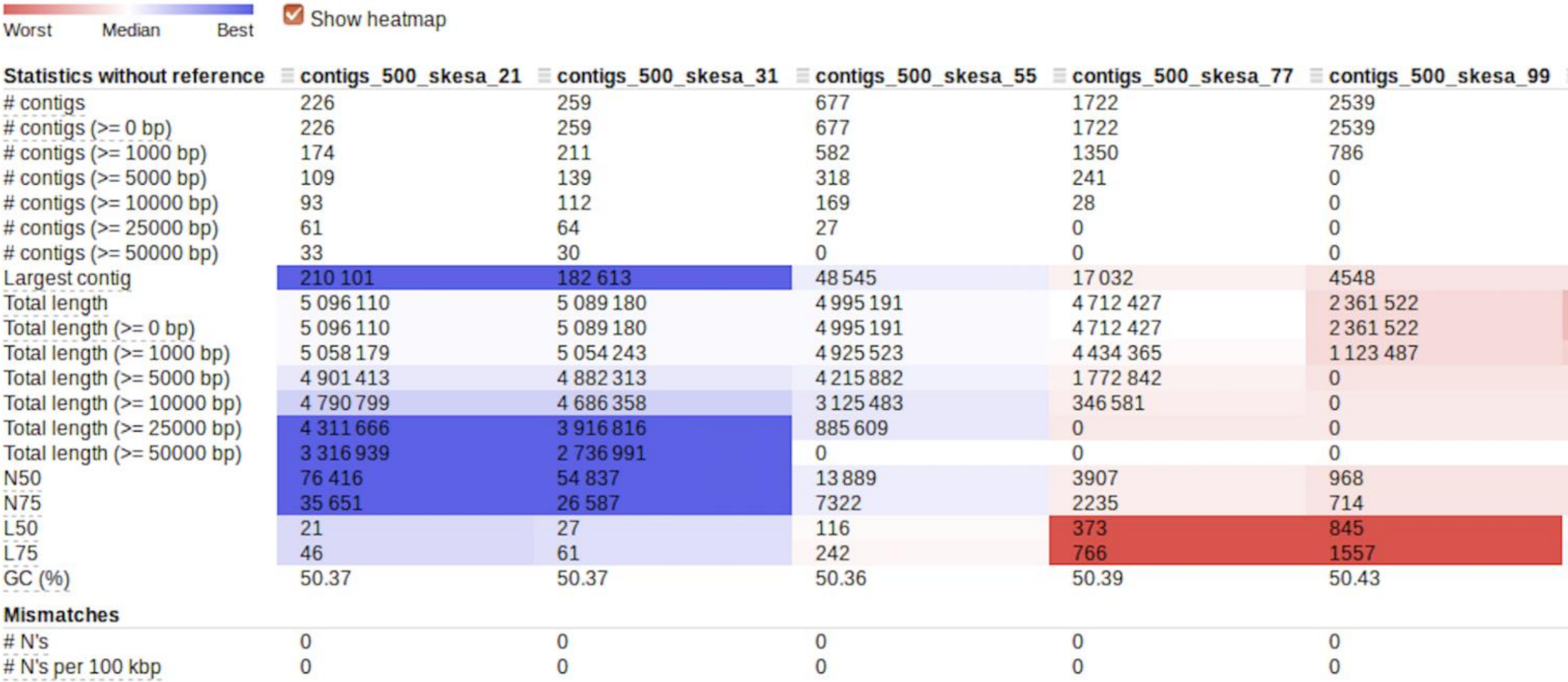
Pros

- It generates k-mers that are longer than mates and up to insert size. This feature allows SKESA to assemble regions accurately that have repeats shorter than insert size but longer than the mate length. To our knowledge, all current assemblers, in contrast, only use k-mers up to the size of mates.
- Extremely fast and produces consistent results for every run

