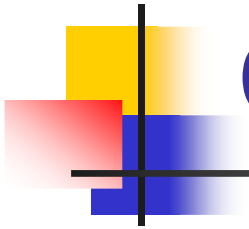
AMOS Assembly Validation and Visualization

Michael Schatz

Center for Bioinformatics and Computational Biology
University of Maryland

August 13, 2006
University of Hawaii



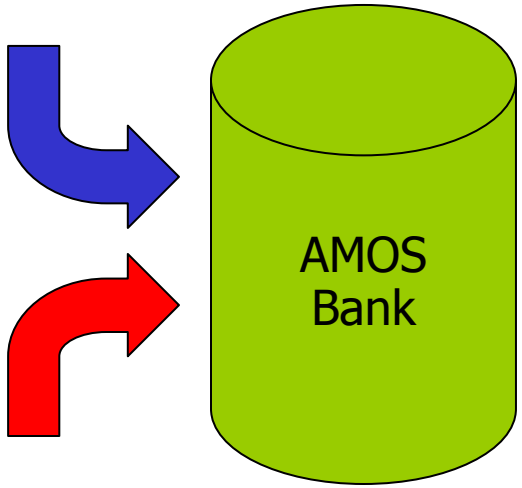
Outline

- AMOS Validation Pipeline
  - Mate-Based Validation
    - C/E Statistic
  - Read Alignment Validation
  - Read Breakpoint Validation
  - Read Depth Validation
- Hawkeye
  - Contigs, Inserts, Histograms, SNP Barcode, Features
  - Misassembly Walkthrough

AMOS Validation Pipeline

- Automatically scan an assembly to locate misassembly signatures for further analysis and correction

- cavalidate prefix (.frg, .asm)
  1. Load CA Assembly Data into Bank
  2. Evaluate Mate Pairs & Libraries
  3. Evaluate Read Alignments
  4. Evaluate Read Breakpoints
  5. Analyze Depth of Coverage
  6. List Surrogates
  7. Load Misassembly Signatures into Bank



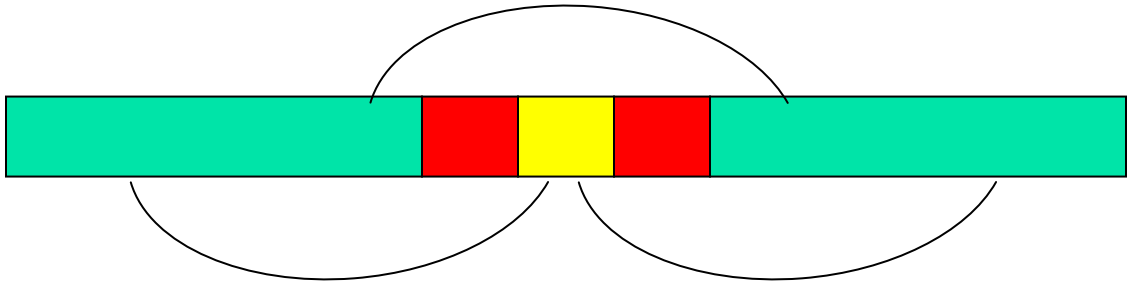
- amosvalidate prefix (.afg)
  - Same as cavalidate, except skips surrogates

Mate-Happiness: asmQC

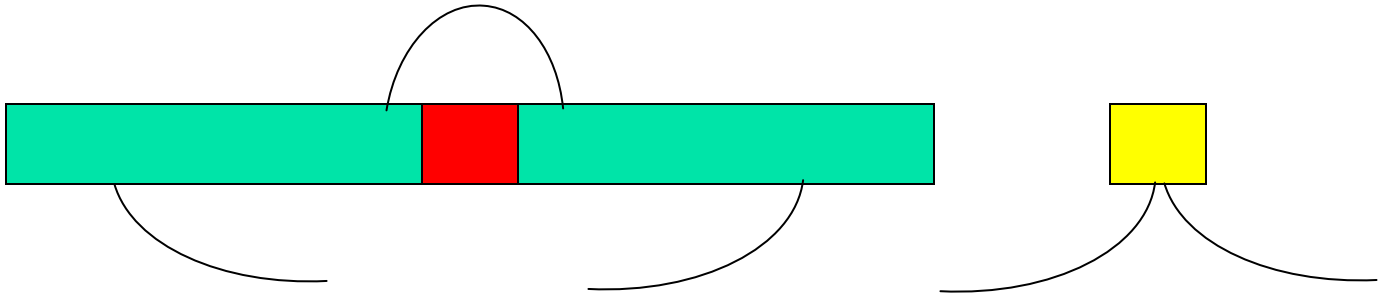
- Evaluate mate "happiness" across assembly
  - Happy = Correct orientation and distance
- Finds regions with multiple:
  - Compressed Mates
  - Expanded Mates
  - Invalid same orientation (→ →)
  - Invalid outie orientation (← →)
  - Missing Mates
    - Linking mates (mate in a different scaffold)
    - Singleton mates (mate is not in any contig)
- Regions with high C/E statistic

Mate-Happiness: asmQC

- Excision: Skip reads between flanking repeats
  - Truth

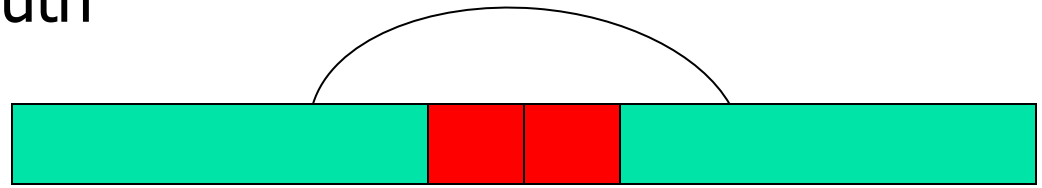


  - Misassembly: Compressed Mates, Missing Mates

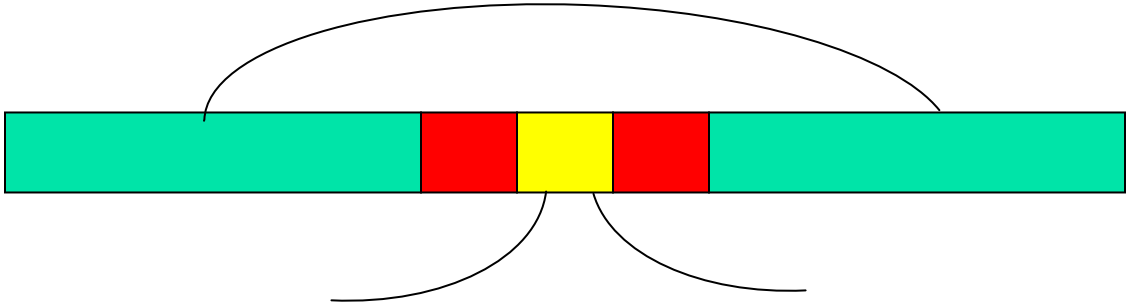


Mate-Happiness: asmQC

- Insertion: Additional reads between flanking repeats
  - Truth

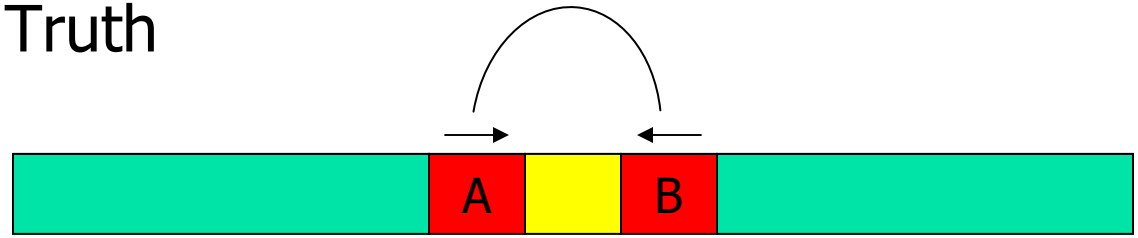


  - Misassembly: Expanded Mates, Missing Mates

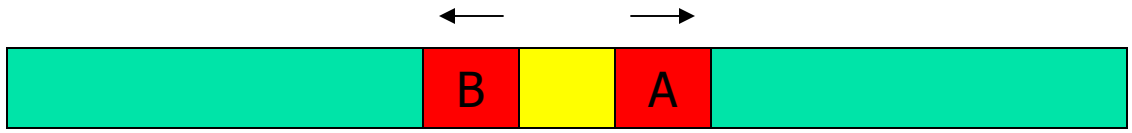


Mate-Happiness: asmQC

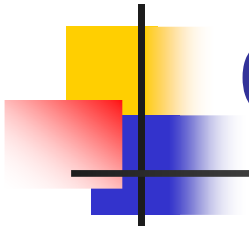
- Rearrangement: Reordering of reads
  - Truth



  - Misassembly: Misoriented Mates



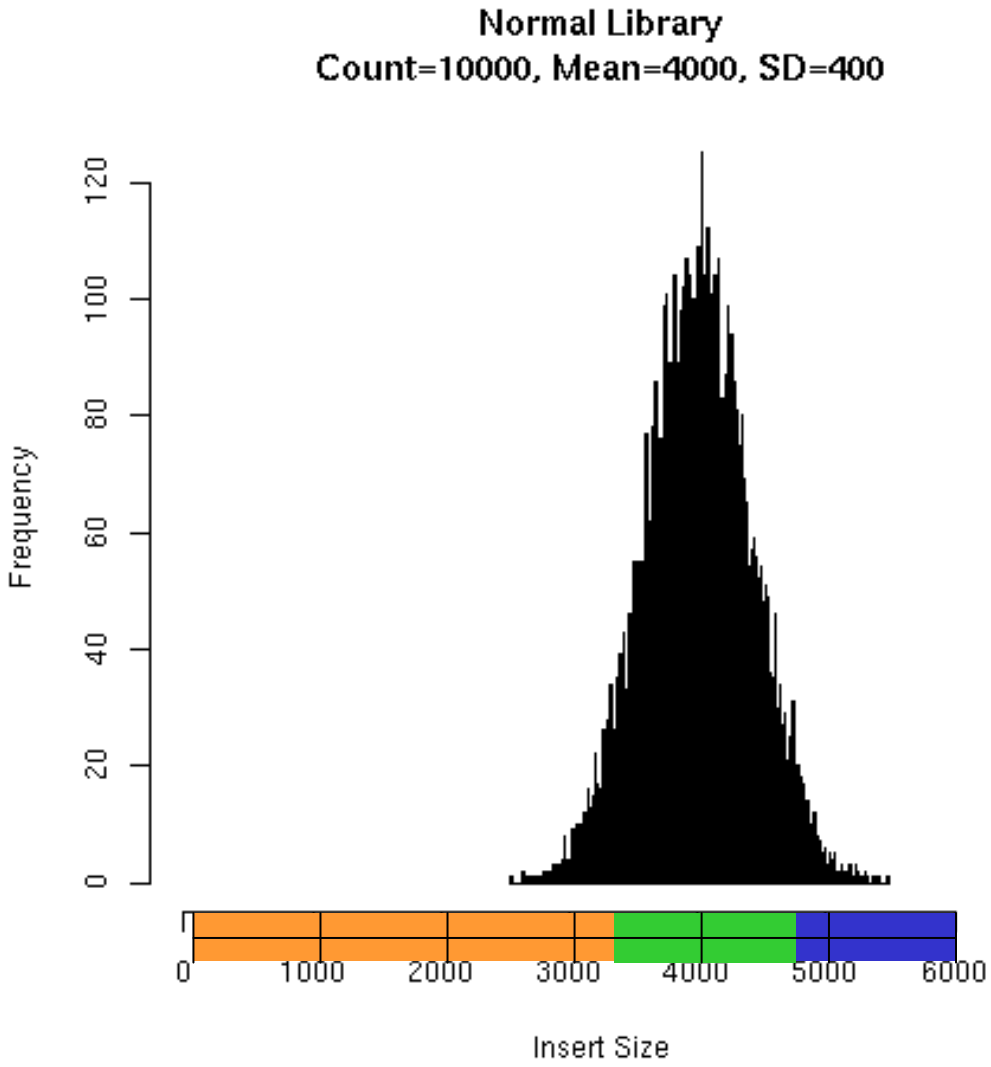
Note: Unhappy mates may also occur for biological or technical reasons.

