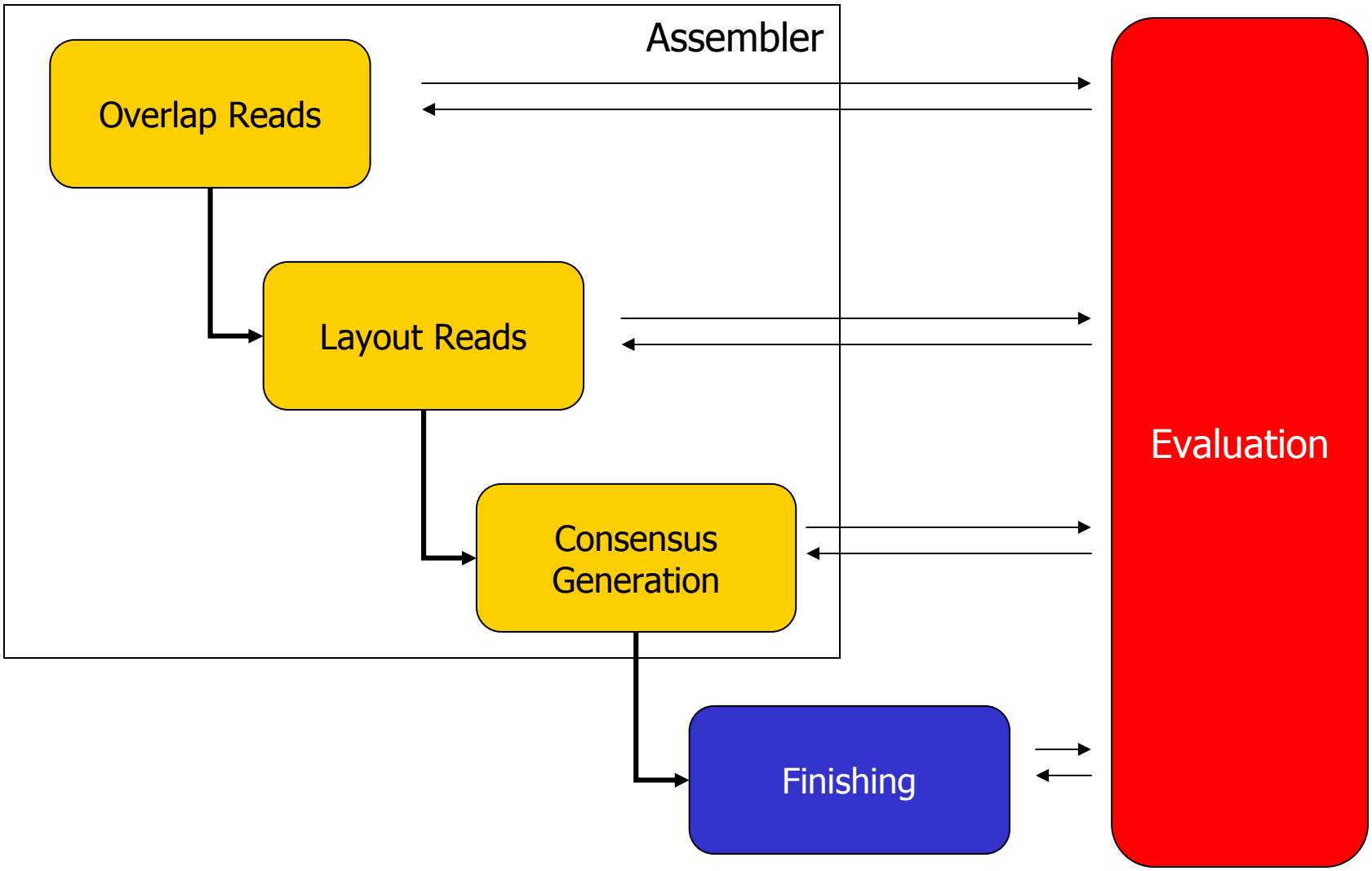
Interactive visual analytic tools for genome assemblies

Michael Schatz

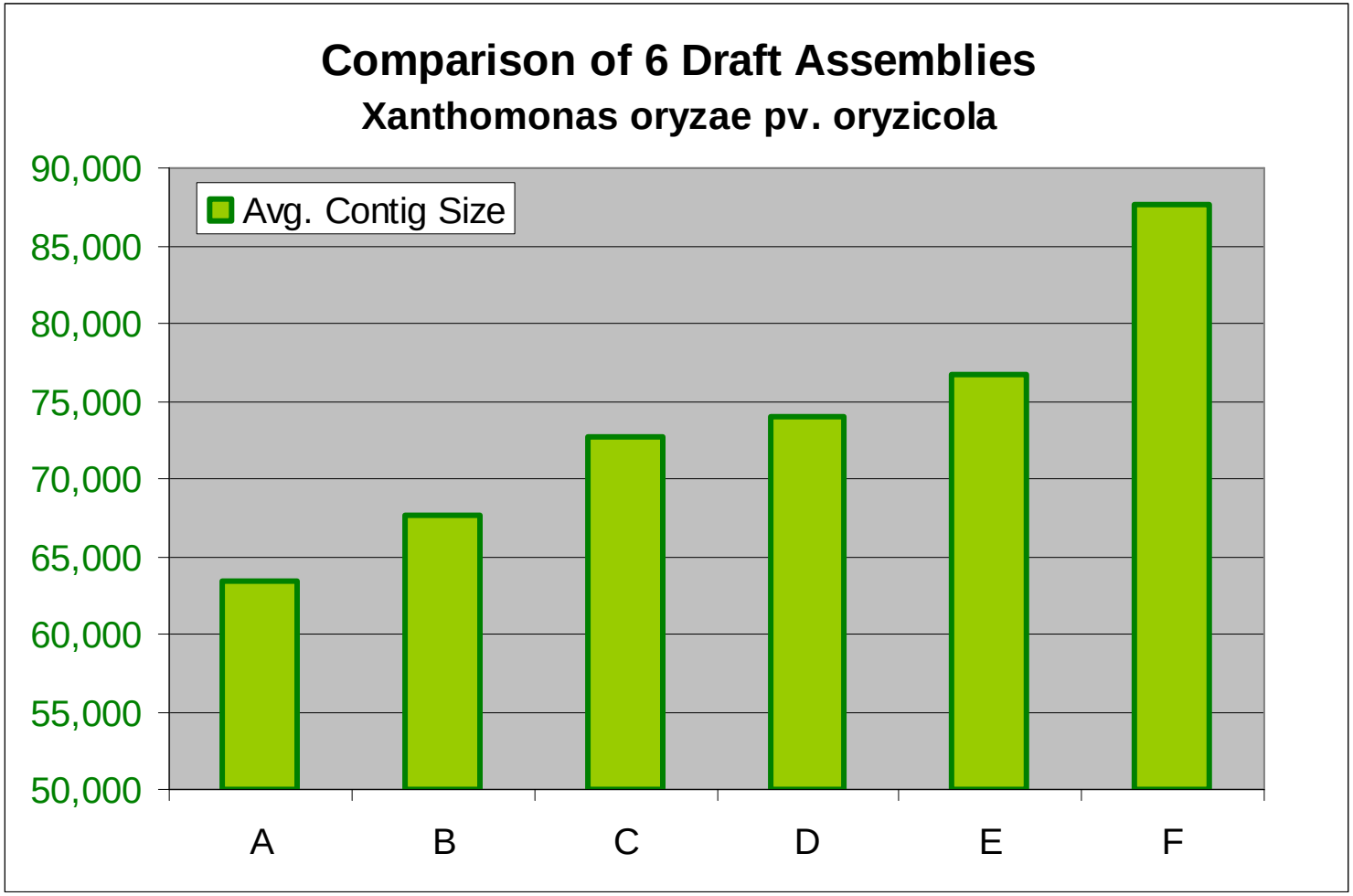
Center for Bioinformatics and Computational Biology
University of Maryland

