



# Improving Genome Assemblies without Sequencing

---

Michael Schatz

April 25, 2005

TIGR Bioinformatics Seminar



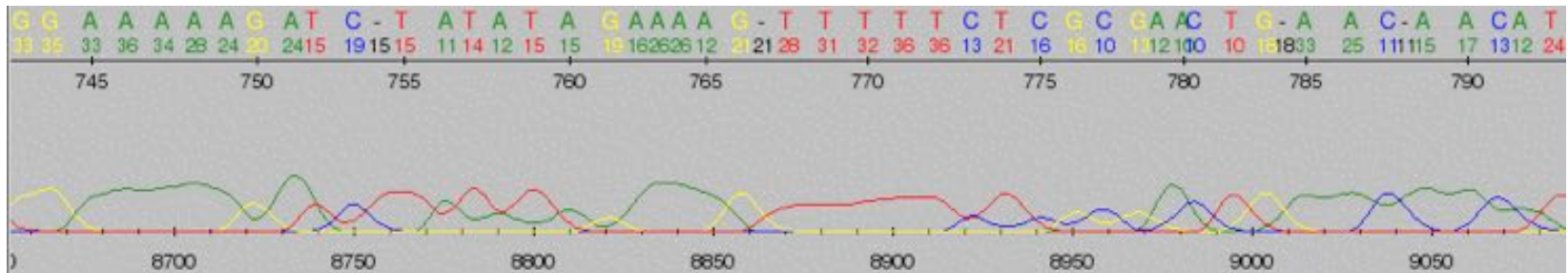
# Assembly Pipeline Overview

---

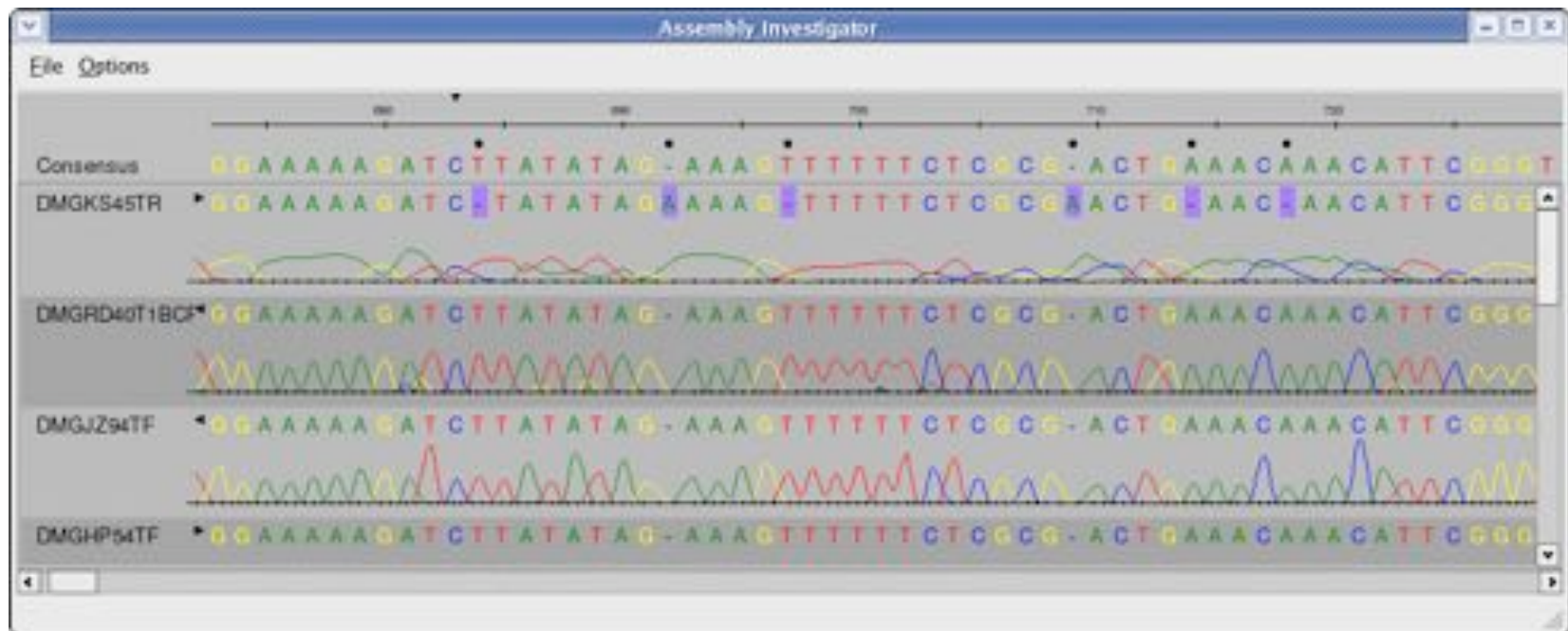
1. Sequence shotgun reads
2. Call Bases *phred/TraceTuner/KB Base Caller*
3. Trim Reads *lucy*
4. Assemble *CA/TA/Arachne*
5. Electronic Finishing
  - Second generation base-caller *AutoEditor*
  - Automatic Gap Closure *AutoJoiner*
  - Research Techniques

# AutoEditor

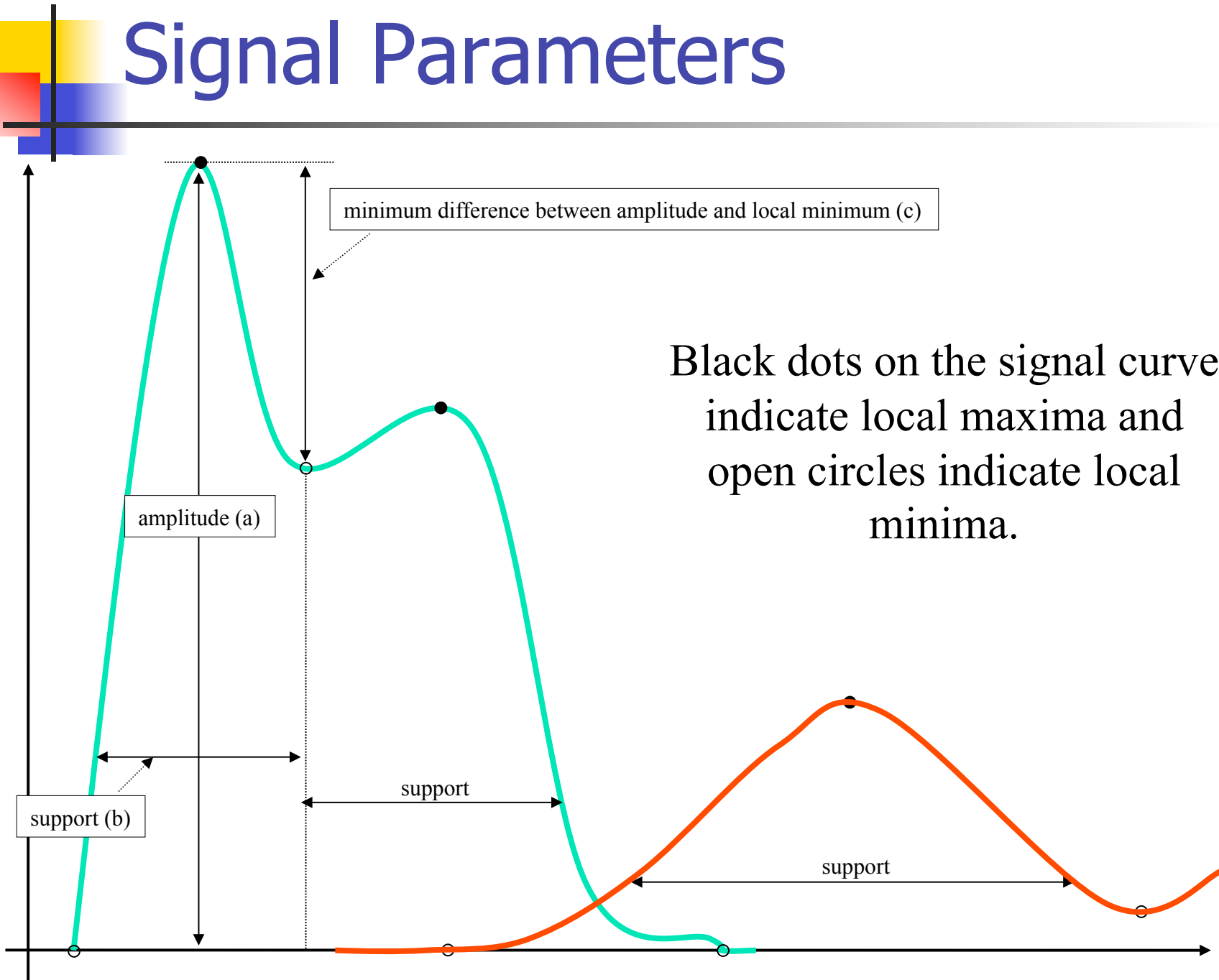
Base-calling in the context of single chromatogram is hard...