Cons

- Does not has a built in scaffolding tool

SKESA



Worst Median Best

☑ Show heatmap

Statistics without reference	contigs_500_skesa_21	contigs_500_skesa_31	contigs_500_skesa_55	contigs_500_skesa_77	contigs_500_skesa_99
# contigs	226	259	677	1722	2539
# contigs (>= 0 bp)	226	259	677	1722	2539
# contigs (>= 1000 bp)	174	211	582	1350	786
# contigs (>= 5000 bp)	109	139	318	241	0
# contigs (>= 10000 bp)	93	112	169	28	0
# contigs (>= 25000 bp)	61	64	27	0	0
# contigs (>= 50000 bp)	33	30	0	0	0
Largest contig	210 101	182 613	48 545	17 032	4548
Total length	5 096 110	5 089 180	4 995 191	4 712 427	2 361 522
Total length (>= 0 bp)	5 096 110	5 089 180	4 995 191	4 712 427	2 361 522
Total length (>= 1000 bp)	5 058 179	5 054 243	4 925 523	4 434 365	1 123 487
Total length (>= 5000 bp)	4 901 413	4 882 313	4 215 882	1 772 842	0
Total length (>= 10000 bp)	4 790 799	4 686 358	3 125 483	346 581	0
Total length (>= 25000 bp)	4 311 666	3 916 816	885 609	0	0
Total length (>= 50000 bp)	3 316 939	2 736 991	0	0	0
N50	76 416	54 837	13 889	3907	968
N75	35 651	26 587	7322	2235	714
L50	21	27	116	373	845
L75	46	61	242	766	1557
GC (%)	50.37	50.37	50.36	50.39	50.43
Mismatches					
# N's	0	0	0	0	0
# N's per 100 kbp	0	0	0	0	0

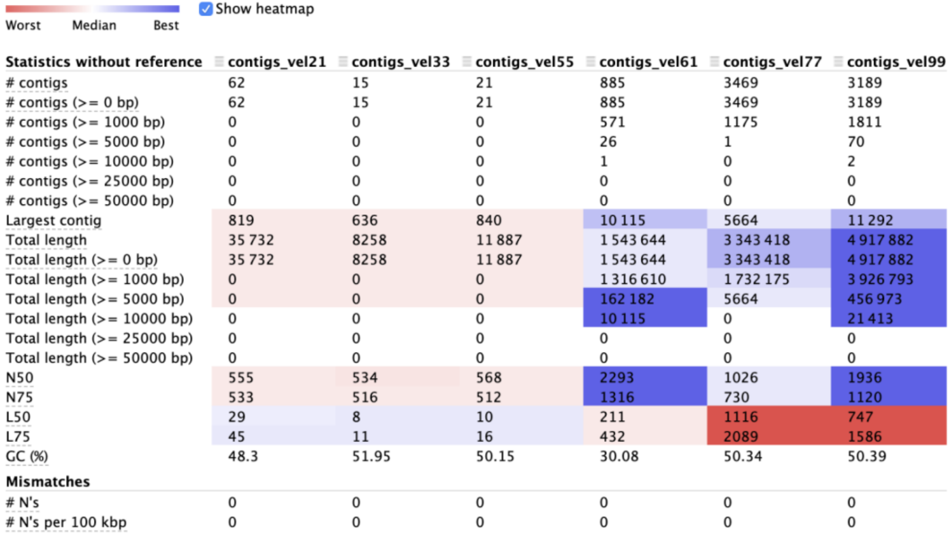
Velvet

Pros:

- Very fast computing time relative to other assembly tools

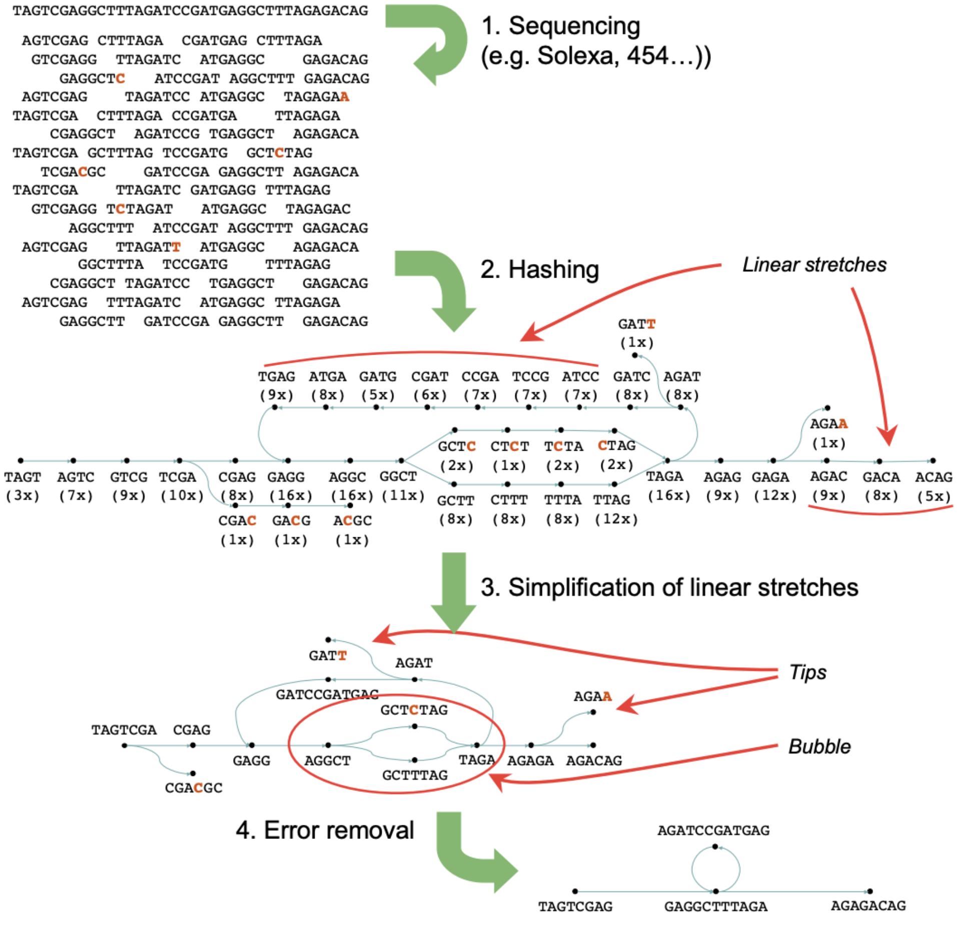
Cons:

- Optimum for high coverage, very short read (25-50 bp) datasets
- Bad at creating assemblies with optimal values for both N50 & L50



