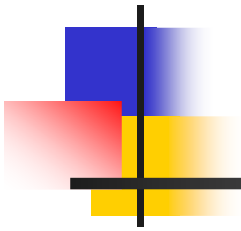
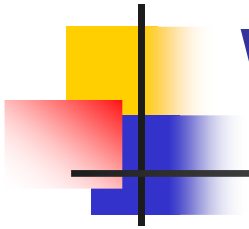


Improving Genome Assemblies without Sequencing

Michael Schatz

August 13, 2006
University of Hawaii

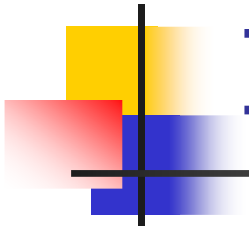


Why draft genomes

1. Size and Cost:
   1x coverage of dog genome ~ $6x10^6$ reads ($ 6 mil)
   8x coverage of 5 Mbp bacterium ~ 60,000 reads ($ 60 k)

2. Reference genome completed:
   multiple strains of an organism
   outbreak strain

3. Speed:
   sequencing 5 Mbp bacterium < 1 month ($ 60 k)
   finishing 5 Mbp bacterium ~ 2 years ($500 k)



Improving the Draft

Key Ideas:

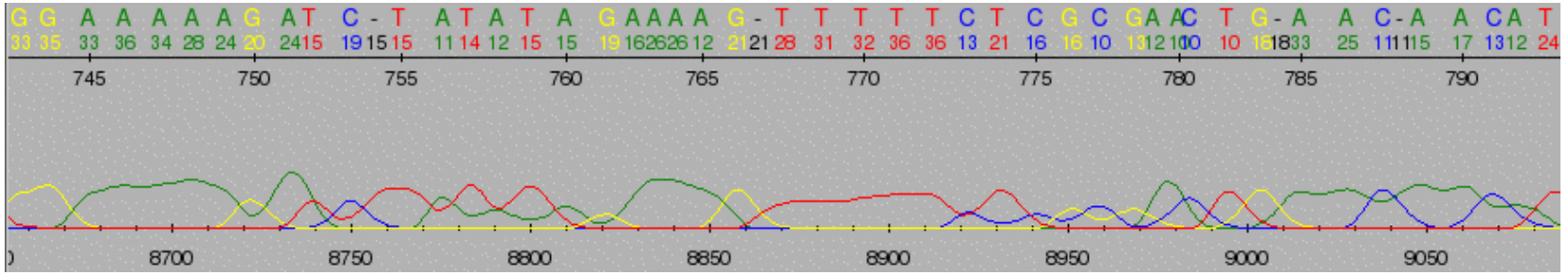
- Have to be relatively conservative at first
- But, there is a lot of additional contextual information available after the initial assembly.

Use this contextual information to revise original

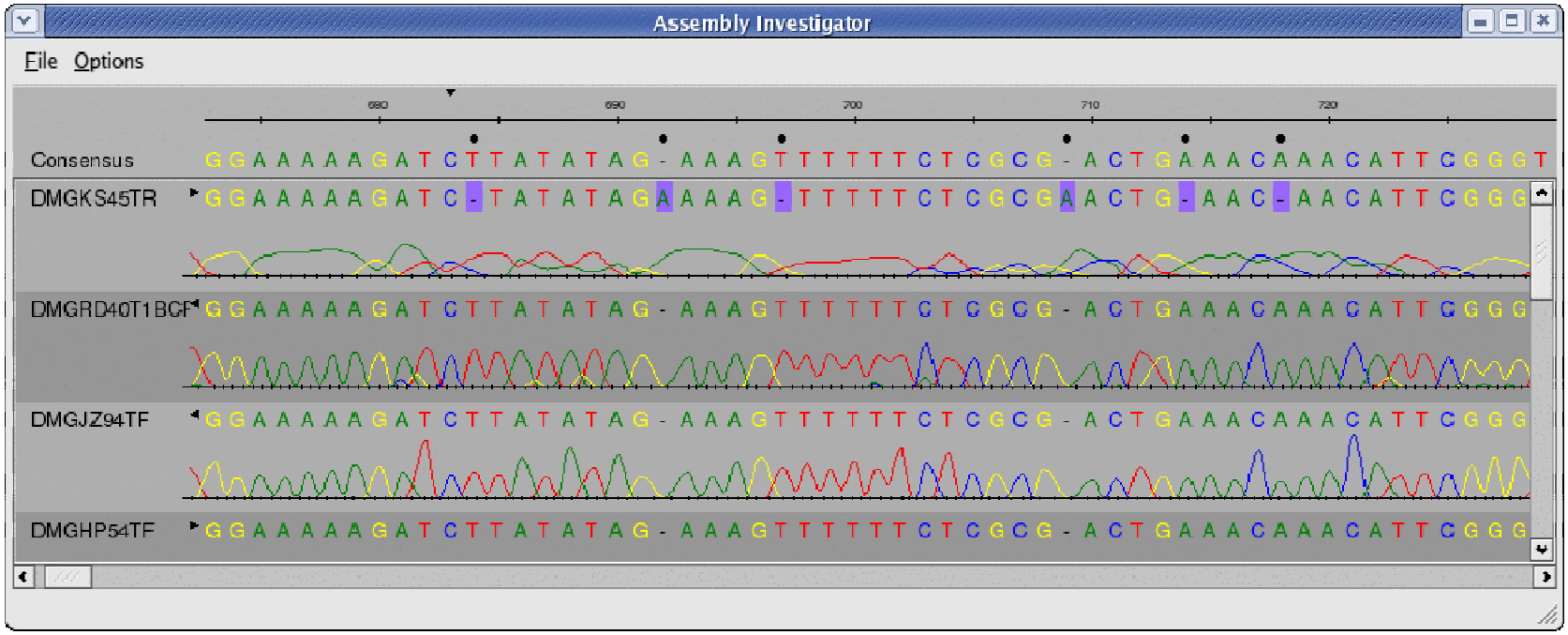
- Base-calling: AutoEditor
- Clear ranges: AutoJoiner