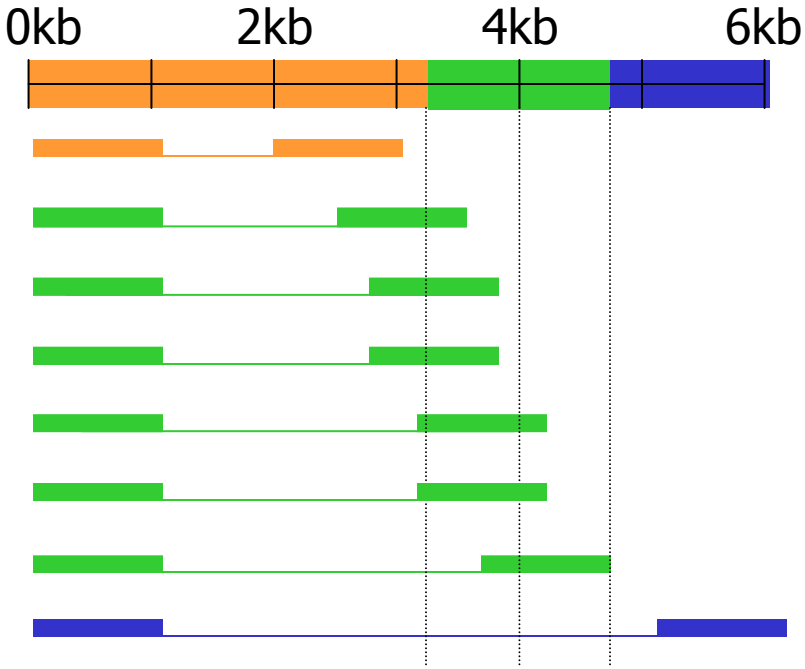


C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.
- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”
- Compute the statistic at all positions
  - (Local Mean – Global Mean) / Scaling Factor
- Introduced by Jim Yorke’s group at UMD

