AMOS Assembly Validation and Visualization

Michael Schatz

Center for Bioinformatics and Computational Biology
University of Maryland

August 13, 2006
University of Hawaii

Outline

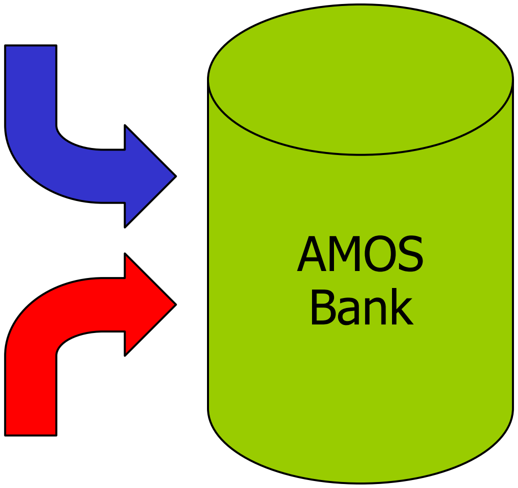
- AMOS Validation Pipeline
  - Mate-Based Validation
    - C/E Statistic
  - Read Alignment Validation
  - Read Breakpoint Validation
  - Read Depth Validation

- 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

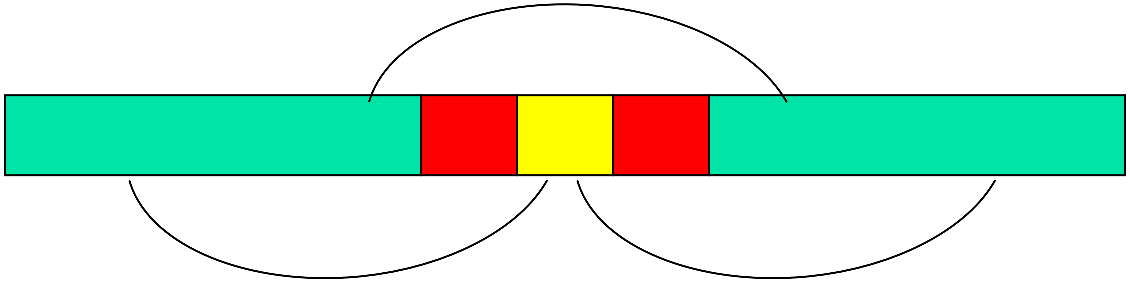
- Evaluate mate “happiness” across assembly
  - Happy = Correct orientation and distance

- 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)

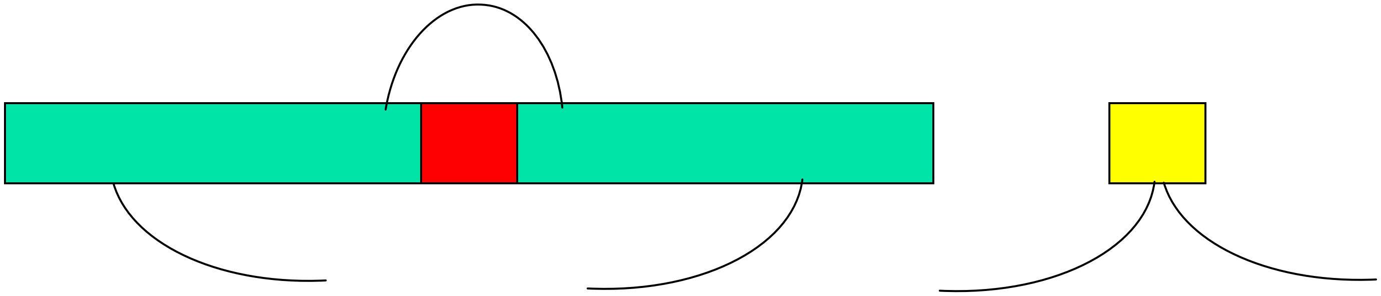
- Regions with high C/E statistic

Mate-Happiness: asmQC

- Excision: Skip reads between flanking repeats
  - Truth

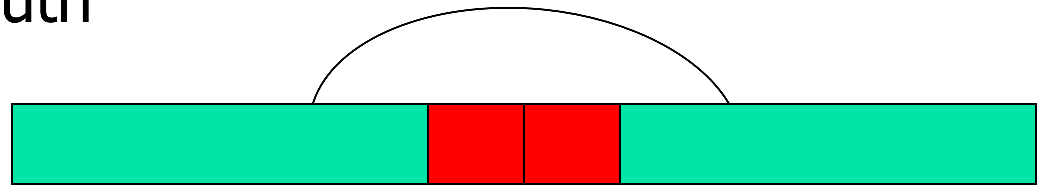


  - Misassembly: Compressed Mates, Missing Mates

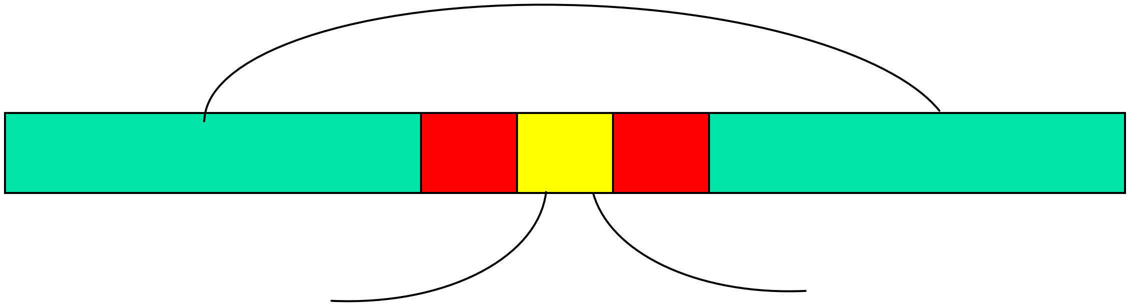


Mate-Happiness: asmQC

- Insertion: Additional reads between flanking repeats
  - Truth

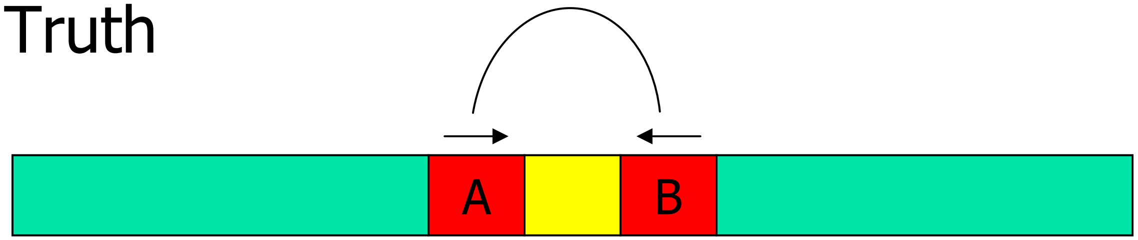


  - Misassembly: Expanded Mates, Missing Mates

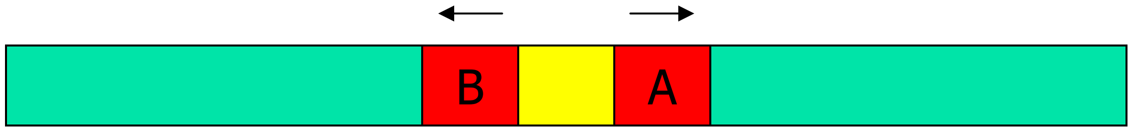


Mate-Happiness: asmQC

- Rearrangement: Reordering of reads
  - Truth



  - Misassembly: Misoriented Mates



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

C/E Statistic

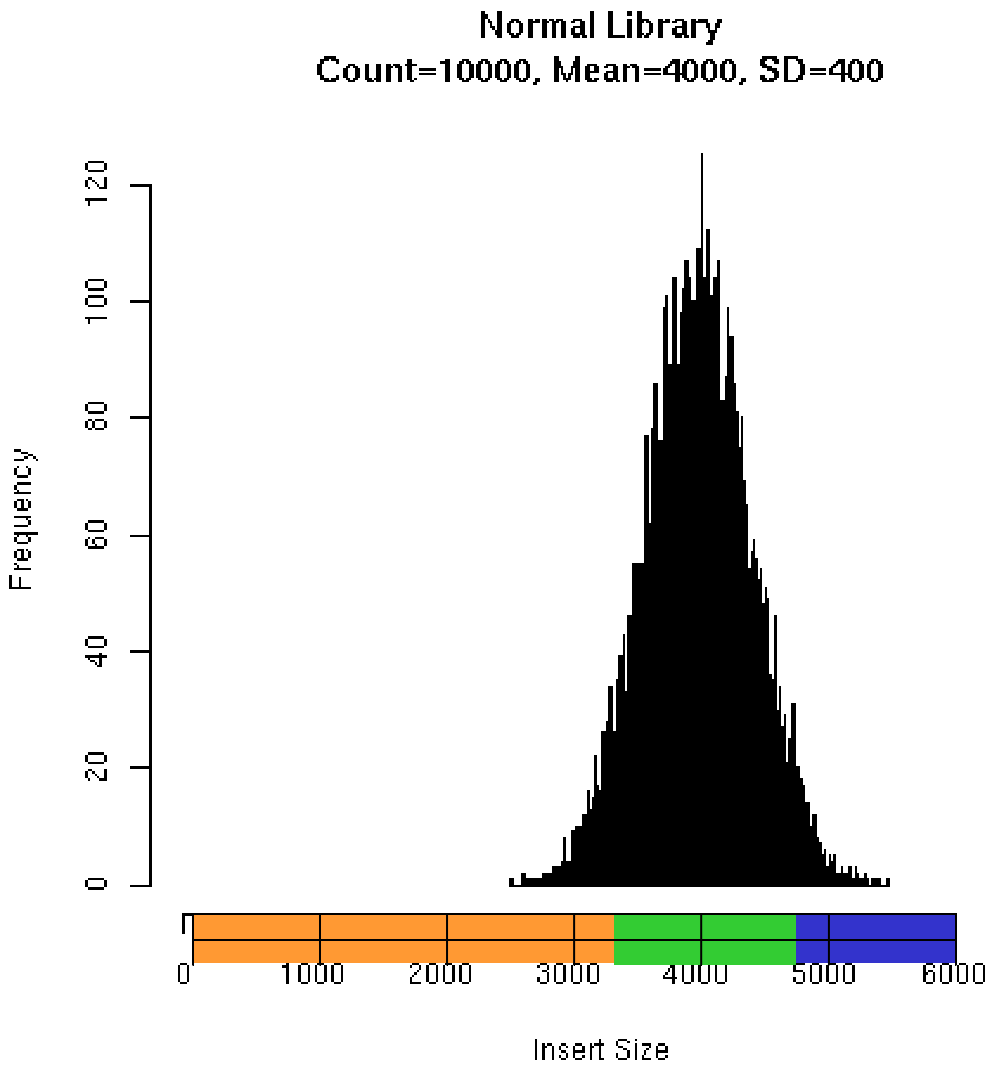
- The presence of individual compressed or expanded mates is rare but expected.

