Genome Assembly: Lab & Preliminary Results

Monday, February 15, 2016

BIOL 7210: Genome Assembly Group

Aroon Chande, Cheng Chen, Alicia Francis, Alli Gombolay, Namrata Kalsi, Ellie Kim, Tyrone Lee, Wilson Martin, Tannishtha Som, Peijue Zhang

Outline

1. Project Data
2. Updated Genome Pipeline
3. Comparison of Assemblers
4. Read Assembly: *de novo*, reference-guided, and hybrid
5. Assembly improvement
6. Assembly quality assessment

Objectives

1. Research classical and new tools to use
2. Evaluate and compare these available tools
3. Choose and combine best assemblers to utilize
4. Create wrapper for pipeline to increase efficiency

Project Data

Project Data

- *Haemophilus influenzae*
  - 36 samples from CDC

- Sequencing data:
  - Paired-end, 250bp reads
  - From Illumina HiSeq2500
  - Already demultiplexed

http://www.denniskunkel.com/DK/Bacteria/97500E.html

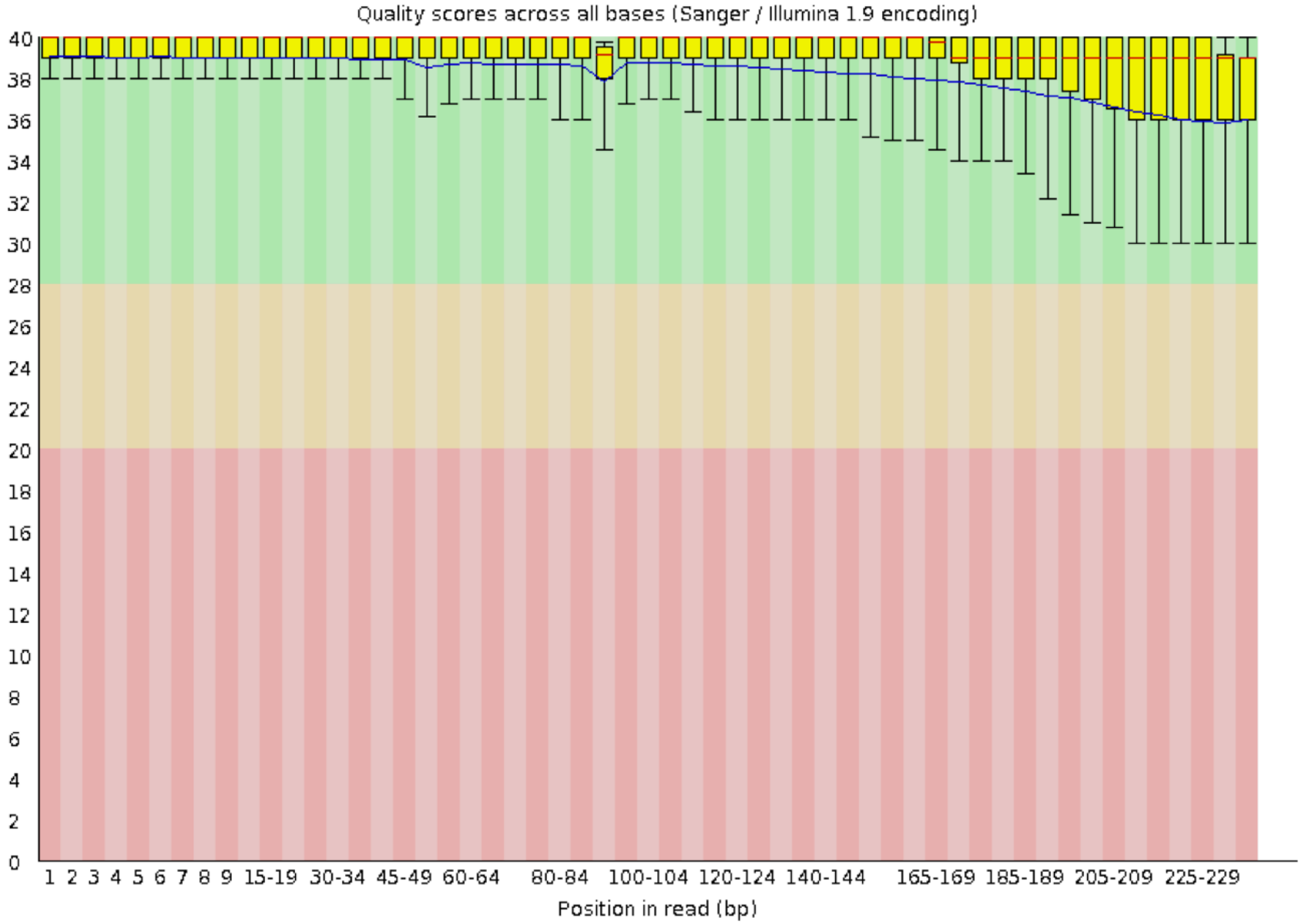
Test Data

```
M05964_HUY4067A110_TCCGGAGA-TAATCTTA_L002_R1_001_val_1.fq.gz
M05964_HUY4067A110_TCCGGAGA-TAATCTTA_L002_R2_001_val_2.fq.gz

M07572_HUY4067A111_TCCGGAGA-CAGGACGT_L002_R1_001_val_1.fq.gz
M07572_HUY4067A111_TCCGGAGA-CAGGACGT_L002_R2_001_val_2.fq.gz

M10540_HUY4067A3_ATTACTCG-CCTATCCT_L001_R1_001_val_1.fq.gz
M10540_HUY4067A3_ATTACTCG-CCTATCCT_L001_R2_001_val_2.fq.gz

M16180_HUY4067A127_GAGATTCC-CAGGACGT_L002_R1_001_val_1.fq.gz
M16180_HUY4067A127_GAGATTCC-CAGGACGT_L002_R2_001_val_2.fq.gz
```