Sampling the Genome





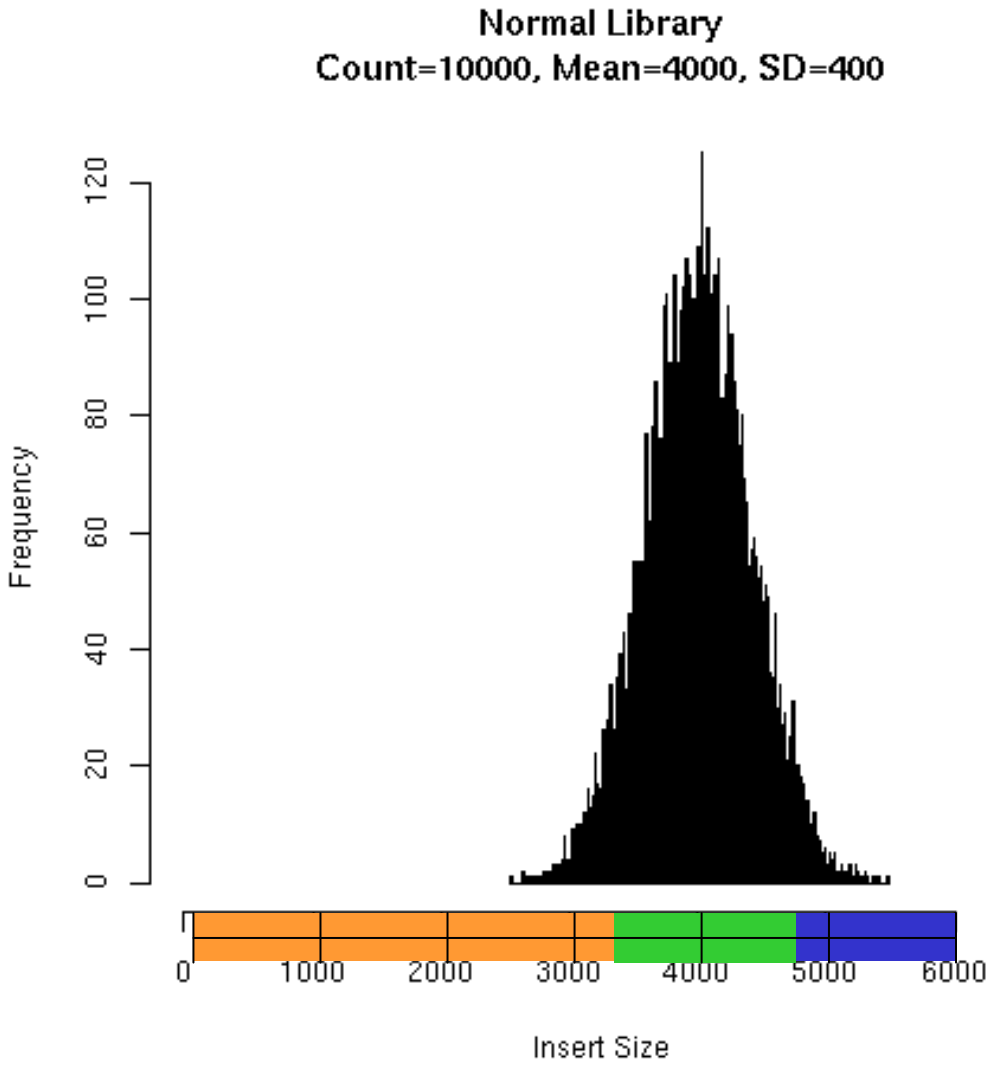
8 inserts: 3kb-6kb

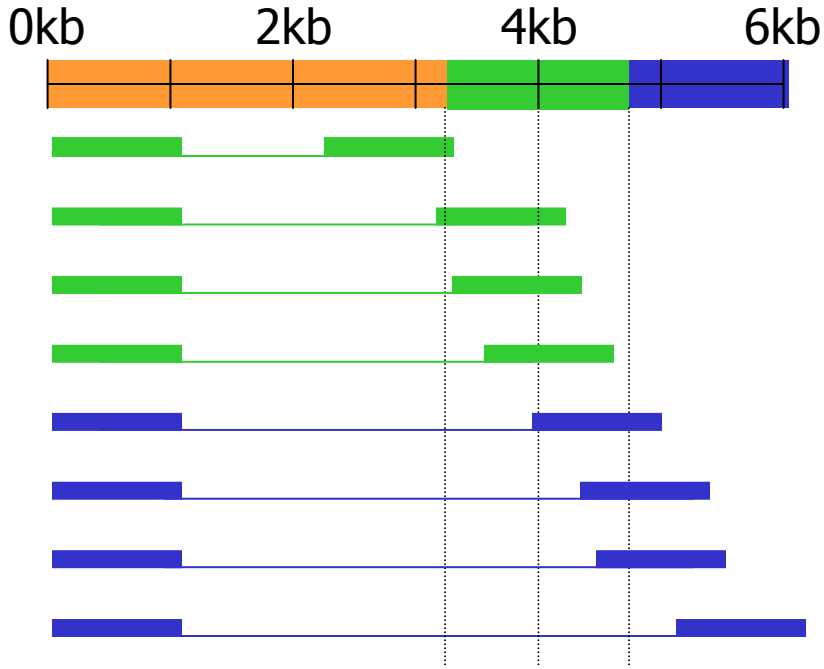
Local Mean: 4048

C/E Stat: $\frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33$

Near 0 indicates overall happiness

C/E-Statistic: Expansion





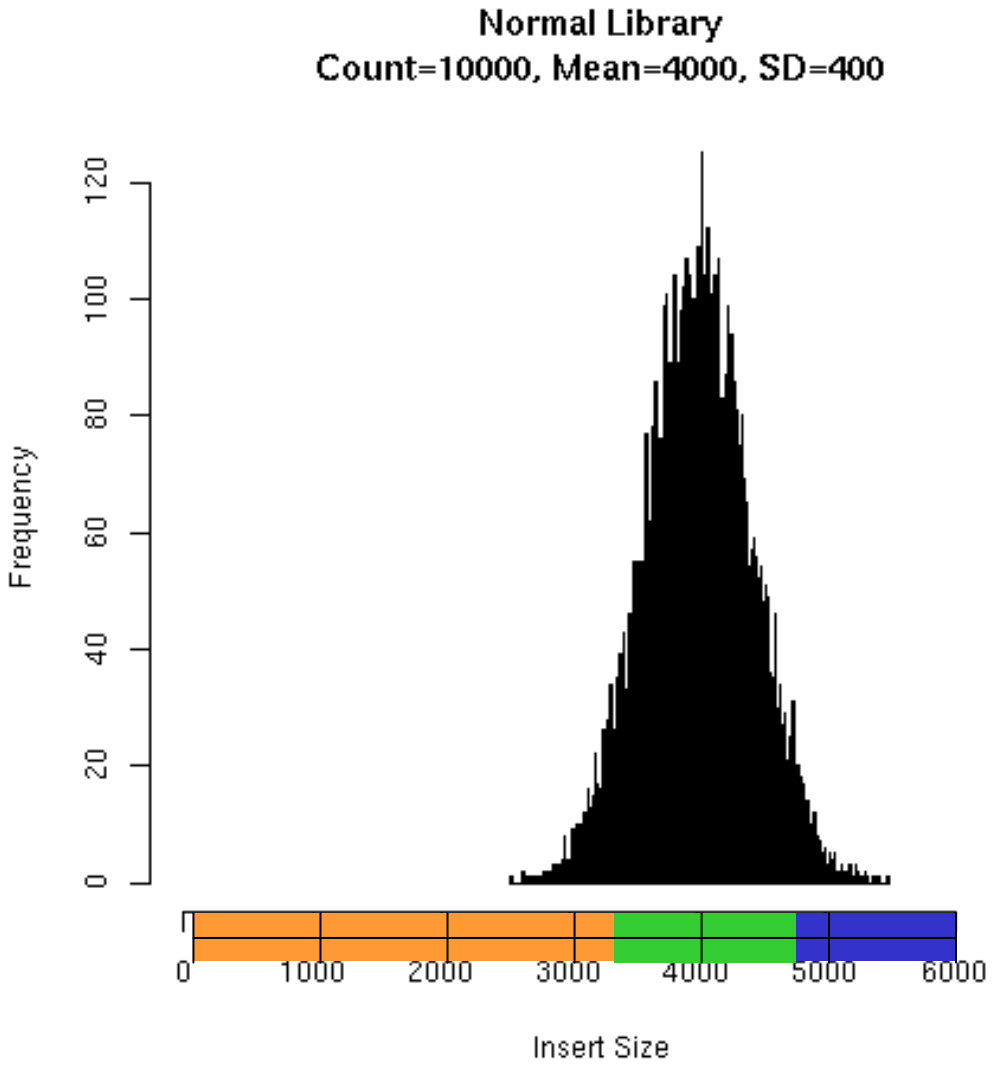
8 inserts: 3.2kb-6kb

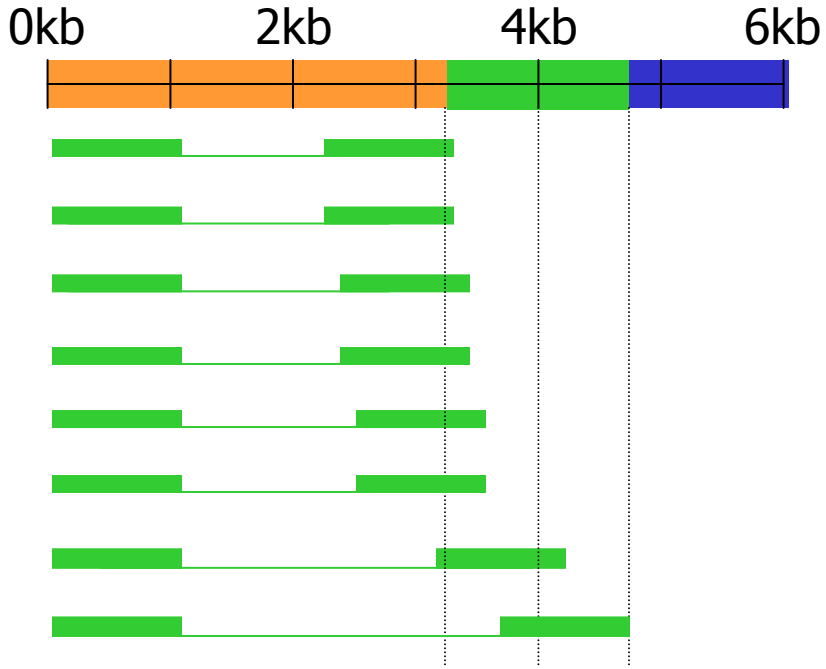
Local Mean: 4461

C/E Stat: $\frac{(4461-4000)}{(400 / \sqrt{8})} = +3.26$

C/E Stat ≥ 3.0 indicates Expansion

C/E-Statistic: Compression





8 inserts: 3.2 kb-4.8kb

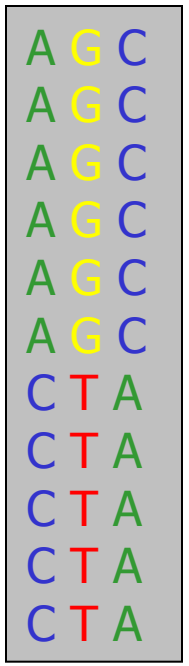
Local Mean: 3488

C/E Stat: $\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62$

C/E Stat ≤ -3.0 indicates Compression

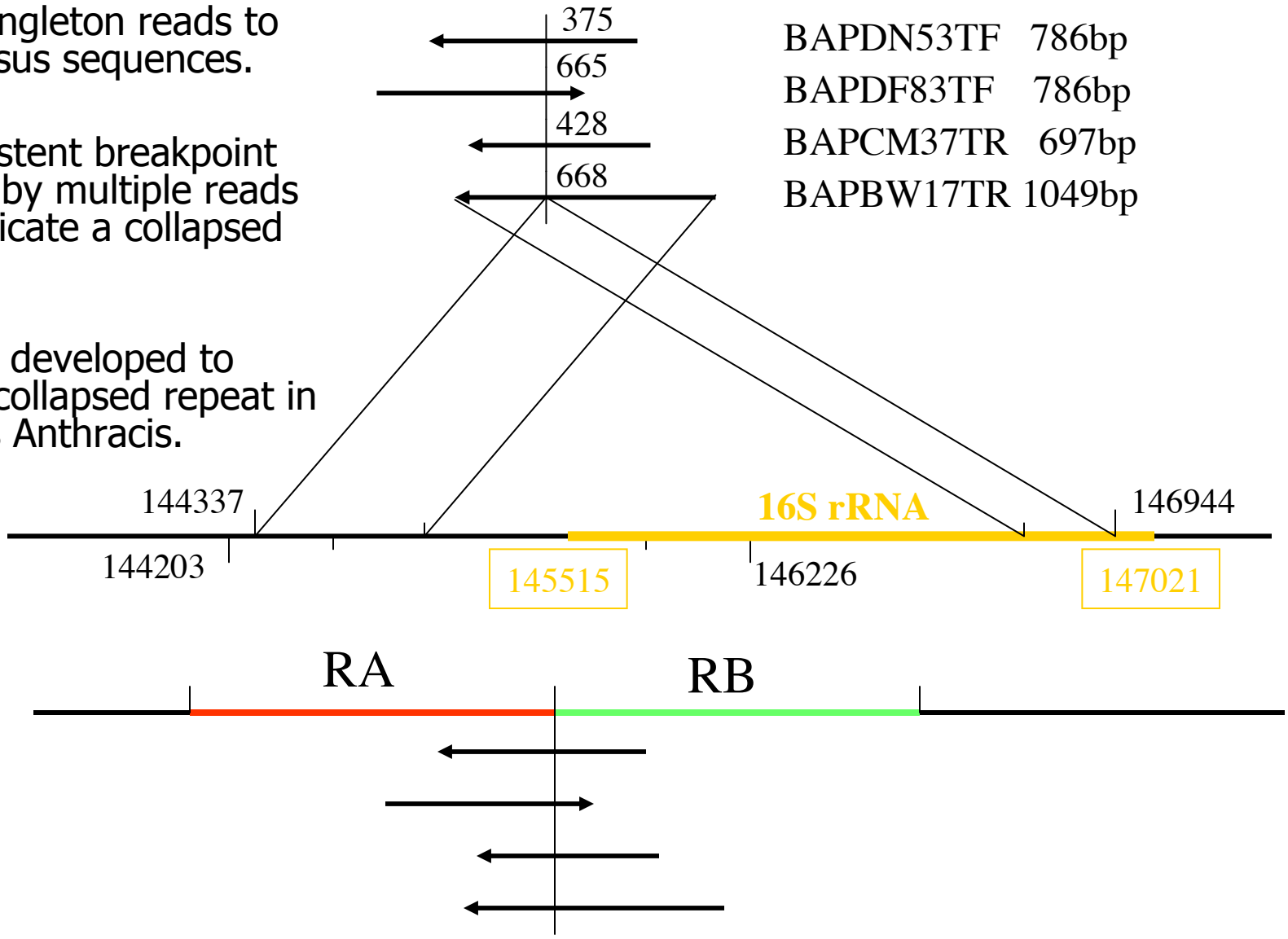
Read Alignment

- Multiple reads with same conflicting base are unlikely
  - 1x QV 30: 1/1000 base calling error
  - 2x QV 30: 1/1,000,000 base calling error
  - 3x QV 30: 1/1,000,000,000 base calling error
- Regions of correlated SNPs are likely to be assembly errors or interesting biological events
  - Highly specific metric
- AMOS Tools: analyzeSNPs & clusterSNPs
  - Locate regions with high rate of correlated SNPs
  - Parameterized thresholds:
    - Multiple positions within 100bp sliding window
    - 2+ conflicting reads
    - Cumulative QV >= 40 (1/10000 base calling error)



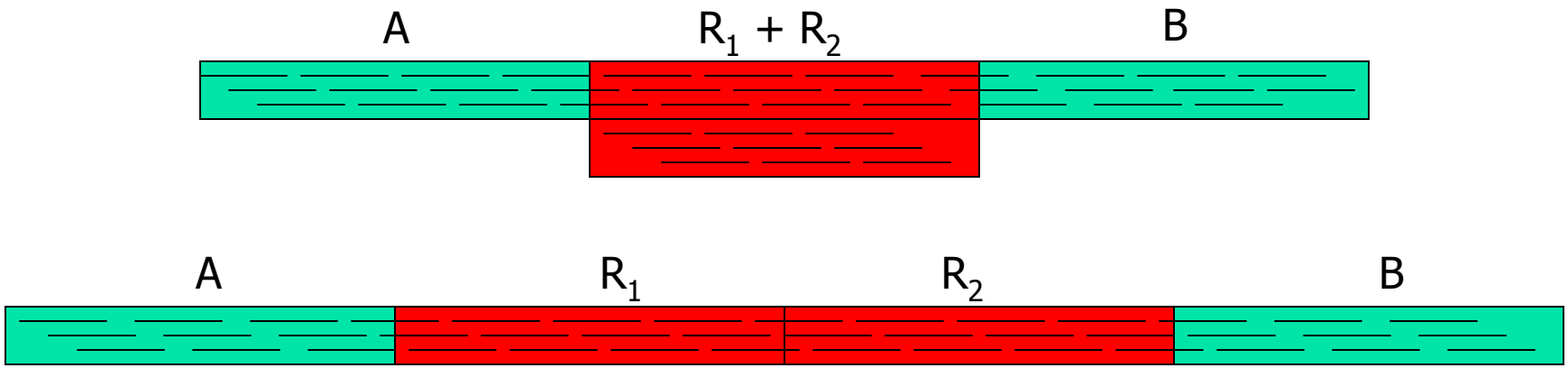
Read Breakpoints

- Align singleton reads to consensus sequences.
- A consistent breakpoint shared by multiple reads can indicate a collapsed repeat.
- Initially developed to detect collapsed repeat in Bacillus Anthracis.

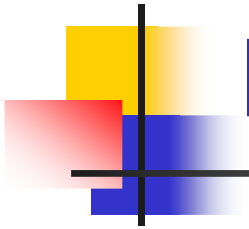


Read Coverage

- Find regions of contigs where the depth of coverage is unusually high
- Collapsed Repeat Signature
  - Can detect collapse of 100% identical repeats
- AMOS Tool: analyzeReadDepth
  - 2.5x mean coverage