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



# Signal Parameters





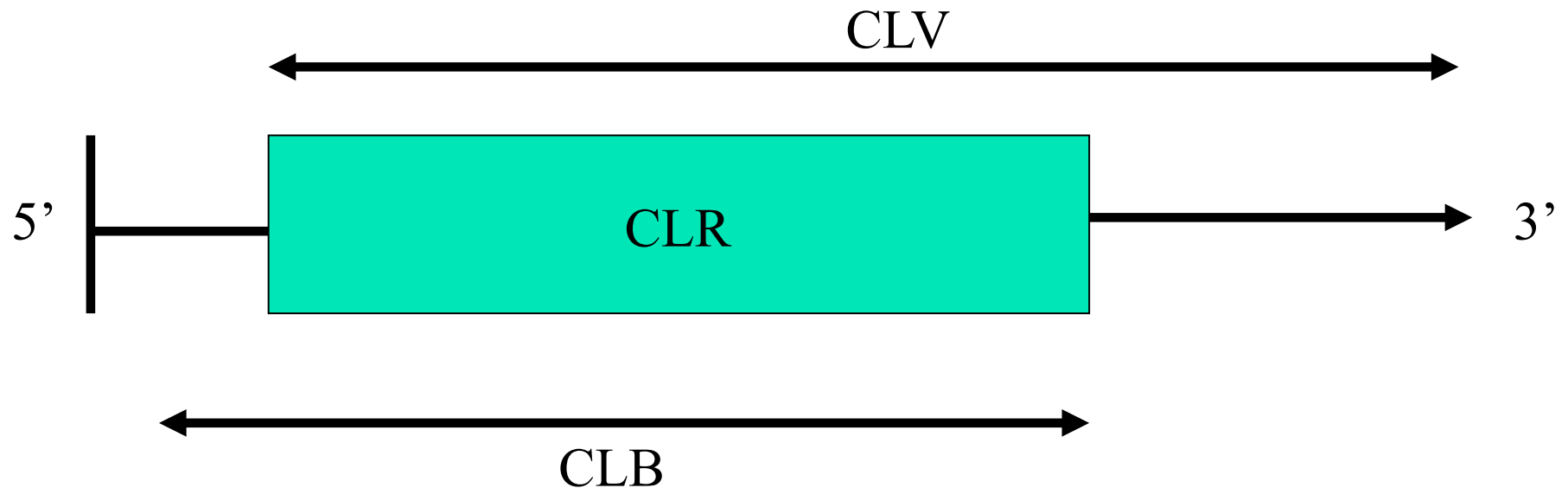
# AutoEditor Results

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

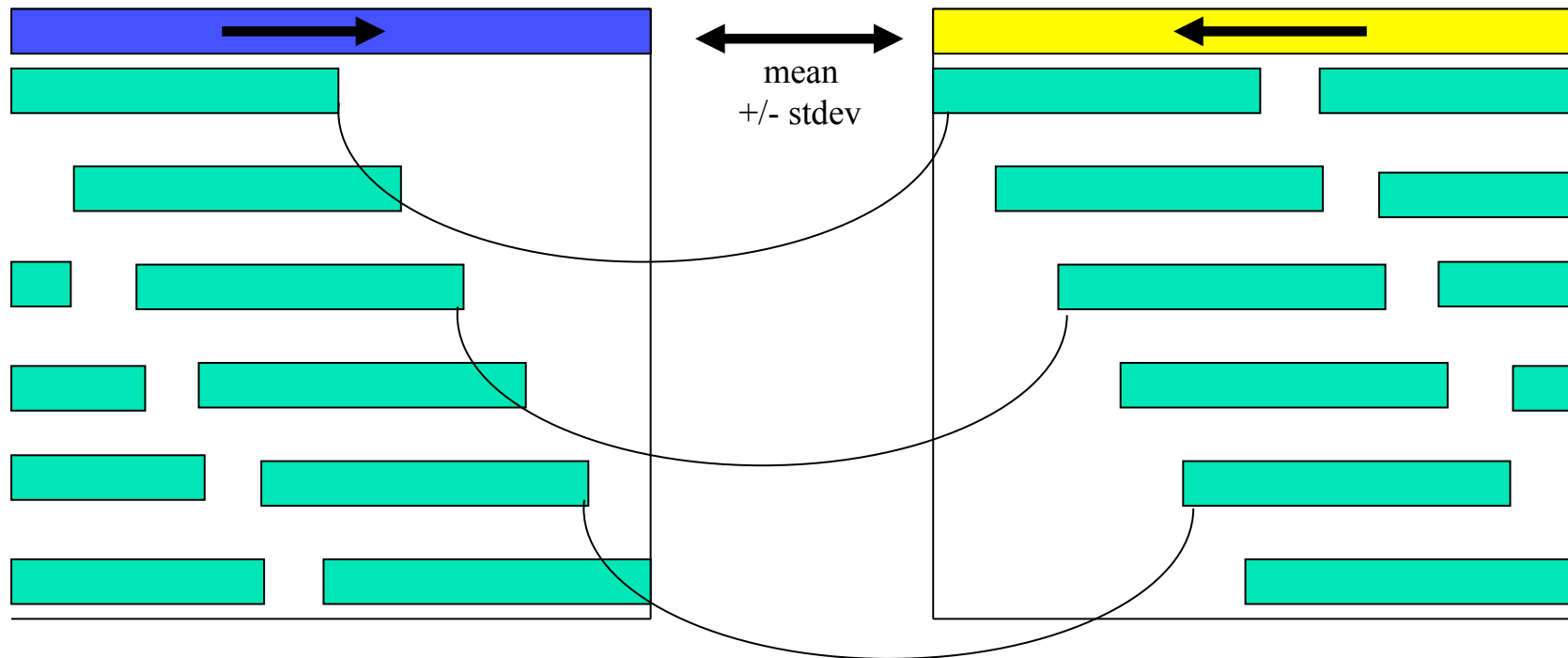
Ask Pawel for more information!

# Quick Trimming Review



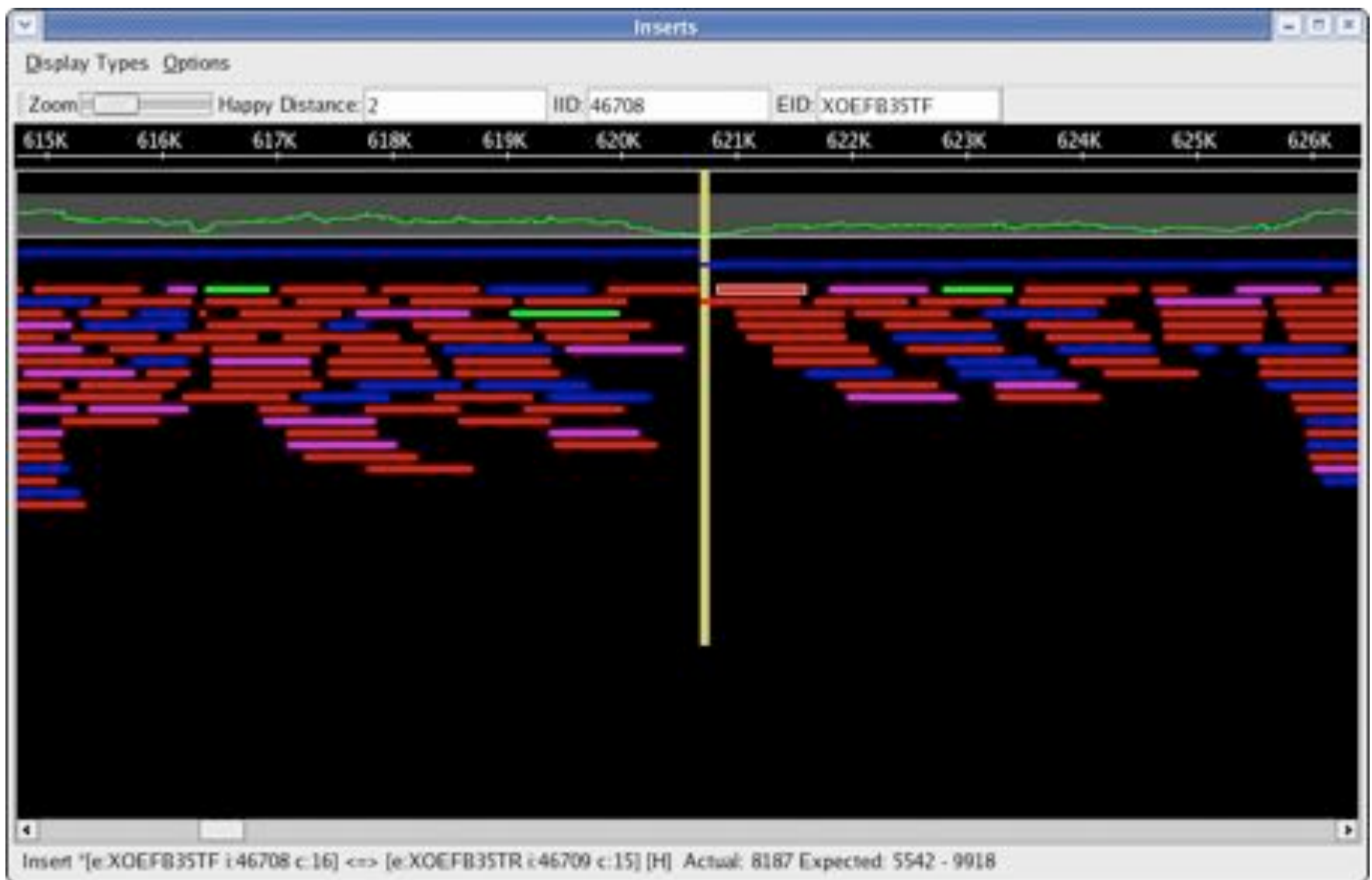
Trimming identifies the regions of good quality for the assembler to use (CLR), as the intersection of the region free of vector (CLV) and the region free of bad quality (CLB).

# Quick Assembly Review



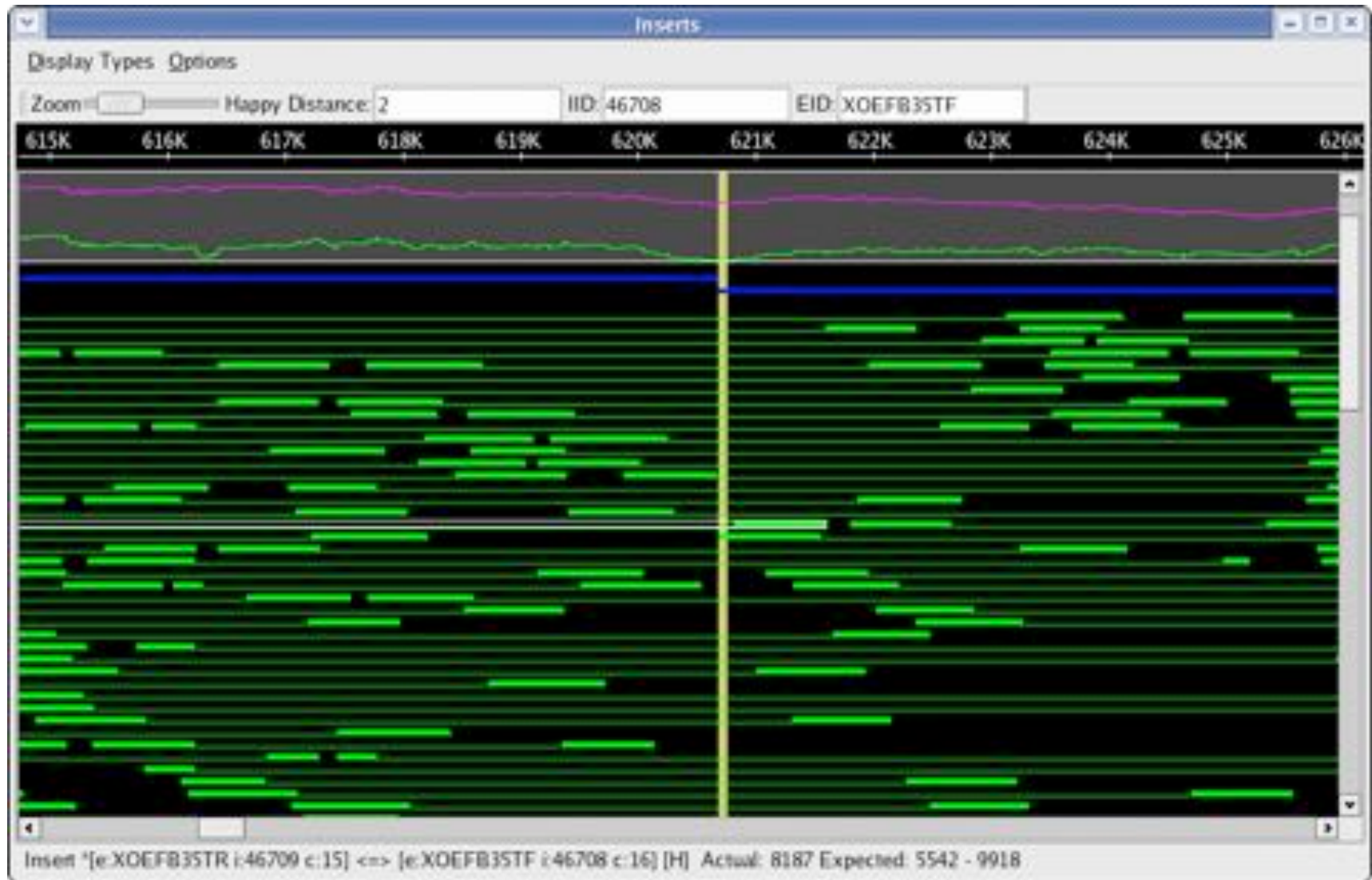
The individual reads (green) have been assembled into 2 contigs (blue & yellow). The mate relationship between the reads allows for the contigs to be oriented and the gap size to be estimated.