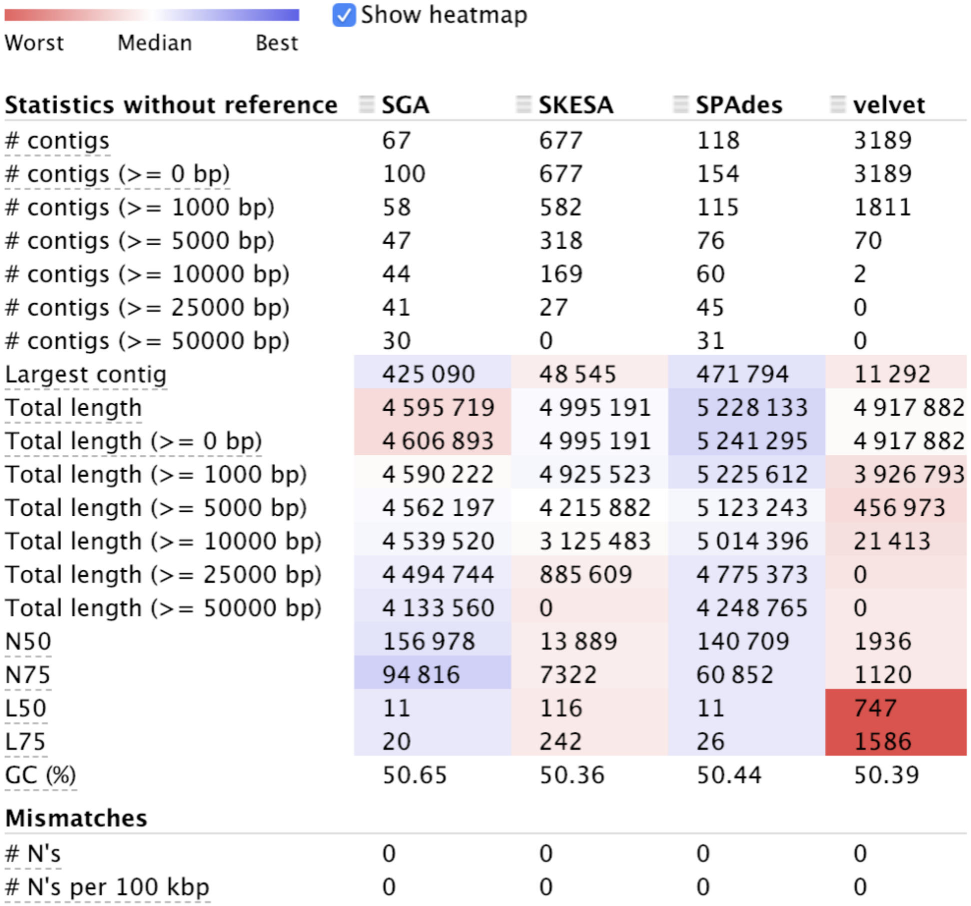
☑ Show heatmap

Statistics without reference	contigs_vel21	contigs_vel33	contigs_vel55	contigs_vel61	contigs_vel77	contigs_vel99
# contigs	62	15	21	885	3469	3189
# contigs (>= 0 bp)	62	15	21	885	3469	3189
# contigs (>= 1000 bp)	0	0	0	571	1175	1811
# contigs (>= 5000 bp)	0	0	0	26	1	70
# contigs (>= 10000 bp)	0	0	0	1	0	2
# contigs (>= 25000 bp)	0	0	0	0	0	0
# contigs (>= 50000 bp)	0	0	0	0	0	0
Largest contig	819	636	840	10 115	5664	11 292
Total length	35 732	8258	11 887	1 543 644	3 343 418	4 917 882
Total length (>= 0 bp)	35 732	8258	11 887	1 543 644	3 343 418	4 917 882
Total length (>= 1000 bp)	0	0	0	1 316 610	1 732 175	3 926 793
Total length (>= 5000 bp)	0	0	0	162 182	5664	456 973
Total length (>= 10000 bp)	0	0	0	10 115	0	21 413
Total length (>= 25000 bp)	0	0	0	0	0	0
Total length (>= 50000 bp)	0	0	0	0	0	0
N50	555	534	568	2293	1026	1936
N75	533	516	512	1316	730	1120
L50	29	8	10	211	1116	747
L75	45	11	16	432	2089	1586
GC (%)	48.3	51.95	50.15	30.08	50.34	50.39
Mismatches						
# N's	0	0	0	0	0	0
# N's per 100 kbp	0	0	0	0	0	0



A. Initial pipeline of the Velvet package.

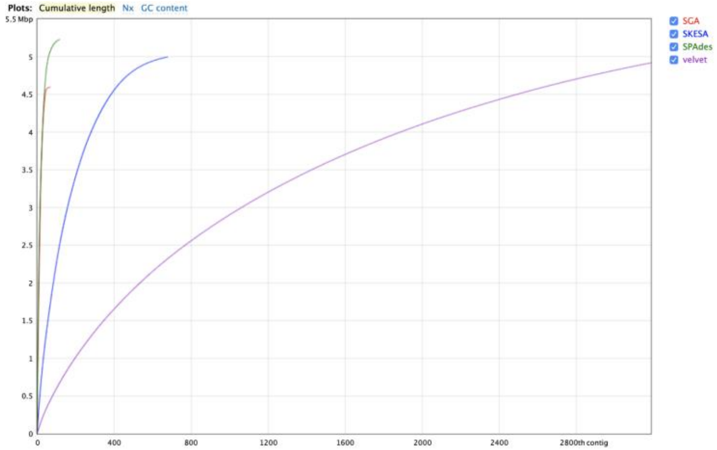
Quality Assessment of Assembly Tools (w/ QUAST)