Hawkeye

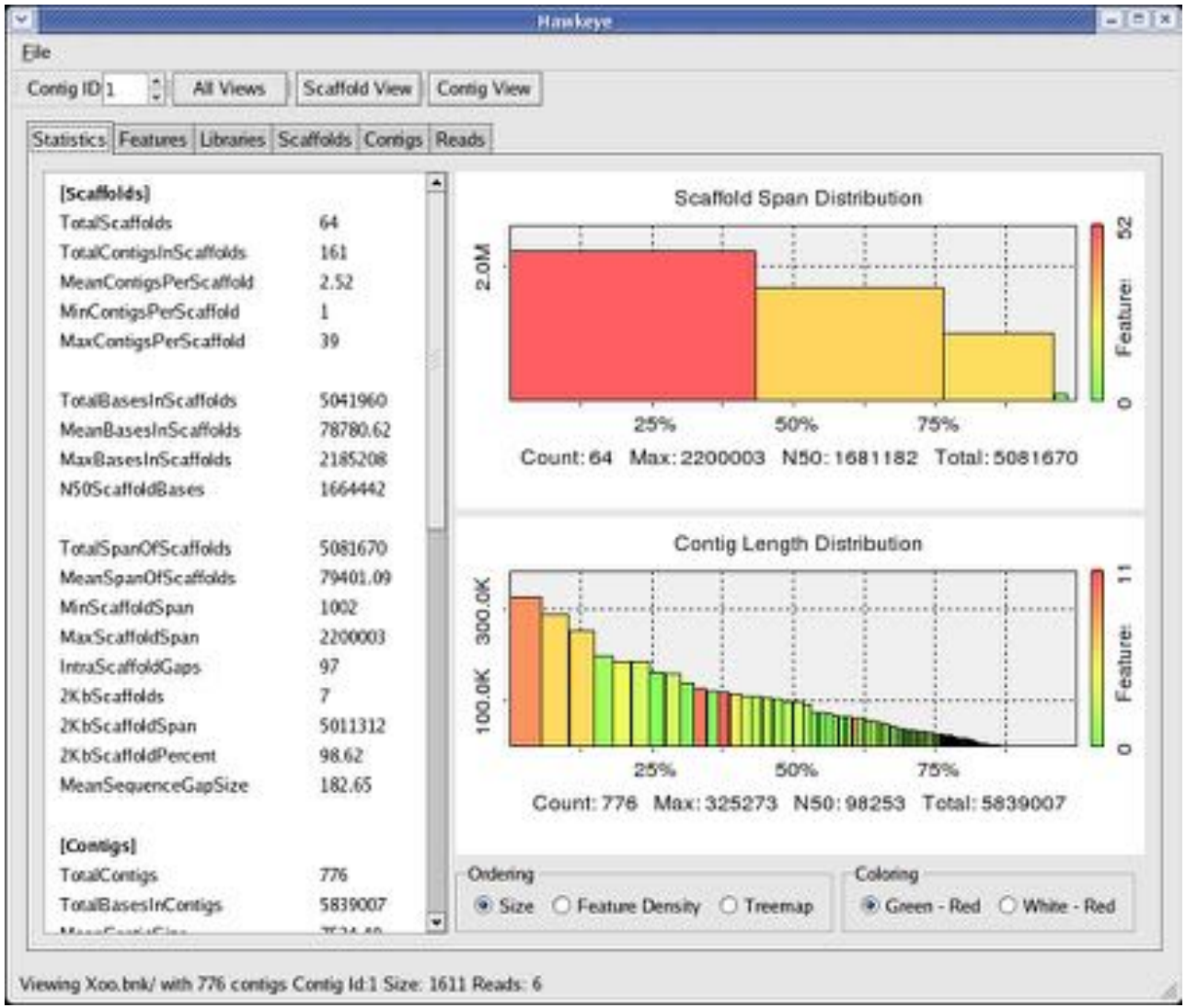


Hawkeye Goals

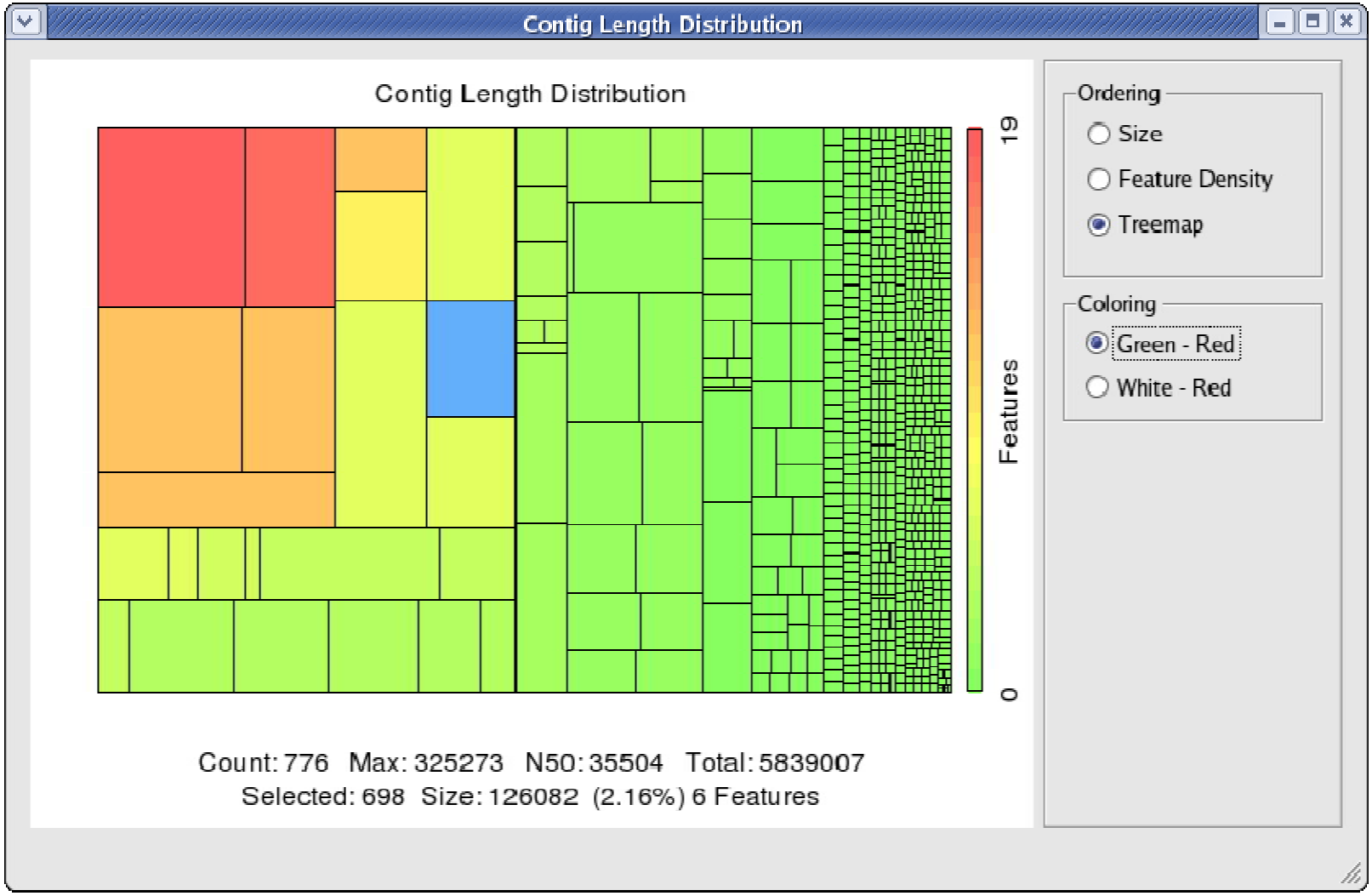
Interactively explore and analyze

- Libraries
  - Insert Sizes, Read Length, Inserts
- Scaffolds & Contigs
  - Sizes, Composition, Sequence, Multiple Alignment, SNP Barcode
- Inserts
  - Happiness, Coverage, CE Statistic
- Reads
  - Clear Range, Quality Values, Chromatograms
- Features
  - Arbitrary regions of interest
  - Including Misassembly Signatures!!!

Launch Pad

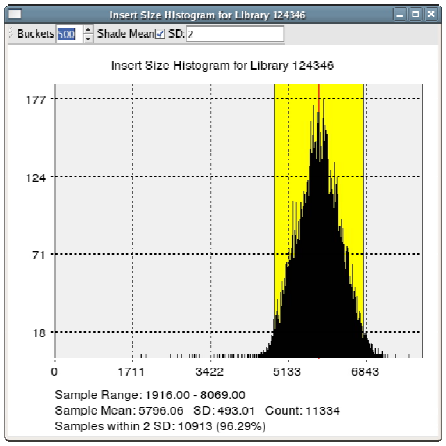

Hawkeye
File
Contig ID 1
All Views
Scaffold View
Contig View
Statistics
Features
Libraries
Scaffolds
Contigs
Reads
[Scaffolds]
TotalScaffolds 64
TotalContigsInScaffolds 161
MeanContigsPerScaffold 2.52
MinContigsPerScaffold 1
MaxContigsPerScaffold 39
TotalBasesInScaffolds 5041960
MeanBasesInScaffolds 78780.62
MaxBasesInScaffolds 2185208
N50ScaffoldBases 1664442
TotalSpanOfScaffolds 5081670
MeanSpanOfScaffolds 79401.09
MinScaffoldSpan 1002
MaxScaffoldSpan 2200003
IntraScaffoldGaps 97
2KbScaffolds 7
2KbScaffoldSpan 5011312
2KbScaffoldPercent 98.62
MeanSequenceGapSize 182.65
[Contigs]
TotalContigs 776
TotalBasesInContigs 5839007
Scaffold Span Distribution
2.0M
25%
50%
75%
Count: 64 Max: 2200003 N50: 1681182 Total: 5081670
Feature: 52 0
Contig Length Distribution
300.0K
100.0K
25%
50%
75%
Count: 776 Max: 325273 N50: 98253 Total: 5839007
Feature: 11 0
Ordering
Size
Feature Density
Treemap
Coloring
Green - Red
White - Red
Viewing Xoo.bnk/ with 776 contigs Contig Id:1 Size: 1611 Reads: 6

Contig Length Distribution

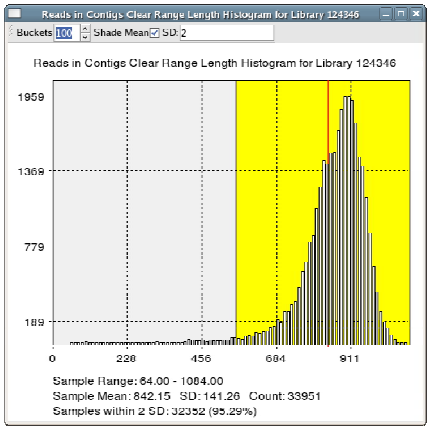


Histograms & Statistics

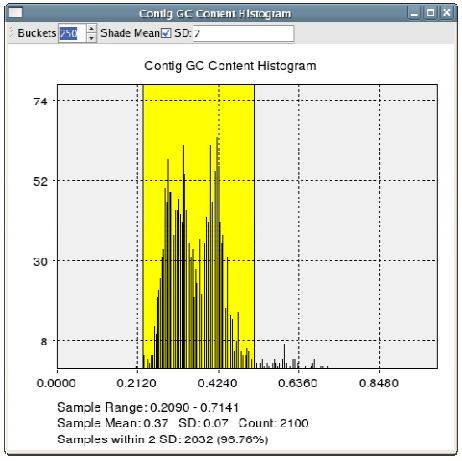
Insert Size



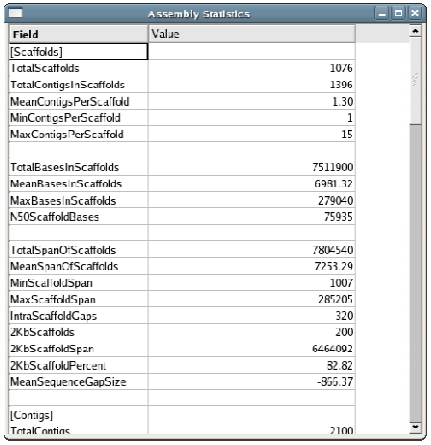
Read Length



GC Content



Overall Statistics



- Bird's eye view of data and assembly quality

Scaffold View

a. Statistical Plots

b. Scaffold

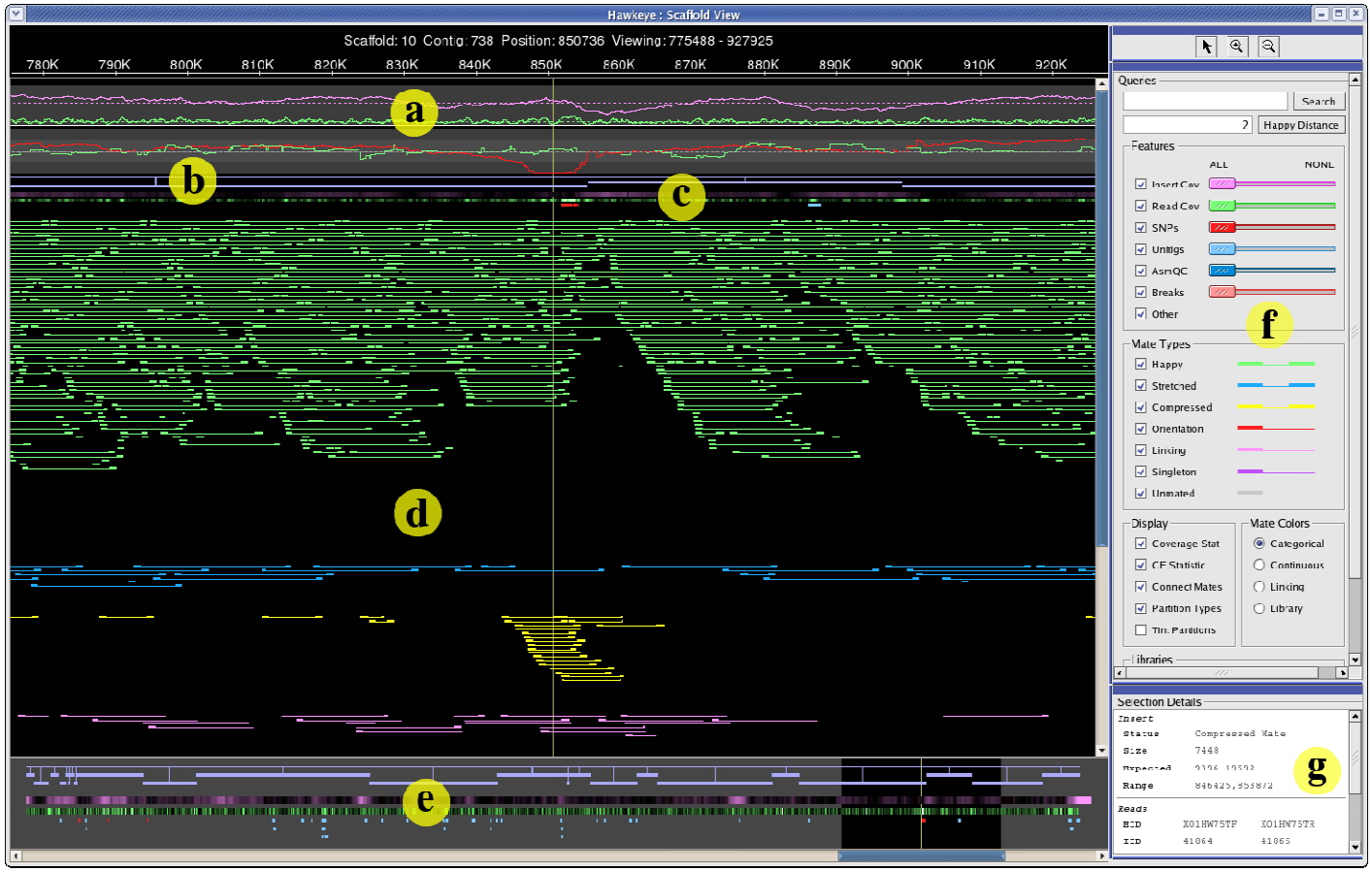
c. Features

d. Inserts

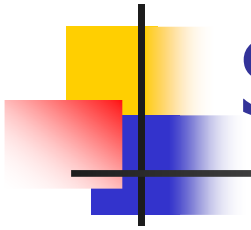
e. Overview

f. Control Panel

g. Details



Standard Feature Types



[B] Breakpoint
Alignment ends at this position

[C] Coverage
Location of unusual mate coverage (asmQC)

[S] SNPs
Location of Correlated SNPs

[U] Unitig
Used to report location of surrogate unitigs in CA assemblies

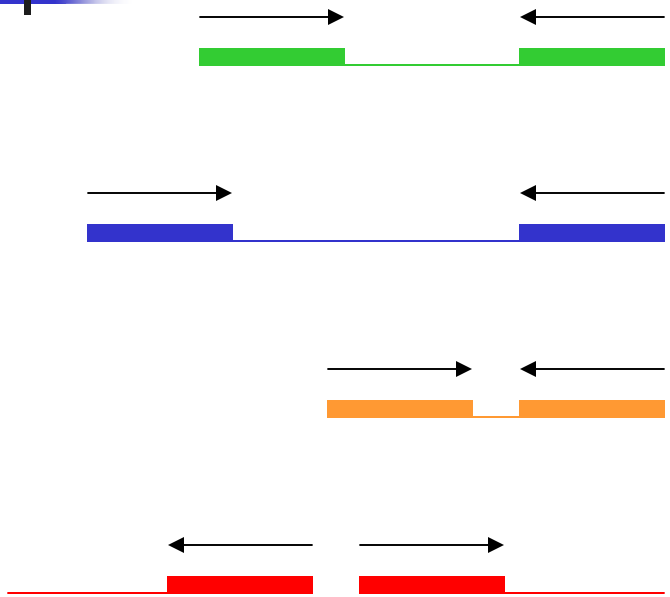
[X] Other
All other Features

Loading Features:
$ loadFeatures bankname featfile

Featfile format:
Contigid type end5 end3 comment

Insert Happiness

Both mates present



Happy

- Oriented Correctly &&
- |Insert Size – Library.mean| <= Happy-Distance * Library.sd