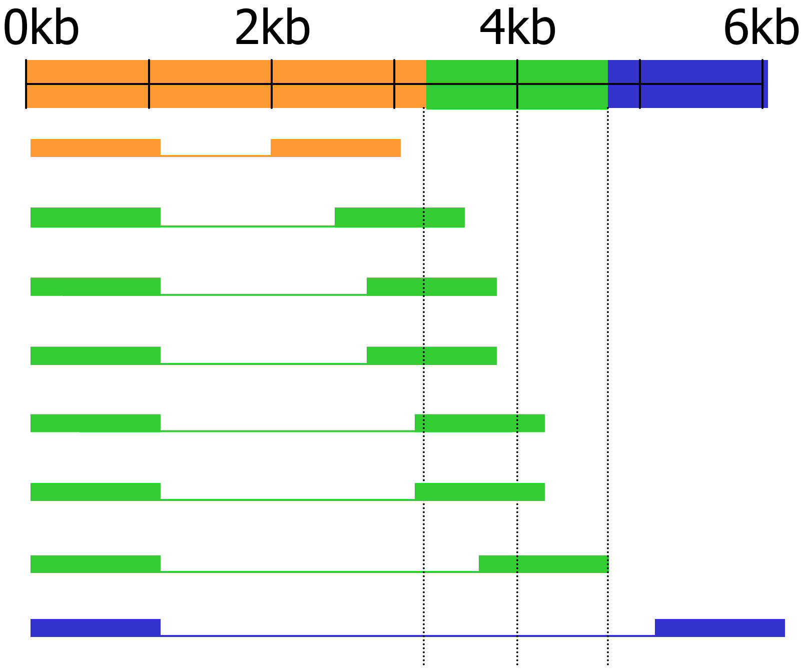
- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”

- Compute the statistic at all positions
  - (Local Mean – Global Mean) / Scaling Factor