AutoEditor

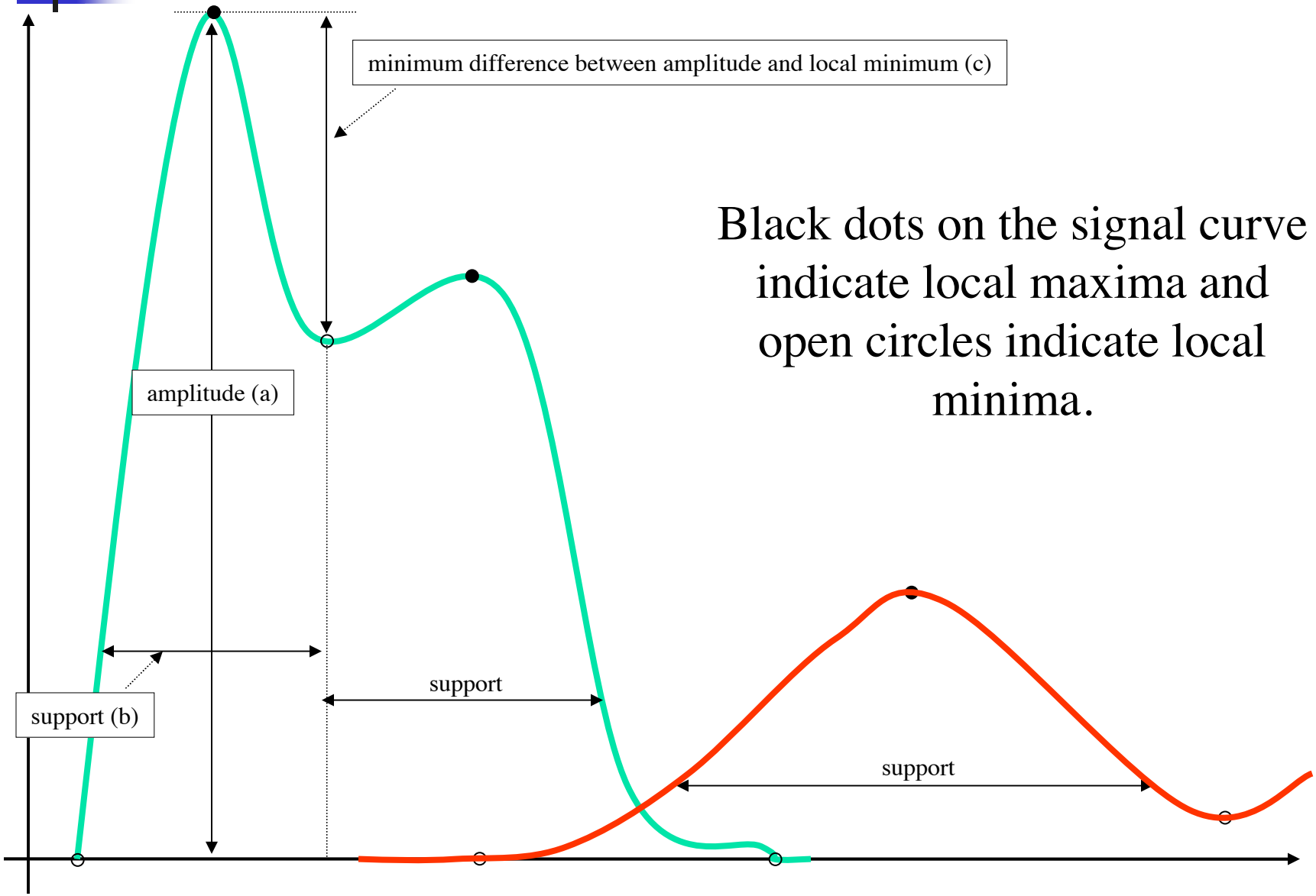
Base-calling in the context of single chromatogram is hard…



but finding base-calling “mistakes” in a multiple alignment is easy.



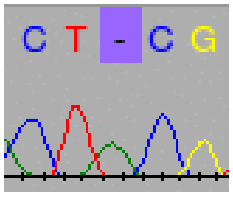
Signal Parameters



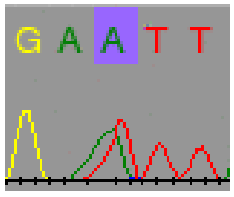
Black dots on the signal curve indicate local maxima and open circles indicate local minima.

AutoEditor Algorithm

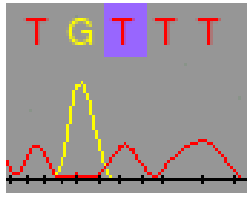
1. Scan Multiple Alignment for Discrepancies
2. For each discrepancy:
    1. Reanalyze Chromatogram signal of discrepant base near discrepancy
        1. If base-calling error:
            1. Edit Read to match consensus
        2. If chromatogram supports discrepancy:
            - Leave Read unchanged