October 29, 2006
9th Annual Computational Genomics Conference

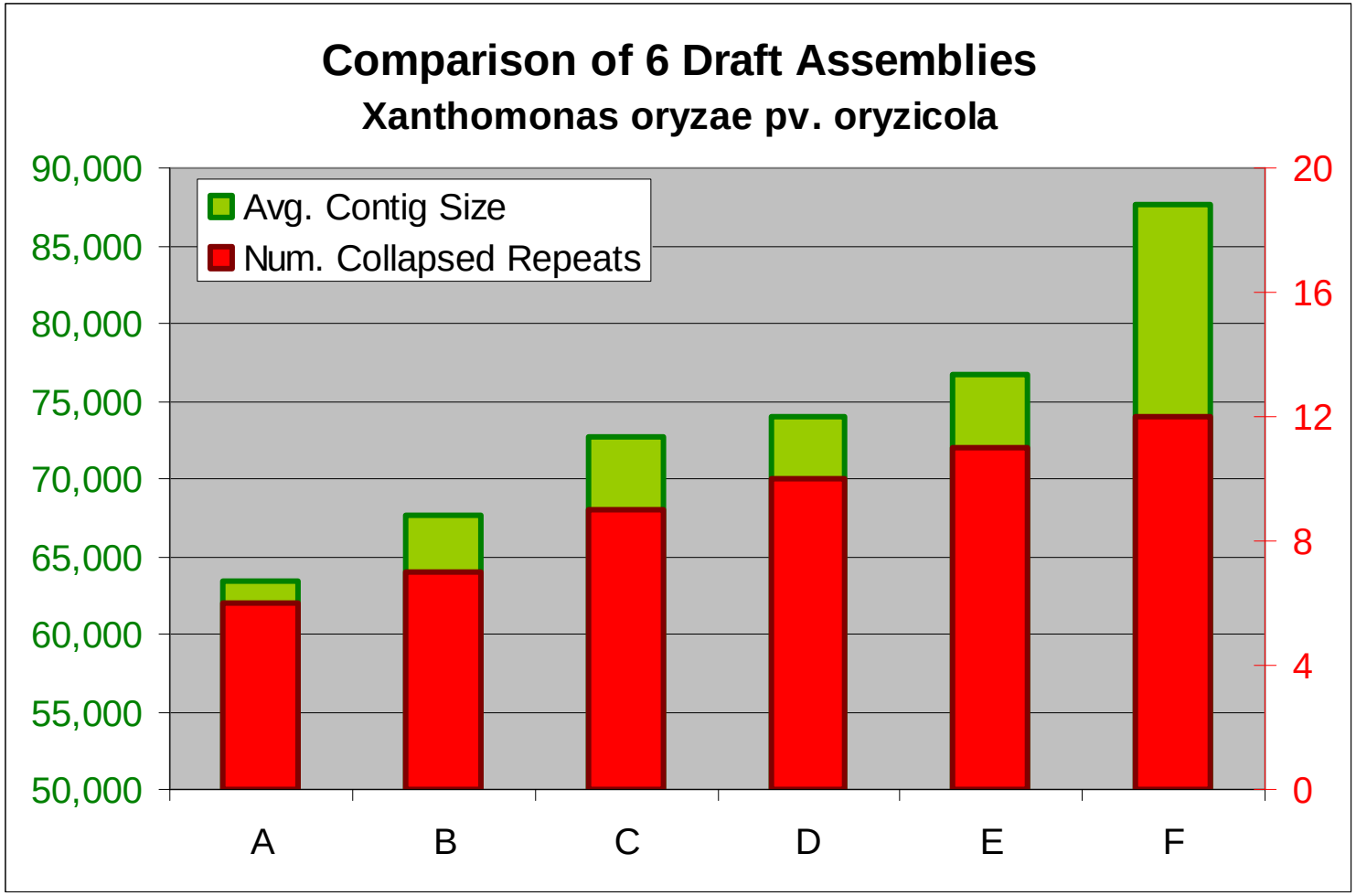
Genome Assembly



Assembly Evaluation

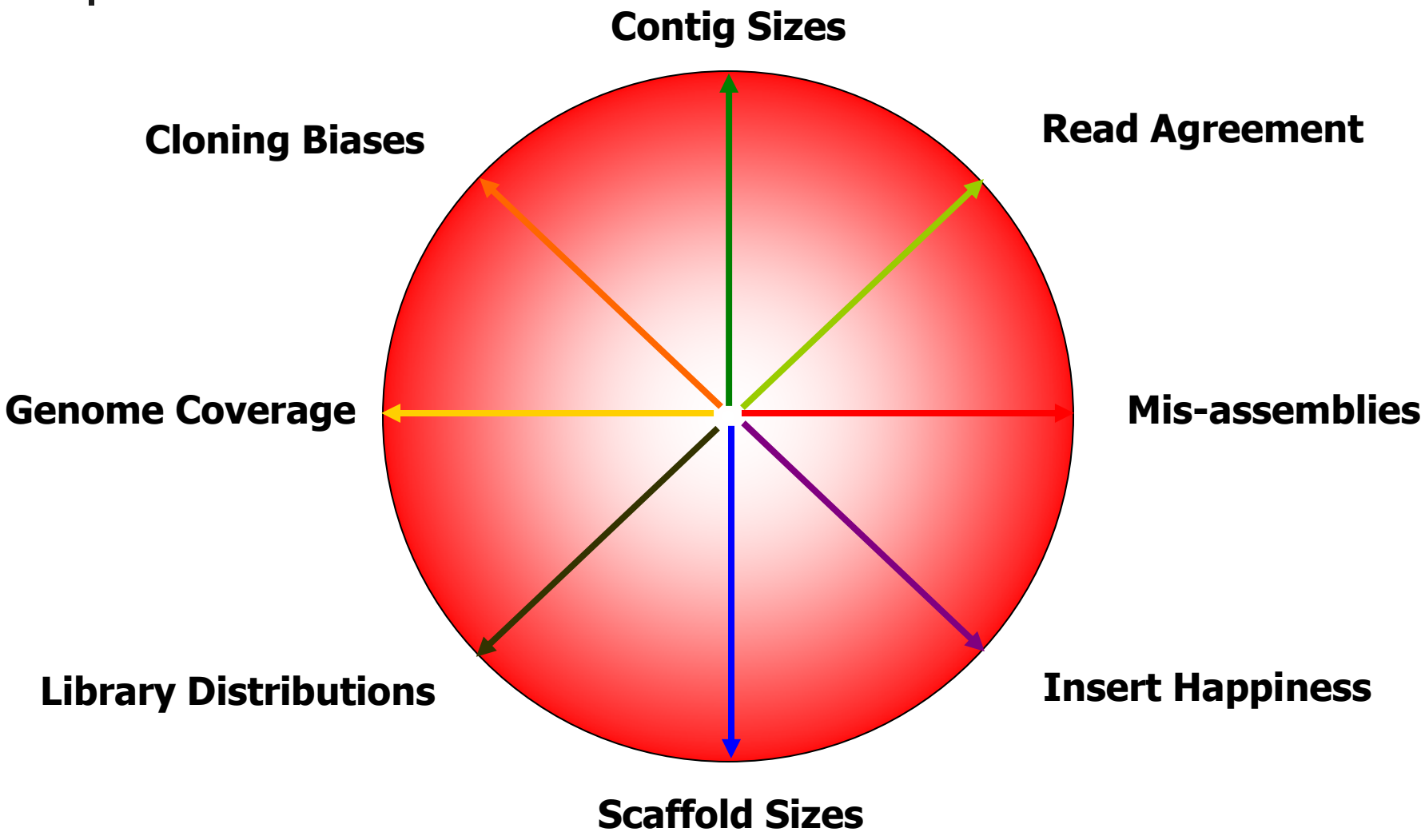


Assembly Evaluation



* Bigger is not always better

Dimensions of Assembly Quality



Is this scaffold correct? **Is this contig correct?** **Is this base correct?**

Hawkeye Goals

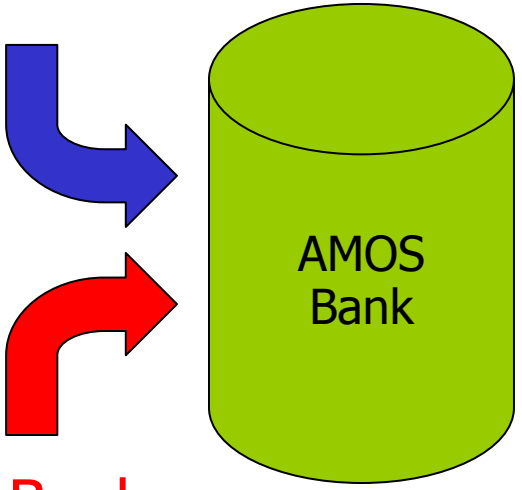
Interactively explore and analyze

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

AMOS Validation Pipeline

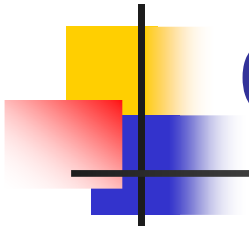
- Computationally scan for mis-assembly signatures in an assembly.