- Introduced by Jim Yorke’s group at UMD

Sampling the Genome





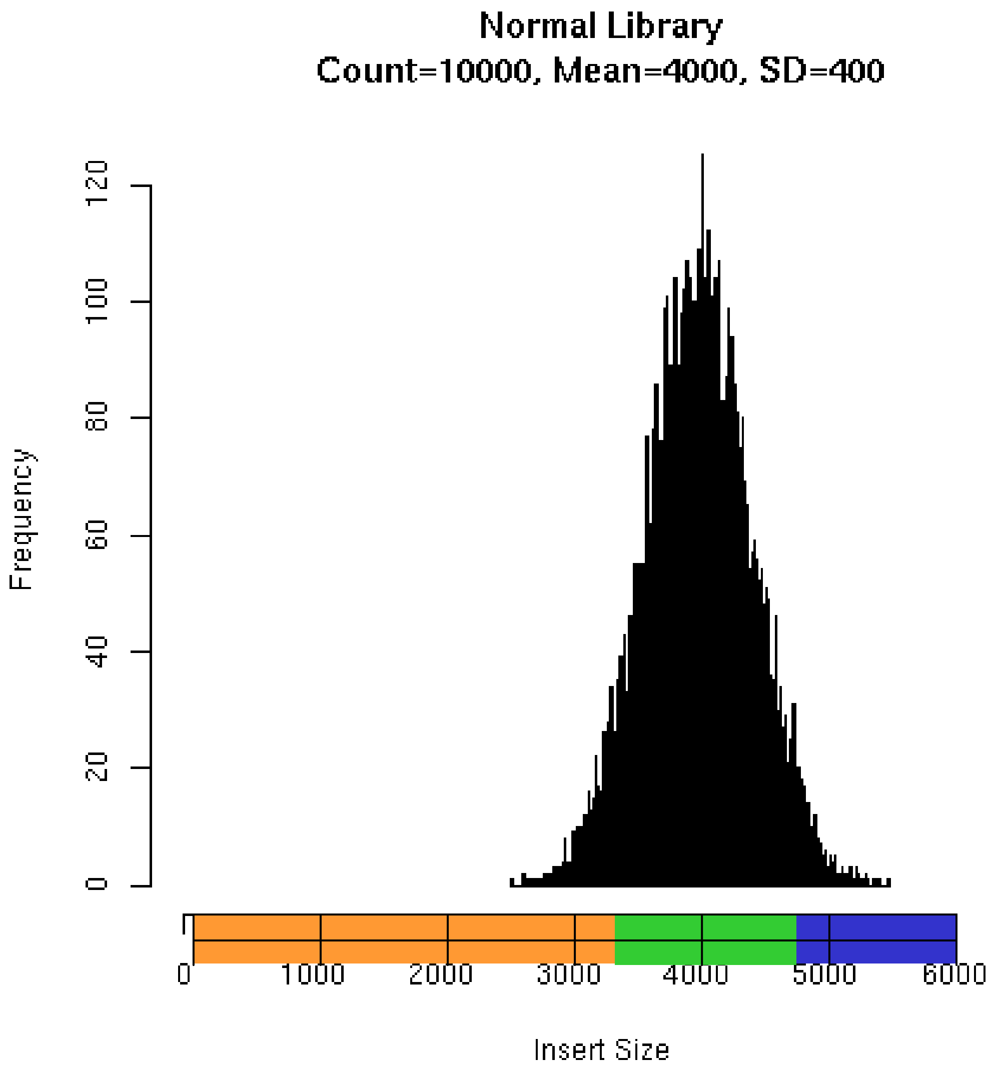
8 inserts: 3kb-6kb

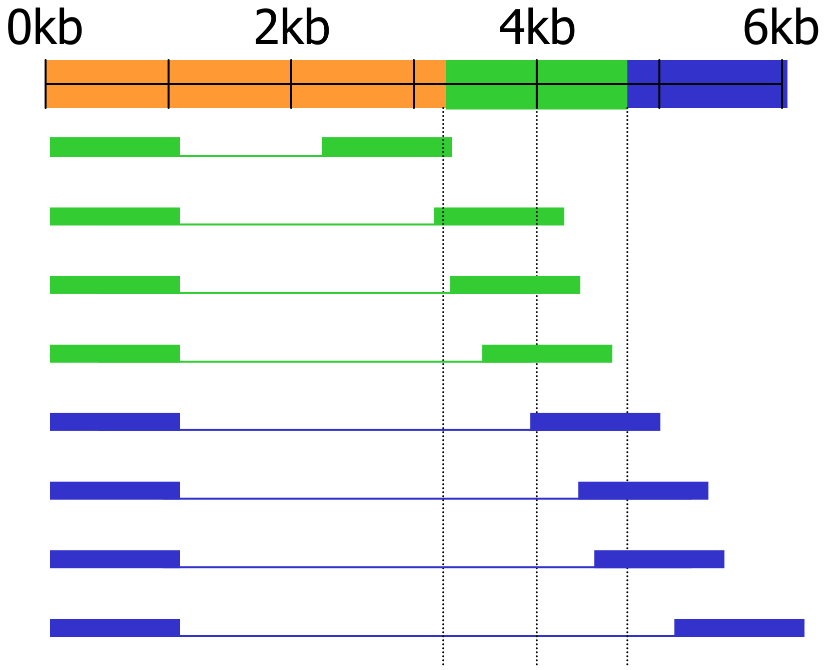
Local Mean: 4048

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

Near 0 indicates overall happiness

C/E-Statistic: Expansion





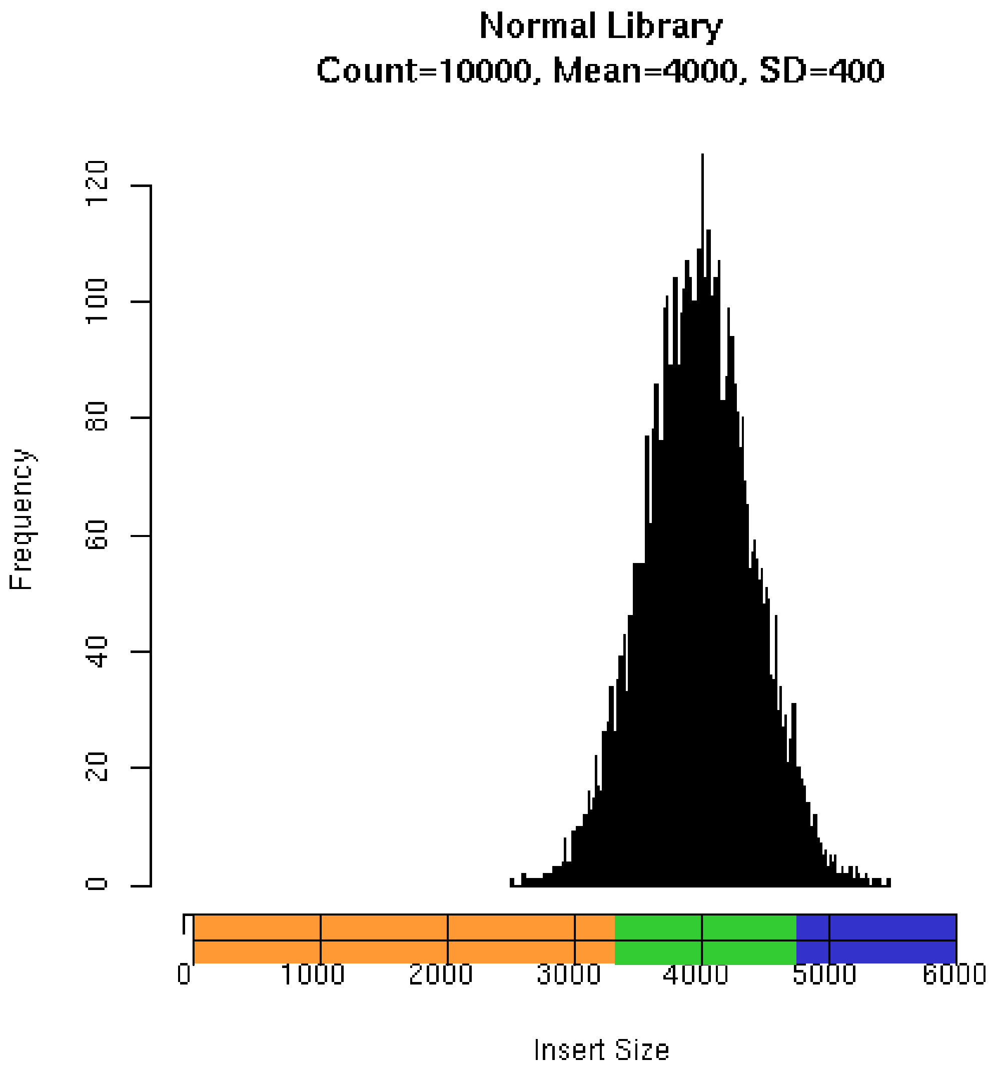
8 inserts: 3.2kb-6kb

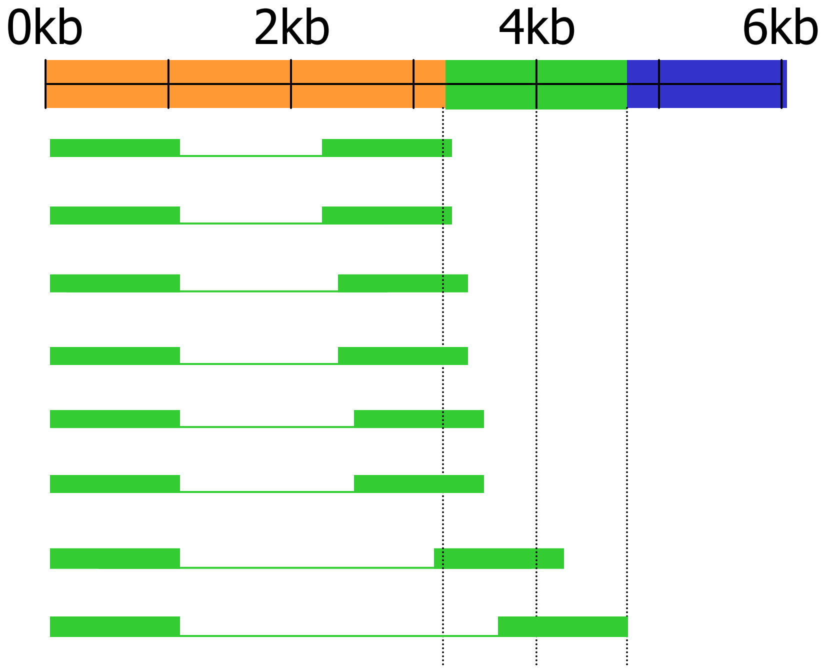
Local Mean: 4461

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

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

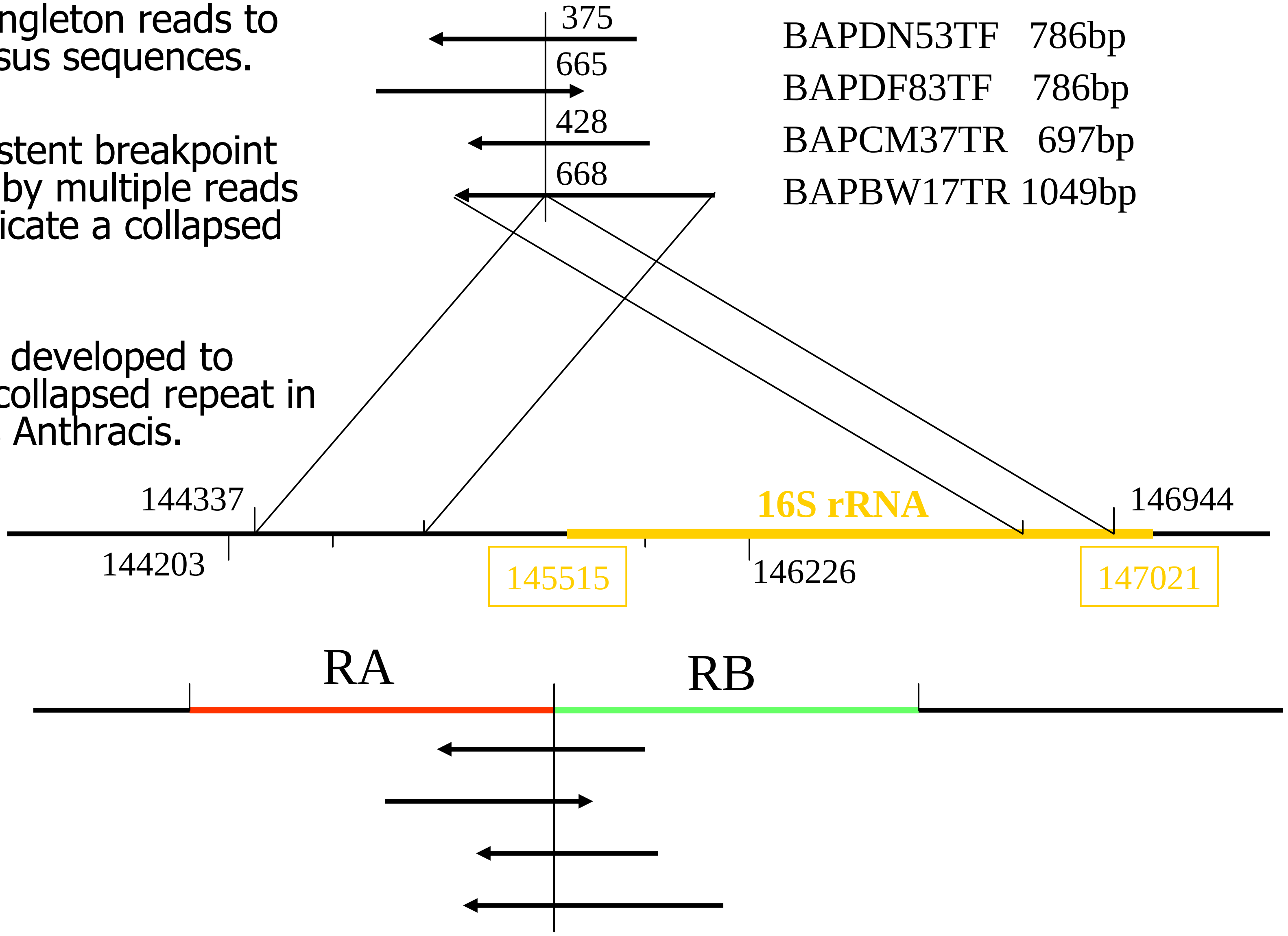
- Regions of correlated SNPs are likely to be assembly errors or interesting biological events
  - Highly specific metric

- 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)

A	G	C
A	G	C
A	G	C
A	G	C
A	G	C
A	G	C
C	T	A
C	T	A
C	T	A
C	T	A
C	T	A

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.



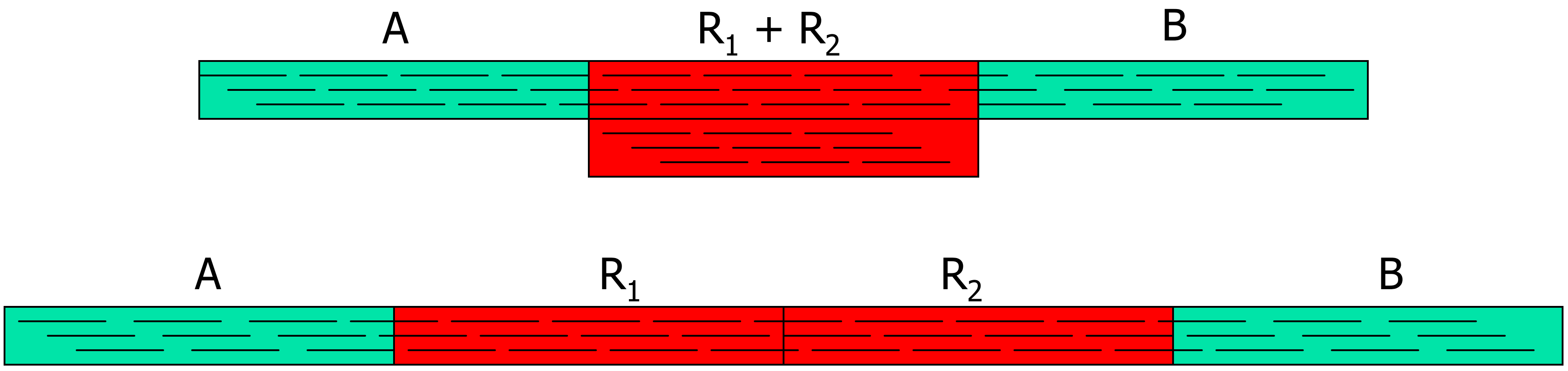
BAPDN53TF	786bp
BAPDF83TF	786bp
BAPCM37TR	697bp
BAPBW17TR	1049bp