# Read Coverage



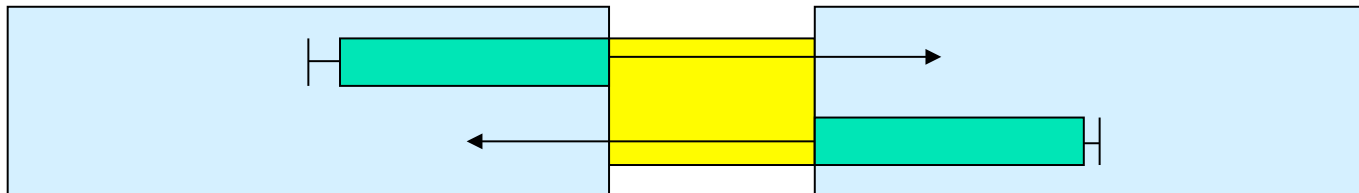
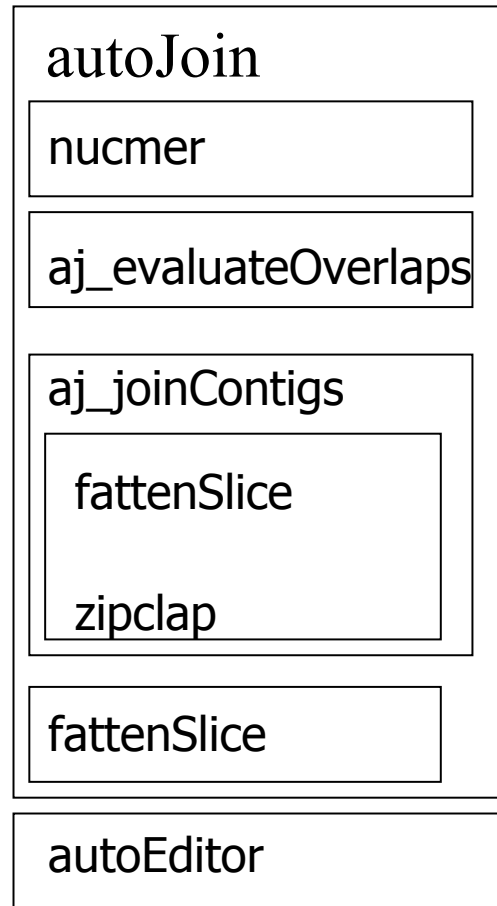


# Sequencing Gap



# AutoJoiner Architecture

1. All-vs-All Alignment
2. Analyze Alignments
3. Extend  
and  
Join Contigs
4. Contig Fattening
5. AutoEdit Result



# All-vs-All Alignment

```
read-alignment.txt + (ifocal/asmg/scratch/...lignment-1047283847434-1047283847435) - GVIM4
File Edit Tools Syntax Buffers Window Help
% show-coords -rc1 out.delta | grep XOEFB35TF
 35001  35998 | 1023  14 | 998   1010 | 95.15 | 53479 1023 | 1047283847434 XOEFB35TF
 53419  53479 | 1023 963 | 61    61 | 100.00 | 53479 1023 | 1047283847434 XOEFB35TF
  1012   1133 | 285 405 | 122   121 | 90.24 | 1163 1023 | XODA905TF   XOEFB35TF
   769    980 | 1023 813 | 212   211 | 99.06 | 980  1023 | XOEK861TR   XOEFB35TF

% show-aligns out.delta 1047283847434 XOEFB35TF
=====
-- Alignments between 1047283847434 and XOEFB35TF

-- BEGIN alignment [ +1 53419 - 53479 | -1 1023 - 963 ]

53419      tegtlogogtctggagaaccagatogacttggogccatcogctgggtccag
1023      tegtlogogtctggagaaccagatogacttggogccatcogctgggtccag

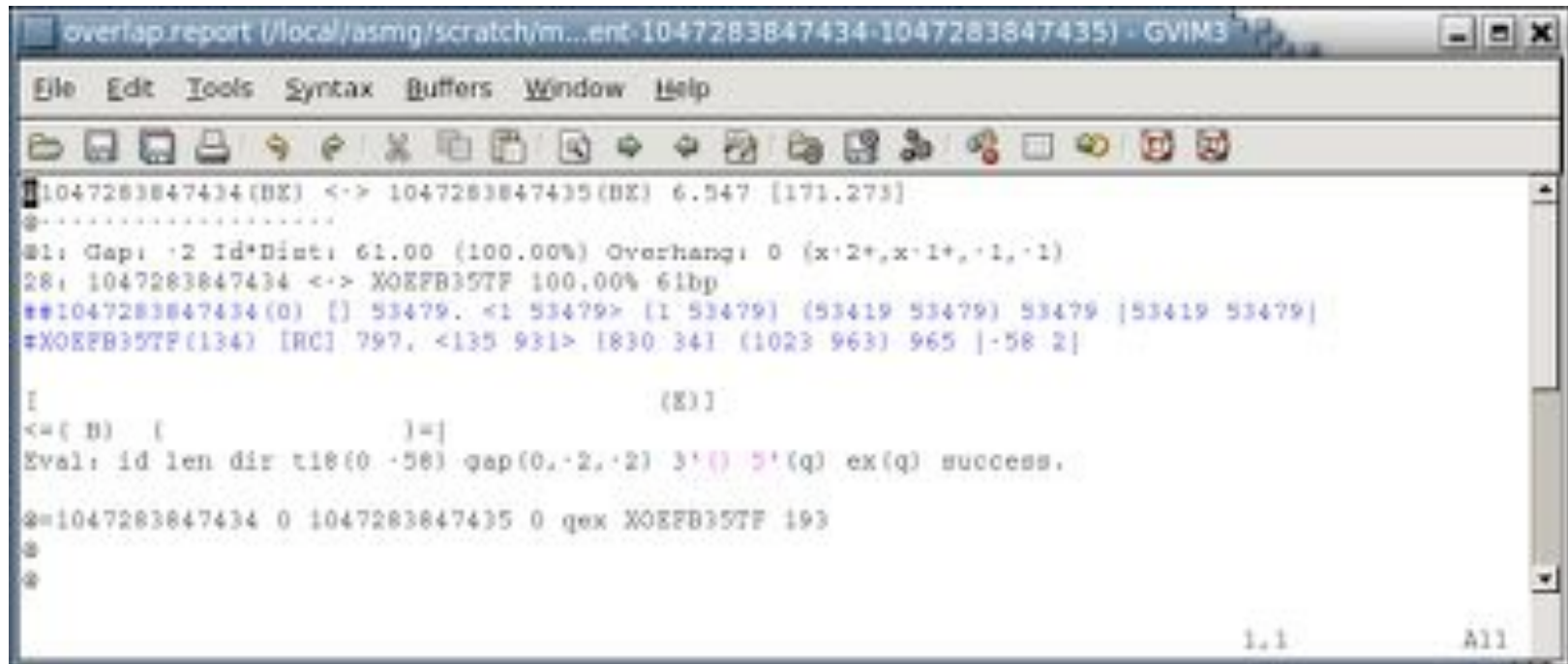
53468      ctgagccatcgg
974      ctgagccatcgg

-- END alignment [ +1 53419 - 53479 | -1 1023 - 963 ]

=====
25,60 Top
```

The first AutoJoiner!

# Alignment Reports



```
overlap.report (/local/asmg/scratch/m...ent-1047283847434-1047283847435) - GVIM3
File Edit Tools Syntax Buffers Window Help
[Icons]
#1047283847434 (BX) <-> 1047283847435 (BX) 6.547 [171.273]
@-----
@1: Gap: -2 Id*Dist: 61.00 (100.00%) Overhang: 0 (x+2+,x-1+,-1,-1)
28: 1047283847434 <-> XOEFB35TF 100.00% 61bp
##1047283847434 (0) [ ] 53479. <1 53479> [1 53479] (53419 53479) 53479 [53419 53479]
#XOEFB35TF(134) [RC] 797. <135 931> 1830 341 (1023 963) 965 [-58 2]

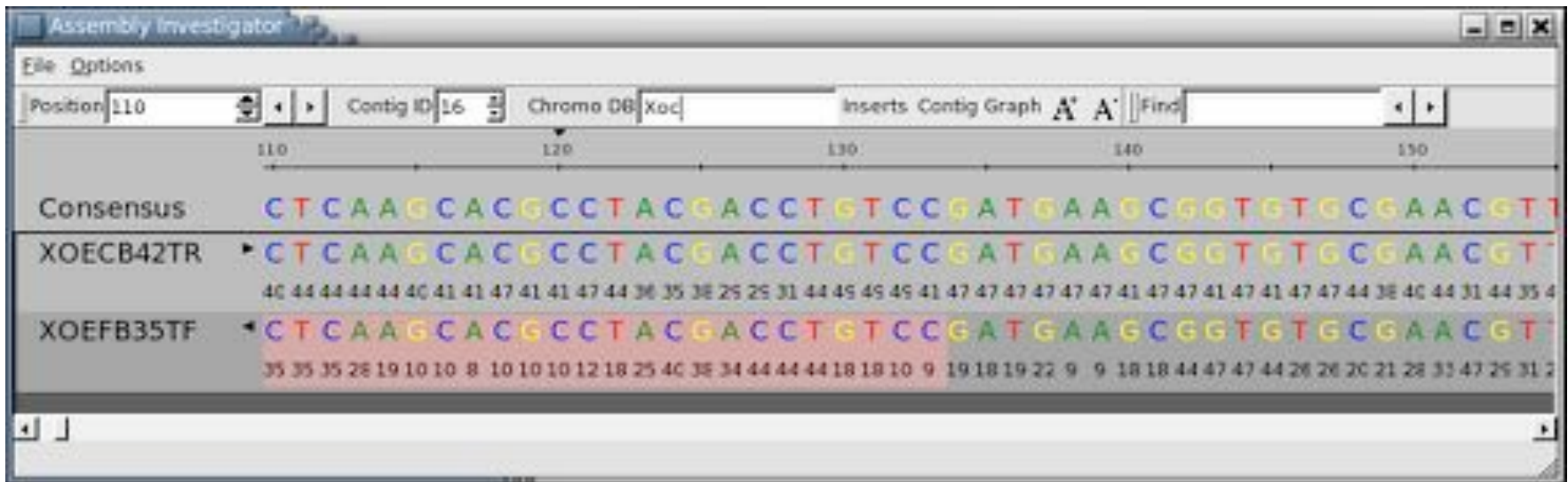
[ (8)]
<=( B) [ ]=|
Eval: id len dir t18(0 -58) gap(0,-2,-2) 3'(q) 5'(q) ex(q) success.

@=1047283847434 0 1047283847435 0 qex XOEFB35TF 193
@
@

1.1 All
```

Why did AutoJoiner make this join?

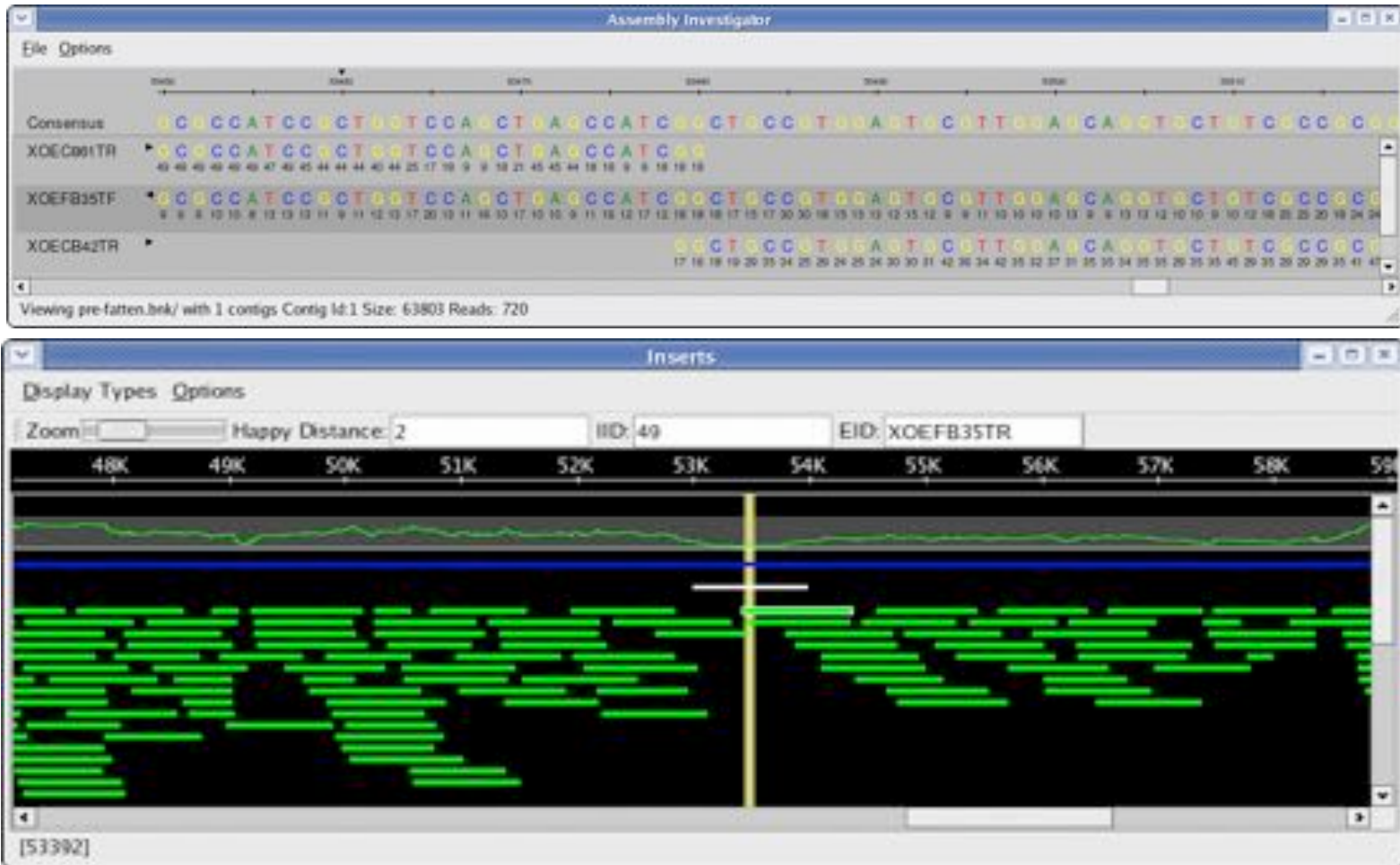
# Contig Extension



## Extension Procedure:

- If necessary, extended selected read by aligning trimmed bases to existing consensus.
- Untrim to desired base, promote untrimmed bases to consensus, shift offsets.