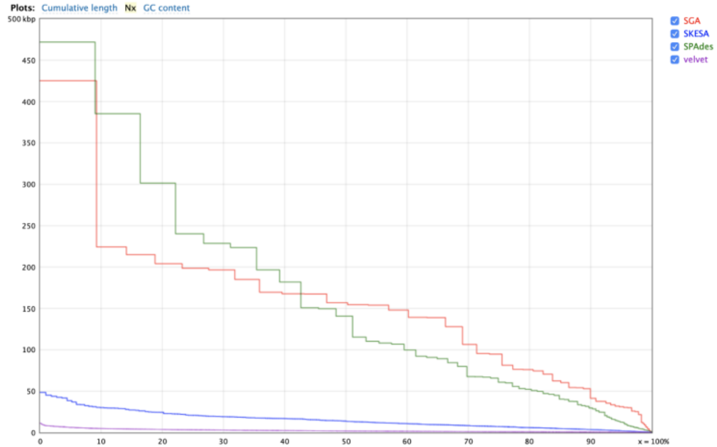


Worst Median Best

☑ Show heatmap

Statistics without reference	SGA	SKESA	SPAdes	velvet
# contigs	67	677	118	3189
# contigs (>= 0 bp)	100	677	154	3189
# contigs (>= 1000 bp)	58	582	115	1811
# contigs (>= 5000 bp)	47	318	76	70
# contigs (>= 10000 bp)	44	169	60	2
# contigs (>= 25000 bp)	41	27	45	0
# contigs (>= 50000 bp)	30	0	31	0
Largest contig	425 090	48 545	471 794	11 292
Total length	4 595 719	4 995 191	5 228 133	4 917 882
Total length (>= 0 bp)	4 606 893	4 995 191	5 241 295	4 917 882
Total length (>= 1000 bp)	4 590 222	4 925 523	5 225 612	3 926 793
Total length (>= 5000 bp)	4 562 197	4 215 882	5 123 243	456 973
Total length (>= 10000 bp)	4 539 520	3 125 483	5 014 396	21 413
Total length (>= 25000 bp)	4 494 744	885 609	4 775 373	0
Total length (>= 50000 bp)	4 133 560	0	4 248 765	0
N50	156 978	13 889	140 709	1936
N75	94 816	7322	60 852	1120
L50	11	116	11	747
L75	20	242	26	1586
GC (%)	50.65	50.36	50.44	50.39
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0





Questions?

References

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Patel RK, Jain M (2012). NGS QC Toolkit: A toolkit for quality control of next generation sequencing data.

Dominguez Del Angel V, Hjerde E, Sterck L *et al.* Ten steps to get started in Genome Assembly and Annotation [version 1; referees: 2 approved]. *F1000Research* 2018, 7(ELIXIR):148 (https://doi.org/10.12688/f1000research.13598.1)

De Novo Sequencing. Illumina. [accessed 2019 Jan 30]. https://www.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing/de-novo-sequencing.html

Zerbino, D.R. & Birney, Ewan (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. (https://genome.cshlp.org/content/18/5/821.short)

Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, Versatile genome assembly evaluation with QUAST-LG, *Bioinformatics* (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266 First published online: June 27, 2018