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



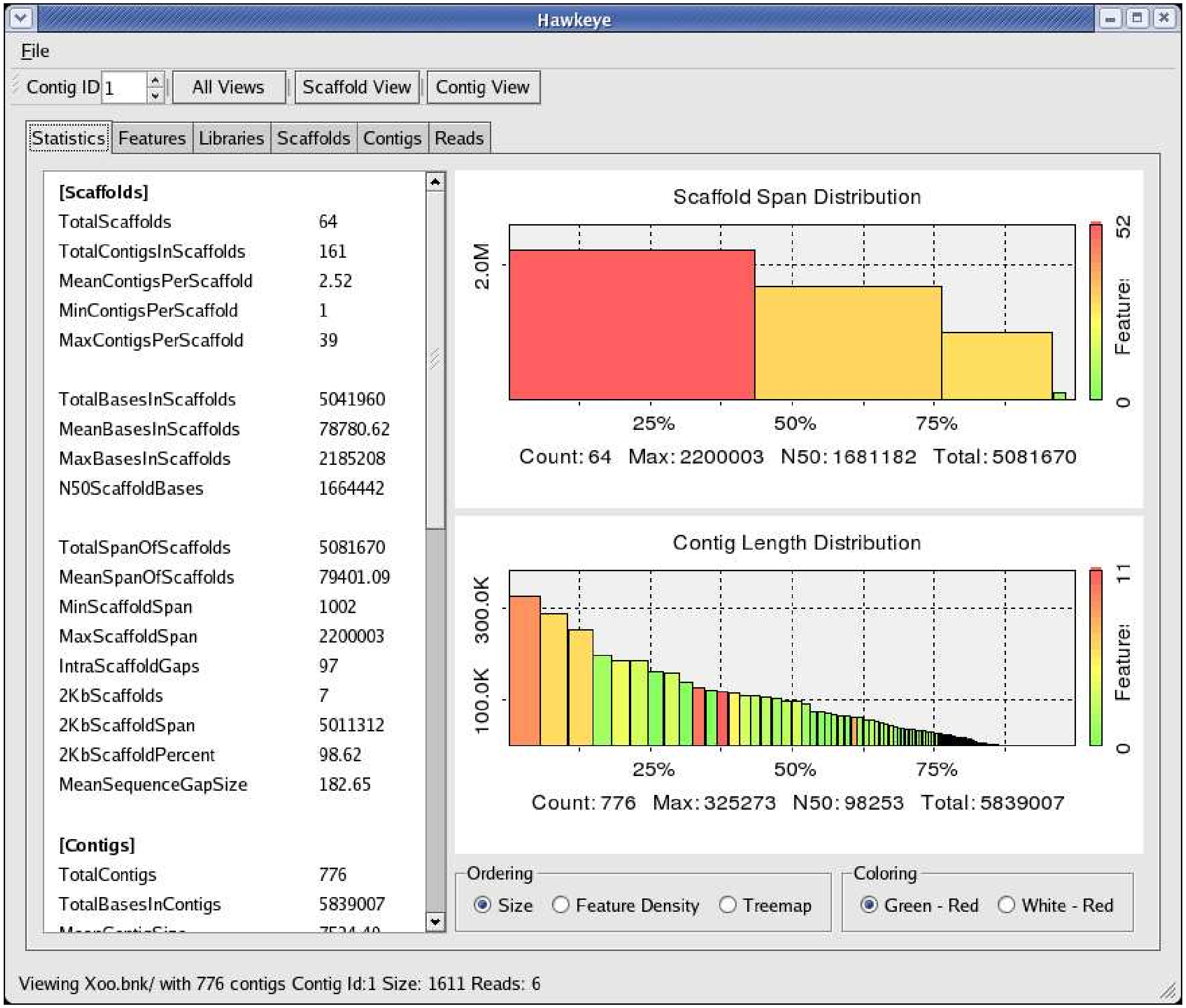
Hawkeye

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



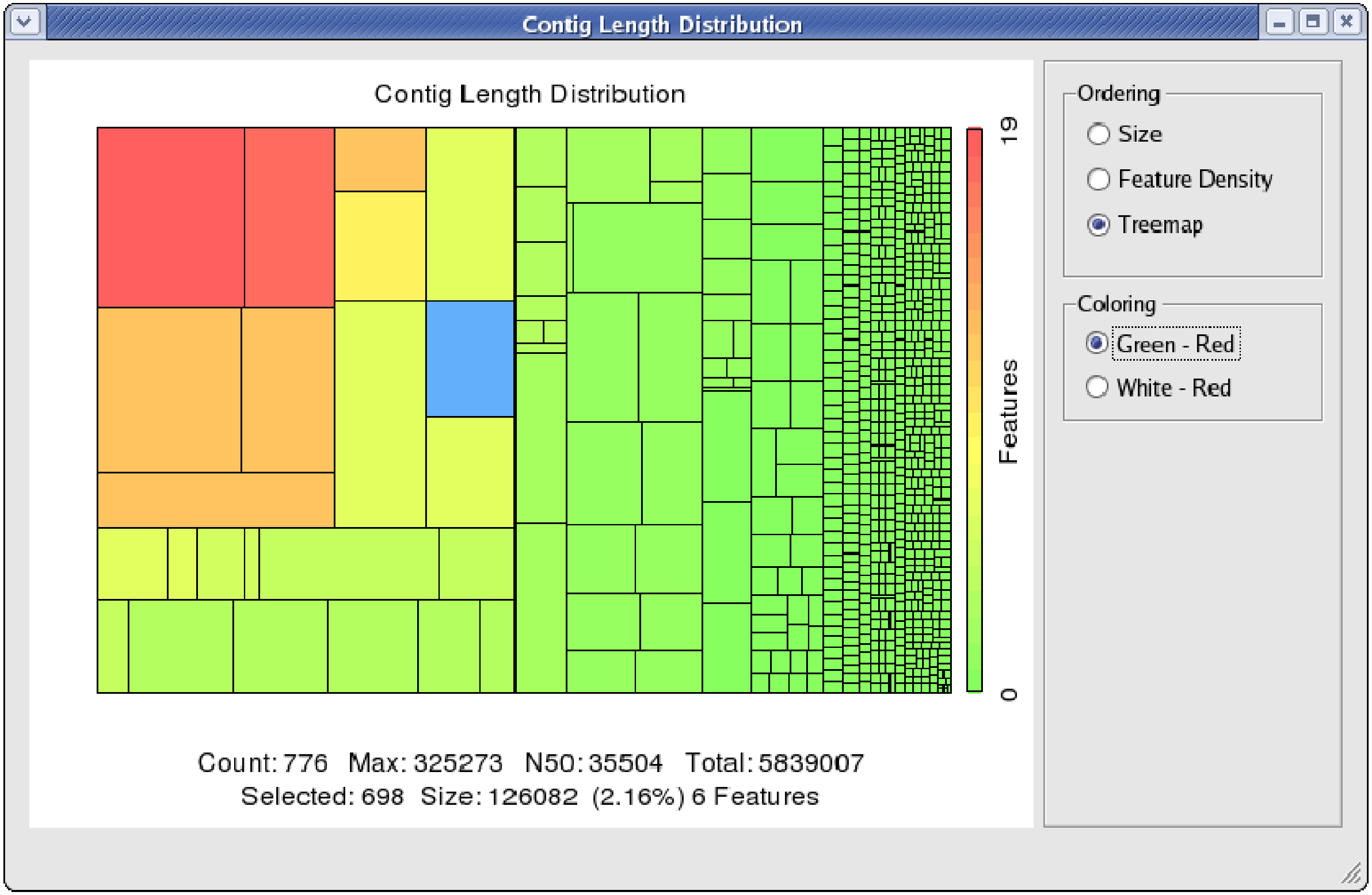
[Scaffolds]	
TotalScaffolds	64
TotalContigsInScaffolds	161
MeanContigsPerScaffold	2.52
MinContigsPerScaffold	1
MaxContigsPerScaffold	39
TotalBasesInScaffolds	5041960
MeanBasesInScaffolds	78780.62
MaxBasesInScaffolds	2185208
N50ScaffoldBases	1664442
TotalSpanOfScaffolds	5081670
MeanSpanOfScaffolds	79401.09
MinScaffoldSpan	1002
MaxScaffoldSpan	2200003
IntraScaffoldGaps	97
2KbScaffolds	7
2KbScaffoldSpan	5011312
2KbScaffoldPercent	98.62
MeanSequenceGapSize	182.65
[Contigs]	
TotalContigs	776
TotalBasesInContigs	5839007

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



Ordering

○ Size

○ Feature Density

● Treemap

Coloring

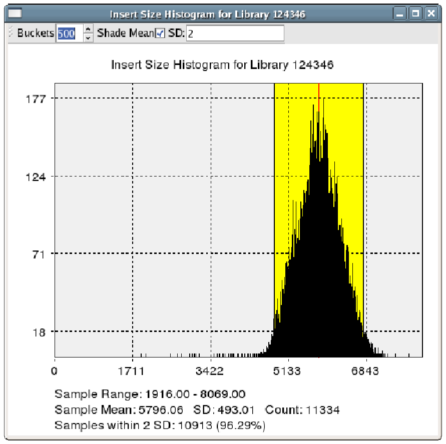
● Green - Red

○ White - Red

Count: 776   Max: 325273   N50: 35504   Total: 5839007
Selected: 698 Size: 126082 (2.16%) 6 Features

Histograms & Statistics

Insert
Size

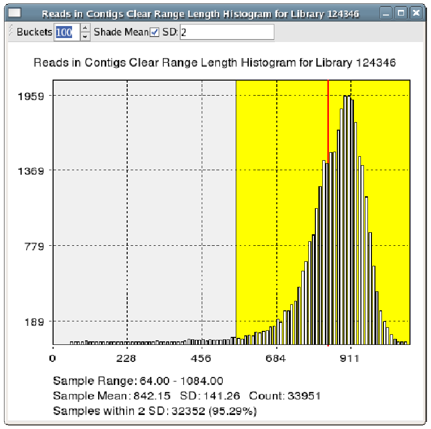


Insert Size Histogram for Library 124346

Buckets 500 Shade Mean ☑ SD: 2

```
Sample Range: 1916.00 - 8069.00
Sample Mean: 5796.06   SD: 493.01   Count: 11334
Samples within 2 SD: 10913 (96.29%)
```

Read
Length

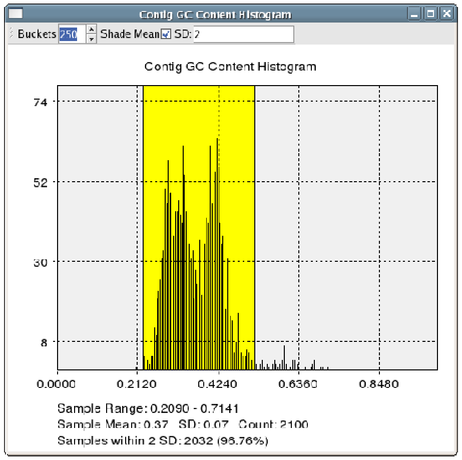

Reads in Contigs Clear Range Length Histogram for Library 124346
1959
1369
779
189
0
228
456
684
911

Reads in Contigs Clear Range Length Histogram for Library 124346

Buckets 100 Shade Mean ☑ SD: 2

```
Sample Range: 64.00 - 1084.00
Sample Mean: 842.15   SD: 141.26   Count: 33951
Samples within 2 SD: 32352 (95.29%)
```

GC
Content

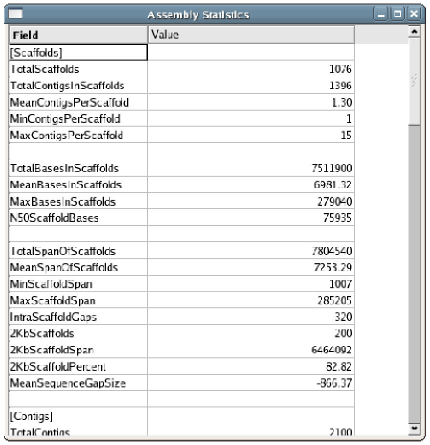


Contig GC Content Histogram

Buckets 250 Shade Mean ☑ SD: 2

```
Sample Range: 0.2090 - 0.7141
Sample Mean: 0.37   SD: 0.07   Count: 2100
Samples within 2 SD: 2032 (96.76%)
```

Overall
Statistics



Assembly Statistics

Field	Value
[Scaffolds]	
TotalScaffolds	1076
TotalContigsInScaffolds	1396
MeanContigsPerScaffold	1.30
MinContigsPerScaffold	1
MaxContigsPerScaffold	15
TotalBasesInScaffolds	7511900
MeanBasesInScaffolds	6981.32
MaxBasesInScaffolds	279040
N50ScaffoldBases	75935
TotalSpanOfScaffolds	7804540
MeanSpanOfScaffolds	7253.29
MinScaffoldSpan	1007
MaxScaffoldSpan	285205
IntraScaffoldGaps	320
2KbScaffolds	200
2KbScaffoldSpan	6464092
2KbScaffoldPercent	82.82
MeanSequenceGapSize	-366.37
[Contigs]	
TotalContigs	2100