- amosvalidate
  1. Load Assembly Data into Bank
  2. Analyze Mate Pairs & Libraries
  3. Analyze Depth of Coverage
  4. Analyze Normalized K-mers
  5. Analyze Read Alignments
  6. Analyze Read Breakpoints
  7. Load Mis-assembly Signatures into Bank



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

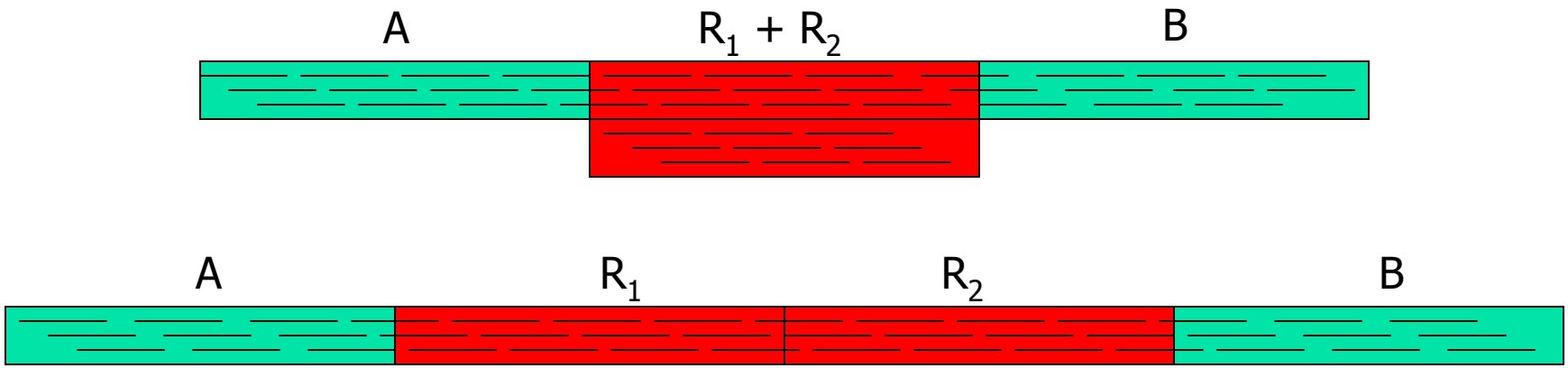


C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.
- Does the distribution of 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
  - > +3 indicates significant expansion
  - < -3 indicates significant compression
- Introduced by Dr. Jim Yorke's group at UMD

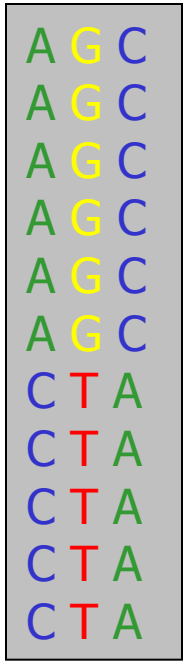
Read Coverage

- 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



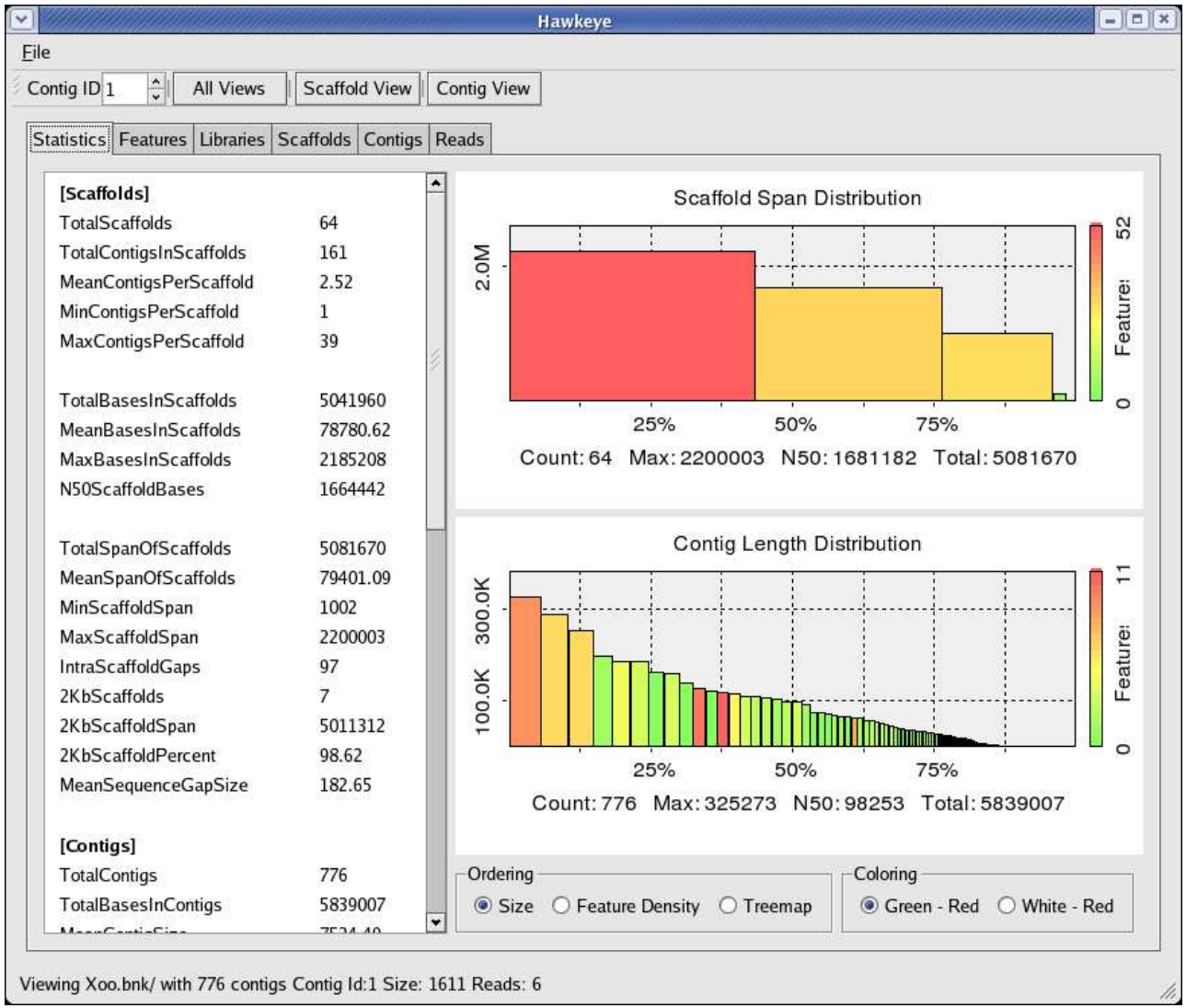
Read Alignment

- 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 for nearly identical repeats
- 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)