Stretched

- Oriented Correctly &&
- Insert Size > Library.mean + Happy-Distance * Library.sd

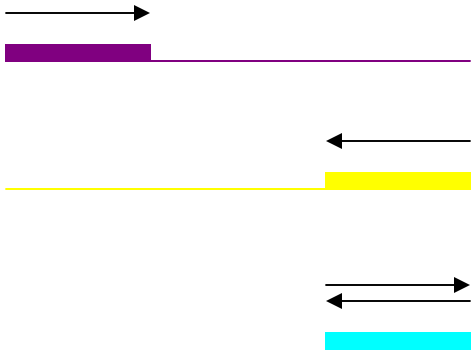
Compressed

- Oriented Correctly &&
- Insert Size < Library.mean - Happy-Distance * Library.sd

Misoriented

- Same or Outies

Only 1 read present



Linking

- Read's mate is in some other scaffold

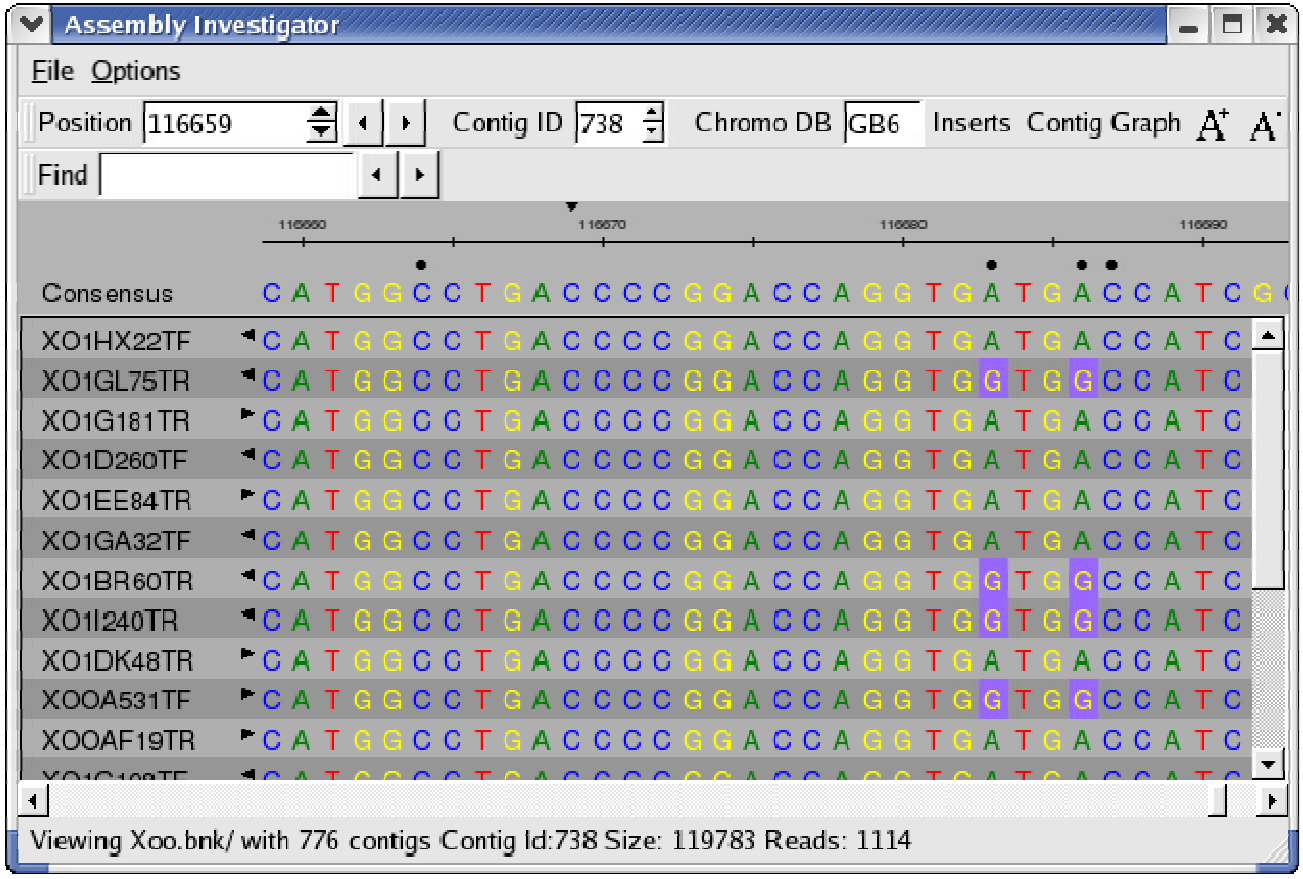
Singleton

- Read's mate is a singleton

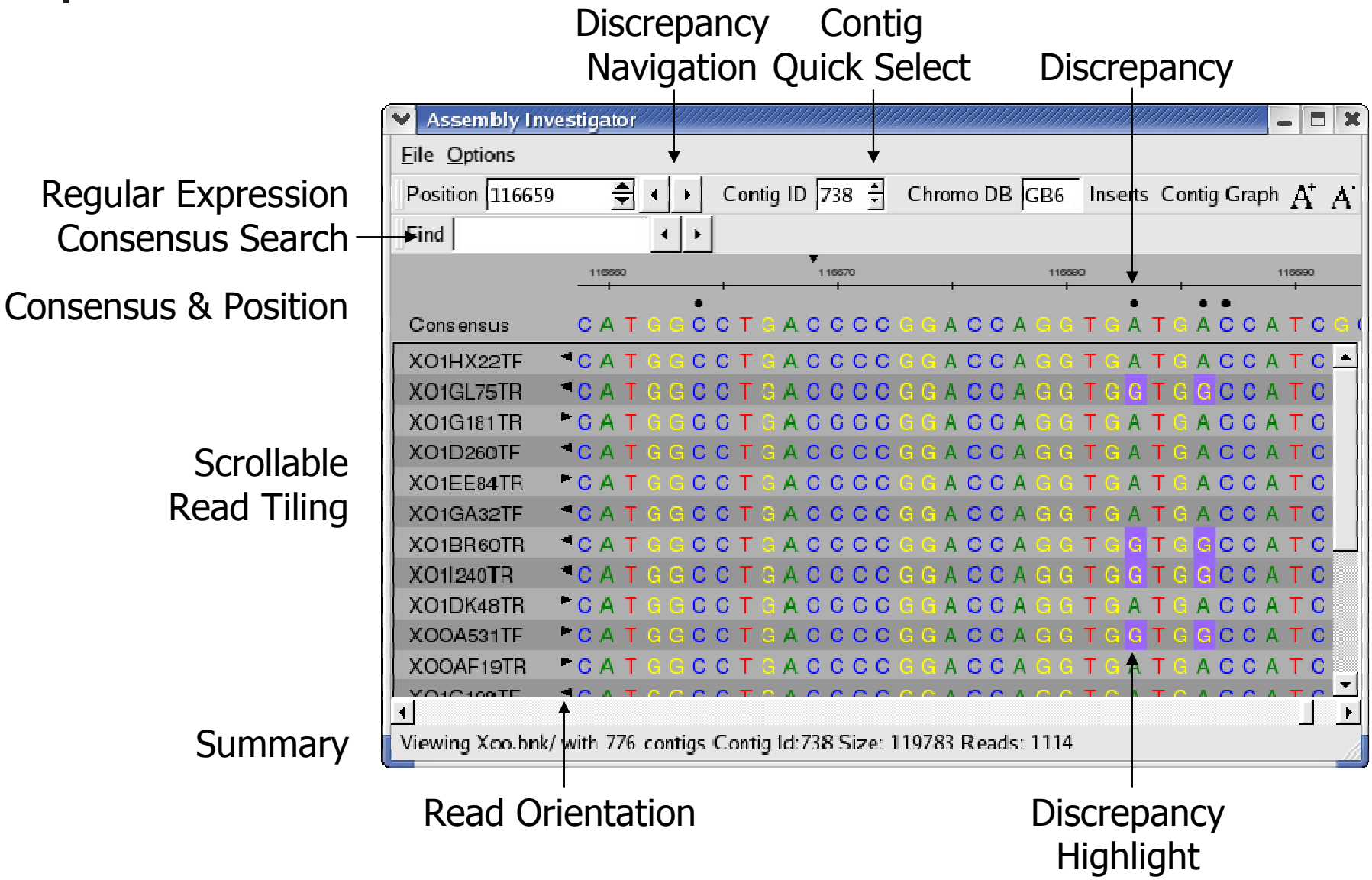
Unmated

- No mate was provided for read

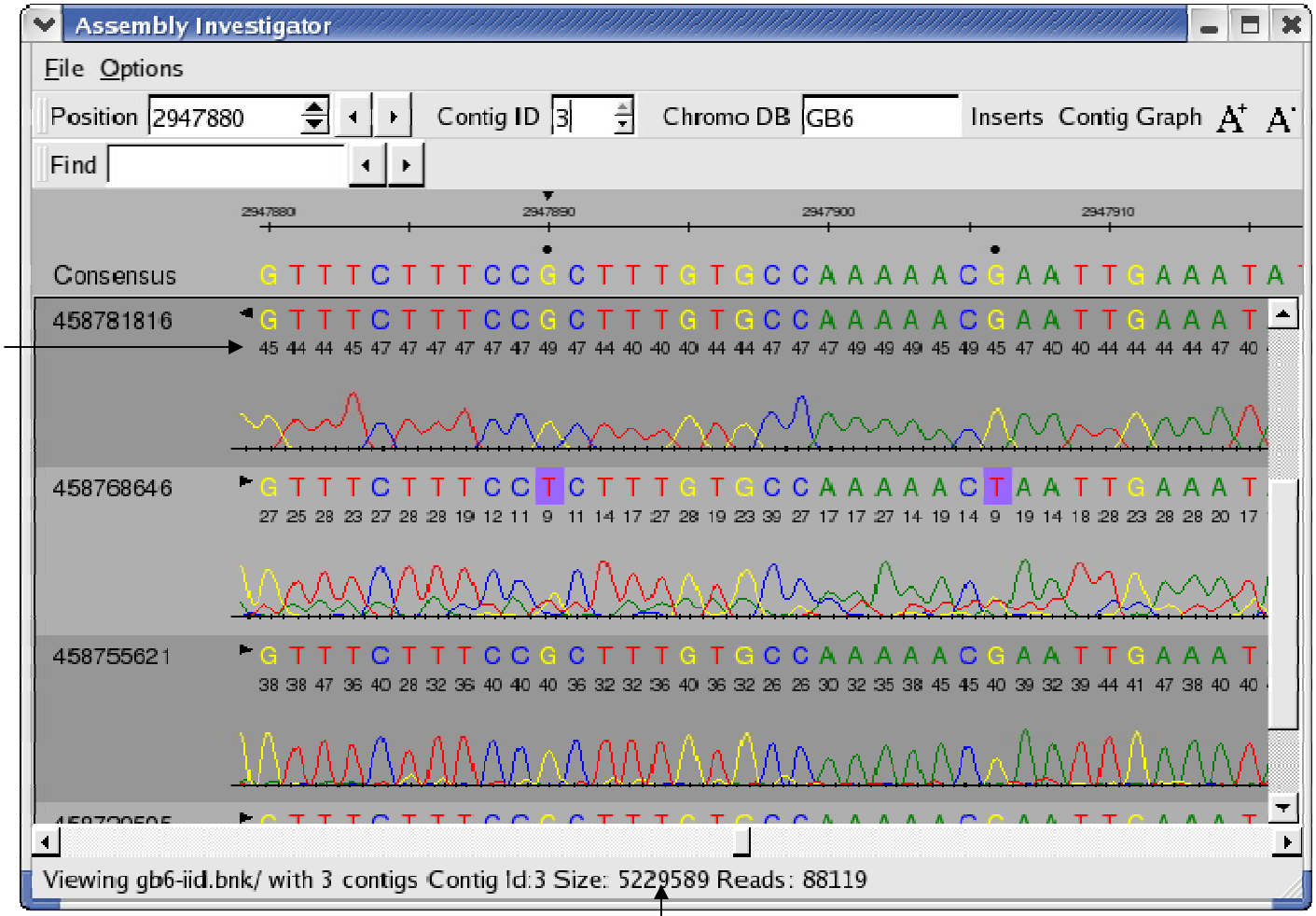
Contig View


Assembly Investigator
File Options
Position 116659
Contig ID 738
Chromo DB GB6
Inserts
Contig Graph
Find
Consensus
XO1HX22TF
XO1GL75TR
XO1G181TR
XO1D260TF
XO1EE84TR
XO1GA32TF
XO1BR60TR
XO1I240TR
XO1DK48TR
XOOA531TF
XOOAF19TR
Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114