- Bird’s eye view of data and assembly quality

Scaffold View

a. Statistical Plots

b. Scaffold

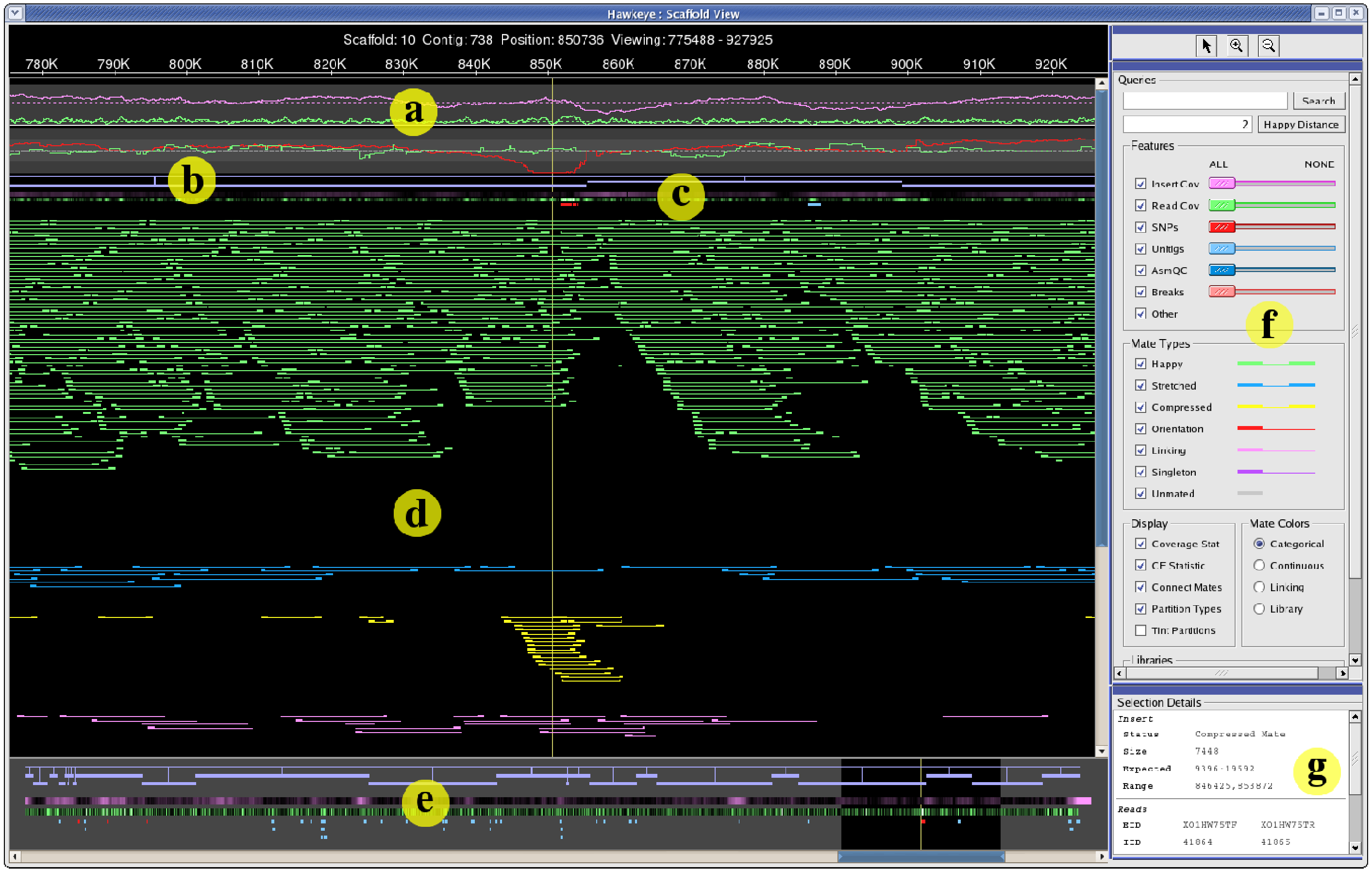
c. Features

d. Inserts

e. Overview

f. Control Panel

g. Details



Scaffold: 10 Contig: 738 Position: 850736 Viewing: 775488 - 927925

Queries: 2

Features:
☑ Insert Cov
☑ Read Cov
☑ SNPs
☑ Unitigs
☑ AsmQC
☑ Breaks
☑ Other

Mate Types:
☑ Happy
☑ Stretched
☑ Compressed
☑ Orientation
☑ Linking
☑ Singleton
☑ Unmated

Display:
☑ Coverage Stat
☑ CE Statistic
☑ Connect Mates
☑ Partition Types
☐ Trim Partitions

Mate Colors:
● Categorical
○ Continuous
○ Linking
○ Library

```
Insert
  Status      Compressed Mate
  Size        7448
  Expected    9396-19592
  Range       846425,853872

Reads
  EID        XO1HW75TF     XO1HW75TR
  IID        41064         41065
```

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

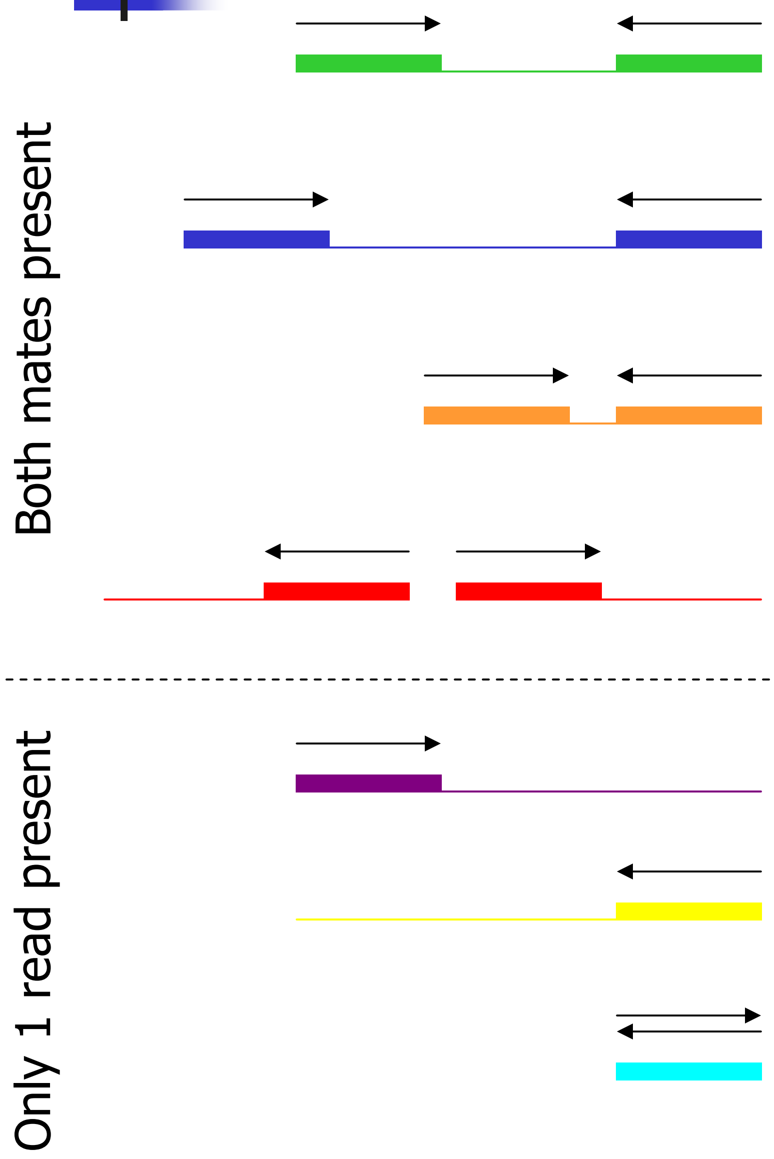
Loading Features:

```
$ loadFeatures bankname featfile
```