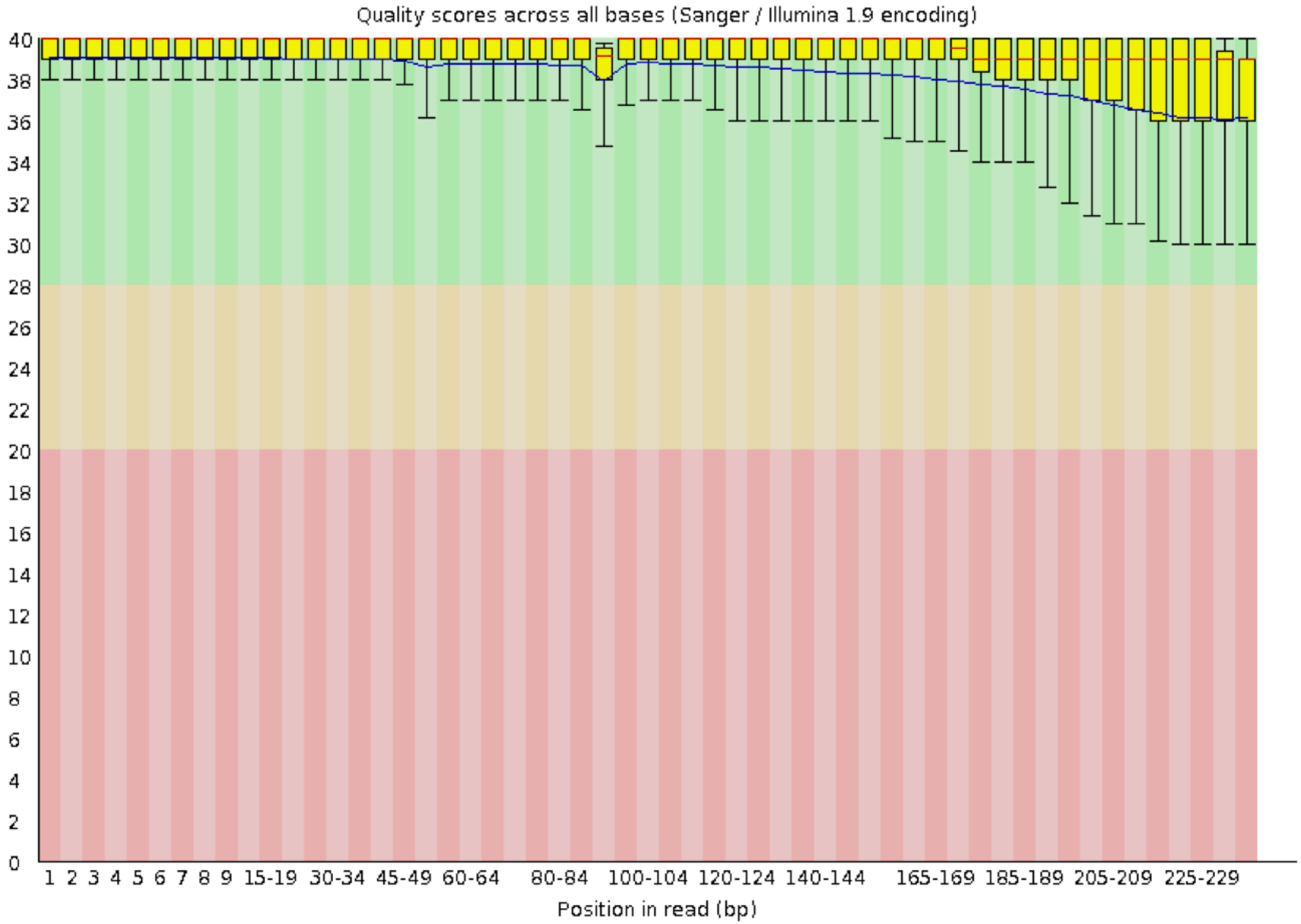
Test Data

	Total Sequences	Sequences flagged as poor quality	Sequence length	%GC
M05964	1660873	0	100-235	37
M07572	1502079	0	100-235	38
M10540	1480071	0	100-235	37
M16180	1857386	0	100-235	38