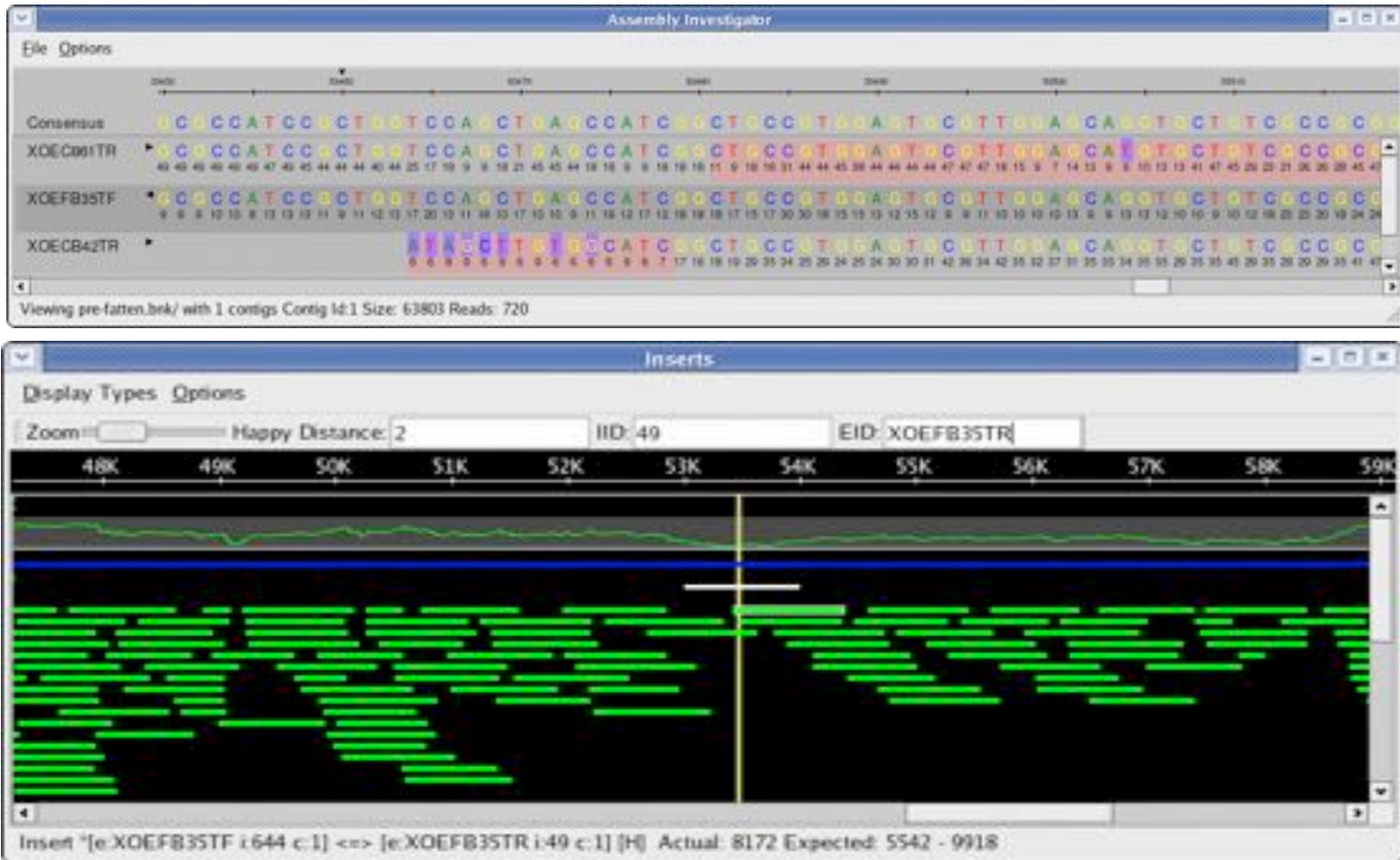
# Contig Joining



“Zip” together contigs by pairwise alignment between consensi.

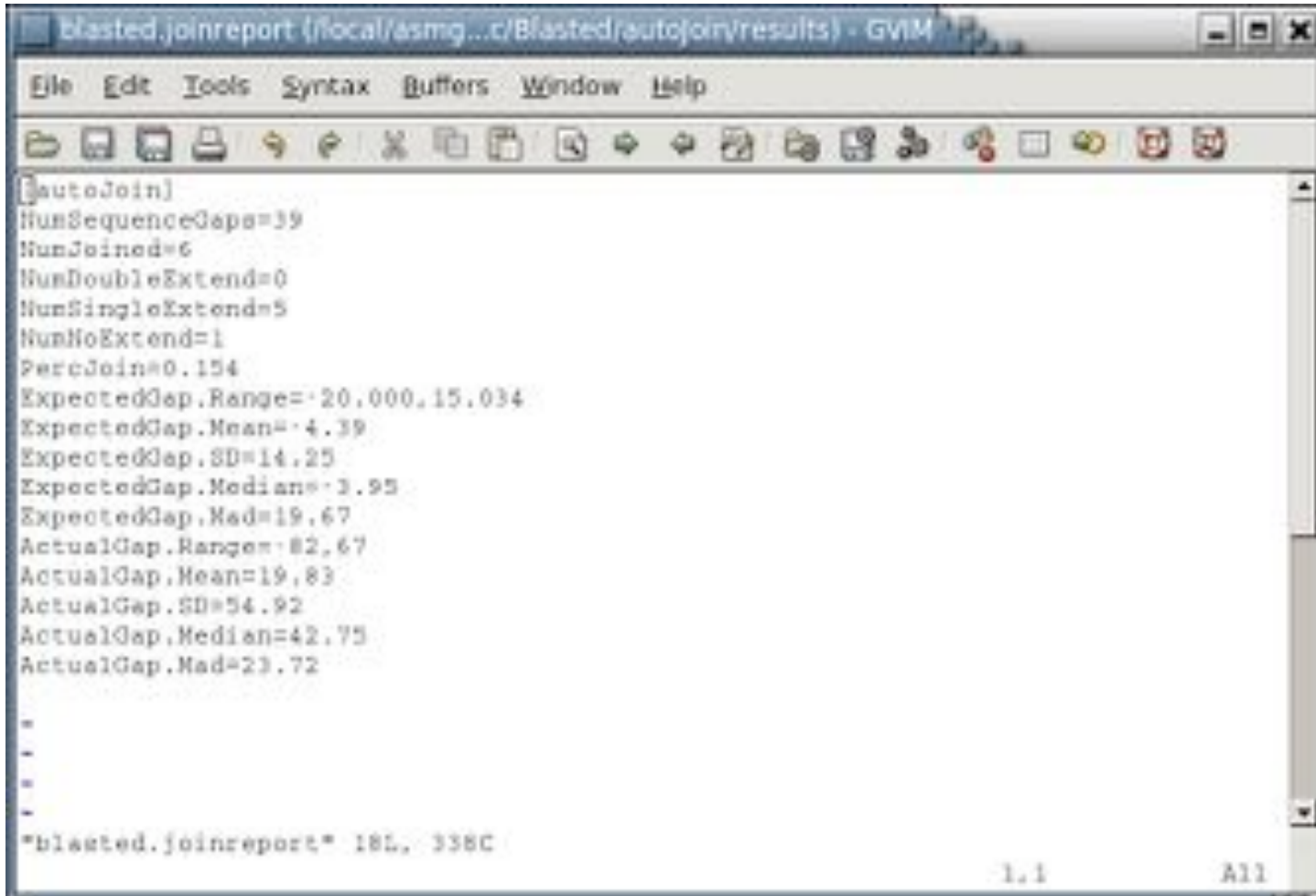


# Contig Fattening



“Fatten” addition reads in the join region to increase coverage.

# Join Report



```
blasted.joinreport (/local/asmg...c/Blasted/autojoin/results) - GVM
File Edit Tools Syntax Buffers Window Help
[autoJoin]
NumSequenceGaps=39
NumJoined=6
NumDoubleExtend=0
NumSingleExtend=5
NumNoExtend=1
PercJoin=0.154
ExpectedGap.Range=-20,000,15,034
ExpectedGap.Mean=-4.39
ExpectedGap.SD=14.25
ExpectedGap.Median=-3.95
ExpectedGap.Mad=19.67
ActualGap.Range=-82,67
ActualGap.Mean=19,83
ActualGap.SD=54.92
ActualGap.Median=42.75
ActualGap.Mad=23.72
=
--
=
-
"blasted.joinreport" 18L, 338C
1.1 All
```

What did AutoJoiner do?