A -> -
Deletion in Read



T -> A
Substitution in Read



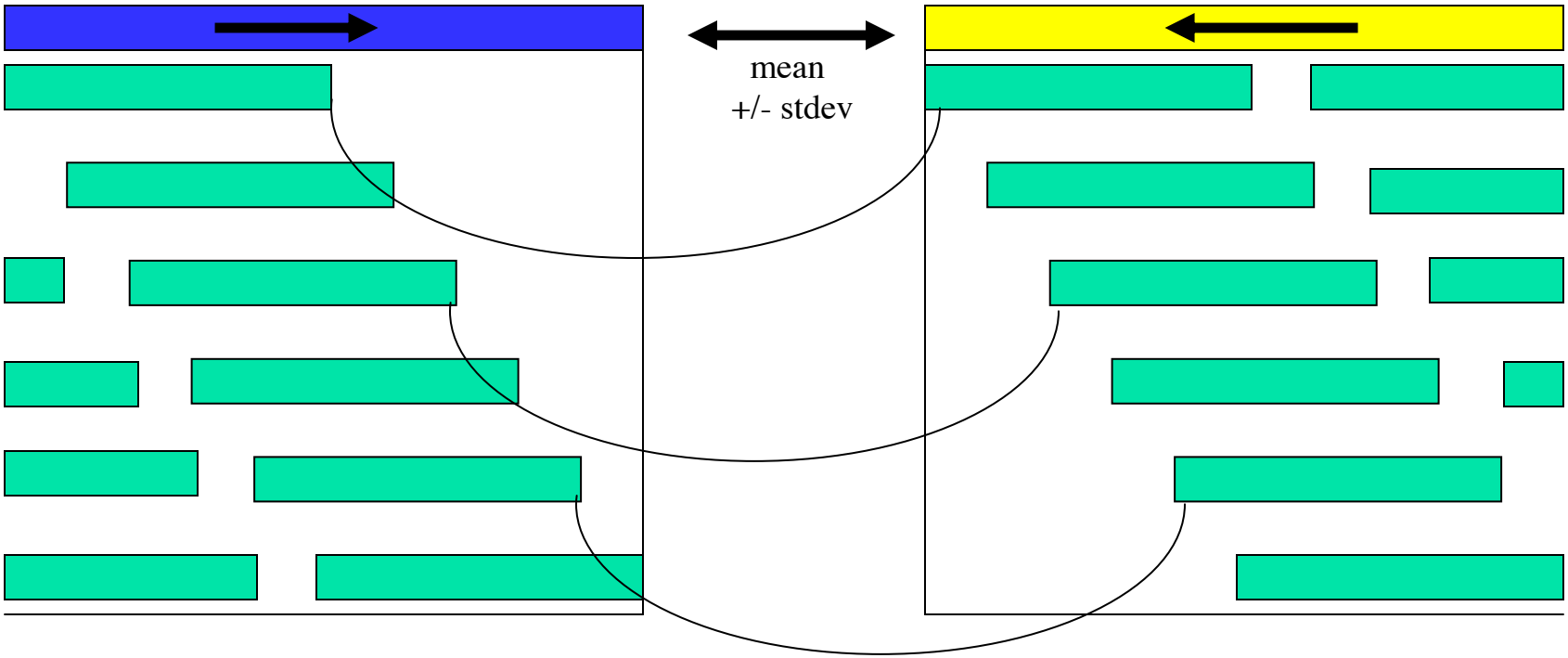
- -> T
Insertion in Read

AutoEditor Results

- Corrects 80% of all discrepant base-calls with an error rate better than 1/8800.
- Increase consensus quality, decrease finishing costs
- Remaining discrepancies highlight assembly problem regions or interesting biological events.

Organism	Read length	Corrections	AE errors
Listeria monocytogenes	37 420 828	145 274	4
Wolbachia sp.	11 446 011	51 163	0
Burkholderia mallei	47 407 080	99 711	28
Brucella suis	26 629 877	112 359	2
Streptococcus agalactiae	23 485 615	105 878	3
Coxiella burnetii	29 135 115	117 232	30
Campylobacter jejuni	15 013 845	792 37	11
Chlamydophila caviae	10 286 694	36 972	6
Dehalococcoides ethenogenes	10 724 521	46 416	12
Neorickettsia sennetsu Miyayama	8 805 232	37 425	0
Fibrobacter succinogenes	46 463 268	196 150	4
Mycoplasma capricolum	9 353 819	15 444	0
Prevotella intermedia	20 084 365	94 162	3
Pseudomonas syringae	50 369 232	177 897	46
Total	346 625 502	1 315 320	149

Quick Assembly Review

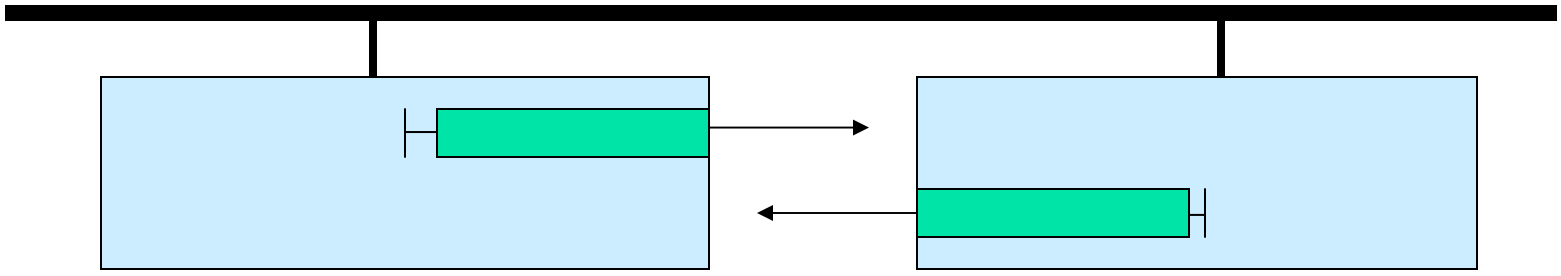


The individual reads (green) have been assembled into 2 contigs (blue & yellow). The mate relationship between the reads allows for the contigs to be oriented and the gap size to be estimated.