Per base sequence quality

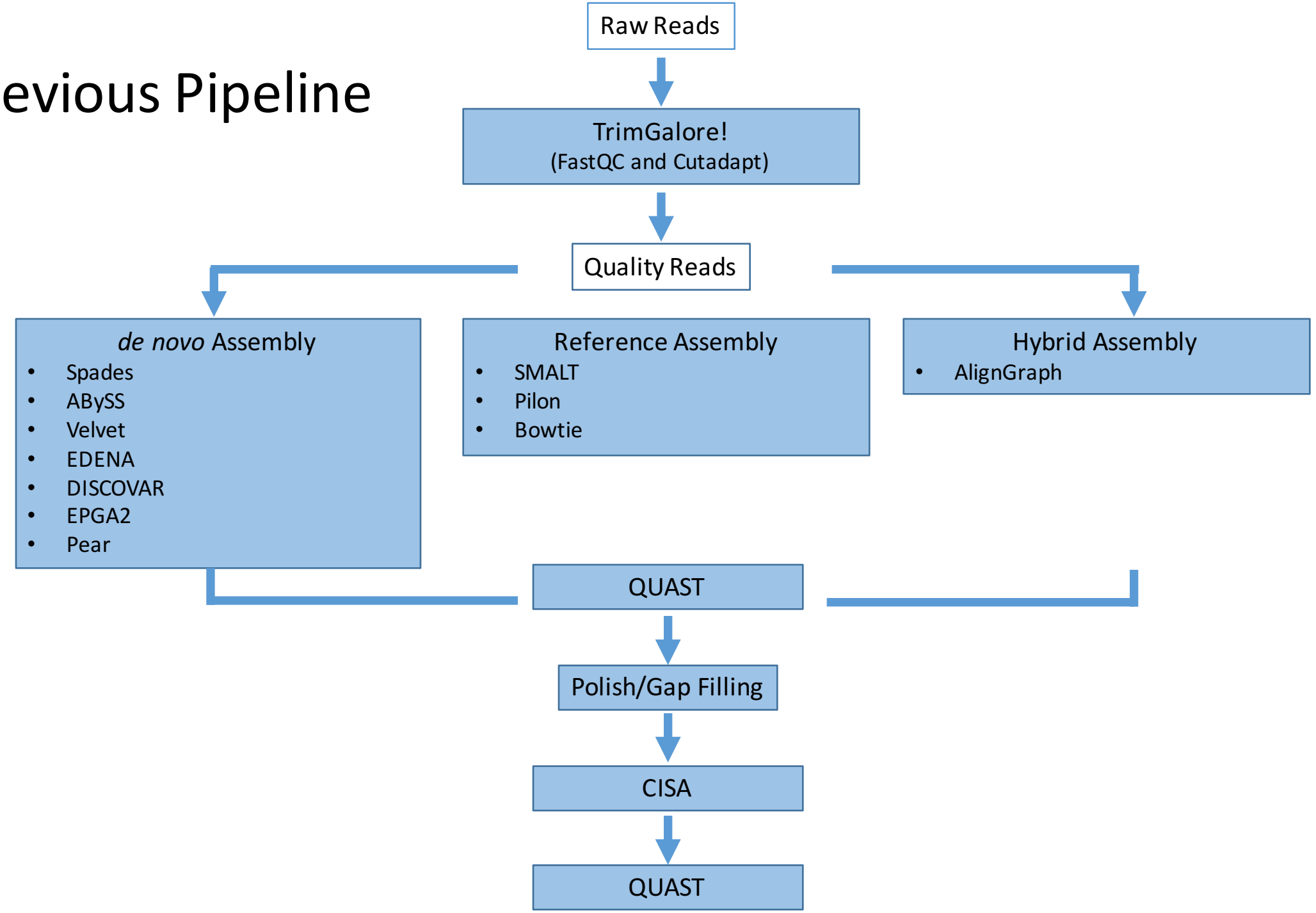


Per base sequence quality



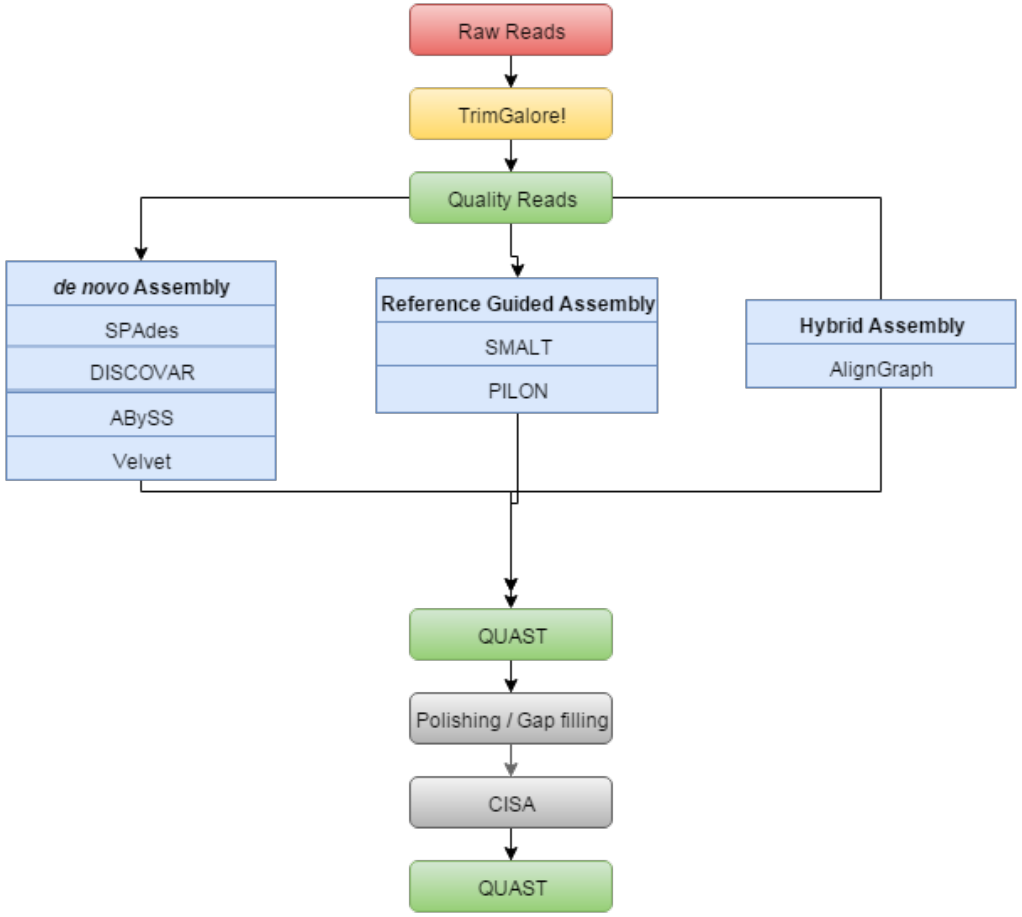
Updated Genome Pipeline

Previous Pipeline



Assemblers	Did it work?	If not, why?
SPAdes	Yes	
ABySS	Yes	
Velvet	Yes	
Edena	No	It can only handle 128 bp.
Discovar	Still running	
EPGA2	No	It can only handle 50 bp.
PEAR	No	does not assemble; only merges the paired-end reads to generate a longer working reads.
Smalt	Yes	
Pilon	Still running	
AlignGraph	Still running	

Updated Genome Pipeline



Comparison of Assemblers

de novo Assembly:	PROS	CONS
SPAdes	PDBGs, bayeshammer read correction, polishing step	Reads correction is costly, kmer size greatly influences results -- no self-determination of kmer
ABySS	Large size of data.Store data in different computers. Parallel computation of the assembly algorithm across a network of commodity computers.	K could be bigger which can increase the improve the assemblies
Velvet	Easy to install, easy to run, can take in long and short reads	May need large amount of RAM for large genomes, no self-determination of kmer or coverage but the wrapper VelvetOptimiser automatically finds the optimal parameter setting
EDENA	Easy to install, computing process is lot faster than other approximate matching based assembly, taking low memory cost	Hard limits on size of read it can handle (128 base pairs)
DISCOVAR	Probably the most sensitive available tool available.	Cant run on busted old Red Hat Linunix distro on the server
EPGA2	Simple installation, shows to be memory efficient and have improved higher coverage for contigs and scaffolds	Hard limits on size of read it can handle (50 base pairs)
PEAR	It does not require preprocessing of the raw data or specifying the fragment size. PEAR neither requires prior information on read length nor target fragment size.	Does not assemble the genome, it only merges the paired-end reads to generate a longer working reads, if there is some overlap between the two reads.

RFERENCE ASSEMBLIES	PROS	CONS
SMALT	Easy to install. The software will perform very well over a range of current sequencing platforms and for a large variety of mapping tasks including plant species. The user can adjust the trade-off between sensitivity and speed by tuning the length and spacing of the hashed words.	The mapping speed is relatively slower. Needs additional tools(like samtools) to process the mapping result.
PILON	Portable Java VM executable. Fully automated, all-in-one tool. Works with many types of sequence data, but is particularly strong when supplied with paired end data	Java is memory inefficient. Pipeline relies on other software to convert reads into sorted and mapped .bam files.