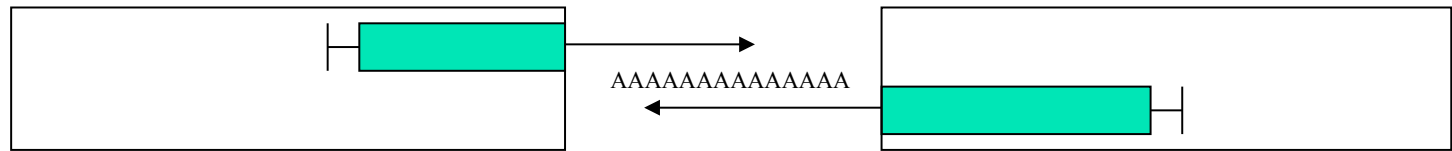
# AutoJoiner Validation

| Project          | Gaps        | Joined     | %             | Invalid  | Degenerate | Gap Size           | Mean          | Join ID             | Mean         |
|------------------|-------------|------------|---------------|----------|------------|--------------------|---------------|---------------------|--------------|
| blm              | 106         | 14         | 13.20%        | 0        | 0          | -200.5:156         | 21.18         | 98.39:100.00        | 99.26        |
| dmg              | 52          | 13         | 25.00%        | 0        | 3          | -666.5:36.5        | -140.73       | 98.99:100.00        | 99.64        |
| gb5              | 32          | 10         | 31.20%        | 0        | 0          | -13:146            | 42.3          | 98.50:99.87         | 99.31        |
| gba              | 110         | 38         | 34.50%        | 0        | 0          | -452:229           | 29.18         | 97.18:100.00        | 99.36        |
| gbm              | 43          | 6          | 14.00%        | 0        | 0          | -17:22             | -4.67         | 99.24:100.00        | 99.78        |
| gbr              | 32          | 11         | 34.40%        | 0        | 0          | -62.5:103.5        | 15.36         | 98.89:99.92         | 99.44        |
| gbs              | 31          | 5          | 16.10%        | 0        | 0          | -5:32.5            | 10.9          | 98.99:98.65         | 99.33        |
| gcb              | 10          | 2          | 20.00%        | 0        | 0          | -37.5:-4.5         | -21           | 99.38:99.81         | 99.59        |
| gcj              | 22          | 11         | 50.00%        | 0        | 0          | -53:139            | 27.45         | 98.71:99.81         | 99.45        |
| gcp              | 25          | 8          | 32.00%        | 0        | 4          | -555:184.5         | -75.31        | 99.03:99.85         | 99.42        |
| gde              | 82          | 17         | 20.70%        | 0        | 1          | -113:203.5         | 22.29         | 97.04:99.93         | 99.14        |
| ges              | 27          | 17         | 63.00%        | 0        | 0          | -779:-302          | -586.71       | 100.00:100.00       | 100          |
| gh               | 131         | 33         | 25.20%        | 0        | 6          | -182.5:212         | 21.79         | 98.81:100.00        | 99.51        |
| gmcap            | 10          | 2          | 20.00%        | 0        | 0          | -11.5:171          | 79.75         | 98.67:99.84         | 99.25        |
| gpi              | 150         | 52         | 34.70%        | 0        | 0          | -231.5:181         | 20.3          | 97.96:99.93         | 99.4         |
| gps              | 162         | 43         | 26.50%        | 0        | 0          | -1069.5:213.5      | -36.13        | 98.76:100.00        | 99.51        |
| gsa              | 262         | 32         | 12.20%        | 0        | 0          | -618:136           | -43.44        | 94.90:100.00        | 99.25        |
| crypt_1          | 20          | 8          | 40.00%        | 0        | 0          | -36:186.5          | 63.62         | 98.74:99.88         | 99.43        |
| crypt_2          | 7           | 5          | 71.40%        | 1        | 0          | -39:148            | 27.8          | 98.63:99.56         | 99.18        |
| crypt_3          | 21          | 8          | 38.10%        | 0        | 0          | -93:67.5           | -3.06         | 97.83:100.00        | 99.31        |
| crypt_4          | 25          | 7          | 28.00%        | 0        | 0          | -90:159            | 45.21         | 98.94:100.00        | 99.52        |
| crypt_5          | 23          | 8          | 34.80%        | 0        | 0          | -111.5:249         | 35.12         | 98.98:99.92         | 99.43        |
| crypt_6          | 14          | 7          | 50.00%        | 0        | 0          | -14:192            | 37.21         | 98.41:99.93         | 99.58        |
| crypt_7          | 17          | 6          | 35.30%        | 0        | 0          | -3.5:230.5         | 66.67         | 99.09:100.00        | 99.62        |
| crypt_8          | 15          | 6          | 40.00%        | 0        | 0          | -19:57.5           | 15            | 99.20:100.00        | 99.74        |
| crypt_9          | 12          | 6          | 50.00%        | 0        | 0          | -423:34            | -120.5        | 99.16:100.00        | 99.82        |
| crypt_10         | 14          | 7          | 50.00%        | 1        | 0          | -777:124           | -91.21        | 95.23:100.00        | 99.04        |
| crypt_11         | 12          | 2          | 16.70%        | 0        | 0          | -6:69.5            | 31.75         | 99.63:99.69         | 99.66        |
| crypt_12         | 10          | 4          | 40.00%        | 0        | 0          | -340:77.5          | -69.88        | 99.77:100.00        | 99.86        |
| crypt_13         | 13          | 7          | 53.80%        | 1        | 0          | -213.5:144         | 19.07         | 99.38:100.00        | 99.7         |
| <b>Composite</b> | <b>1490</b> | <b>395</b> | <b>26.51%</b> | <b>3</b> | <b>14</b>  | <b>-1069.5:249</b> | <b>-25.89</b> | <b>94.90:100.00</b> | <b>99.45</b> |

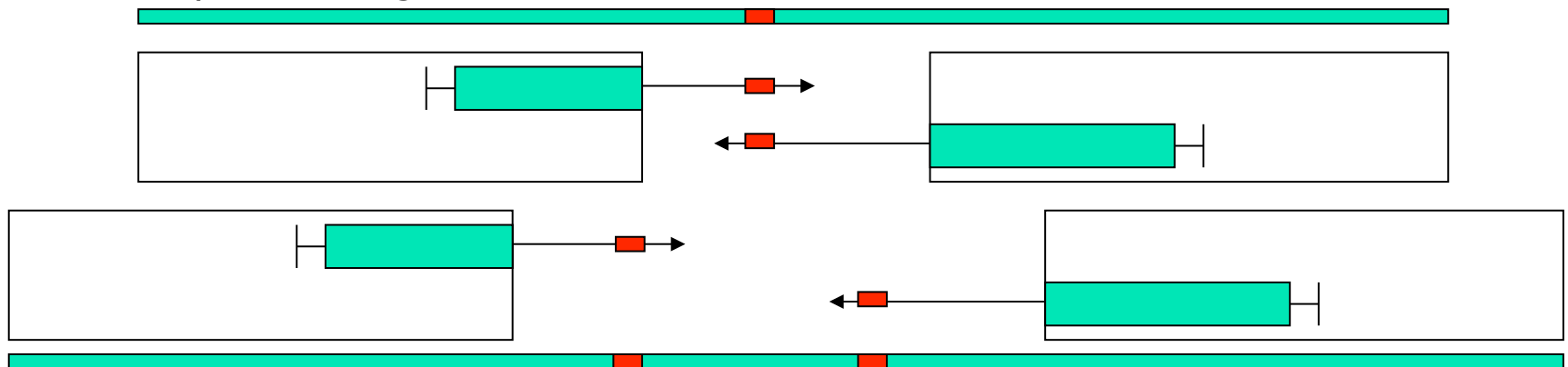
25%+ of all sequencing gaps closed with 3 mistakes.

# Complicating Issues

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



- Undetected repeats
  - Require strict agreement with scaffold

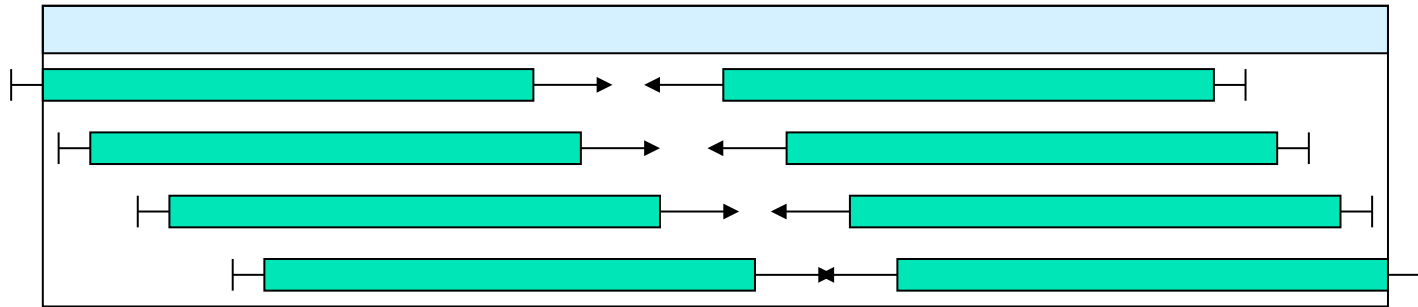


- Chimeric reads / Hard Stops
  - Good: Require high alignment similarity.
  - Better: Recognize hard stops by coverage gradients, other clues.
  - Best: Recognize unreliable sequence at chromatogram level.

# Pre-Production Techniques

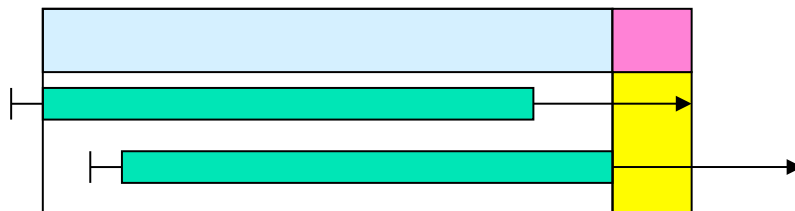
## ■ Contig Fattening

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



## ■ Contig Growing

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





# Measuring Assembly Quality

---

- **Gross Status** QC file
  - scaffold & contig sizes
- **Connectedness Status** Cloe, AssemblyViewer,  
getCoverage, cvgChop, asmQC
  - read & clone coverage
- **Insert Status** QC file, Assembly Viewer, asmQC
  - mate happiness,
  - library randomness
- **Consensus Status** Cloe, getCoverage, getqc
  - Quality Class
  - Consensus Quality Value
- **Read Status** findTcovSnps
  - Correlated SNPs



# Finding Suspicious Regions

---

```
% ls
```

```
blasted.frg blasted.asm
```

```
% /local/asmg/Linux/bin/cavalidate blasted
```

```
Doing step 10: toAmos
```

```
Doing step 20: bank-transact
```

```
Doing step 30: asmQC
```

```
Doing step 40: bank2contig
```

```
Doing step 50: getCoverage
```

```
Doing step 60: findTcovSnps
```

```
Doing step 70: ClusterSnps
```

```
Doing step 80: Load SNP Features
```

```
Doing step 90: Find Surrogates
```

```
Doing step 100: Load Surrogates
```

```
Doing step 1000: Dump Features
```

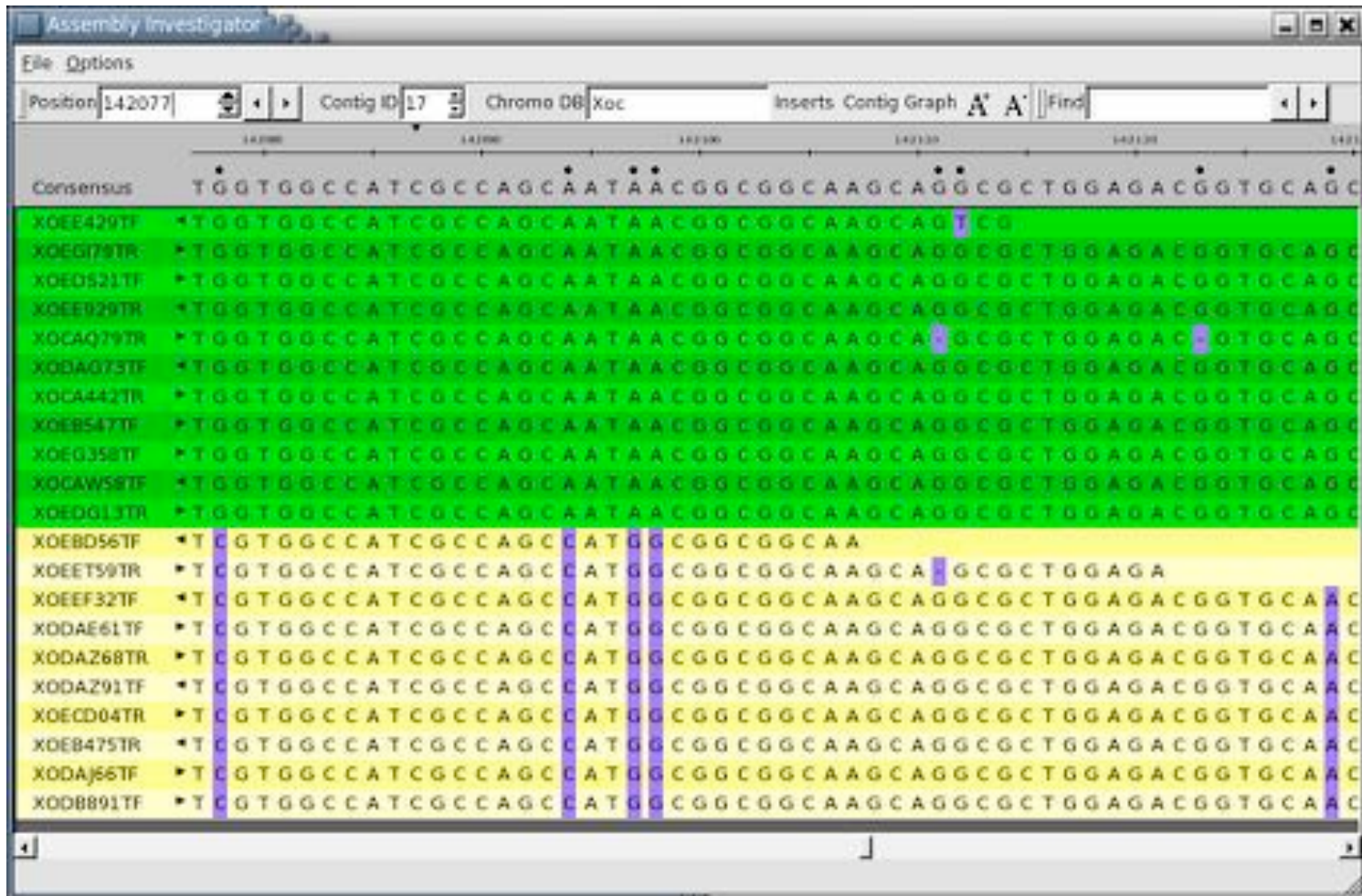
```
Doing step 1010: Get Suspicious Features
```

```
Doing step 1020: Create Suspicious Regions
```

```
% sort -nrk 6 blasted.snp.feats | head
```