AutoJoiner Architecture

Automatic Gap Closure

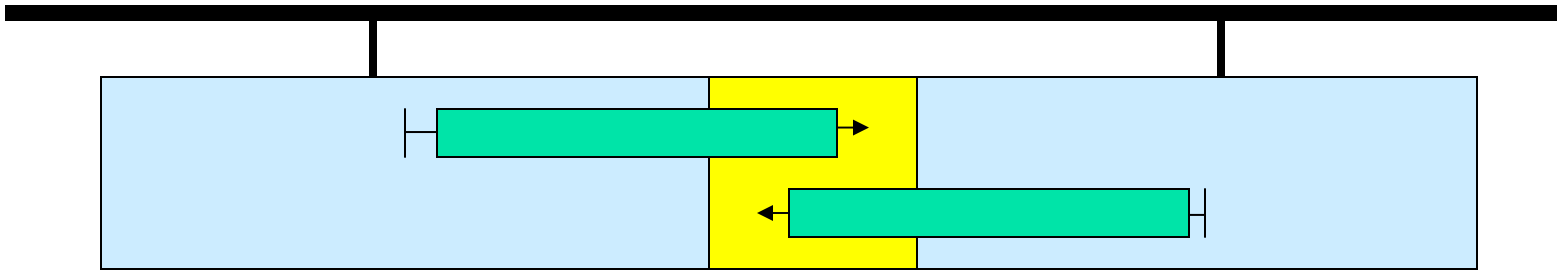
- All-vs-All Alignment
- Analyze Alignments
- Extend Contigs
- Join Contigs
- Contig Fattening



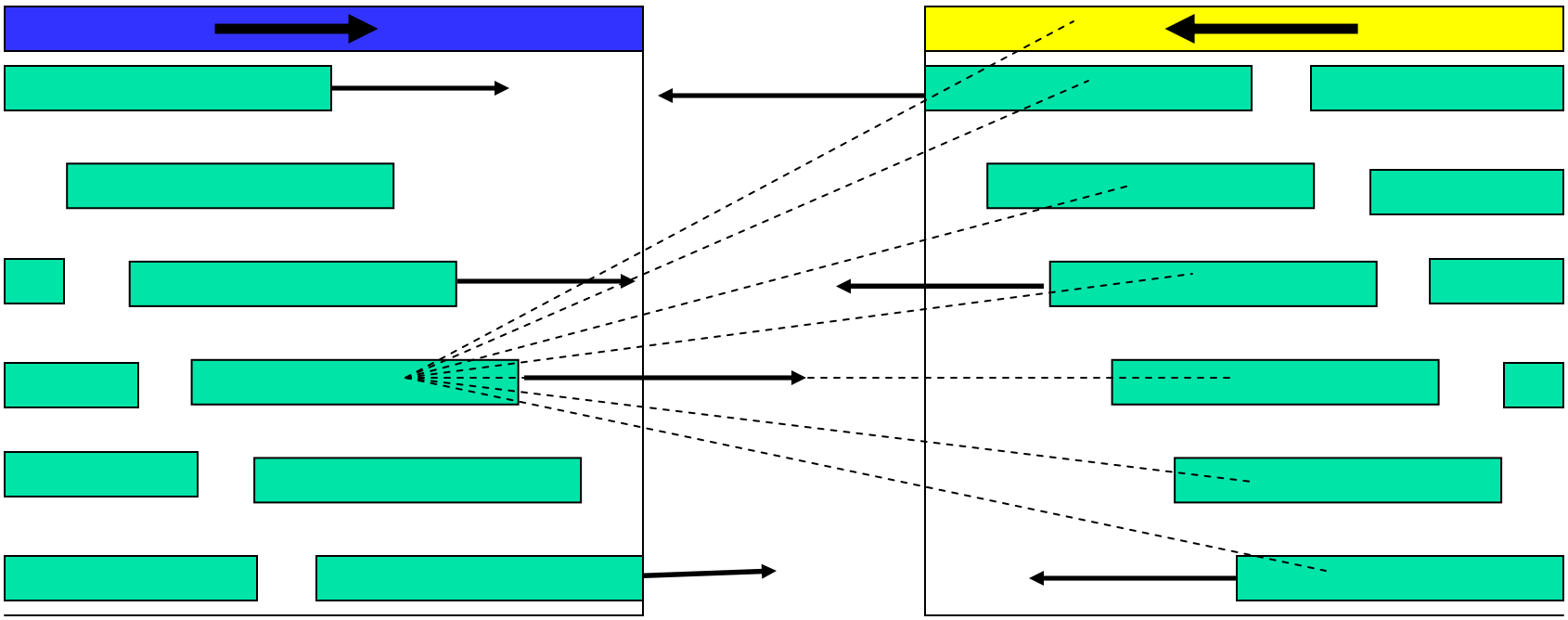
AutoJoiner Architecture

Automatic Gap Closure

- All-vs-All Alignment
- Analyze Alignments
- Extend Contigs
- Join Contigs
- Contig Fattening