Featfile format:

```
Contigid type end5 end3 comment
```

Insert Happiness



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

Linking

- Read’s mate is in some other scaffold

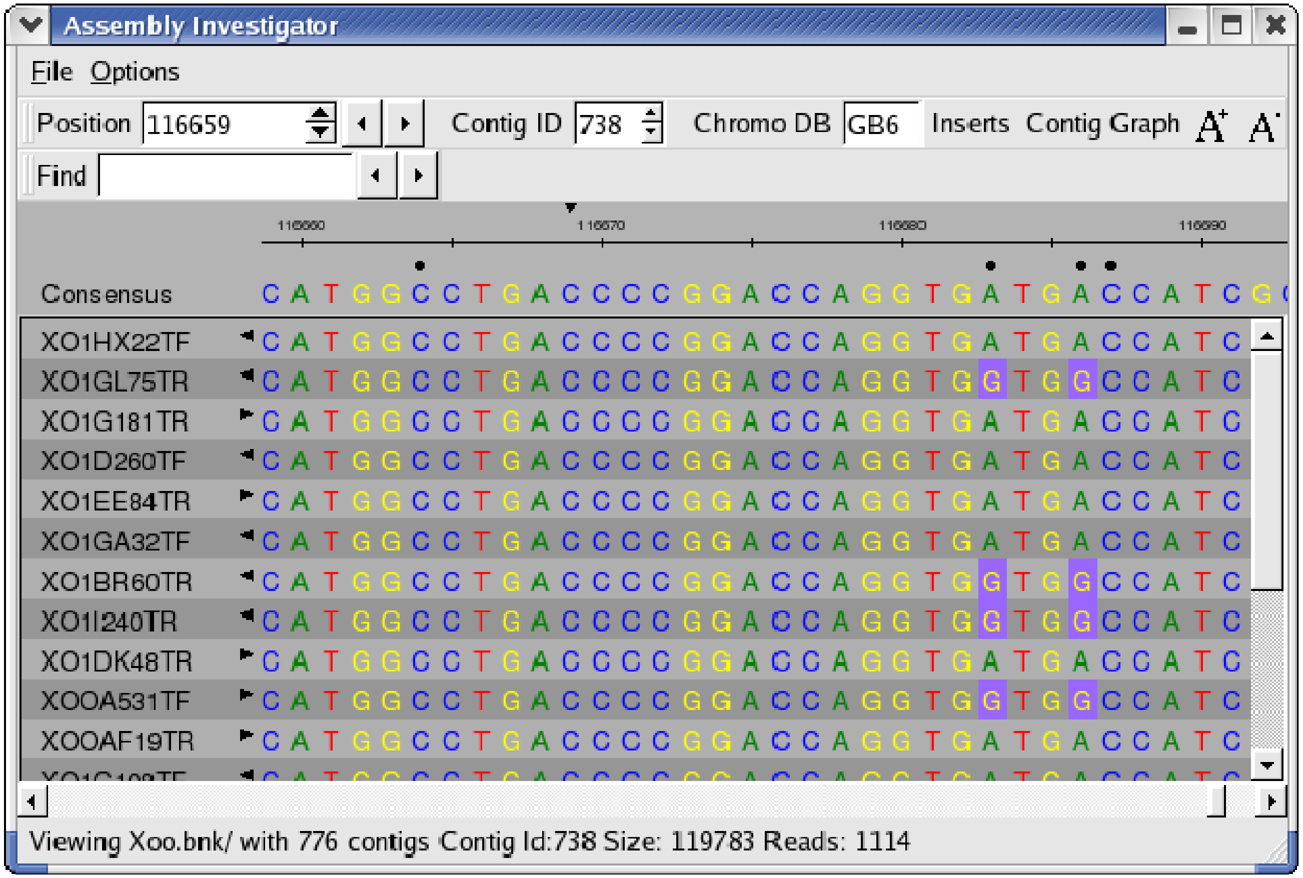
Singleton

- Read’s mate is a singleton

Unmated

- No mate was provided for read

Contig View

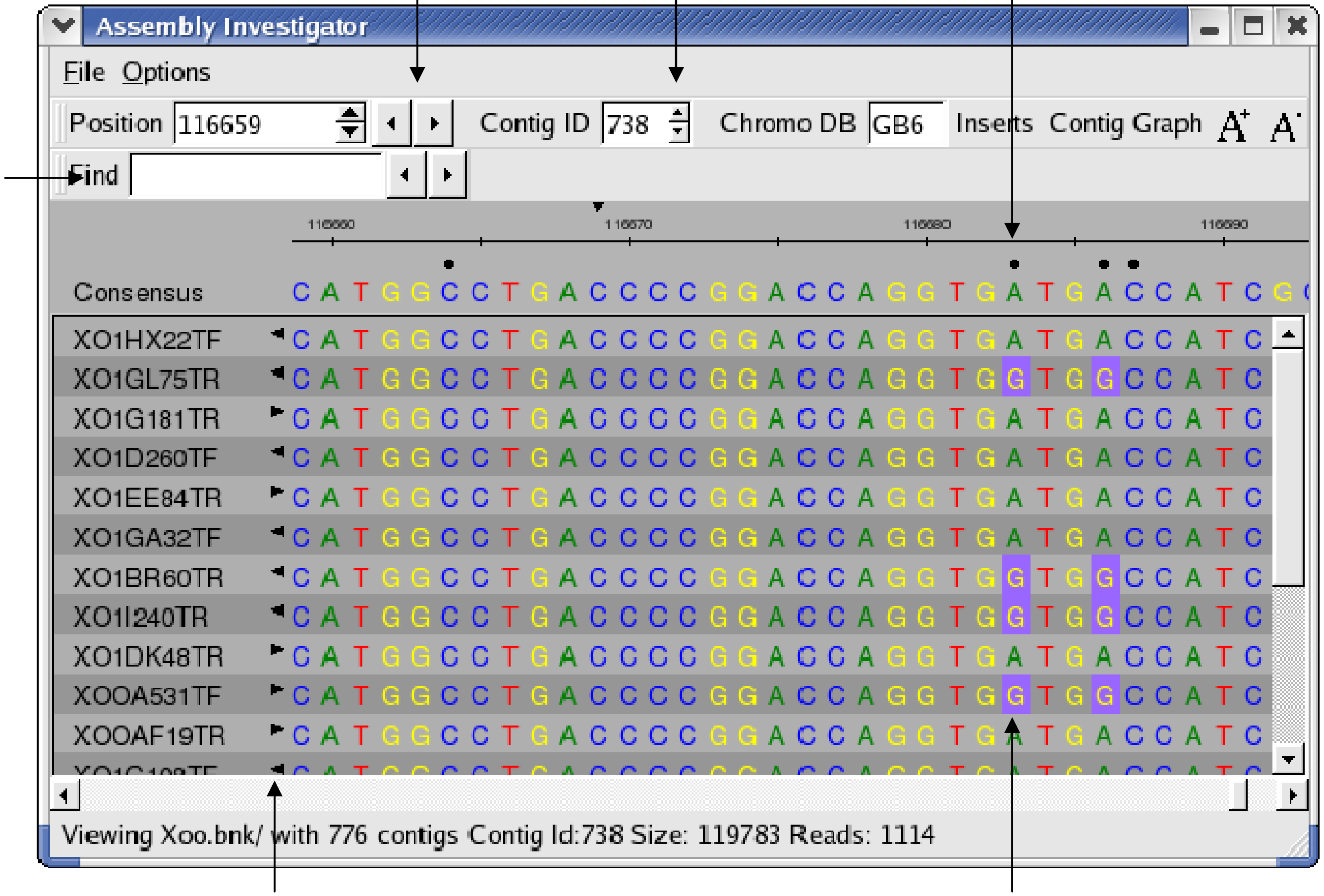


Position 116659 Contig ID 738 Chromo DB GB6

```
            116660              116670              116680              116690
Consensus   C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C G
XO1HX22TF  ◀C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XO1GL75TR  ◀C A T G G C C T G A C C C C G G A C C A G G T G G T G G C C A T C
XO1G181TR  ▶C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XO1D260TF  ◀C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XO1EE84TR  ▶C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XO1GA32TF  ◀C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XO1BR60TR  ◀C A T G G C C T G A C C C C G G A C C A G G T G G T G G C C A T C
XO1I240TR  ◀C A T G G C C T G A C C C C G G A C C A G G T G G T G G C C A T C
XO1DK48TR  ▶C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
XOOA531TF  ▶C A T G G C C T G A C C C C G G A C C A G G T G G T G G C C A T C
XOOAF19TR  ▶C A T G G C C T G A C C C C G G A C C A G G T G A T G A C C A T C
```

Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114

Contig View



Position 116659

Contig ID 738

Chromo DB GB6

```
Consensus   CATGGCCTGACCCCGGACCAGGTGATGACCATCG
XO1HX22TF  ◄CATGGCCTGACCCCGGACCAGGTGATGACCATC
XO1GL75TR  ◄CATGGCCTGACCCCGGACCAGGTGGTGGCCATC
XO1G181TR  ►CATGGCCTGACCCCGGACCAGGTGATGACCATC
XO1D260TF  ◄CATGGCCTGACCCCGGACCAGGTGATGACCATC
XO1EE84TR  ►CATGGCCTGACCCCGGACCAGGTGATGACCATC
XO1GA32TF  ◄CATGGCCTGACCCCGGACCAGGTGATGACCATC
XO1BR60TR  ◄CATGGCCTGACCCCGGACCAGGTGGTGGCCATC
XO1I240TR  ◄CATGGCCTGACCCCGGACCAGGTGGTGGCCATC
XO1DK48TR  ►CATGGCCTGACCCCGGACCAGGTGATGACCATC
XOOA531TF  ►CATGGCCTGACCCCGGACCAGGTGGTGGCCATC
XOOAF19TR  ►CATGGCCTGACCCCGGACCAGGTGATGACCATC
```

Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114

Discrepancy Navigation

Contig Quick Select

Discrepancy

Regular Expression Consensus Search

Consensus & Position

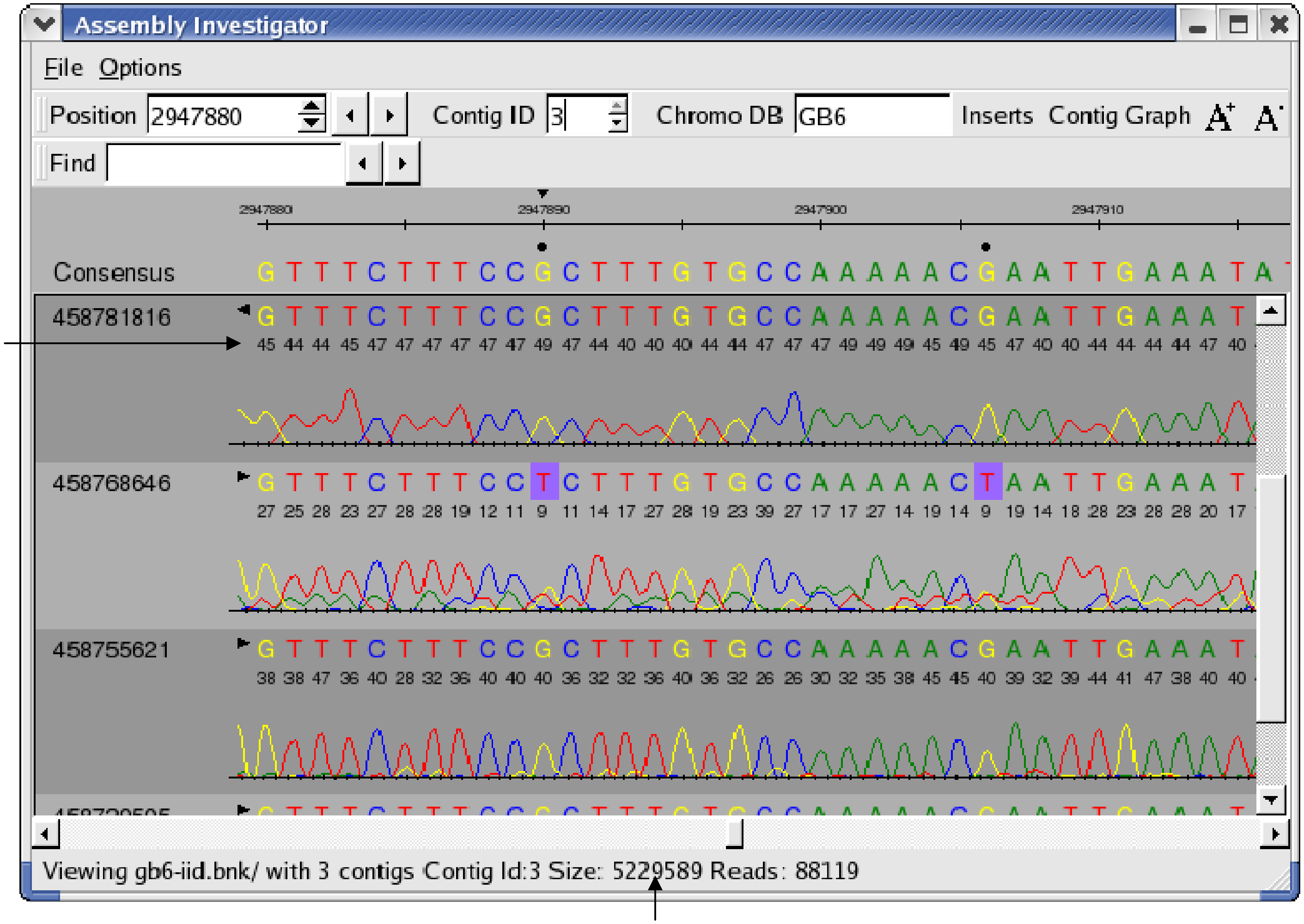
Scrollable Read Tiling

Summary

Read Orientation

Discrepancy Highlight

Contig View Expanded



```
          2947880                       2947890                       2947900                       2947910
Consensus  G  T  T  T  C  T  T  T  C  C  G  C  T  T  T  G  T  G  C  C  A  A  A  A  A  C  G  A  A  T  T  G  A  A  A  T  A
458781816  G  T  T  T  C  T  T  T  C  C  G  C  T  T  T  G  T  G  C  C  A  A  A  A  A  C  G  A  A  T  T  G  A  A  A  T
          45 44 44 45 47 47 47 47 47 47 49 47 44 40 40 40 44 44 47 47 47 49 49 49 45 49 45 47 40 40 44 44 44 44 47 40
458768646  G  T  T  T  C  T  T  T  C  C  T  C  T  T  T  G  T  G  C  C  A  A  A  A  A  C  T  A  A  T  T  G  A  A  A  T
          27 25 28 23 27 28 28 19 12 11  9 11 14 17 27 28 19 23 39 27 17 17 27 14 19 14  9 19 14 18 28 23 28 28 20 17
458755621  G  T  T  T  C  T  T  T  C  C  G  C  T  T  T  G  T  G  C  C  A  A  A  A  A  C  G  A  A  T  T  G  A  A  A  T
          38 38 47 36 40 28 32 36 40 40 40 36 32 32 36 40 36 32 26 26 30 32 35 38 45 45 40 39 32 39 44 41 47 38 40 40
```