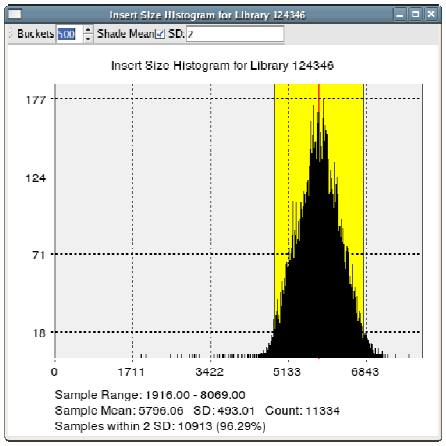
Hawkeye

Launch Pad

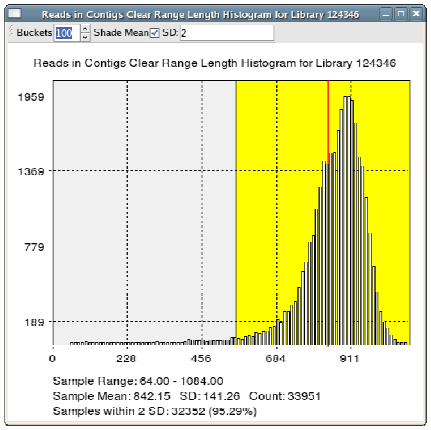

Hawkeye
File
Contig ID 1
All Views
Scaffold View
Contig View
Statistics
Features
Libraries
Scaffolds
Contigs
Reads
[Scaffolds]
TotalScaffolds 64
TotalContigsInScaffolds 161
MeanContigsPerScaffold 2.52
MinContigsPerScaffold 1
MaxContigsPerScaffold 39
TotalBasesInScaffolds 5041960
MeanBasesInScaffolds 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
Contig Length Distribution
300.0K
100.0K
25%
50%
75%
Count: 776 Max: 325273 N50: 98253 Total: 5839007
Ordering
Size
Feature Density
Treemap
Coloring
Green - Red
White - Red
Viewing Xoo.bnk/ with 776 contigs Contig Id:1 Size: 1611 Reads: 6