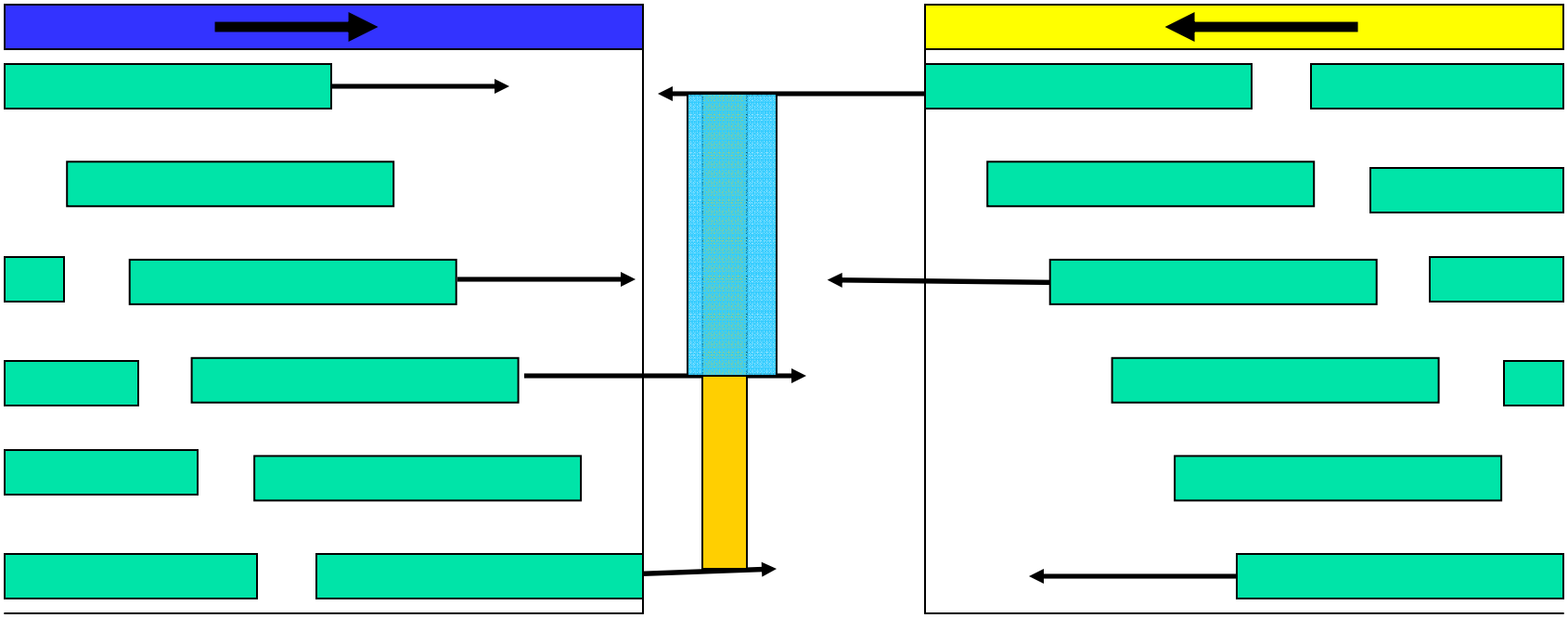


All-vs-all Alignment



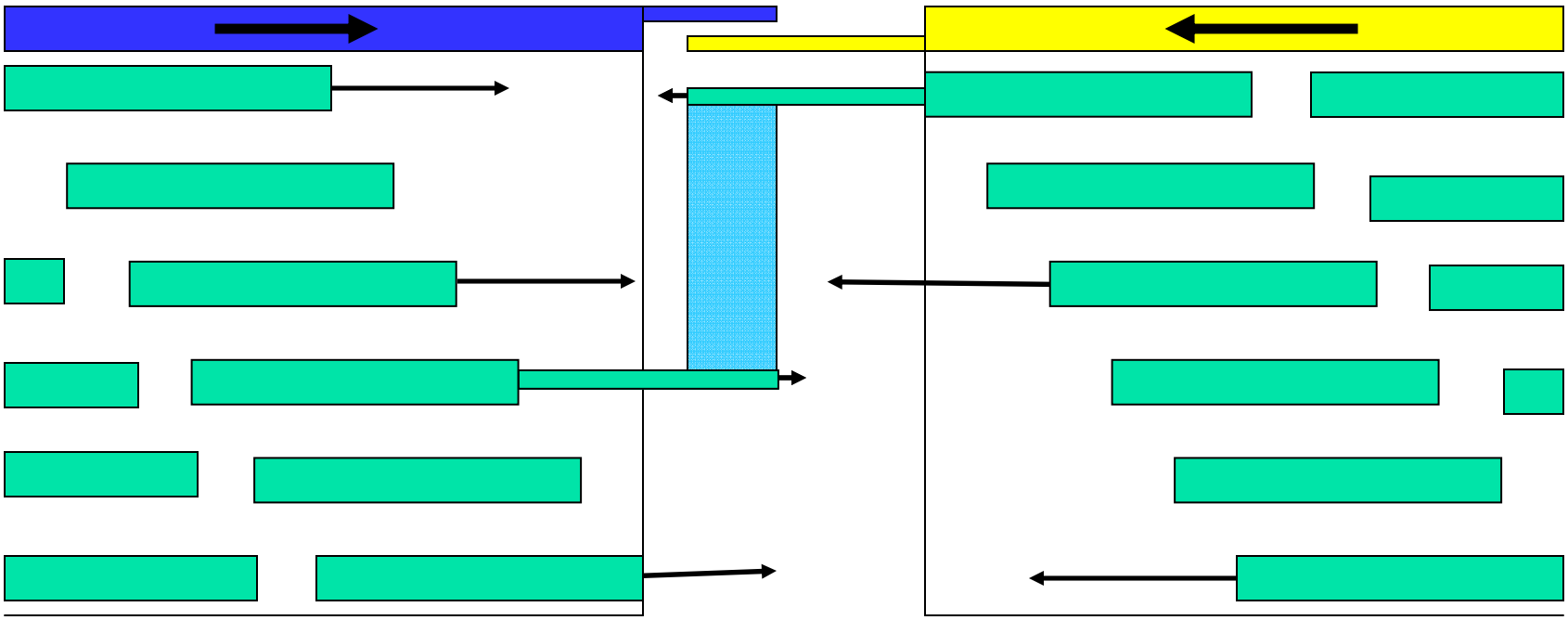
1. An all-vs-all pairwise alignment between the full range sequences from the flanking contigs is computed.