Viewing gb6-iid.bnk/ with 3 contigs Contig Id:3 Size: 5229589 Reads: 88119

Quality Values

Normalized Chromatogram

No size restrictions

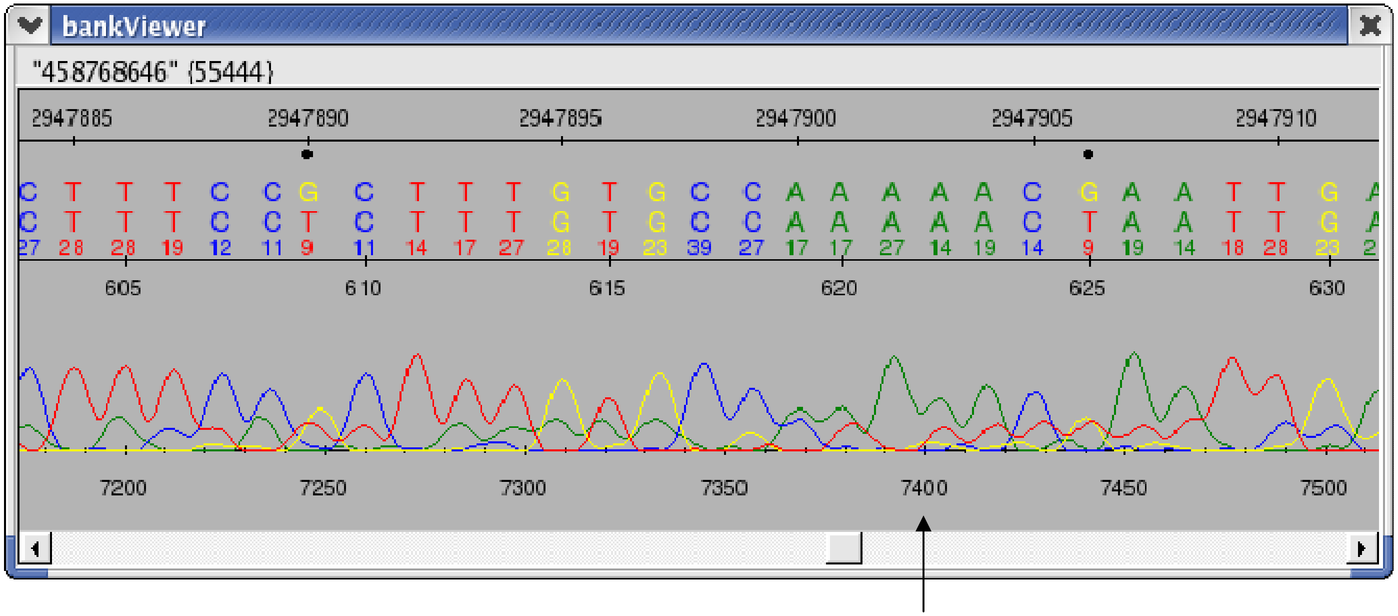
Chromatogram View

Read EID, IID

Consensus

Read

Raw Chromatogram

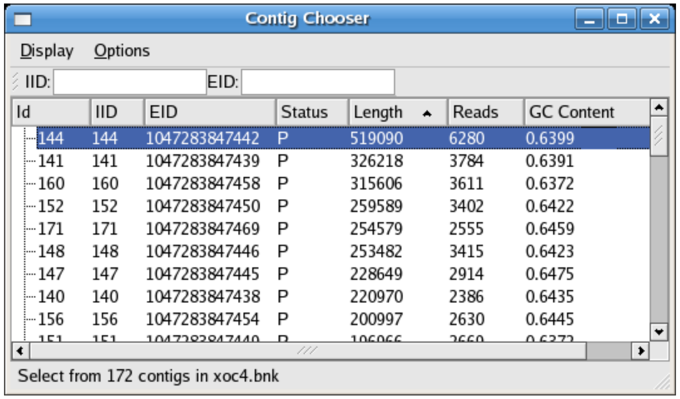


Chromatogram Position

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

Assembly Reports

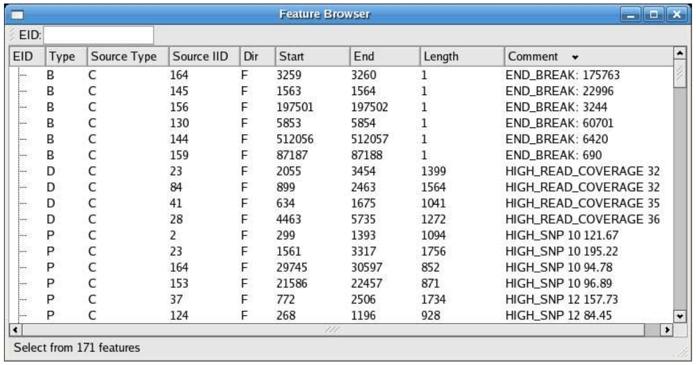
Contigs



Id	IID	EID	Status	Length	Reads	GC Content
144	144	1047283847442	P	519090	6280	0.6399
141	141	1047283847439	P	326218	3784	0.6391
160	160	1047283847458	P	315606	3611	0.6372
152	152	1047283847450	P	259589	3402	0.6422
171	171	1047283847469	P	254579	2555	0.6459
148	148	1047283847446	P	253482	3415	0.6423
147	147	1047283847445	P	228649	2914	0.6475
140	140	1047283847438	P	220970	2386	0.6435
156	156	1047283847454	P	200997	2630	0.6445

Select from 172 contigs in xoc4.bnk

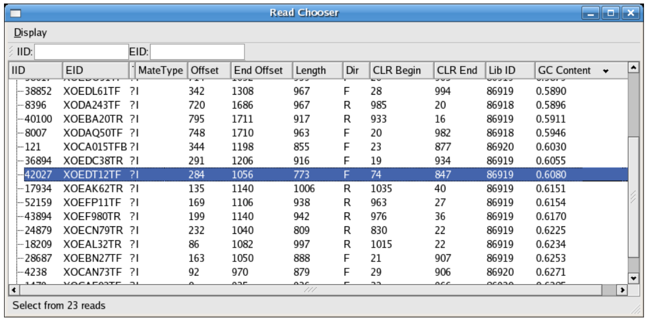
Features



EID	Type	Source Type	Source IID	Dir	Start	End	Length	Comment
	B	C	164	F	3259	3260	1	END_BREAK: 175763
	B	C	145	F	1563	1564	1	END_BREAK: 22996
	B	C	156	F	197501	197502	1	END_BREAK: 3244
	B	C	130	F	5853	5854	1	END_BREAK: 60701
	B	C	144	F	512056	512057	1	END_BREAK: 6420
	B	C	159	F	87187	87188	1	END_BREAK: 690
	D	C	23	F	2055	3454	1399	HIGH_READ_COVERAGE 32
	D	C	84	F	899	2463	1564	HIGH_READ_COVERAGE 32
	D	C	41	F	634	1675	1041	HIGH_READ_COVERAGE 35
	D	C	28	F	4463	5735	1272	HIGH_READ_COVERAGE 36
	P	C	2	F	299	1393	1094	HIGH_SNP 10 121.67
	P	C	23	F	1561	3317	1756	HIGH_SNP 10 195.22
	P	C	164	F	29745	30597	852	HIGH_SNP 10 94.78
	P	C	153	F	21586	22457	871	HIGH_SNP 10 96.89
	P	C	37	F	772	2506	1734	HIGH_SNP 12 157.73
	P	C	124	F	268	1196	928	HIGH_SNP 12 84.45

Select from 171 features

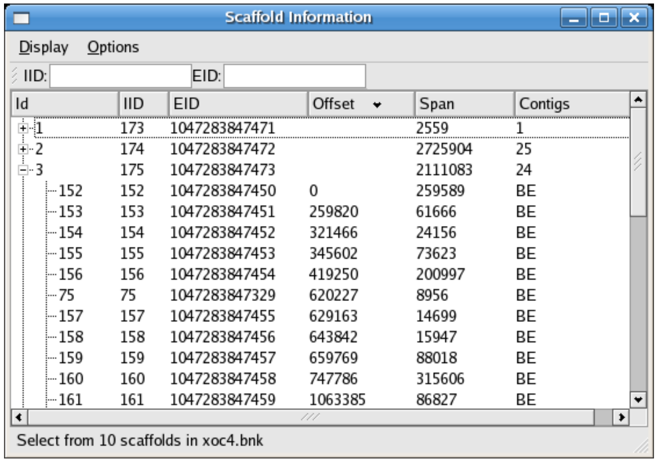
Reads



IID	EID		MateType	Offset	End Offset	Length	Dir	CLR Begin	CLR End	Lib ID	GC Content
38852	XOEDL61TF	?	I	342	1308	967	F	28	994	86919	0.5890
8396	XODA243TF	?	I	720	1686	967	R	985	20	86918	0.5896
40100	XOEBA20TR	?	I	795	1711	917	R	933	16	86919	0.5911
8007	XODAQ50TF	?	I	748	1710	963	F	20	982	86918	0.5946
121	XOCA015TFB	?	I	344	1198	855	F	23	877	86920	0.6030
36894	XOEDC38TR	?	I	291	1206	916	F	19	934	86919	0.6055
42027	XOEDT12TF	?	I	284	1056	773	F	74	847	86919	0.6080
17934	XOEAK62TR	?	I	135	1140	1006	R	1035	40	86919	0.6151
52159	XOEFP11TF	?	I	169	1106	938	R	963	27	86919	0.6154
43894	XOEF980TR	?	I	199	1140	942	R	976	36	86919	0.6170
24879	XOECN79TR	?	I	232	1040	809	R	830	22	86919	0.6225
18209	XOEAL32TR	?	I	86	1082	997	R	1015	22	86919	0.6234
28687	XOEBN27TF	?	I	163	1050	888	F	21	907	86919	0.6253
4238	XOCAN73TF	?	I	92	970	879	F	29	906	86920	0.6271

Select from 23 reads

Scaffolds



Id	IID	EID	Offset	Span	Contigs
1	173	1047283847471		2559	1
2	174	1047283847472		2725904	25
3	175	1047283847473		2111083	24
152	152	1047283847450	0	259589	BE
153	153	1047283847451	259820	61666	BE
154	154	1047283847452	321466	24156	BE
155	155	1047283847453	345602	73623	BE
156	156	1047283847454	419250	200997	BE
75	75	1047283847329	620227	8956	BE
157	157	1047283847455	629163	14699	BE
158	158	1047283847456	643842	15947	BE
159	159	1047283847457	659769	88018	BE
160	160	1047283847458	747786	315606	BE
161	161	1047283847459	1063385	86827	BE

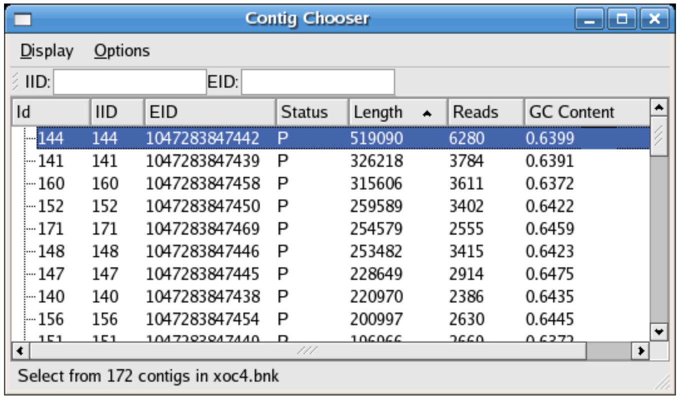
Select from 10 scaffolds in xoc4.bnk

- Full Integration: “Double click takes you there”