HYBRID ASSEMBLY	PROS	CONS
AlignGraph	The positional information generated from the alignments and paired-end reads allows the initial assemblies to be extended without incorrect extensions and early terminations. Performance tests show AlignGraph is able to considerably improve the contigs and scaffolds from several assemblers (Velvet, ABySS, etc.)	Quite slow to run

Assembly Improvement

Assembly Improvement –with pilon

- Diagnostics
  - Summarize alignment pileups (for every locus for every library)
    - Individual base stacks
      - Allele counts
      - Allele quality sums
    - Coverage information
      - Sequence coverage
      - Physical coverage
    - Indels
      - Insertion events
      - Deletion events
    - Pairing information
      - Valid pair coverage
      - Invalid pair coverage
      - Local implied insert size
- Using Diagnostics, classifies each base pair as
  - –correct
  - -incorrect
  - -ambiguous
  - -uncallable
- Then pilon will fix “incorrect” bases
- Pilon takes assembly and morphs it to conform to evidence in read data

Assembly Improvement –with pilon

======
M07572_contigs_velvet.fa
======
Before
Total length of sequence: 1960411 bp
Total number of sequences: 62
N25 stats: 25% of total sequence length is contained in the 2 sequences >= 277366 bp
N50 stats: 50% of total sequence length is contained in the 4 sequences >= 200322 bp
N75 stats: 75% of total sequence length is contained in the 8 sequences >= 79487 bp
Total GC count: 749441 bp
GC %: 38.23 %

after:
Total length of sequence: 1810882 bp
Total number of sequences: 19
N25 stats: 25% of total sequence length is contained in the 1 sequences >= 1150446 bp
N50 stats: 50% of total sequence length is contained in the 1 sequences >= 1150446 bp
N75 stats: 75% of total sequence length is contained in the 2 sequences >= 311437 bp
Total GC count: 686344 bp
GC %: 37.90 %

Assembly Improvement –with pilon

NODE_13_length_42754_cov_197.760910:309 NODE_13_length_42754_cov_197.760910_pilon:309 T A

NODE_38_length_117_cov_141.606842:117 NODE_38_length_117_cov_141.606842_pilon:117 A G

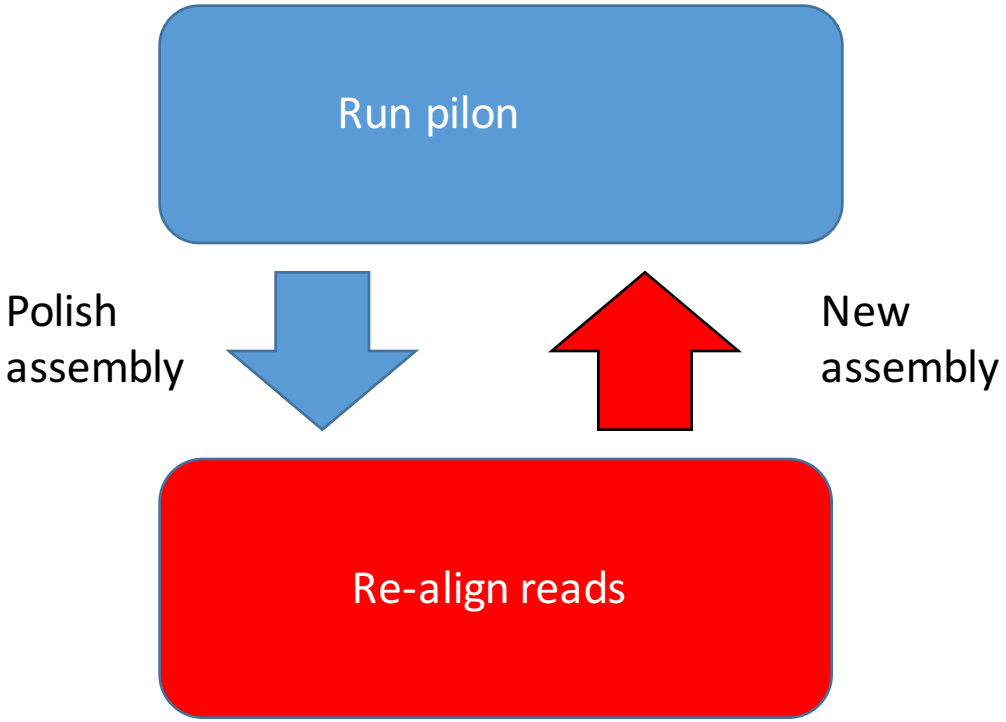
NODE_90_length_8800_cov_222.210907:282-357 NODE_90_length_8800_cov_222.210907_pilon:282-1338
NNNNNNNNNNTTGGTTGGGCTGCATAAGCAACACCTGCCGTTAGCCCAAGCATAACGGAATAGGCANNNNNNNNNN
TTTTGTATCACTATTTTCGGTAGAACCAGATACATTAATTTGCTCTAGTTGTTCAGAAAGATTACTATTTTGGTTGGTTGGTTGGTTGGTTGGTTGGTTGGTTGGTTGGTTN
NNNNNNNNN