Alignment Analysis



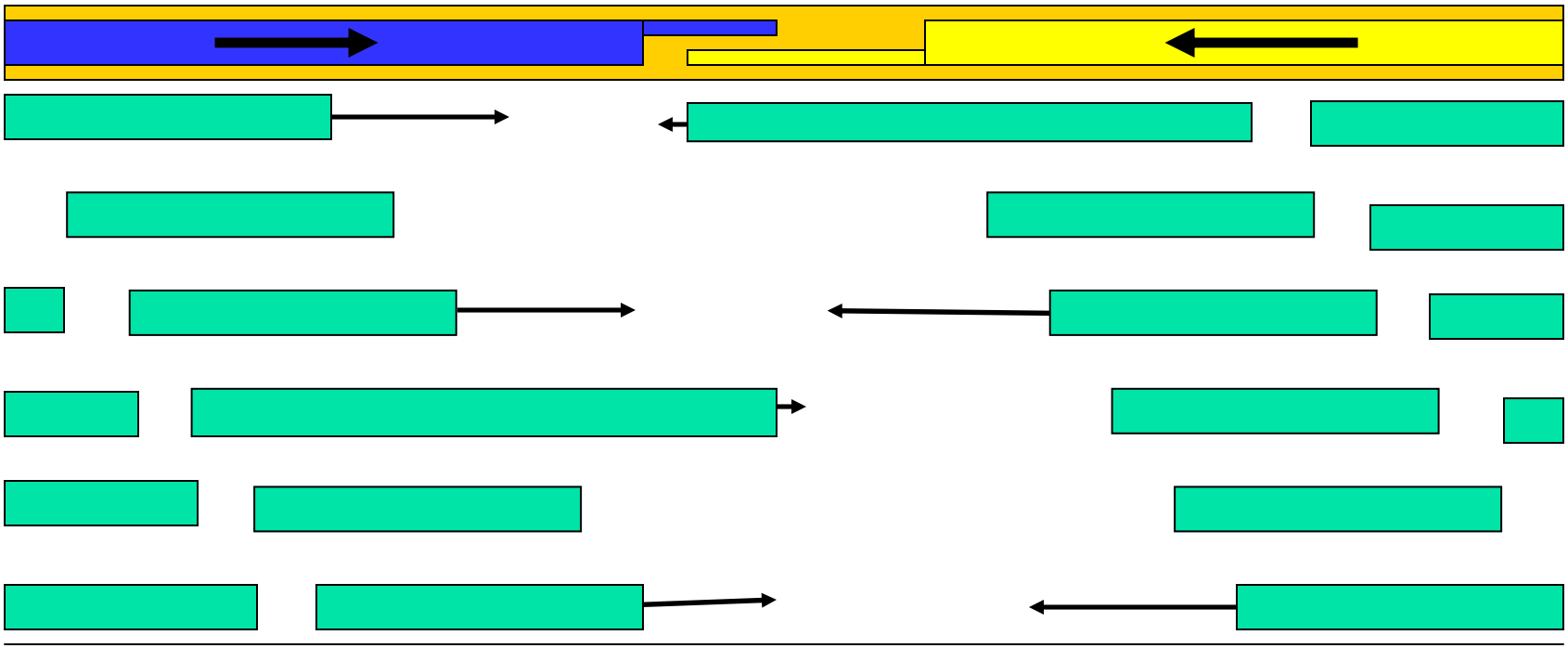
2. The alignments are tested for consistency with the scaffold and for being of sufficient quality. If any alignments satisfy the requirements, the best alignment (blue) is selected for joining the contigs.