Contig View


Discrepancy Navigation
Contig Quick Select
Discrepancy
Regular Expression Consensus Search
Consensus & Position
Scrollable Read Tiling
Summary
Read Orientation
Discrepancy Highlight
Assembly Investigator
File Options
Position 116659
Contig ID 738
Chromo DB GB6
Inserts
Contig Graph
Find
Consensus
XO1HX22TF
XO1GL75TR
XO1G181TR
XO1D260TF
XO1EE84TR
XO1GA32TF
XO1BR60TR
XO1I240TR
XO1DK48TR
XOOA531TF
XOOAF19TR
Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114

Contig View Expanded



Quality Values

Normalized Chromatogram

No size restrictions

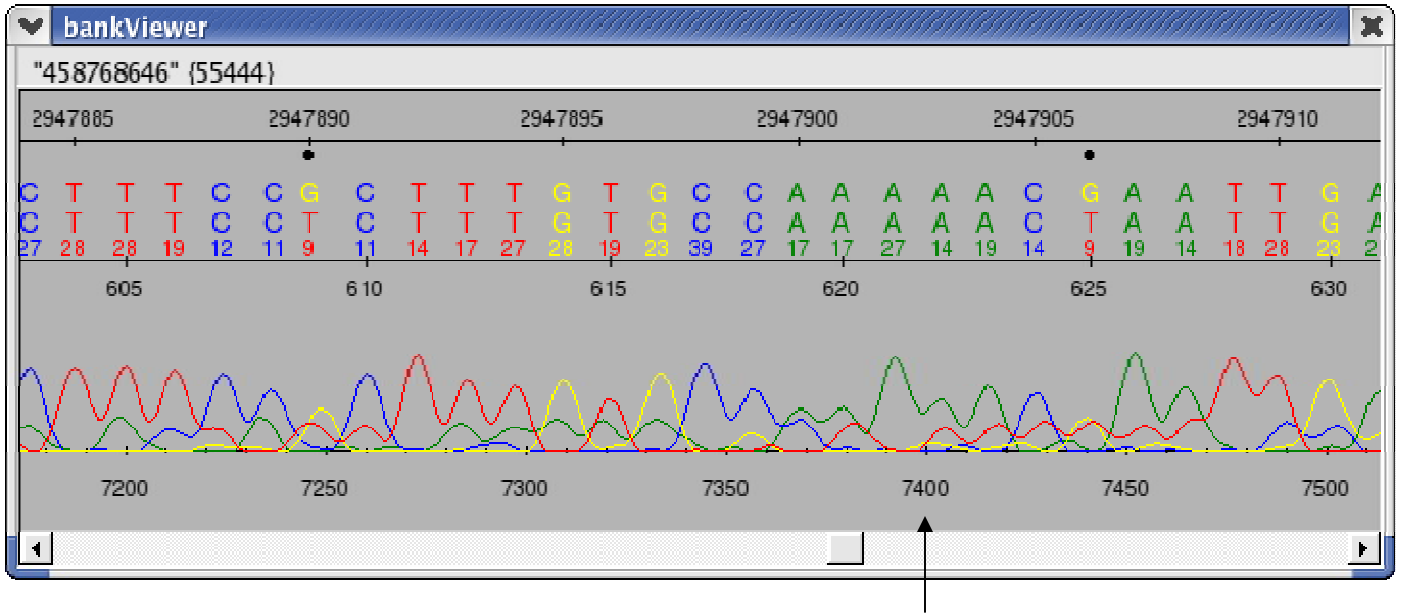
Chromatogram View

Read EID, IID

Consensus

Read

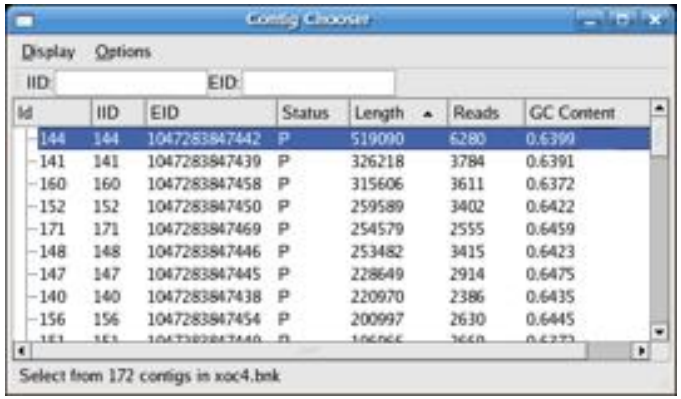
Raw Chromatogram



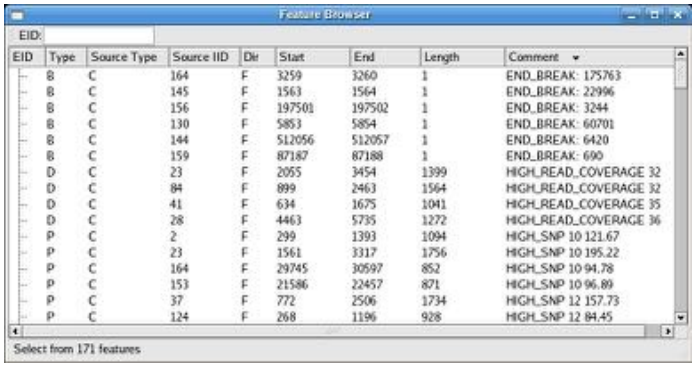
Chromatograms are loaded from specified directories, or on demand from Trace Archive.

Assembly Reports

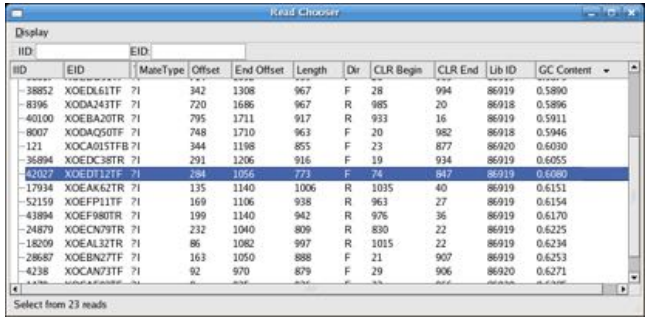
Contigs



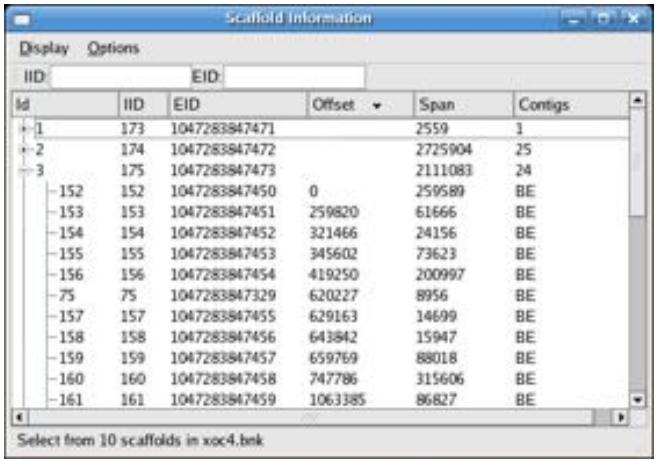
Features



Reads



Scaffolds

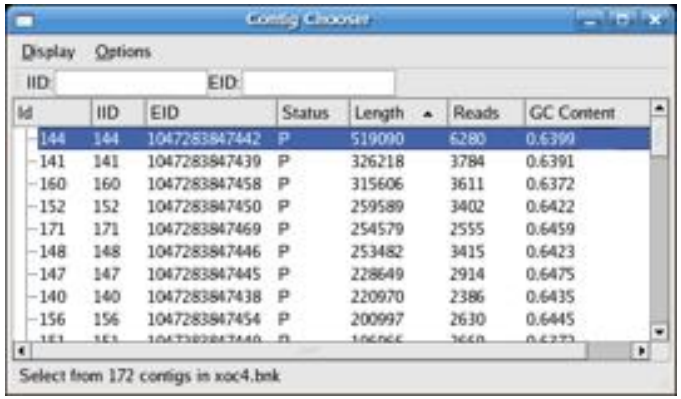


- Full Integration: “Double click takes you there”