CTCTCGTTTTTATTTTTTGTTTTGAAAAAGTAATTTTTAACGTATCCCAAAATGGCGTTTGAGAGAAATTTTCATAGCTAAAGGAAATATTACGTCTCTTTGTTTTATCATTGGTGTGTCTA
GAATCATCCTCGAGTAAATCAGGAAGAGTTTTTAGATATTTTAATGTATAGGATAAATCTTGCCCACGAGAACGATGTTCATATAAATCTGCTGCAAGGGTAAAACGATGATTTTCTGTA
GGATTGAAAGATAATTTTAATAATGTACTGTCTTGCTCAATTTTGTATGGATCTGCTTTTTCTCTTTTTTTACCTTGAATTTTATCGTTATAATTTTTATAACCATAGTTCTCAAGTTCATGCC
CATTTCTGCTTGTTGTAACCACTAGCACATCAAACTTTTTATAACGTCCAGCAAGAGTCAGCGTATTGAATGATTGATTATTTTCTGTAGCGTATCCCTTTTTGTAGCTTACATAGTAATCC
TTGTTAAGGAGATAATCTCTCGCATCTTTTGTTTTATAAATTACAGATCCACCTAAGGAACCACTACCACTTTTGATTGAATTTGCCCCTTTTGTAATATTTACTTCTTTTAAAGTTTCAATT
TCTGCACCATTACGCGTATTATTGAAGTTACCATAACCCTCAAAAAGCTCTTTAAAGCCTTGAGAAGATAATGTTTCAGCTTGACGTAATCCATCAATATTAATCGCTACTCGATTTTCATC
TACACCACGAATGGCAAAACCGCTTTGCCCAAAACGTCCAGCTTCAACAACAGTAACGCCAGTCTCGTATTTAACGATGTCTTTAATATTGTTTGCTTGTTCTCTTTCCAGCGTTTTAGC
CGTTTTTACCGTTTCAGCAATTTTTGGCGGTGTTTTTGTATCACTATTTTCGGTAGAACCAGATACATTAATTTG

NODE_117_length_2306_cov_143.446655:235 NODE_117_length_2306_cov_143.446655_pilon:235-857 .
AGCGTCACCCGAAATATCGCCTACAGCAAGAGTTTCTCCGGTTGCAGTAAGTGTTACAGAGCCGGAAGTGGATTCAACTTTACCCTTAATATCACCTGTTTTAGTGTTTACTTTCACGT
TGCCGGTGGTTGCATTAATGGTTGTGCCCTCTTTGGTGGCAACTTTTTCTGATGCGGTGATATTCACTGCCTTGTGAGAAGTAATATTGTTGTTTACTGTTACATTTTTTGCATTAATAGC
TAAGCCGGTATTATTGTCACCATTATTTTCAGTACTGCTACCGCCATTAGATATTTCCACTTTGCTATTTAGTGTCACATTATGGTCGCCAGCTGAGATTTTTGAATCTTTAACATTGTTAAA
AGTTACTGCTTTGGCTTCGGCACCGCTGTTACCGTCATTACTGTTGCCAATAGTTAAATCACTACCATTTTTGGCTGTAATCTCTGCTTTATCAAAACCTGAAATGTTTAGGTCATTTACC
AGTTTCAATTCTTTGGTTTTAATGGTTAGATTGGCTTGACTTTCCGTACTTGAATCAGAGCTCTCTCCGTTAACACCCTTCTTGATTGTTATCTGTTTGGTAATATTGATTTTATCGGAAGA
AATCGTGAGATTGCCTT

NODE_117_length_2306_cov_143.446655:1809 NODE_117_length_2306_cov_143.446655_pilon:2432 T .