Contig Extension



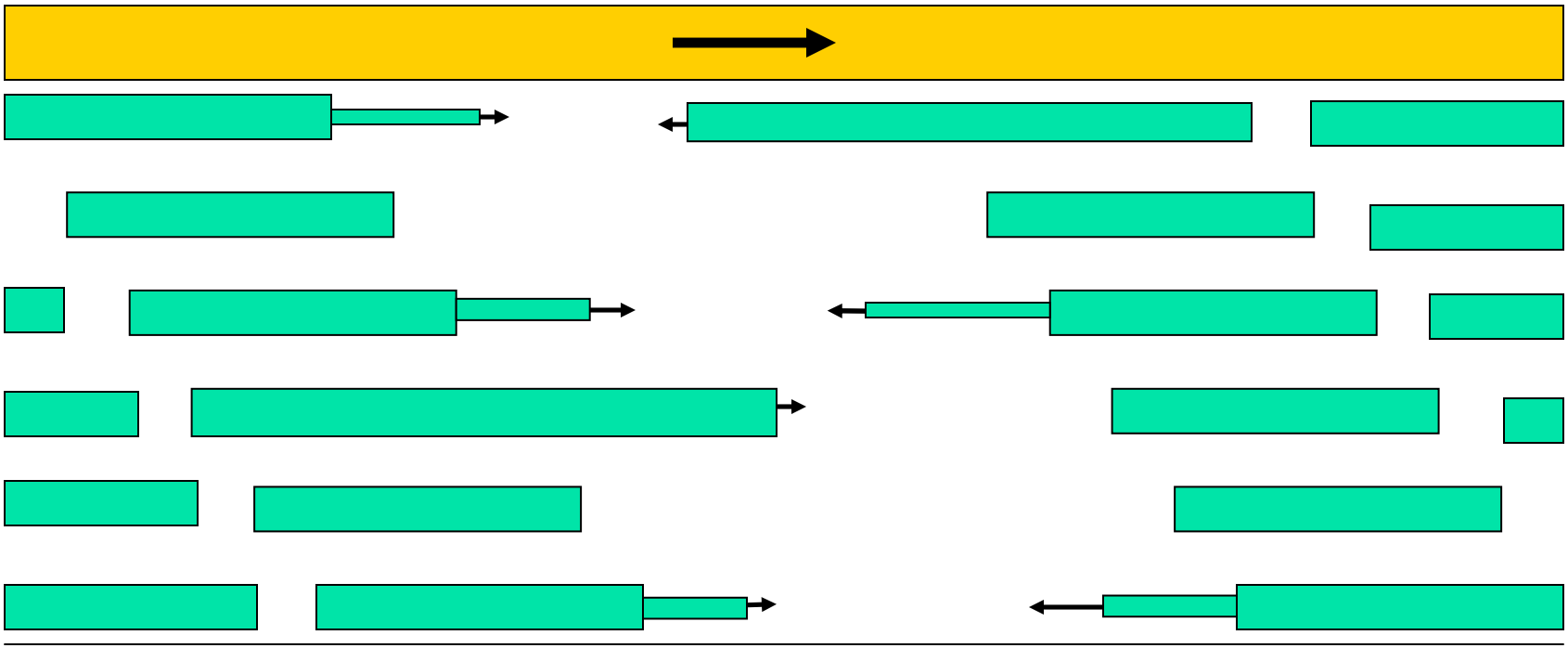
3. The contigs are extended by extending the selected reads beyond their original clear range to the desired position. If necessary, the reads are first aligned to the existing consensus.

Contig Joining



4. The contigs are joined by aligning the newly extended consensus sequences. The joined contig (orange) replaces the original two in the scaffold.

Contig Fattening



5. The join region is fattened to increase the depth of coverage and enhance the consensus quality.

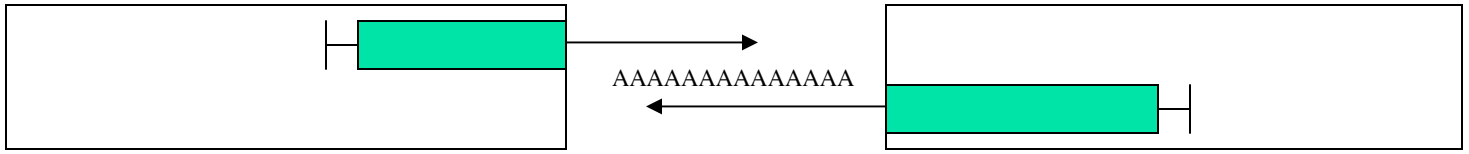
AutoJoiner Validation

- Tested against assemblies of 30 finished genomes and chromosomes.
- Over 25% of gaps closed
- Only 3 invalid joins.

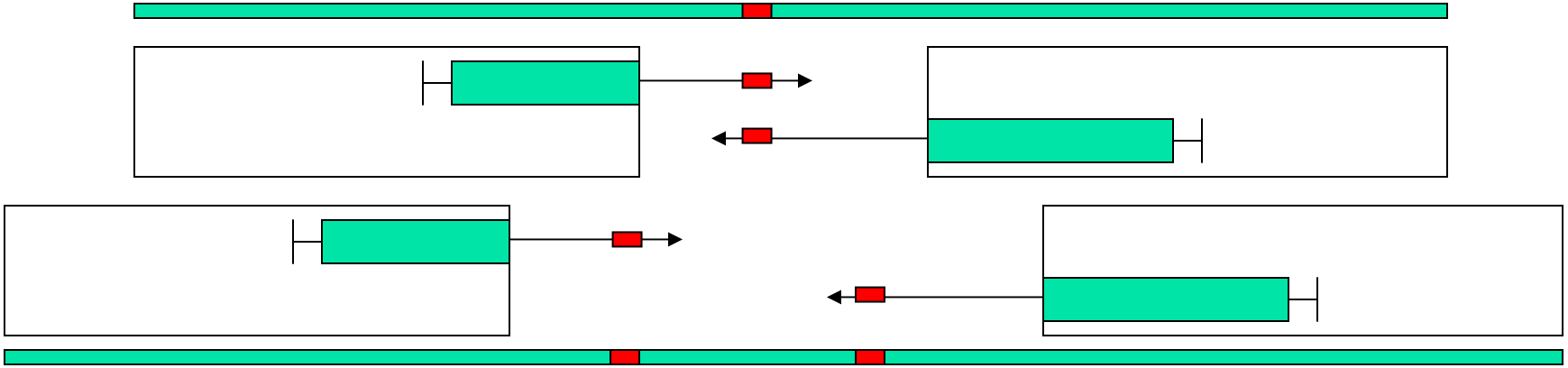
Organism	Genome Size (Mbp)	Gaps	Joined	%	False Joins	Gap Size	Mean
Bacillus anthracis Ames	5.22	110	38	34.5%	0	-452:229	29.18
Bacillus anthracis Ames Ancestor	5.22	32	10	31.2%	0	-13:146	42.3
Brucella suis	3.31	32	11	34.4%	0	-62.5:103.5	15.36
Burkholderia mallei	5.83	43	6	14.0%	0	-17:22	-4.67
Campylobacter jejuni	1.78	22	11	50.0%	0	-53:139	27.45
Chlamydophila caviae	1.17	25	8	32.0%	0	-555:184.5	-75.31
Coxiella burnetii	1.99	10	2	20.0%	0	-37.5:-4.5	-21
Cryptococcus neoformans 1	2.3	20	8	40.0%	0	-36:186.5	63.62
*Cryptococcus neoformans 2**	1.63	7	5	71.4%	1	-39:148	27.8
Cryptococcus neoformans 3	2.11	21	8	38.1%	0	-93:67.5	-3.06
Cryptococcus neoformans 4	2.04	25	7	28.0%	0	-90:159	45.21
Cryptococcus neoformans 5	1.78	23	8	34.8%	0	-111.5:249	35.12
Cryptococcus neoformans 6	1.51	14	7	50.0%	0	-14:192	37.21
Cryptococcus neoformans 7	1.44	17	6	35.3%	0	-3.5:230.5	66.67
Cryptococcus neoformans 8	1.35	15	6	40.0%	0	-19:57.5	15
Cryptococcus neoformans 9	1.18	12	6	50.0%	0	-423:34	-120.5
Cryptococcus neoformans 10	1.09	14	7	50.0%	1	-777:124	-91.21
Cryptococcus neoformans 11	1.02	12	2	16.7%	0	-6:69.5	31.75
Cryptococcus neoformans 12	0.79	10	4	40.0%	0	-340:77.5	-69.88
Cryptococcus neoformans 13	0.76	13	7	53.8%	1	-213.5:144	19.07
Dehalococcoides ethenogenes	1.47	82	17	20.7%	0	-113:203.5	22.29
Fibrobacter succinogenes	3.84	131	33	25.2%	0	-182.5:212	21.79
Listeria monocytogenes	2.9	106	14	13.2%	0	-200.5:156	21.18
Mycoplasma capricolum	1.15	10	2	20.0%	0	-11.5:171	79.75
Neorickettsia sennetsu Miyayama	0.86	27	17	63.0%	0	-779:-302	-586.71
Prevotella intermedia	2.68	150	52	34.7%	0	-231.5:181	20.3
Pseudomonas syringae	6.53	162	43	26.5%	0	-1069.5:213.5	-36.13
Staphylococcus aureus	2.8	262	32	12.2%	0	-618:136	-43.44
Streptococcus agalactiae	2.16	31	5	16.1%	0	-5:32.5	10.9
Wolbachia sp.	1.27	52	13	25.0%	0	-666.5:36.5	-140.73
Composite	**69.18**	**1490**	**395**	**26.5%**	**3**	**-1069.5:249**	**-25.89**