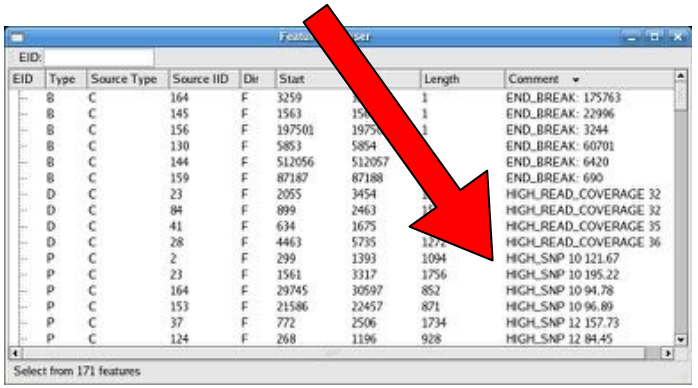
Assembly Reports

Misassembly Walkthough:
Correlated SNPs

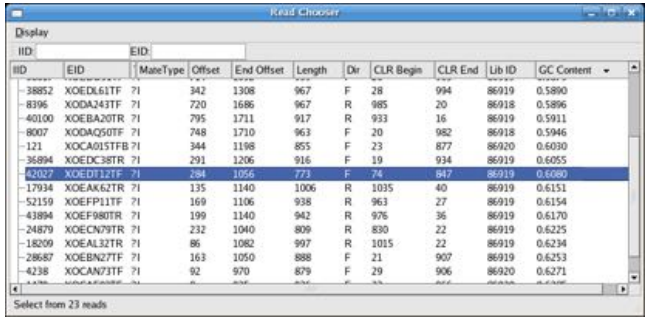
Contigs



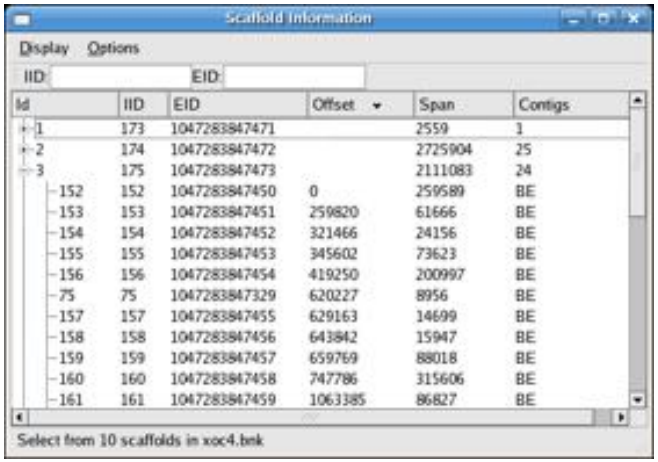
Features



Reads



Scaffolds

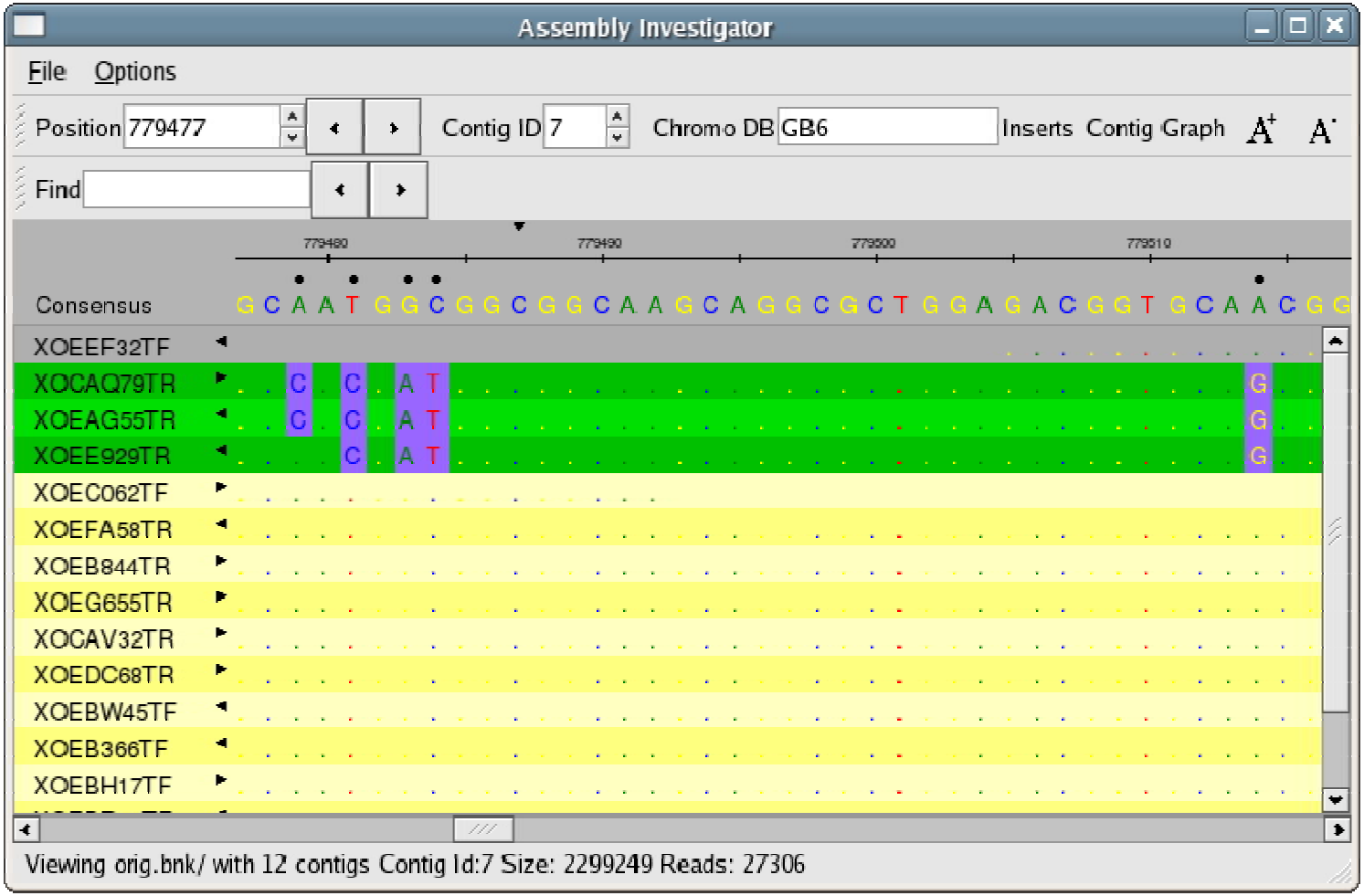


- Full Integration: “Double click takes you there”