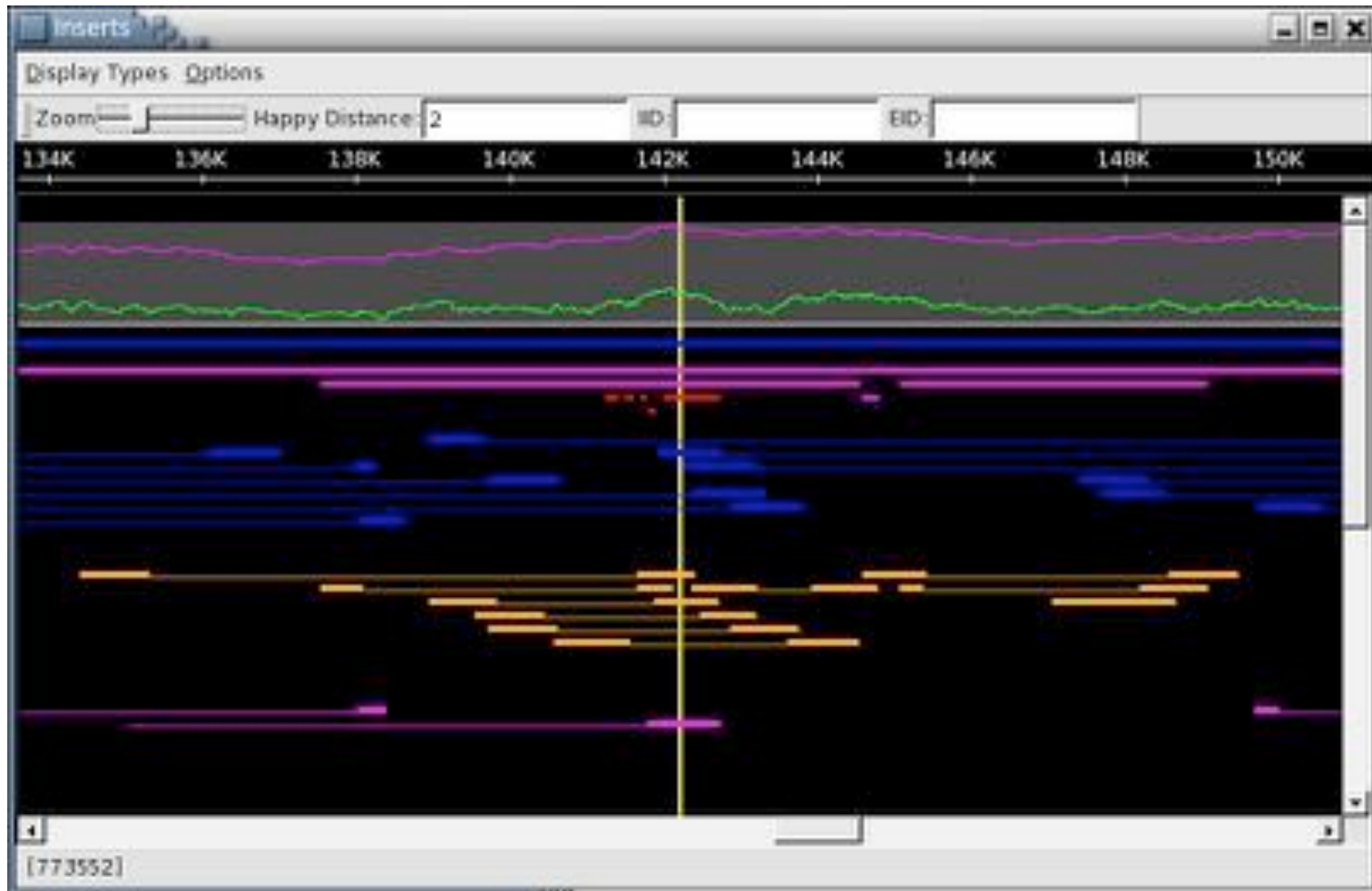
```
1047283847436 P HIGH_SNP 141991 142712 46 15.67
```

# Overcollapsed Repeat?



For a bacterial sample, correlated discrepancies strongly suggest a repeat has been collapsed.

# Mate View

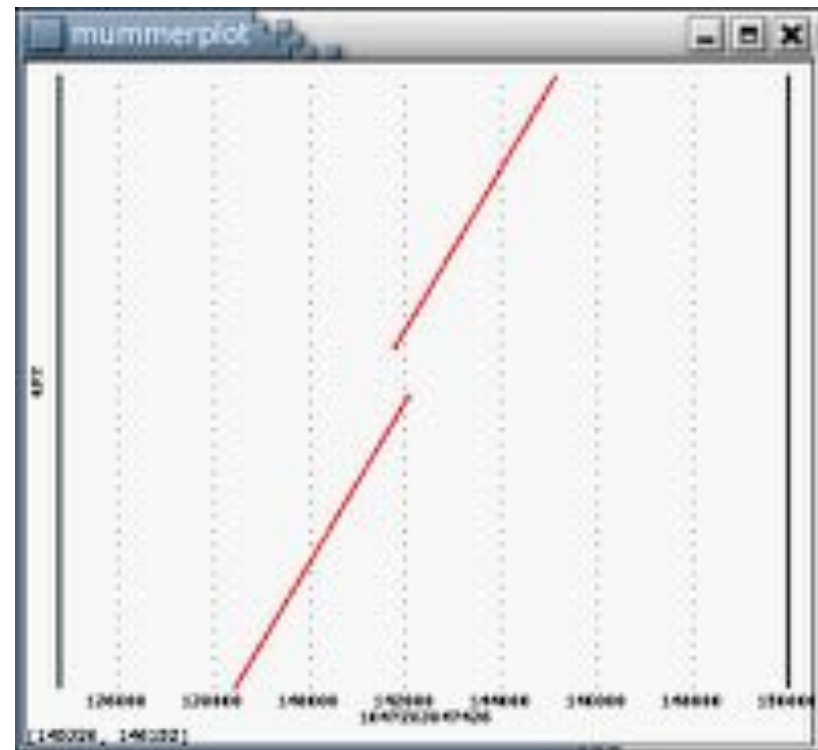
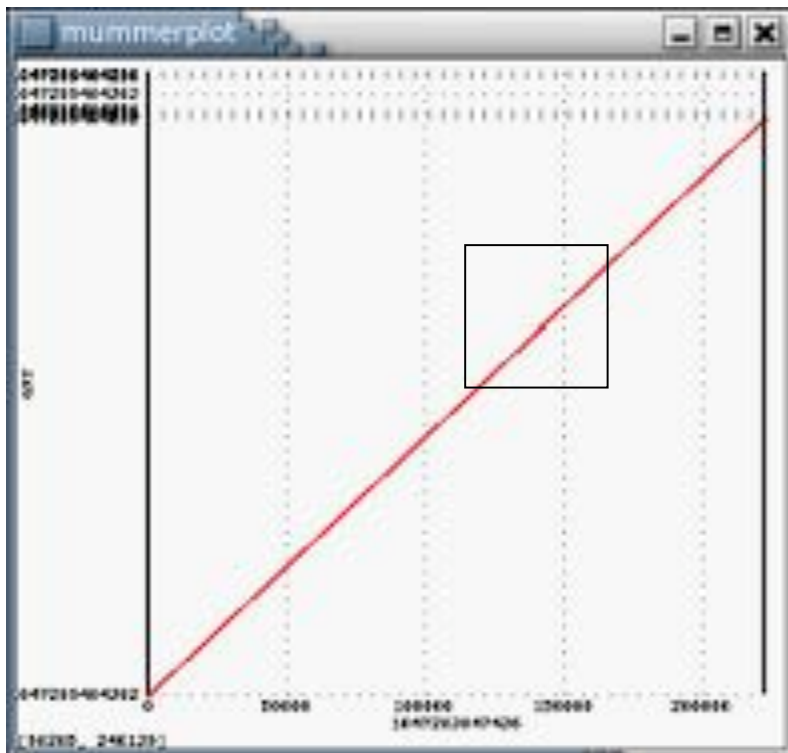


The shrunken mates (orange) suggest the assembly has a deletion from the true sequence.



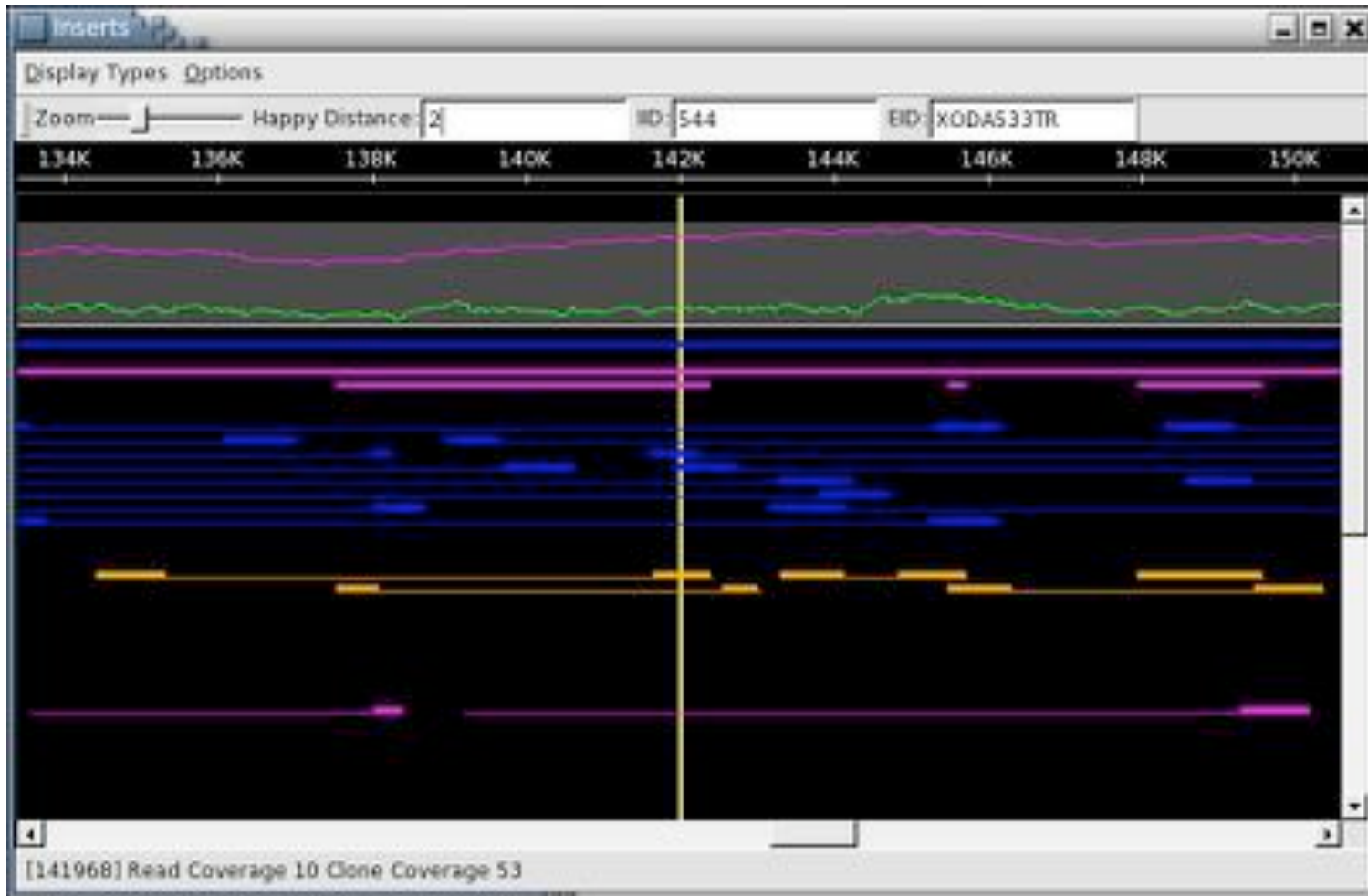
# Local Assembly

```
% run_CA -local -noedit -noupload local.frg -dir ca-0.003 -e 0.003
% nucmer 1047283847436.fasta ca-0.003/local.fasta
% /local/asmg/Linux/bin/mummerplot out.delta -R 1047283847436.fasta -Q ca-0.003/
  local.fasta -layout -filter
```