Alexandre Souvorov, Richa Agarwala and David J. Lipman SKESA: strategic k-mer extension for scrupulous assemblies

https://doi.org/10.1186/s13059-018-1540-z

Supplementary Slides

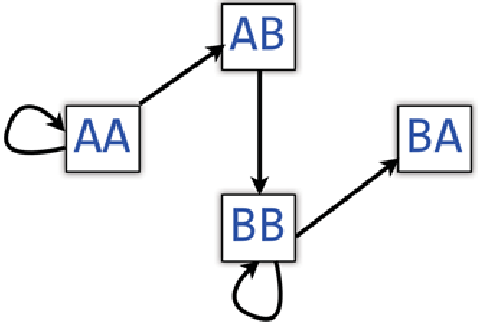
De Bruijn Graphs and Eulerian Walk

AAABBBA

take all 3-mers: AAA, AAB, ABB, BBB, BBA

form L/R 2-mers: AA, AA, AA, AB, AB, BB, BB, BB, BB, BA

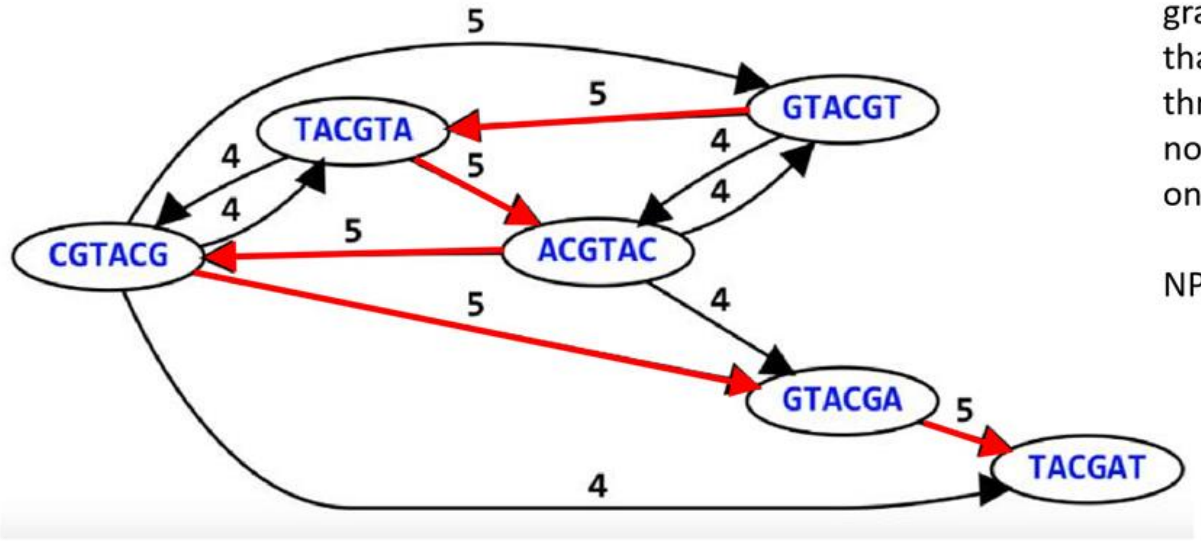
L R L R L R L R L R



Layout – graph traversal for assembly

Nodes: all 6-mers from GTACGTACGAT

Edges: overlaps of length ≥4



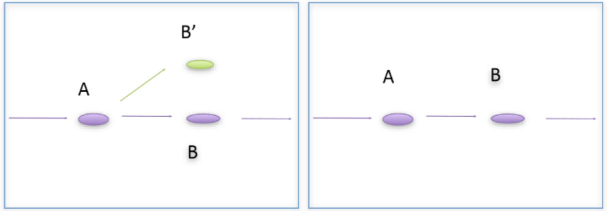
Hamiltonian path

graph traversal that passes through each node (read) only once

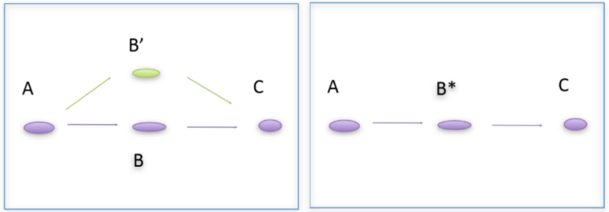
NP complete

Error Correction

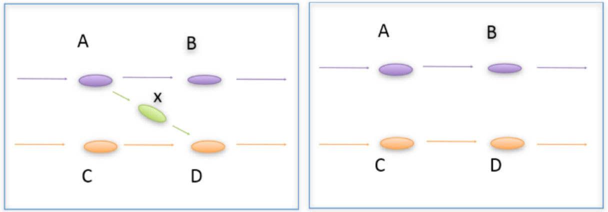
- Errors at end of read
  - Trim off 'dead-end' tips