SNP View

SNP Sorted Reads

Polymorphism View

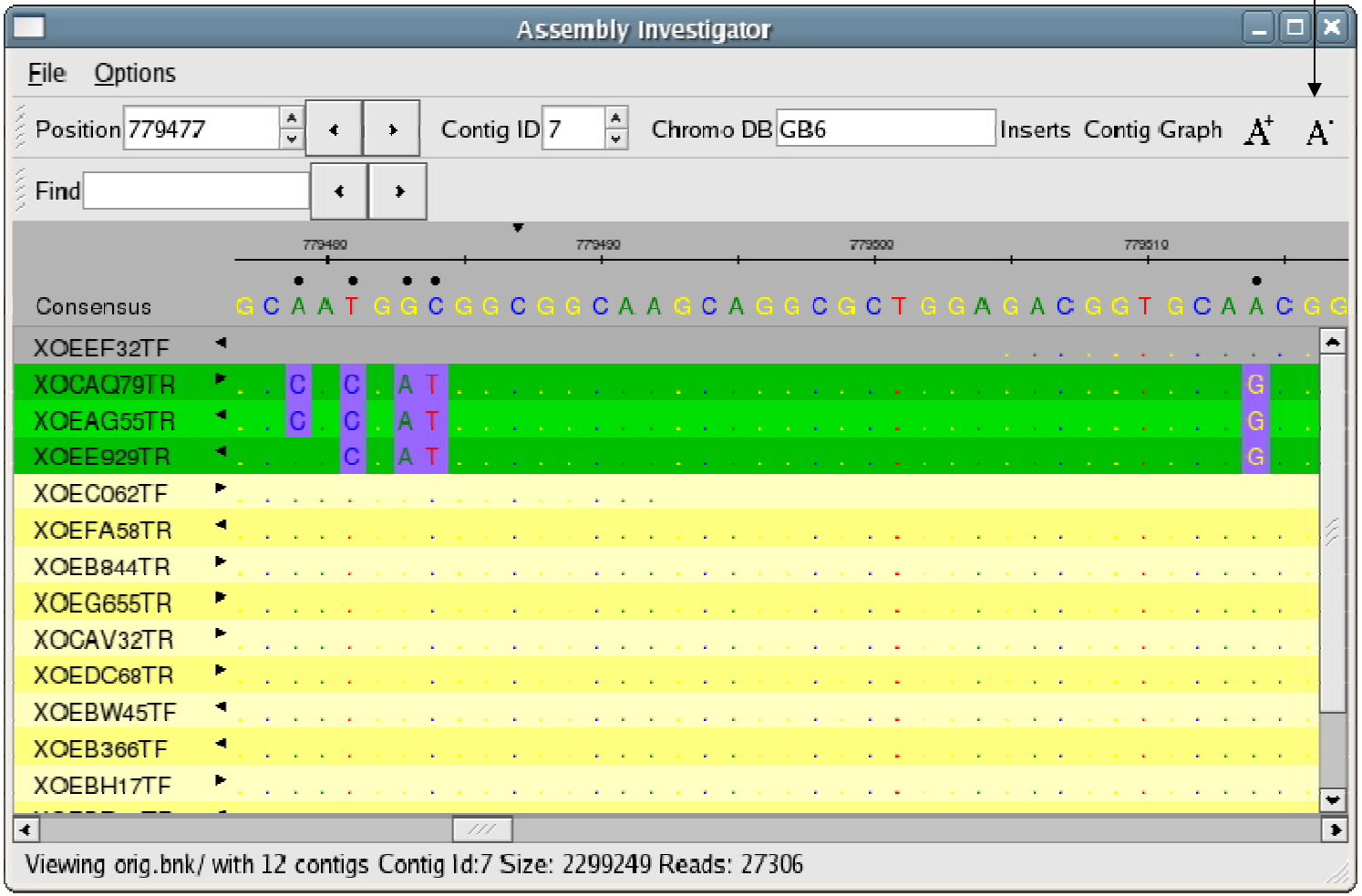


SNP View

Zoom Out

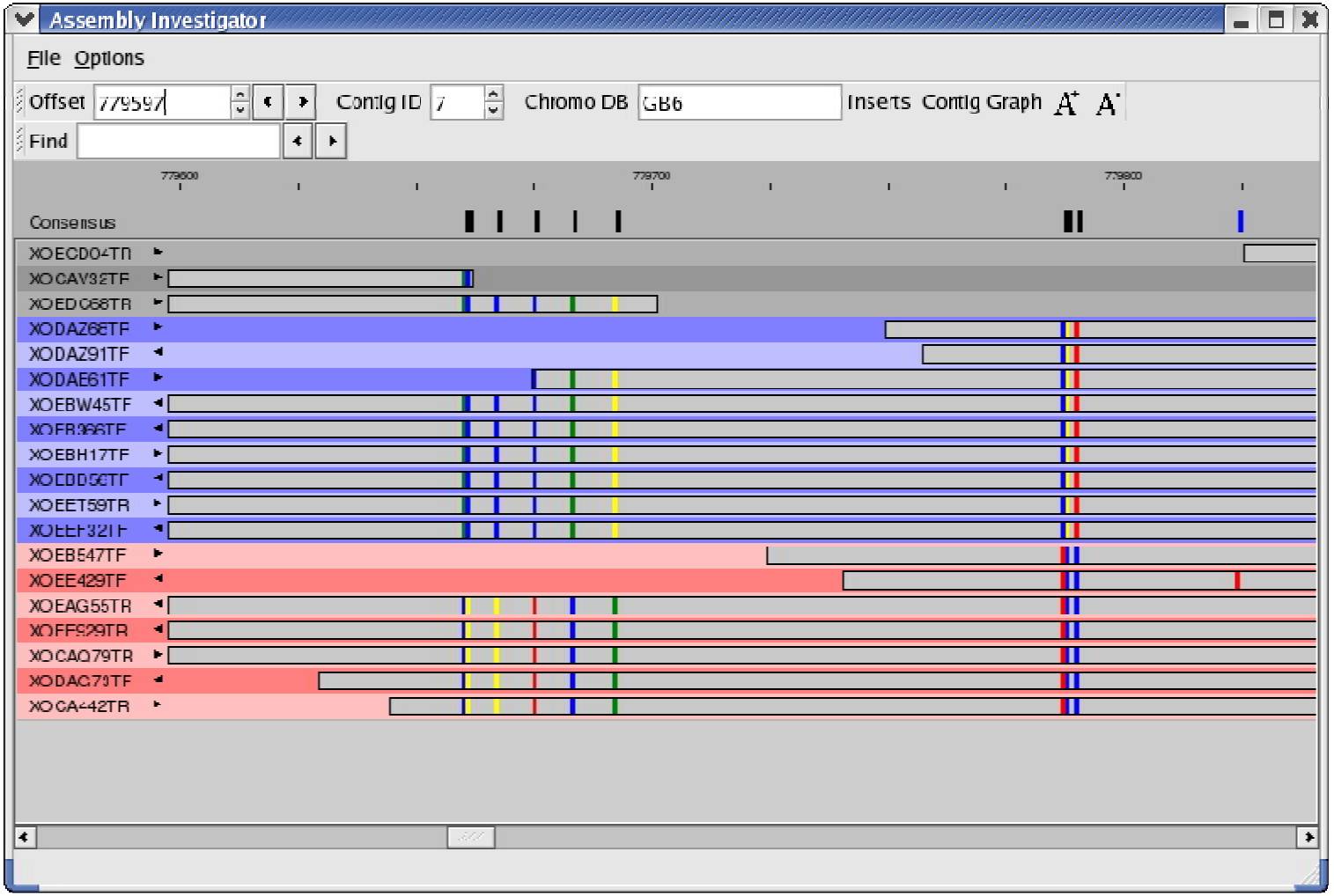
SNP Sorted Reads

Polymorphism View



SNP Barcode

SNP Sorted Reads



Colored Rectangle indicate the positions and composition of the SNPs

Scaffold View

Coverage

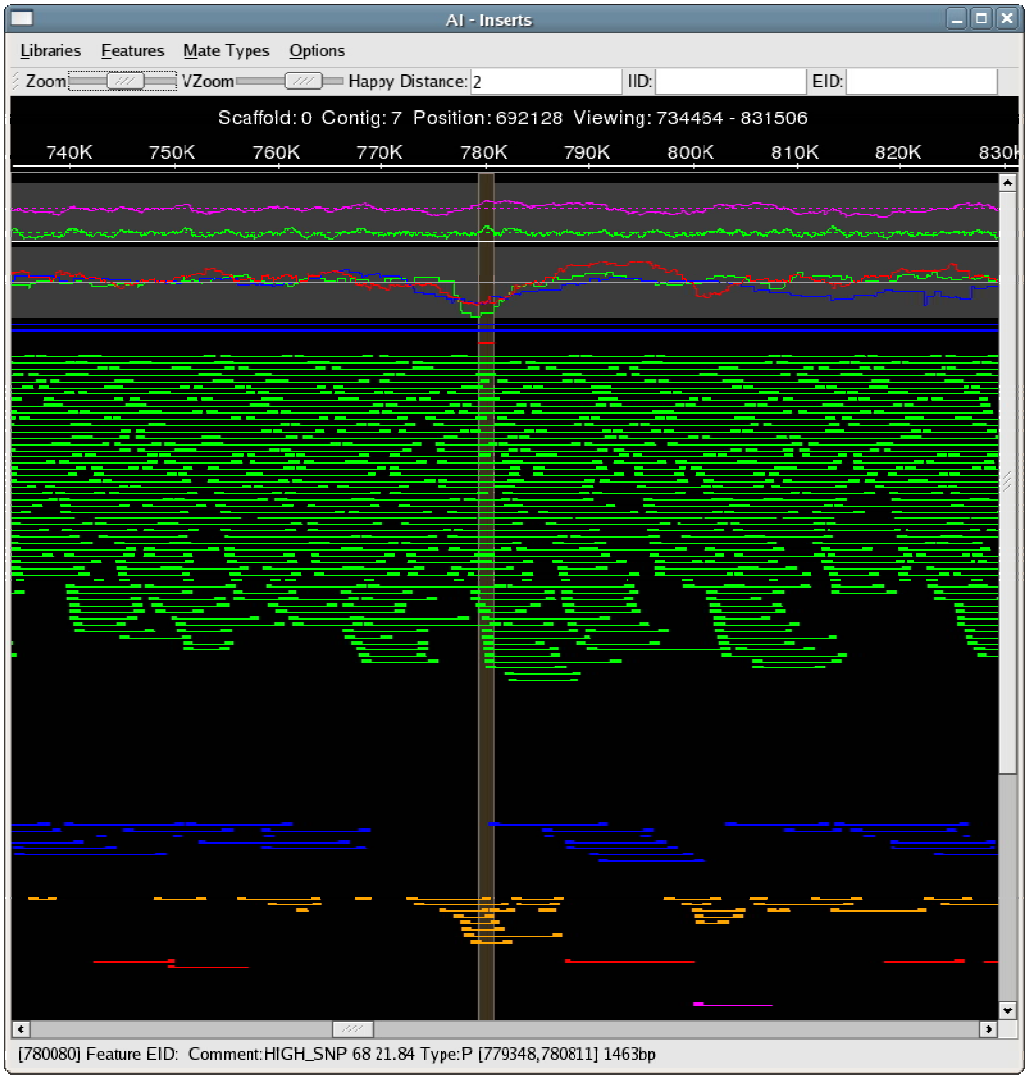
CE Statistic

Happy

Stretched

Compressed

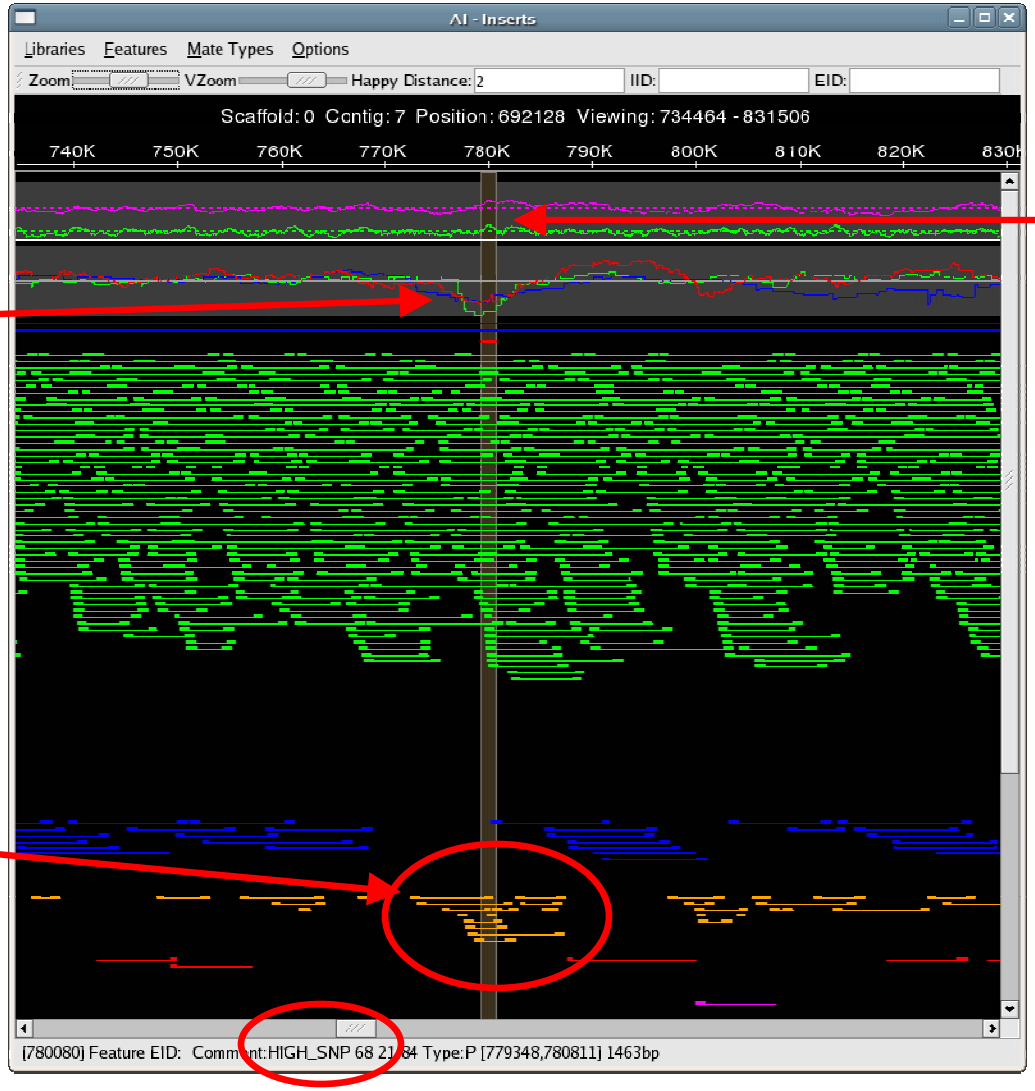
Misoriented



SNP Feature

Linking

Collapsed Repeat



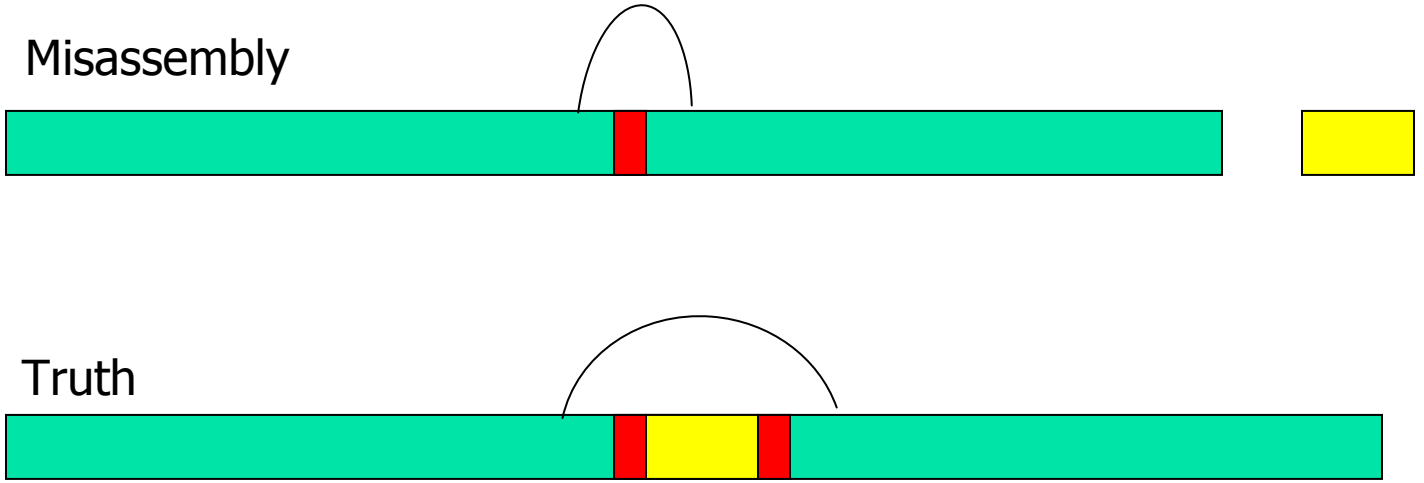
-5.5 CE Dip

Read Coverage Spike

Compressed Mates Cluster

68 Correlated SNPs

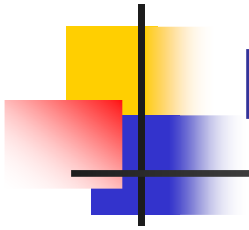
Confirmed Misassembly



Collapsed repeat

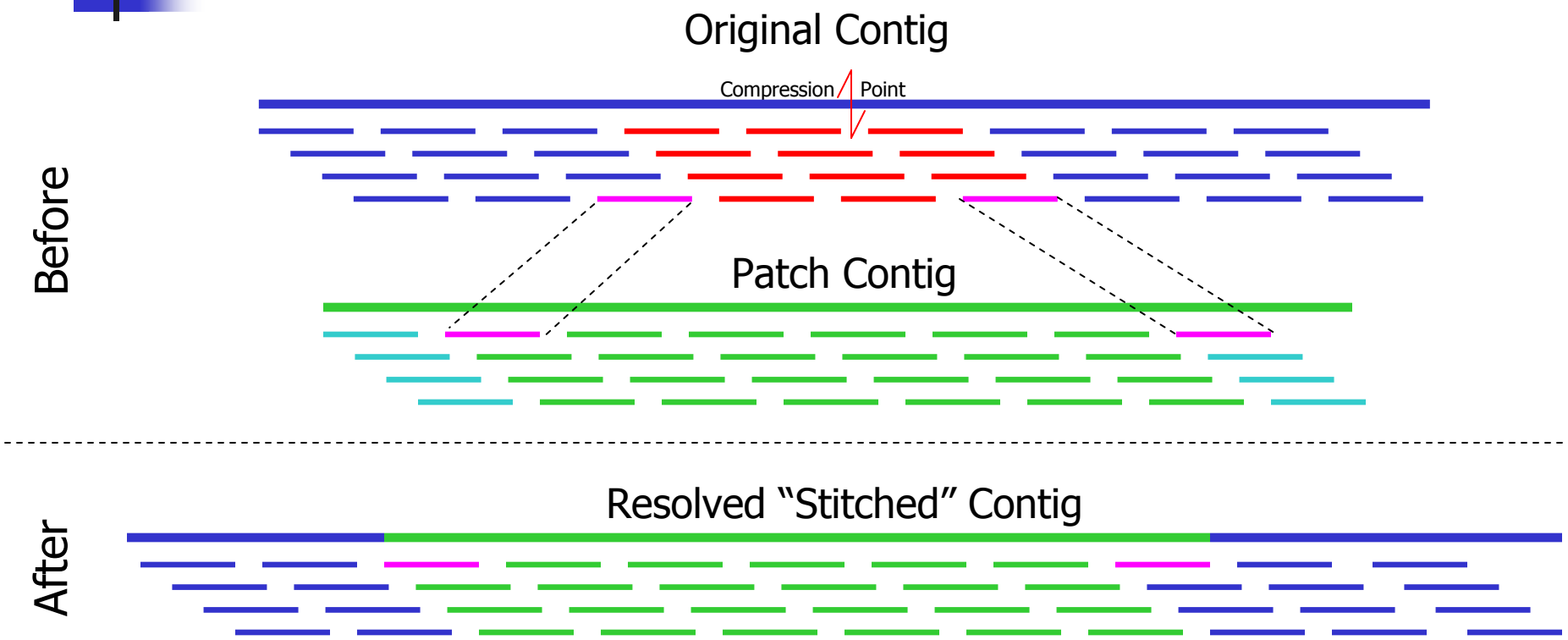
- Compressed mates (-5.5 CE Stat)
- Correlated SNPs (68 Positions within 1400bp)
- Spike in Read Coverage

Fixing collapsed repeats with AMOS



1. Select reads and mates in region of collapse.
   - AMOS: findMissingMates, select-reads
2. Reassemble those reads with stricter parameters.
   - AMOS: minimus
3. Inspect new assembly to ensure misassembly was corrected.
   - AMOS: amosvalidate, Hawkeye
4. Patch the collapsed region of the original assembly with corrected version.
   - AMOS: stitchContigs

stitchContigs



- Replace the reads between the stitch reads in the original contig with corresponding region in the patch contig.
- Can also close gaps or fix contig ends

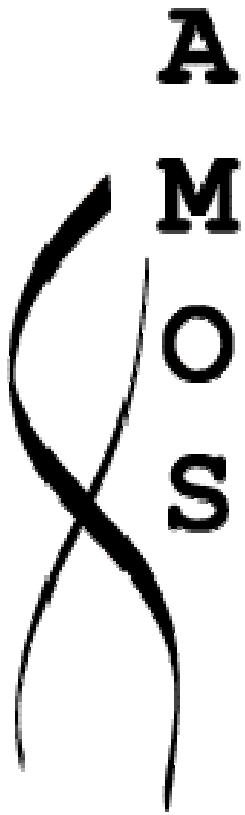
Potential Assembly Problems

- Library Construction
  - Insert Size Histogram
- Contaminate Sequences:
  - GC Content Histogram
- Read Trimming:
  - Missing Mates
  - SNP Barcode
- Coverage Levels
  - Coverage Plot
- A-stat problems / Degenerate Contigs
  - Summary Statistics
  - Scaffold View
- Local Mis-assembly
  - Scaffold, Contig Views, Features

Current Research

- Misassembly signature detection
  - Singleton / Missing mate analysis
  - Integrated & Dynamic Thresholds of detection
- Automated assembly improvement
  - Automatic contig patching
  - Automatic repeat separation
  - Automatic parameter tuning
- Exotic Assembly
  - Multiple haplotypes
  - Metagenomic assembly
  - 454 & Sanger Sequencing Hybrids

More Information



- Contact AMOS
  - http://amos.sourceforge.net
  - amos-help [at] lists.sourceforge.net
- Hawkeye Webpage:
  - http://amos.sourceforge.net/hawkeye
- Acknowledgements
  - Adam Phillippy
  - Ben Shneiderman
  - Steven Salzberg
  - Mihai Pop