Histograms & Statistics

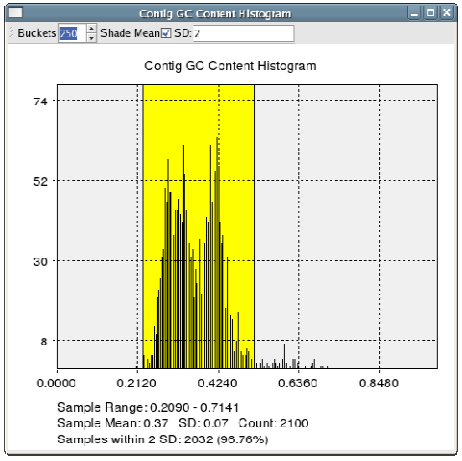
Insert Size



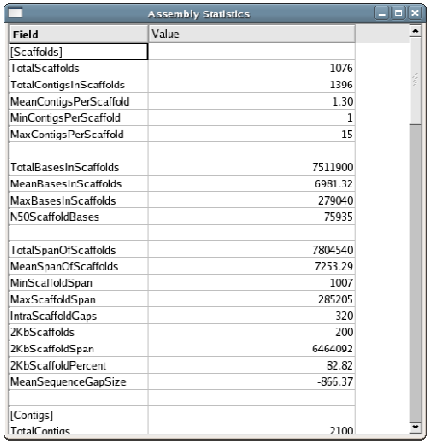
Read Length



GC Content



Overall Statistics



- Bird's eye view of data and assembly quality

Scaffold View

a. Statistical Plots

b. Scaffold

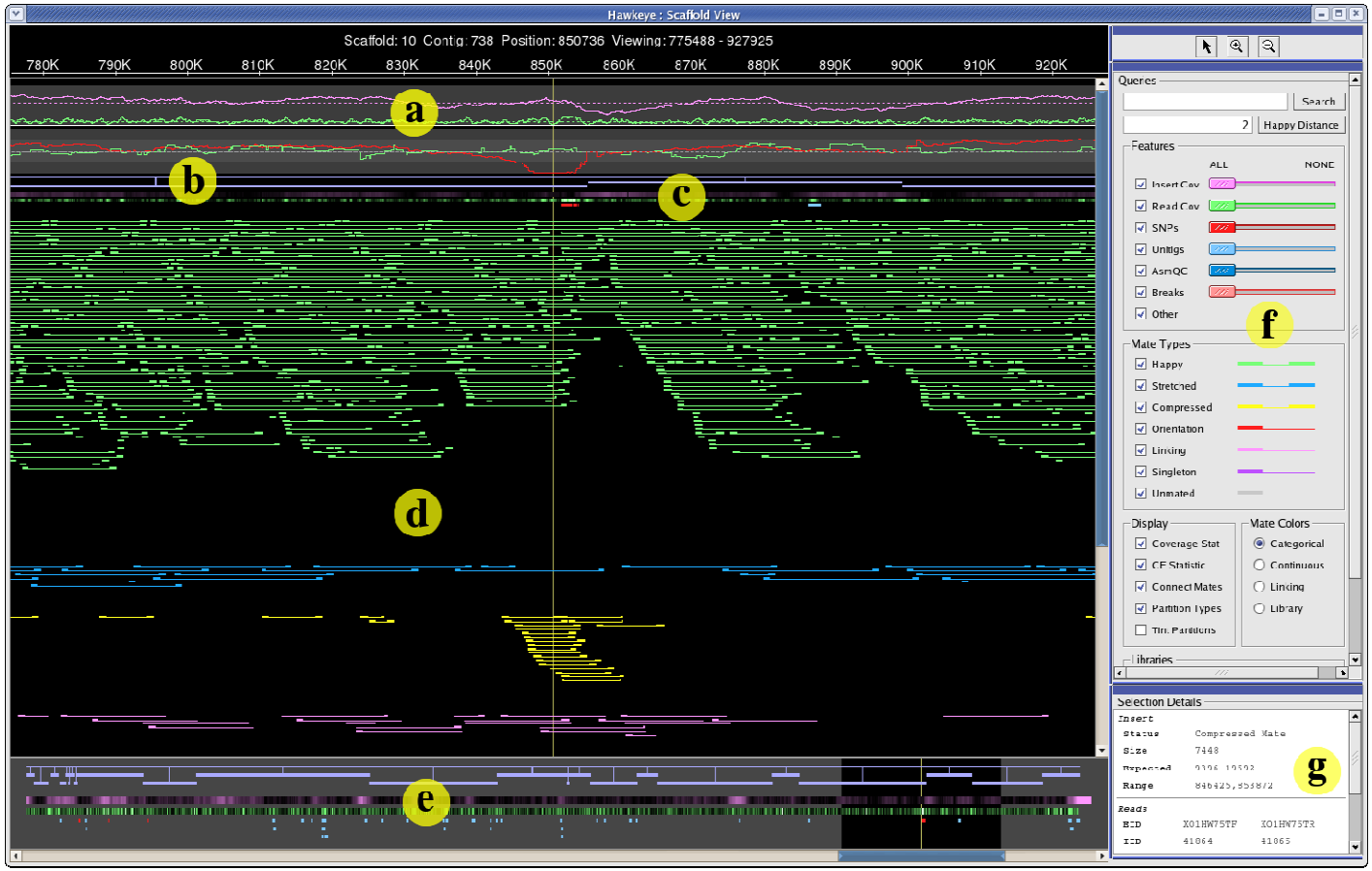
c. Features

d. Inserts

e. Overview

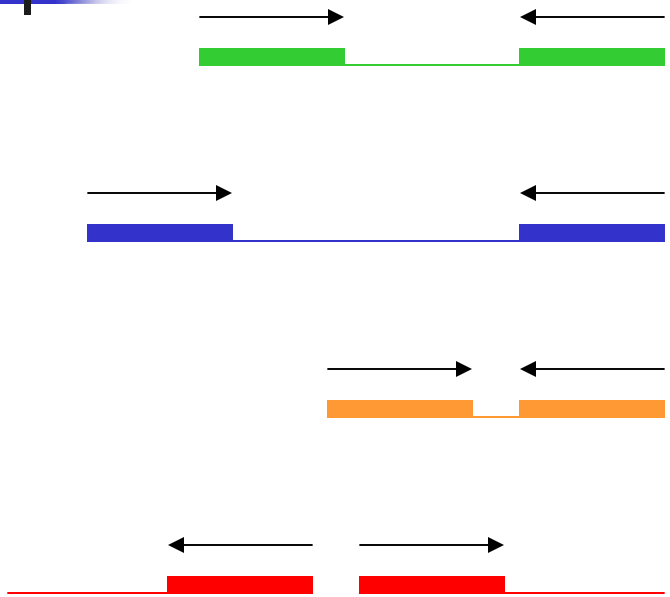
f. Control Panel

g. Details