Complicating Issues

- Poly-monomer tails
  - Use dust to filter low complexity sequence



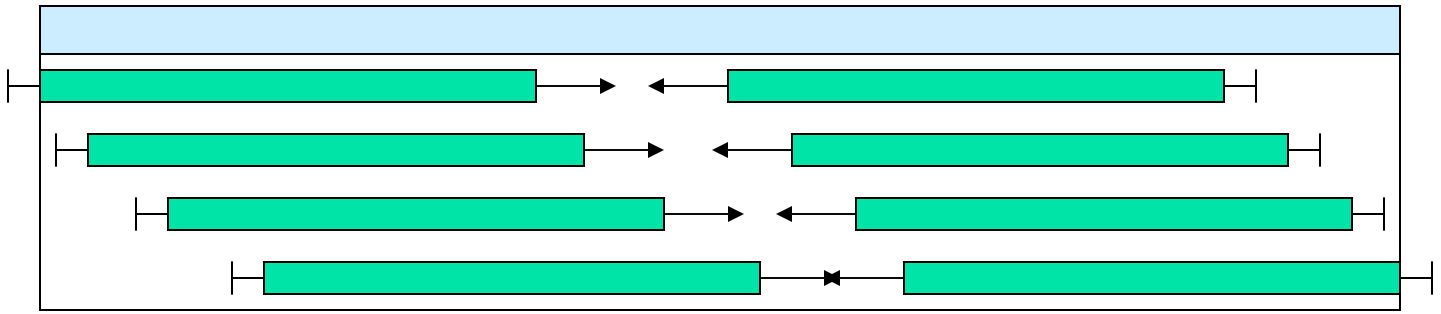
- Undetected repeats
  - Require strict agreement with scaffold



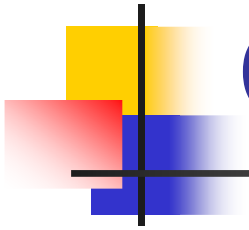
- Chimeric reads / Hard Stops
  - Good: Require high alignment similarity.
  - Better: Recognize hard stops by coverage gradients, other clues.
  - Best: Recognize unreliable sequence at chromatogram level.

Pre-Production Techniques

- Contig Fattening
  - TVG coverage increased from 5.83X to 6.10X (mean extension: 80.5bp)

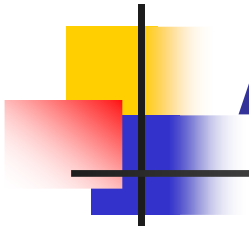


- Contig Growing
  - Extended 6144 edges in TVG (mean extension: 59.0bp)



Conclusions

- Assembly is complicated by genome structure, repeat characteristics, data quality, data management- one size does not fit all.
- Overriding strategy: Start conservatively, and iteratively build as more information becomes available.
- Be aware of potential size/quality tradeoffs, though.
- State-of-the-art assembly is still a craft- lots of room for innovation and better algorithms.



Acknowledgements

CBCB

- Steven Salzberg
- Art Delcher
- Adam Phillippy
- Mihai Pop
- Dan Sommer

TIGR

- Pawel Gajer
- Martin Shumway
- Jason Miller