# Resolved Repeat



Unfortunately, size violated mates are only a clue.  
Ask Mihai for current research techniques.



# Final Results

|                                | Original Assembly         | Reassembled              |
|--------------------------------|---------------------------|--------------------------|
| <b>TotalScaffolds</b>          | 21                        | 5                        |
| <b>TotalSpanOfScaffolds</b>    | 4770228                   | 4819528                  |
| <b>IntraScaffoldGaps</b>       | 51                        | 33                       |
| <b>MeanSequenceGapSize</b>     | 410.18                    | 96.94                    |
| <b>[Top5Scaffolds]</b>         |                           |                          |
| 1                              | 25:2156009:2164583:357.25 | 34:4810208:4813407:96.94 |
| 2                              | 13:826284:830667:365.25   | 1:2558:2558:0.00         |
| 3                              | 3:437076:439209:1066.50   | 1:1473:1473:0.00         |
| 4                              | 3:333768:334012:122.00    | 1:1056:1056:0.00         |
| 5                              | 5:310971:311756:196.25    | 1:1034:1034:0.00         |
| <b>TotalContigsInScaffolds</b> | 72                        | 38                       |
| <b>N50ContigBases</b>          | 151430                    | 253084                   |
| <b>TotalDegenContigs</b>       | 125                       | 1                        |
| <b>DegenContigLength</b>       | 129182                    | 959                      |
| <b>MeanDegenContigSize</b>     | 1033.46                   | 959                      |
| <b>[Top5Contigs]</b>           |                           |                          |
| 1                              | 5939:536280               | 6291:516881              |
| 2                              | 2751:259499               | 6203:436501              |
| 3                              | 3005:238048               | 3805:385244              |
| 4                              | 2199:220720               | 4292:364829              |
| 5                              | 2509:196450               | 4025:355502              |

Expected Genome Size: 4.85Mb - 5.04Mb



# Research Directions

---

- AutoEditor 2.0: Better results, better engineering
- Context Based trimming
  - Partial Overlaps
  - Reference sequence
- Advanced CA Techniques
  - Contained Stones bug fixes
  - Blasting Degenerate and Surrogate Unitigs
  - Assembling in the gap
- Arachne & Other Assemblers
- Assembler Reconciliation
- AMOS Framework
- Assembly Forensics
- Assembly Visualization / Navigation



# Conclusions

---

- Overriding strategy: Start conservatively, and iteratively build as more information becomes available.
- 95.5% - 99.2% of genome in a single scaffold not typical yet, but it could be.
  - Be aware of potential size/quality tradeoffs, though.
- Assembly is complicated by genome structure, repeat characteristics, data quality, data management- one size does not fit all, ask for help.
  - Use Data Support!

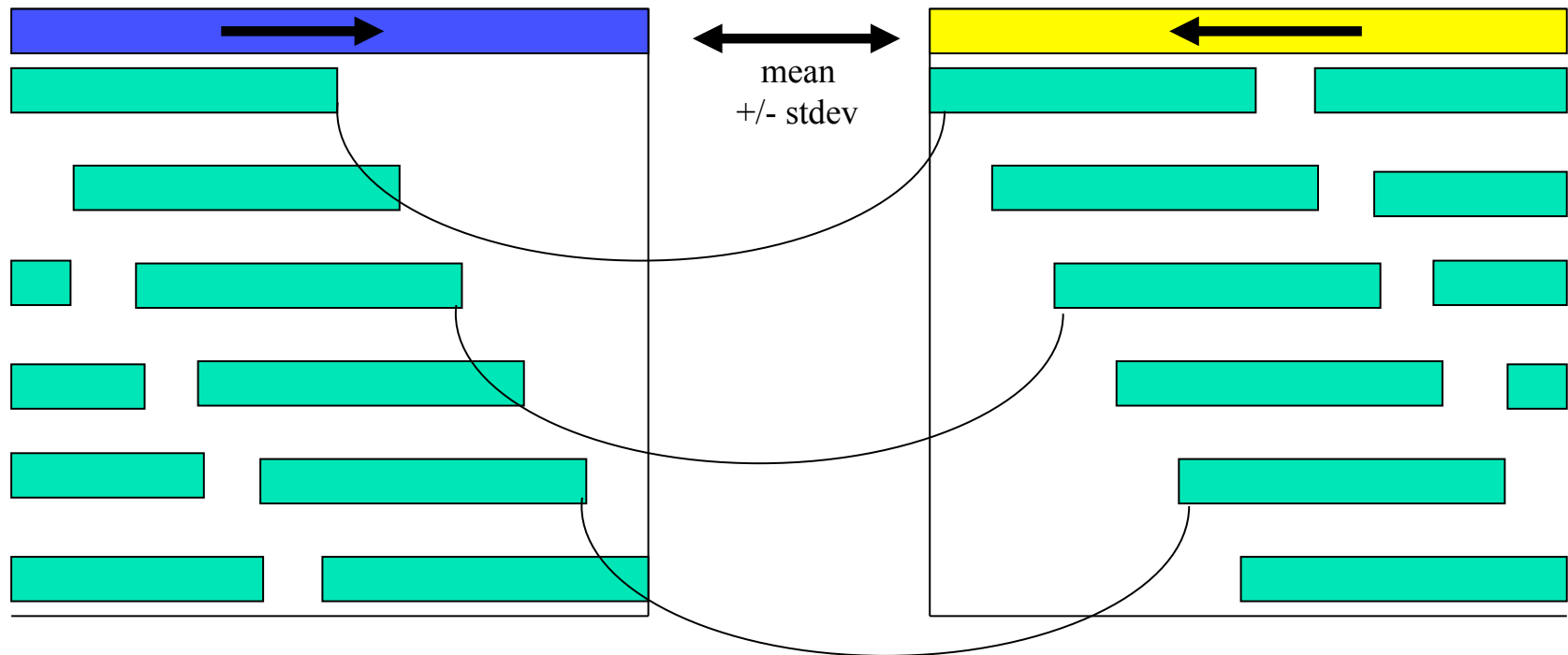


# Acknowledgements

---

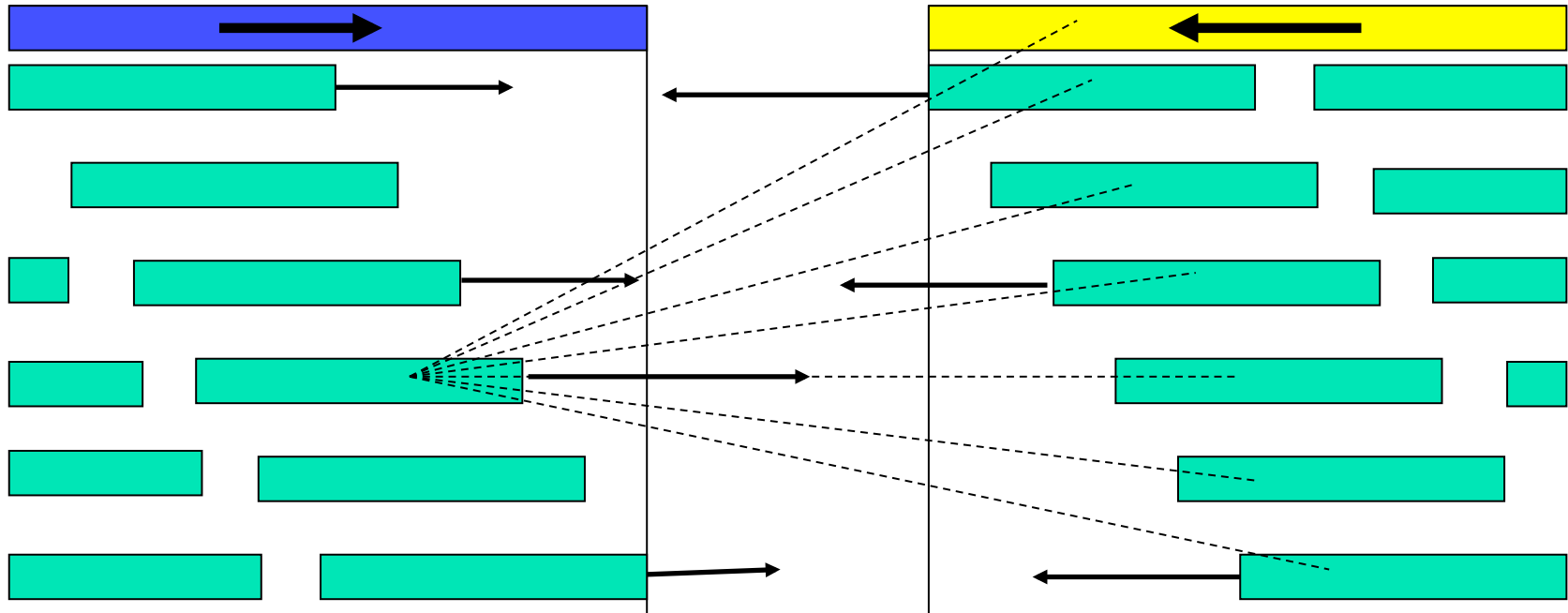
- Steven Salzberg
- Martin Shumway
- Jason Miller
- Pawel Gajer
- Art Delcher
- Mihai Pop
- Adam Phillippy
- WGA
- SE
- Data Support
- Jane Carlton
- Vish Nene