Assembly Improvement –iterative improvement



- Some issues might only be revealed by other fixes
- Data might not have the “reach” to identify or fully fix things in one pass
- Pilon is not a panacea for fundamental flaws in assemblies or data generation.

Assembly Quality Assessment

Assembly Quality Assessment

- Run everything through QUAST
- Objectives
  - Find the best reference to use
  - Rank our *de novo* assemblies and their assemblers
- Metrics of interest
  - N50 – The size of the contig (L50) that makes up at least 50% of the assembly
    L50 – Number of contigs to cover at least 50% of the assembly
  - Total length (Contigs >1000bp) – Length of assembly, useful sanity check
  - GC% – G+C percentage, useful sanity check
  - # of Contigs >1000bp

QUAST – L50/N50

	ABySS-pdbg-99	ABySS-pe-99	ABySS-pdbg-115	ABySS-pe-115	DISCOVAR	Velvet	SPAdes
L50	4.75	4	4.5	3.5	19.75	147.5	4
N50	140928	190232	150455	184770	86846	370281	390527
Total	1924970	1909846	1915018	1907896	3762037	1560932	1931966

QUAST – Choosing a reference

	ABySS-pdbg-99	ABySS-pe-99	ABySS-pdbg-115	ABySS-pe-115	DISCOVAR	Velvet	SPAdes
Mean % mapped	78.13	78.35	78.2	78.04	77.59	69.46	78.43
Min	57.45	57.47	57.49	57.44	41.34	52.08	57.71
Max	86.71	86.64	86.58	86.83	97.32	85.85	86.07
Mean (No M10540)	85.03	85.3	85.1	84.91	84.51	75.26	85.34

M10540 maps poorly to 16/17 NCBI *Hi* reference genomes. Maps very well with one reference. Suggestive that perhaps M10540 is an *Hhaem* as is the NCBI reference, NC_022356

QUAST – One assembler to rule them all?

Contigs >1000bp

	ABySS-pdbg-99	ABySS-pe-99	ABySS-pdbg-115	ABySS-pe-115	DISCOVAR	Velvet	SPAdes	avg by sample
M10540	16	17	24	22	64	11	17	**24.43**
M05964	37	36	31	29	102	22	144	**57.29**
M16180	38	28	25	27	64	457	65	**100.57**
M07572	43	42	46	46	346	37	59	**88.43**
avg by tool	**33.5**	**30.75**	**31.5**	**31**	**144**	**131.75**	**71.25**	

References

- ABySS: http://www.bcgsc.ca/platform/bioinfo/software/abyss

- AlignGraph:

Bao E, et al. (2014) AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references, Bioinformatics, 30, i319-i328. (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058956/pdf/btu291.pdf)

https://github.com/baoe/AlignGraph

- DISCOVAR:

Weisenfeld NI et al. Comprehensive variation discovery in single human genomes. Nat Genet. 2014 Dec; 46(12):1350-5.

- EDENA: http://www.genomic.ch/edena.php

- EPGA2: https://github.com/bioinfomaticsCSU/EPGA2

- PEAR: https://github.com/xflouris/PEAR

References

- Pilon:

Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014 Nov 19; 9(11):e112963.

- SPAdes: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342519/

- SMALT: http://wiki.hpc.ufl.edu/doc/SMALT

- Trim Galore!: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

- Velvet:

1. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.*Genome Research*. 2008;18(5):821-829. doi:10.1101/gr.074492.107.

2. https://en.m.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/De_novo_assembly

- Images:

http://www.denniskunkel.com/DK/Bacteria/97500E.html

Thanks! Questions?