- Errors in middle of read
  - Pop Bubbles

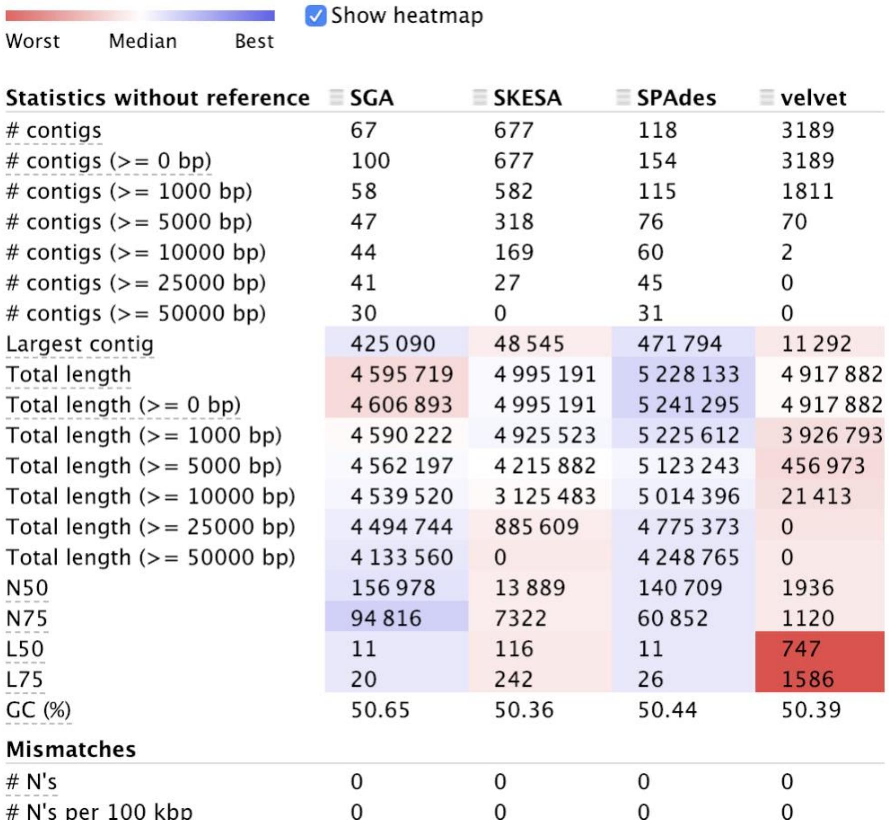


- Chimeric Edges
  - Clip short, low coverage nodes



Quality Assessment of Assembly Tools (w/ QUAST)