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

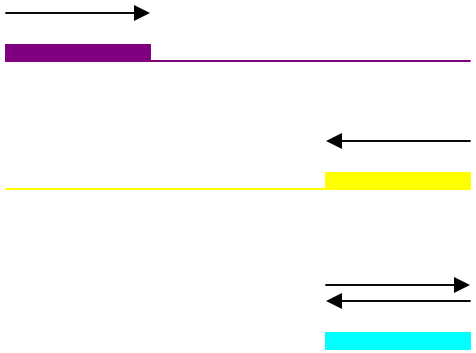
Compressed

- 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

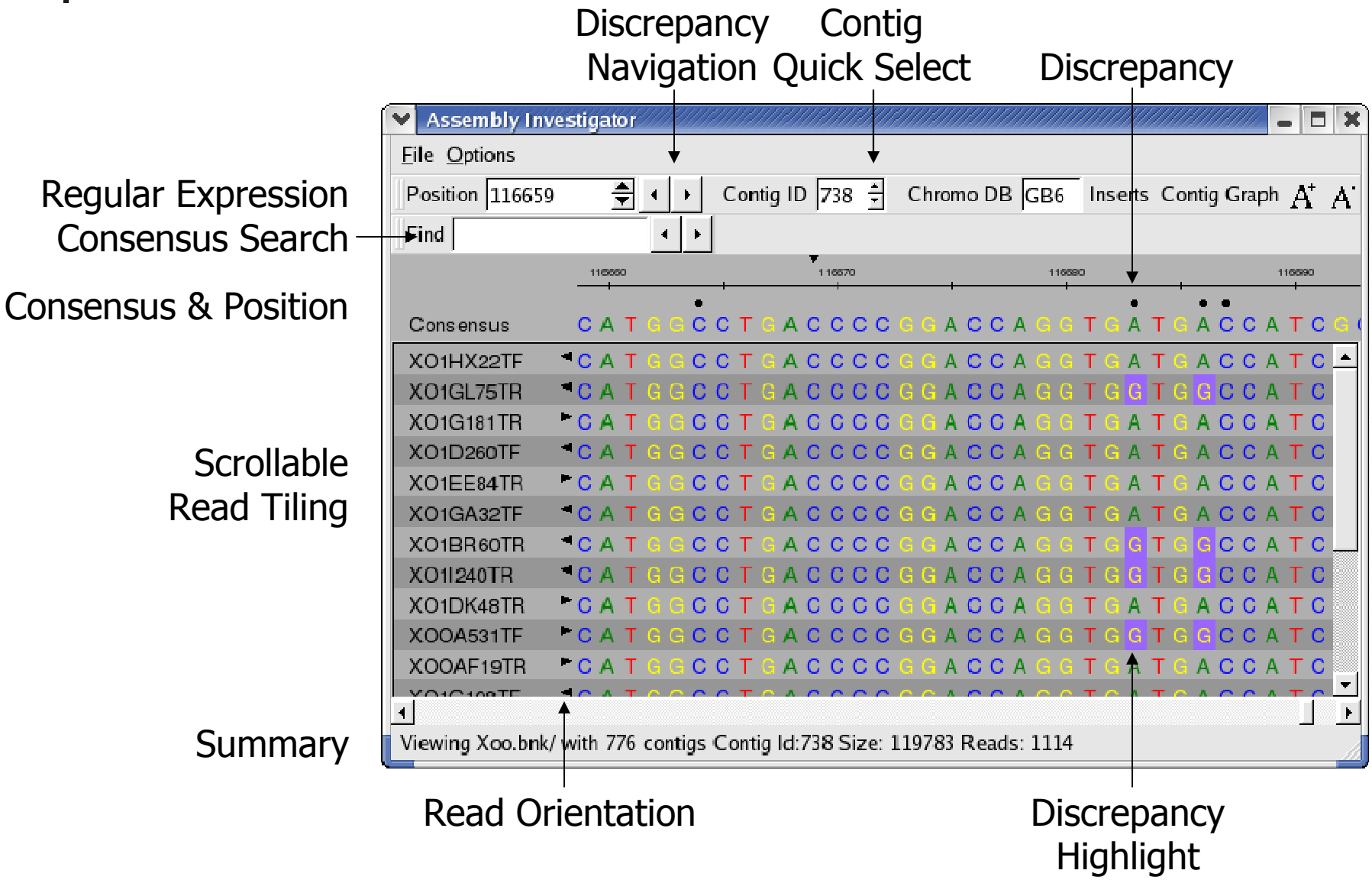
Singleton

- Read's mate is a singleton

Unmated

- No mate was provided for read

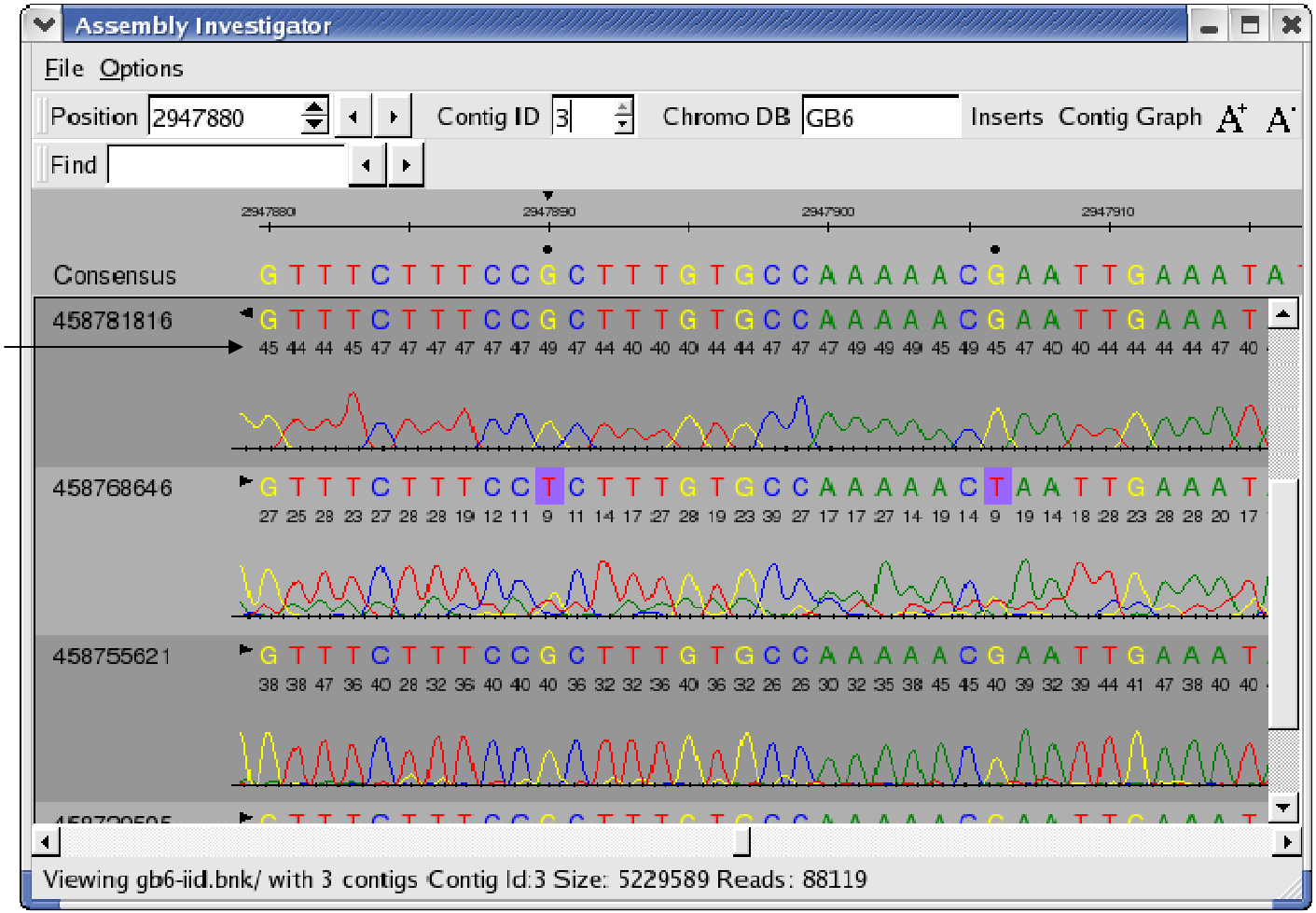
Contig View


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

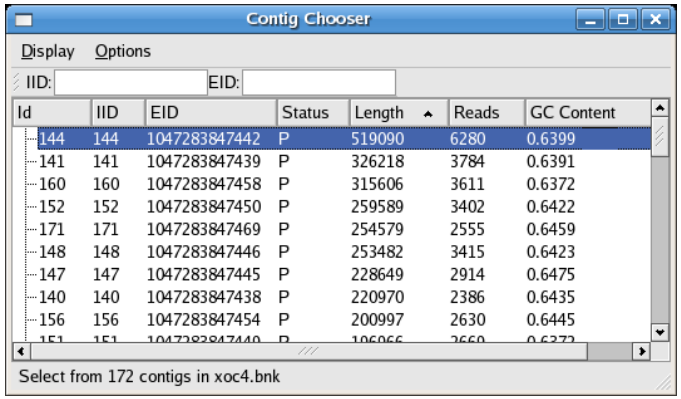