# Sequencing Gap



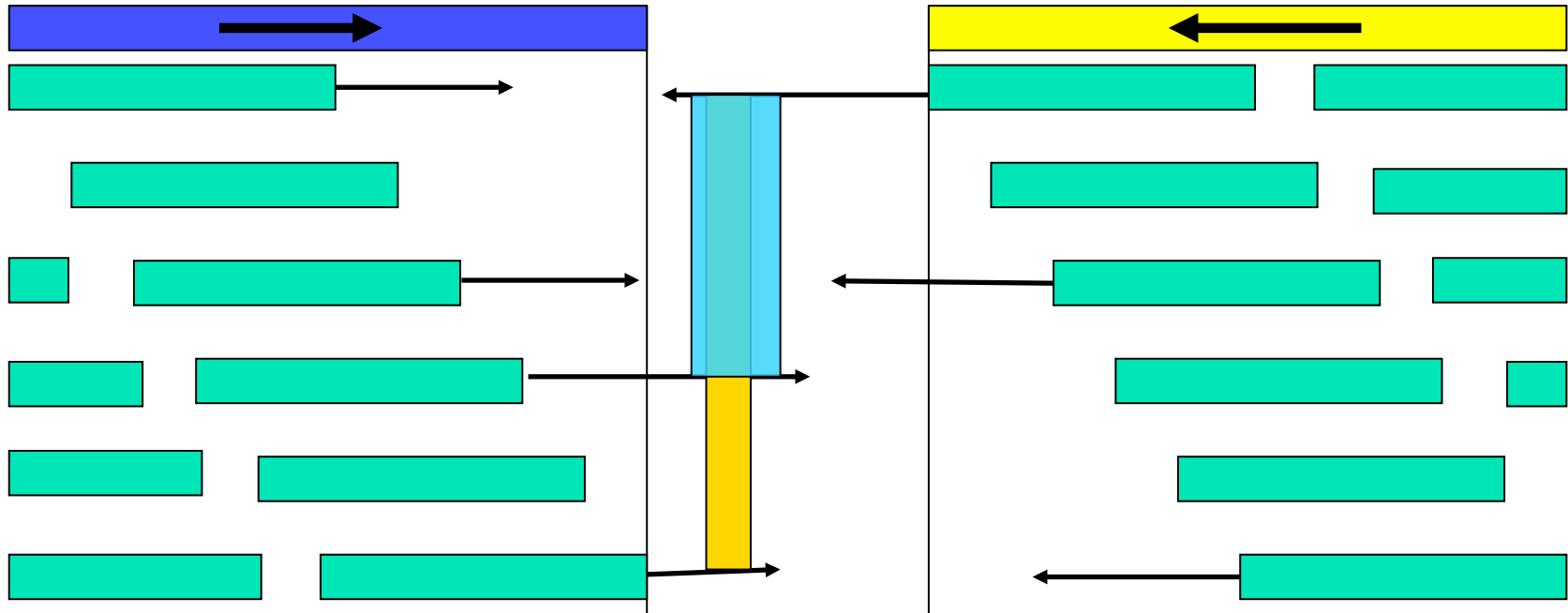
The individual reads (green) have been assembled into 2 contigs (blue & yellow). The mate relationship between the reads allows for the contigs to be oriented and the gap size to be estimated.

# All-vs-all Alignment



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

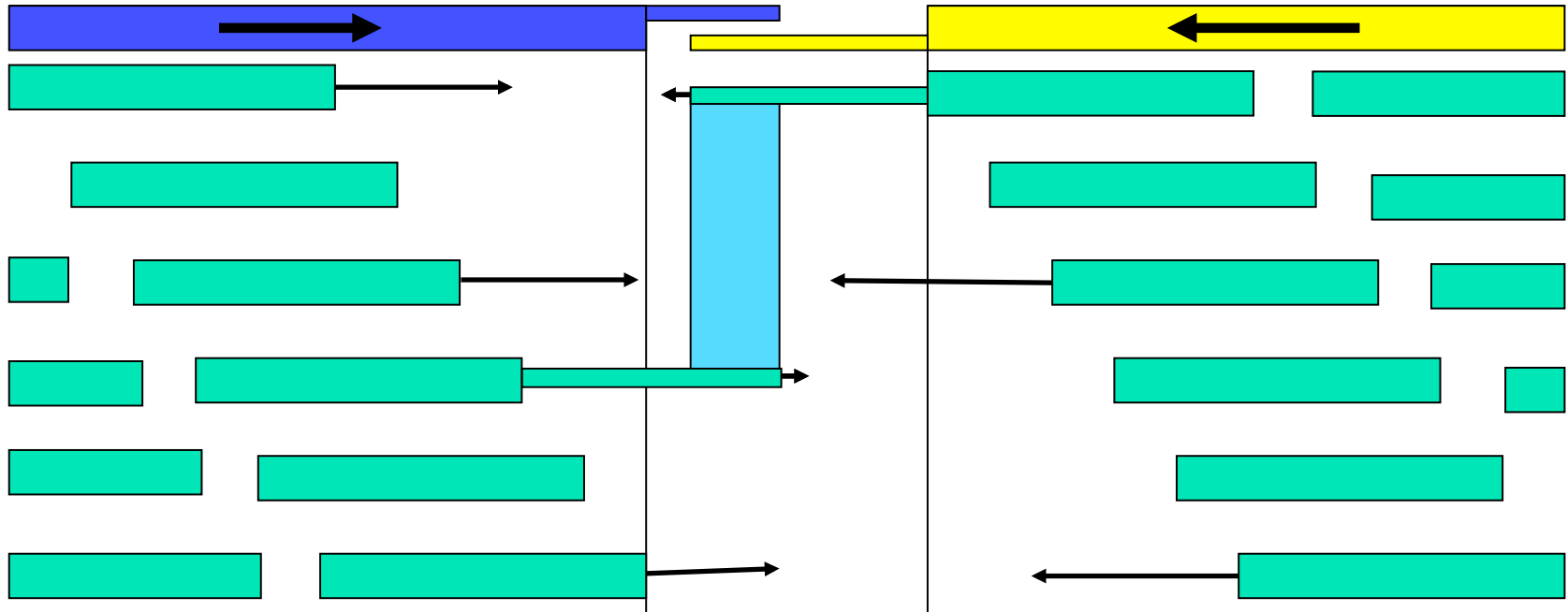
# Alignment Analysis



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.

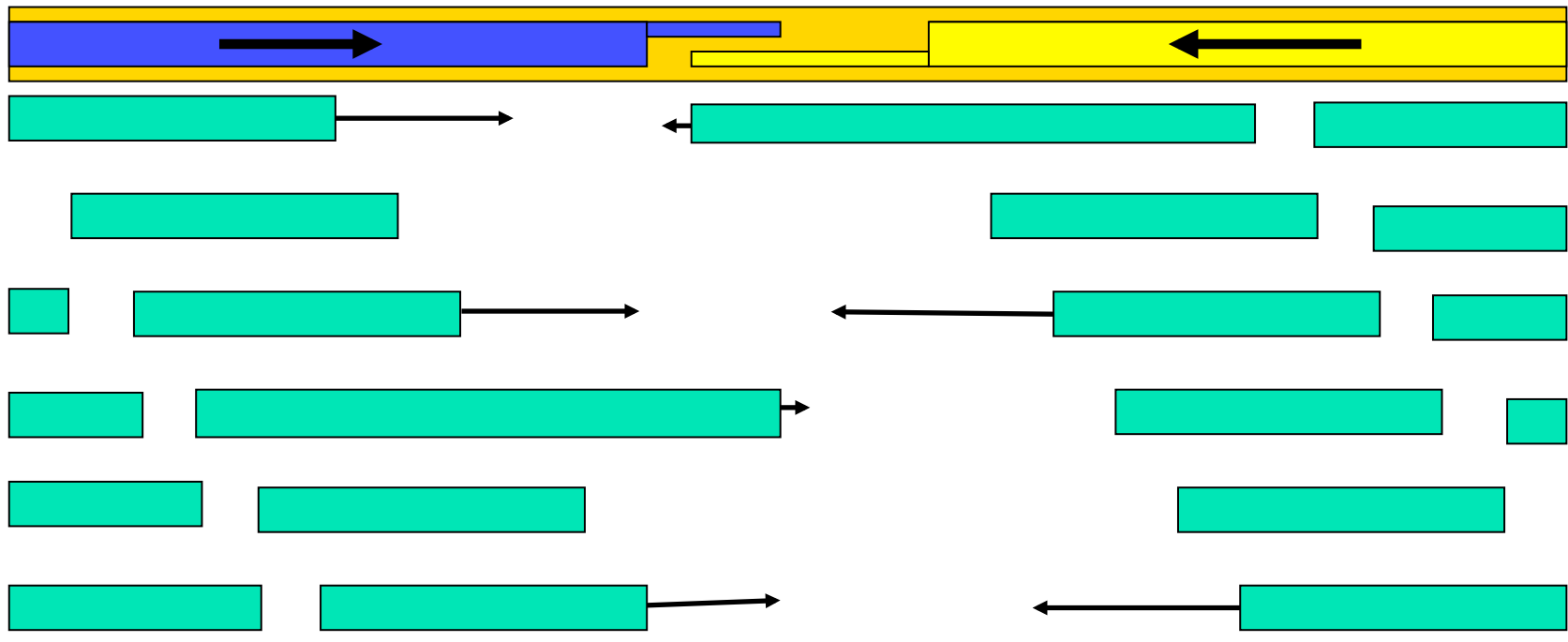


# Contig Extension



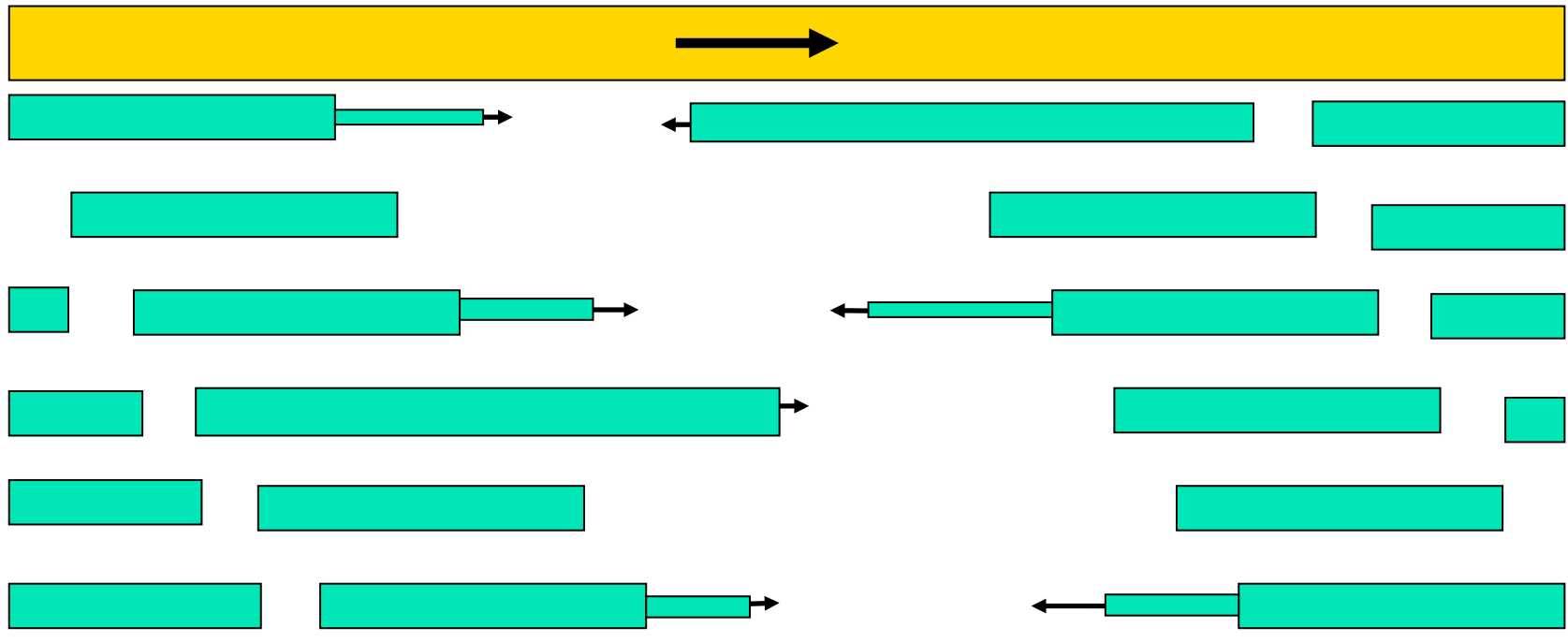
3. The contigs are extended by extending the selected reads beyond their original clear range to the desired position. If necessary, the reads are first aligned to the existing consensus.

# Contig Joining



4. The contigs are joined by aligning the newly extended consensi. Alignment gaps inserted into the conseni are promoted into the appropriate positions in the underlying multiple alignment. 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.