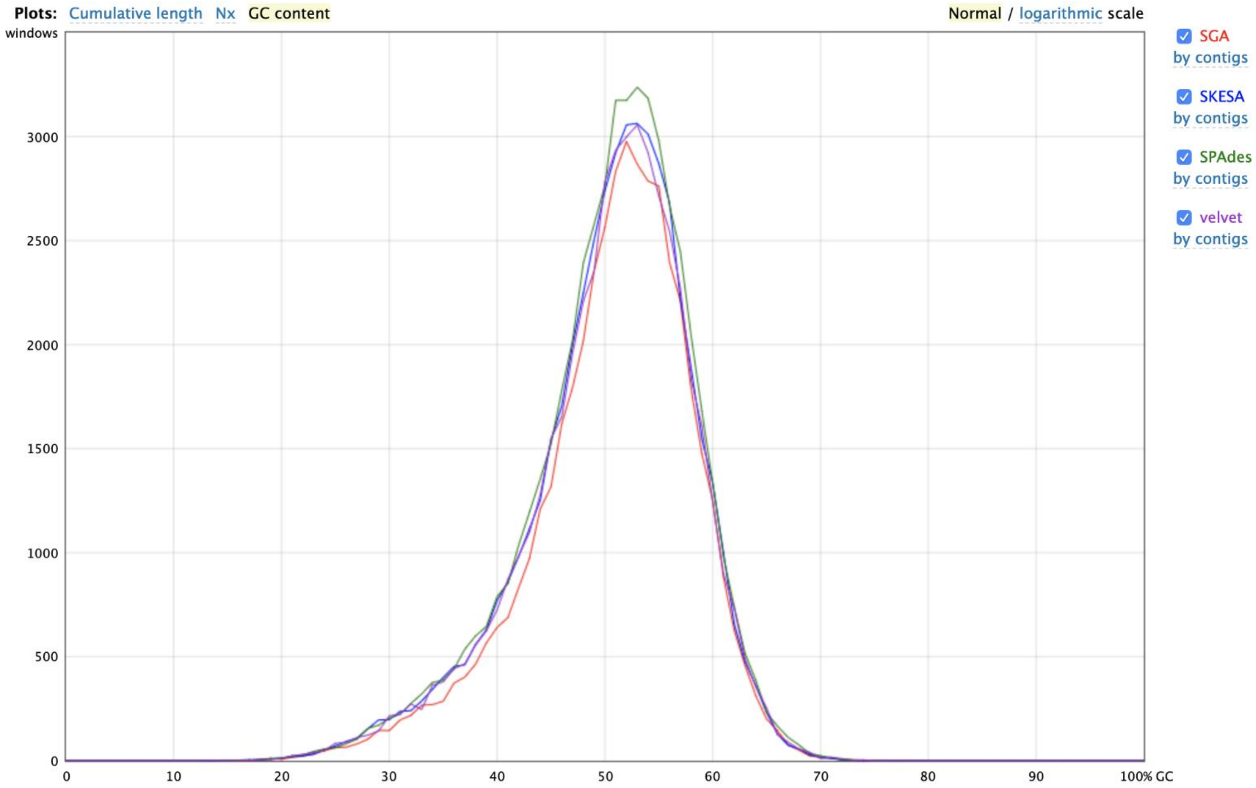
- Can evaluate assembly quality from multiple assemblers without the need of a reference genome
- This tool was created from a combination of previously used methods & quality metrics
- Easy to read and evaluate results from HTML reports that include plots
- Contains Icarus, a tool used to visualize & browse through contig alignment from each read simultaneously
- Contains tools to help with gene prediction for downstream analysis



☑ Show heatmap

Statistics without reference	SGA	SKESA	SPAdes	velvet
# contigs	67	677	118	3189
# contigs (>= 0 bp)	100	677	154	3189
# contigs (>= 1000 bp)	58	582	115	1811
# contigs (>= 5000 bp)	47	318	76	70
# contigs (>= 10000 bp)	44	169	60	2
# contigs (>= 25000 bp)	41	27	45	0
# contigs (>= 50000 bp)	30	0	31	0
Largest contig	425 090	48 545	471 794	11 292
Total length	4 595 719	4 995 191	5 228 133	4 917 882
Total length (>= 0 bp)	4 606 893	4 995 191	5 241 295	4 917 882
Total length (>= 1000 bp)	4 590 222	4 925 523	5 225 612	3 926 793
Total length (>= 5000 bp)	4 562 197	4 215 882	5 123 243	456 973
Total length (>= 10000 bp)	4 539 520	3 125 483	5 014 396	21 413
Total length (>= 25000 bp)	4 494 744	885 609	4 775 373	0
Total length (>= 50000 bp)	4 133 560	0	4 248 765	0
N50	156 978	13 889	140 709	1936
N75	94 816	7322	60 852	1120
L50	11	116	11	747
L75	20	242	26	1586
GC (%)	50.65	50.36	50.44	50.39
Mismatches				
# N's	0	0	0	0
# N's per 100 kbp	0	0	0	0

Quality Assessment of Assembly Tools (w/ QUAST)



Similar GC content plots in all 4 assembly tools (w/ slightly more obvserved in SPAdes)