Chromatograms are loaded from specified directories, or on demand from Trace Archive.

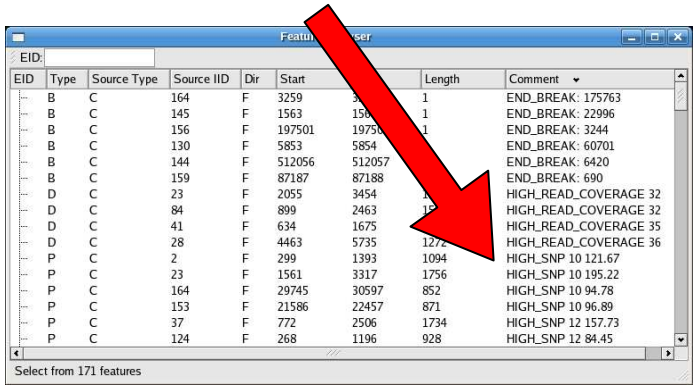
Assembly Reports

Misassembly Walkthough:
Correlated SNPs

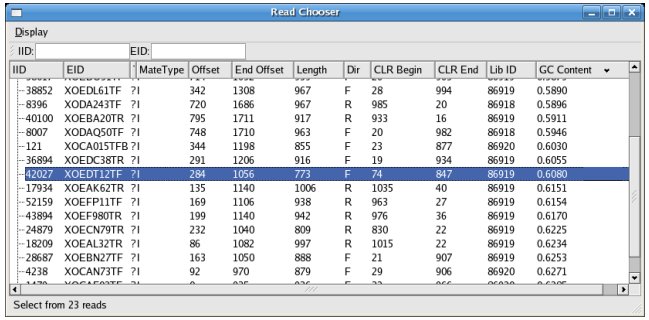
Contigs



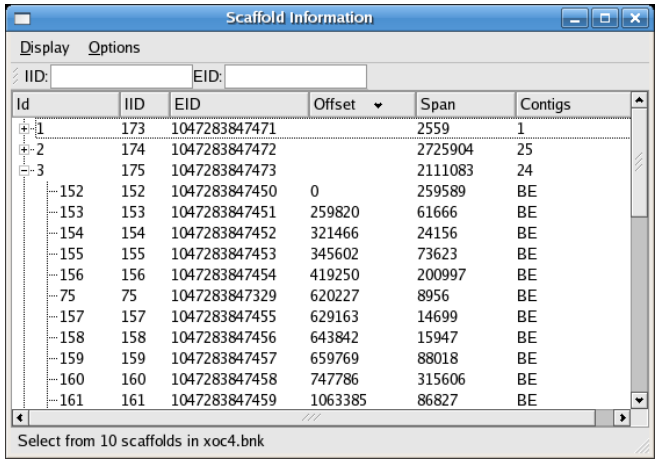
Features



Reads



Scaffolds



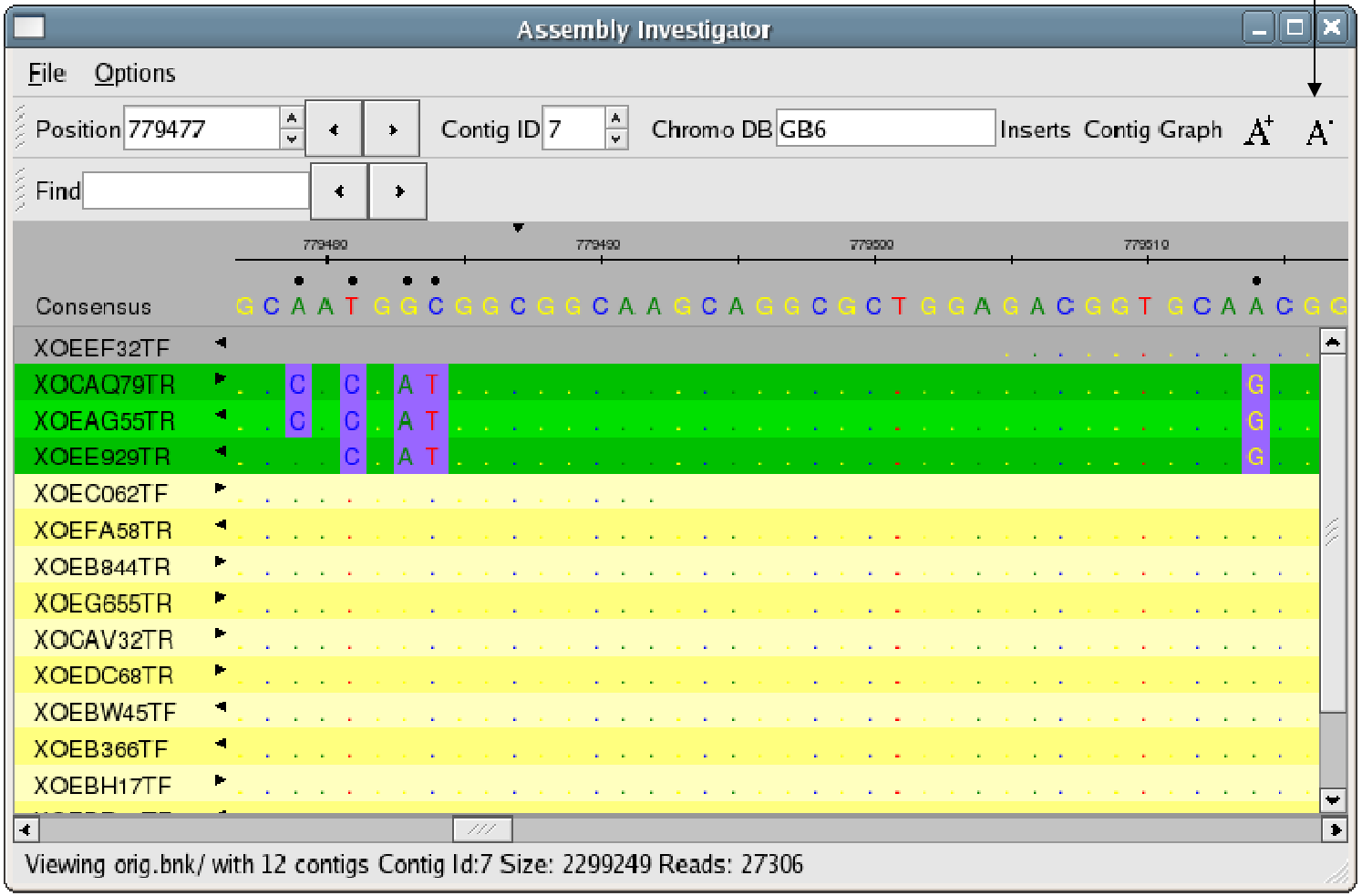
- Full Integration: “Double click takes you there”

SNP View

Zoom Out

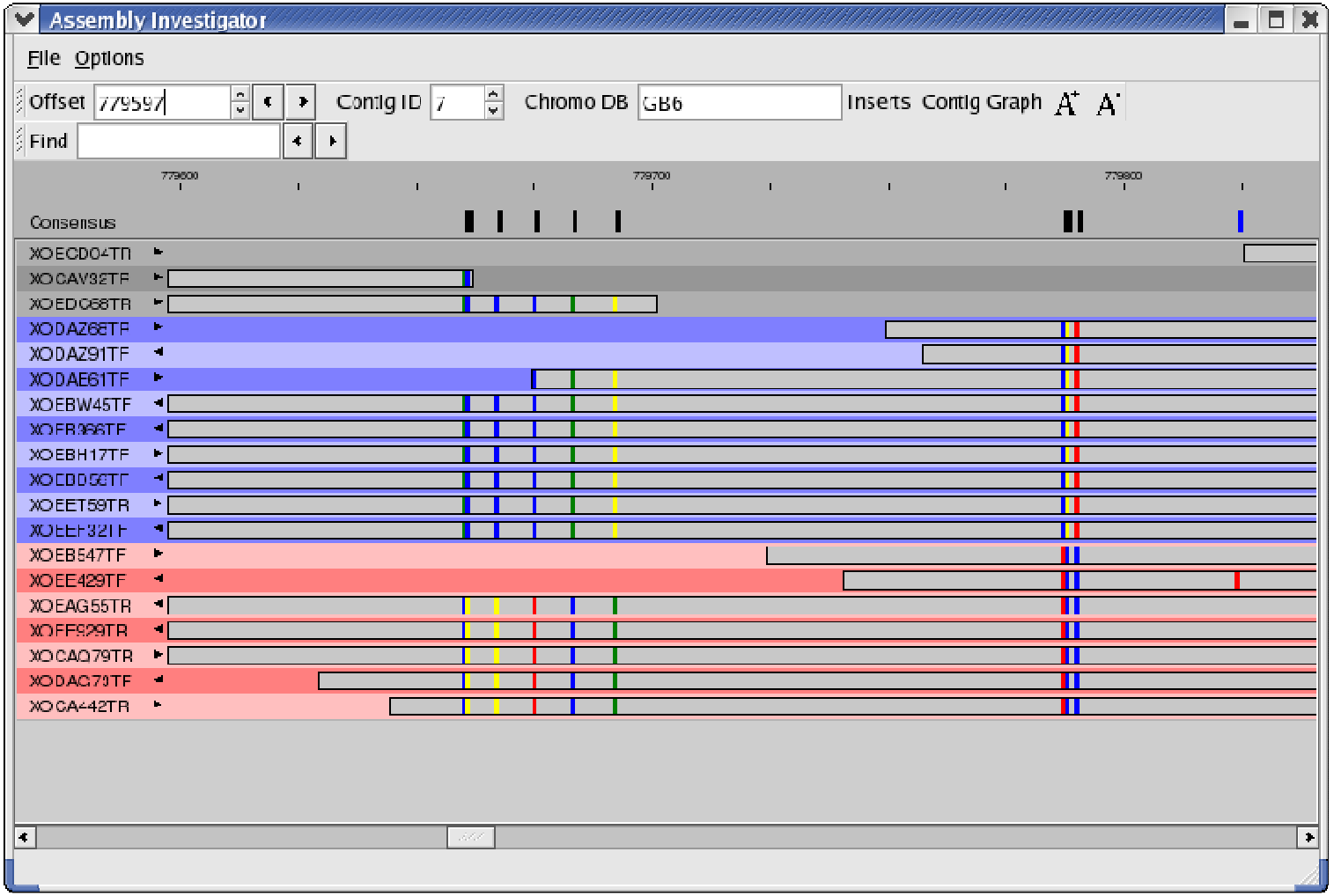
SNP Sorted Reads

Polymorphism View



SNP Barcode

SNP Sorted Reads



Colored Rectangle indicate the positions and composition of the SNPs

Scaffold View

Coverage

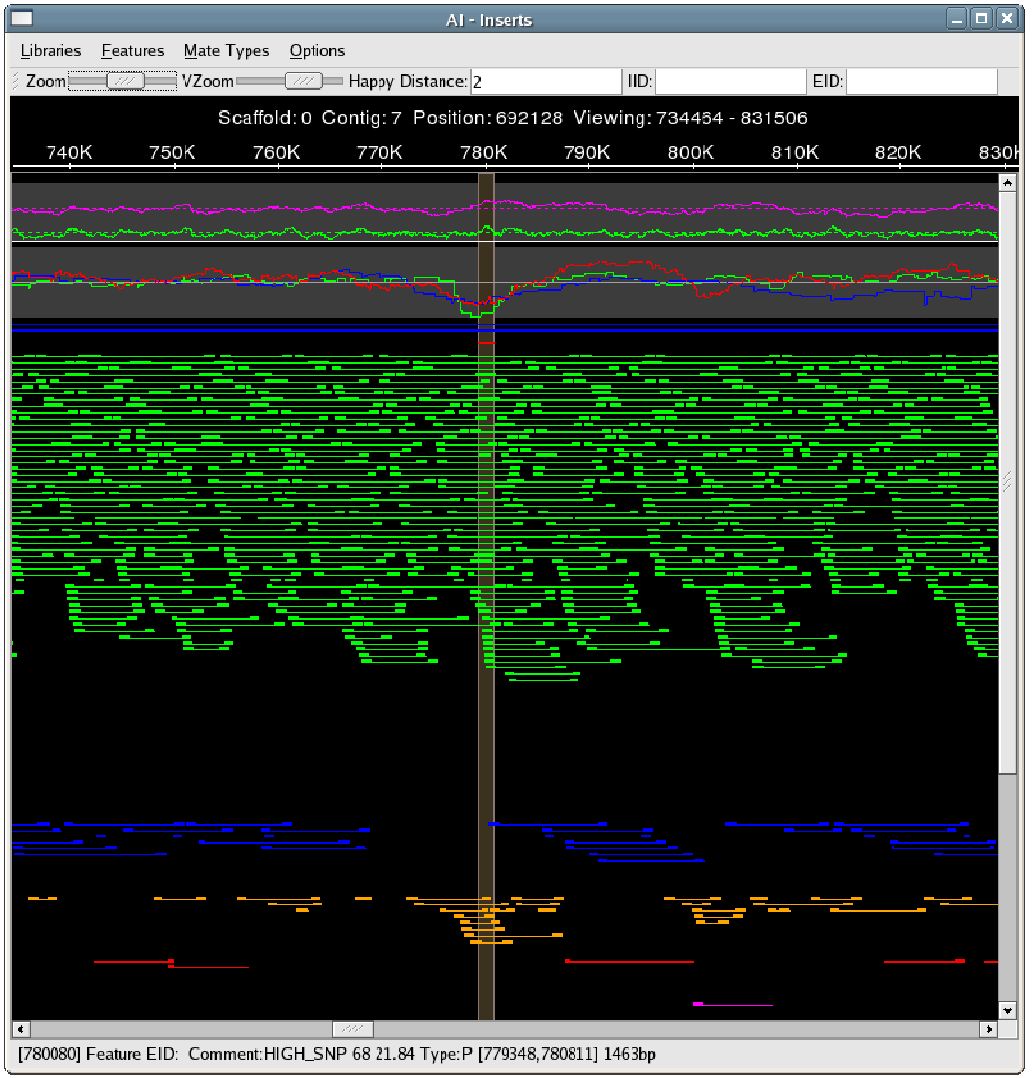
CE Statistic

Happy

Stretched

Compressed

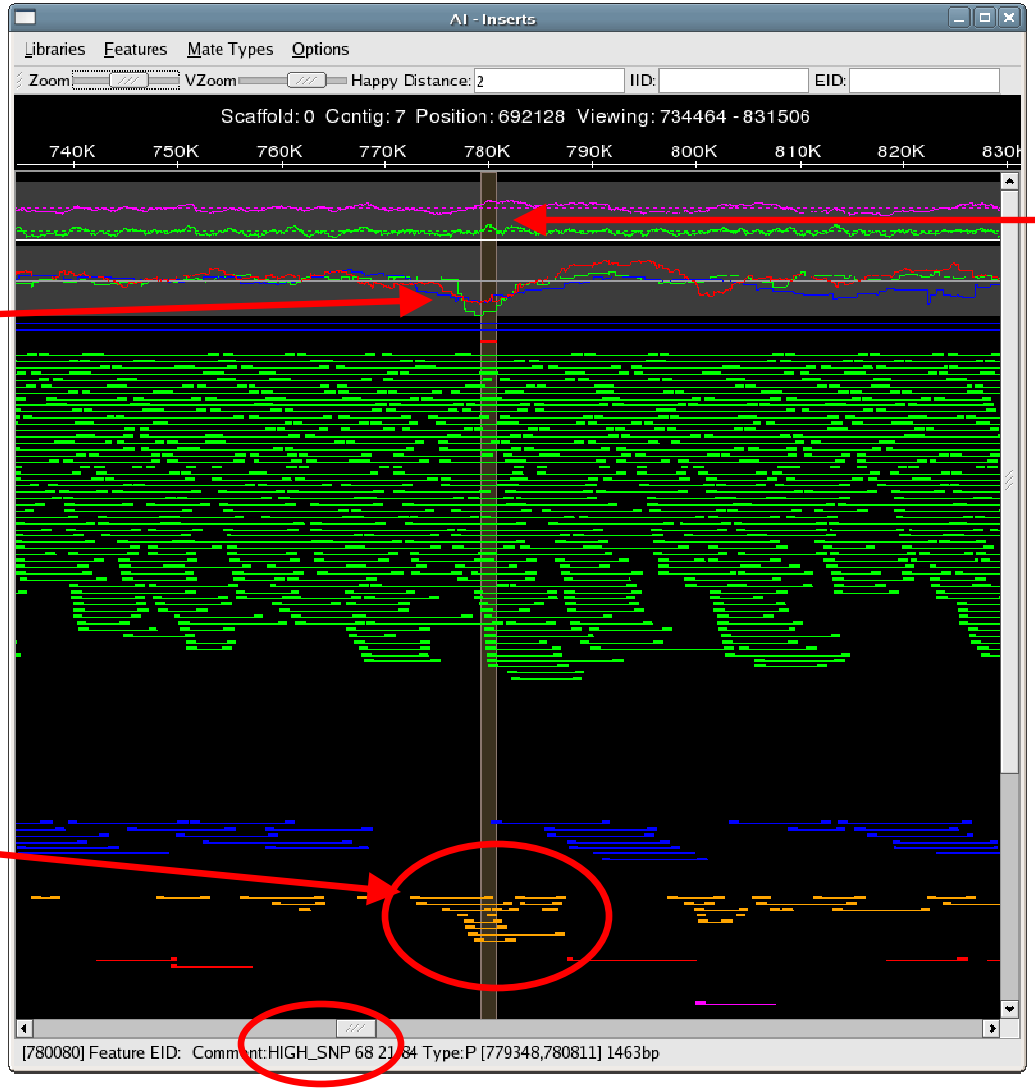
Misoriented



SNP Feature

Linking

Collapsed Repeat



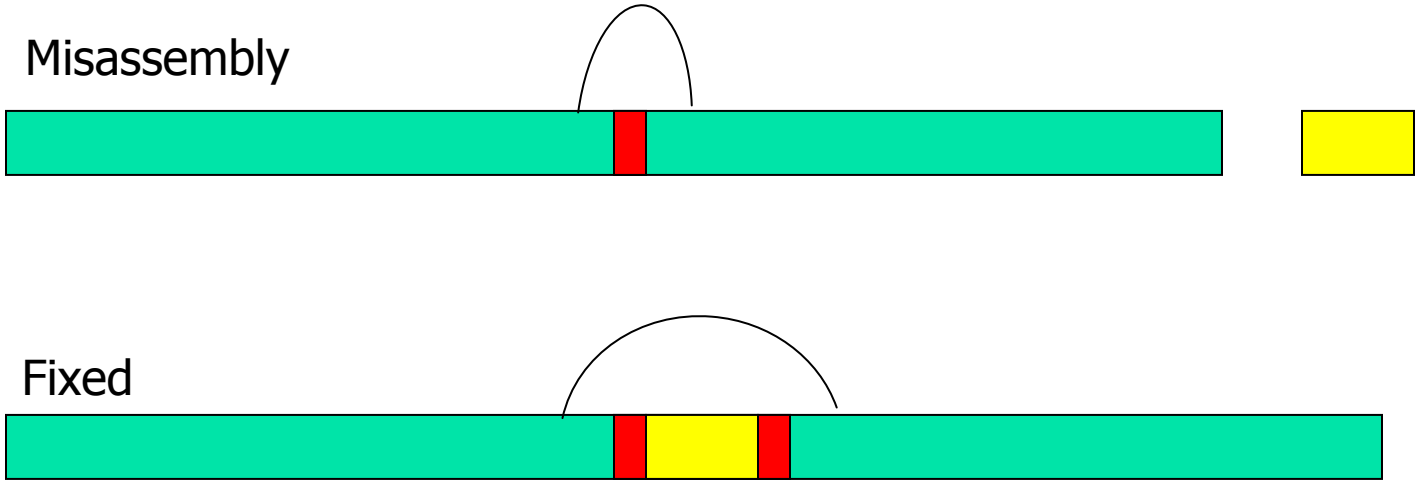
Read Coverage Spike

-5.5 CE Dip

Compressed Mates Cluster

68 Correlated SNPs

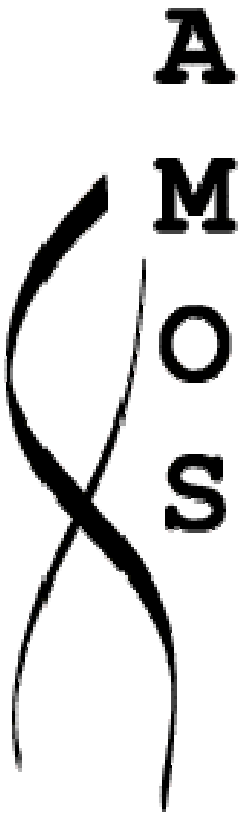
Confirmed Misassembly



Collapsed repeat

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

More Information



- Hawkeye Webpage:
  - http://amos.sourceforge.net/hawkeye
- Contact AMOS
  - amos-help [at] lists.sourceforge.net
- Acknowledgements

Adam Phillippy

Ben Shneiderman

Steven Salzberg

Mihai Pop

Art Delcher