Assembly Reports

Misassembly Walkthough:
Correlated SNPs

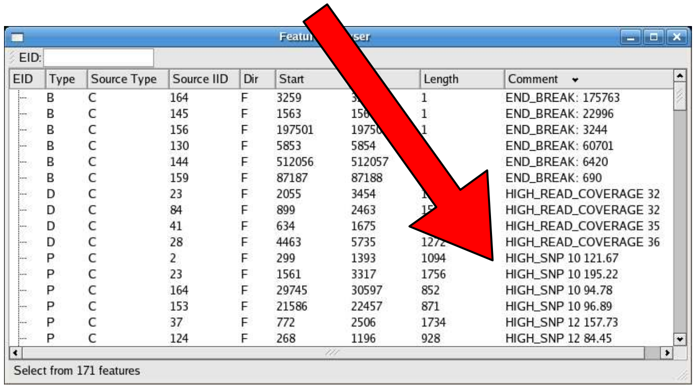
Contigs



Id	IID	EID	Status	Length	Reads	GC Content
144	144	1047283847442	P	519090	6280	0.6399
141	141	1047283847439	P	326218	3784	0.6391
160	160	1047283847458	P	315606	3611	0.6372
152	152	1047283847450	P	259589	3402	0.6422
171	171	1047283847469	P	254579	2555	0.6459
148	148	1047283847446	P	253482	3415	0.6423
147	147	1047283847445	P	228649	2914	0.6475
140	140	1047283847438	P	220970	2386	0.6435
156	156	1047283847454	P	200997	2630	0.6445

Select from 172 contigs in xoc4.bnk

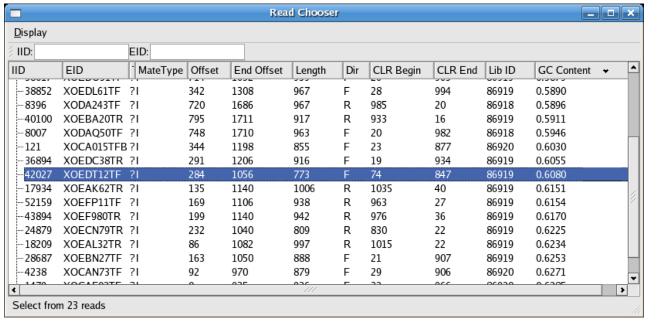
Features



EID	Type	Source Type	Source IID	Dir	Start		Length	Comment
	B	C	164	F	3259	[illegible]	1	END_BREAK: 175763
	B	C	145	F	1563	[illegible]	1	END_BREAK: 22996
	B	C	156	F	197501	[illegible]	1	END_BREAK: 3244
	B	C	130	F	5853	5854	[illegible]	END_BREAK: 60701
	B	C	144	F	512056	512057	[illegible]	END_BREAK: 6420
	B	C	159	F	87187	87188	[illegible]	END_BREAK: 690
	D	C	23	F	2055	3454	[illegible]	HIGH_READ_COVERAGE 32
	D	C	84	F	899	2463	[illegible]	HIGH_READ_COVERAGE 32
	D	C	41	F	634	1675	[illegible]	HIGH_READ_COVERAGE 35
	D	C	28	F	4463	5735	1272	HIGH_READ_COVERAGE 36
	P	C	2	F	299	1393	1094	HIGH_SNP 10 121.67
	P	C	23	F	1561	3317	1756	HIGH_SNP 10 195.22
	P	C	164	F	29745	30597	852	HIGH_SNP 10 94.78
	P	C	153	F	21586	22457	871	HIGH_SNP 10 96.89
	P	C	37	F	772	2506	1734	HIGH_SNP 12 157.73
	P	C	124	F	268	1196	928	HIGH_SNP 12 84.45

Select from 171 features

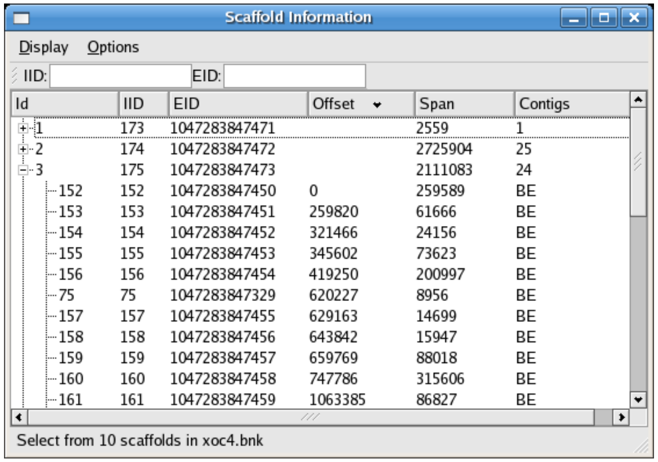
Reads



IID	EID	MateType	Offset	End Offset	Length	Dir	CLR Begin	CLR End	Lib ID	GC Content
38852	XOEDL61TF	?I	342	1308	967	F	28	994	86919	0.5890
8396	XODA243TF	?I	720	1686	967	R	985	20	86918	0.5896
40100	XOEBA20TR	?I	795	1711	917	R	933	16	86919	0.5911
8007	XODAQ50TF	?I	748	1710	963	F	20	982	86918	0.5946
121	XOCA015TFB	?I	344	1198	855	F	23	877	86920	0.6030
36894	XOEDC38TR	?I	291	1206	916	F	19	934	86919	0.6055
42027	XOEDT12TF	?I	284	1056	773	F	74	847	86919	0.6080
17934	XOEAK62TR	?I	135	1140	1006	R	1035	40	86919	0.6151
52159	XOEFP11TF	?I	169	1106	938	R	963	27	86919	0.6154
43894	XOEF980TR	?I	199	1140	942	R	976	36	86919	0.6170
24879	XOECN79TR	?I	232	1040	809	R	830	22	86919	0.6225
18209	XOEAL32TR	?I	86	1082	997	R	1015	22	86919	0.6234
28687	XOEBN27TF	?I	163	1050	888	F	21	907	86919	0.6253
4238	XOCAN73TF	?I	92	970	879	F	29	906	86920	0.6271

Select from 23 reads

Scaffolds



Id	IID	EID	Offset	Span	Contigs
1	173	1047283847471		2559	1
2	174	1047283847472		2725904	25
3	175	1047283847473		2111083	24
152	152	1047283847450	0	259589	BE
153	153	1047283847451	259820	61666	BE
154	154	1047283847452	321466	24156	BE
155	155	1047283847453	345602	73623	BE
156	156	1047283847454	419250	200997	BE
75	75	1047283847329	620227	8956	BE
157	157	1047283847455	629163	14699	BE
158	158	1047283847456	643842	15947	BE
159	159	1047283847457	659769	88018	BE
160	160	1047283847458	747786	315606	BE
161	161	1047283847459	1063385	86827	BE

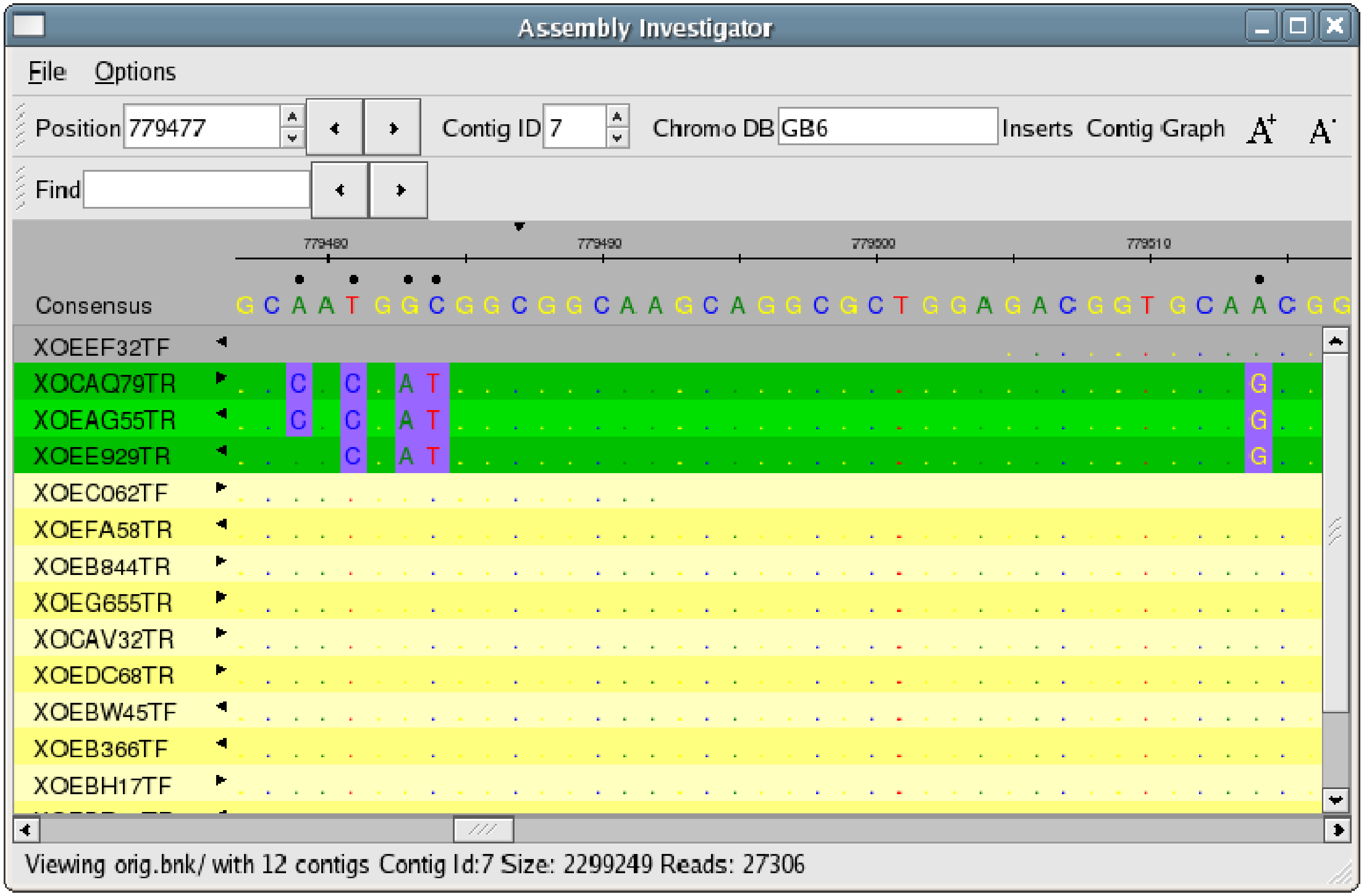
Select from 10 scaffolds in xoc4.bnk

- Full Integration: “Double click takes you there”

SNP View

SNP Sorted Reads

Polymorphism View



Position 779477

Find

Contig ID 7

Chromo DB GB6

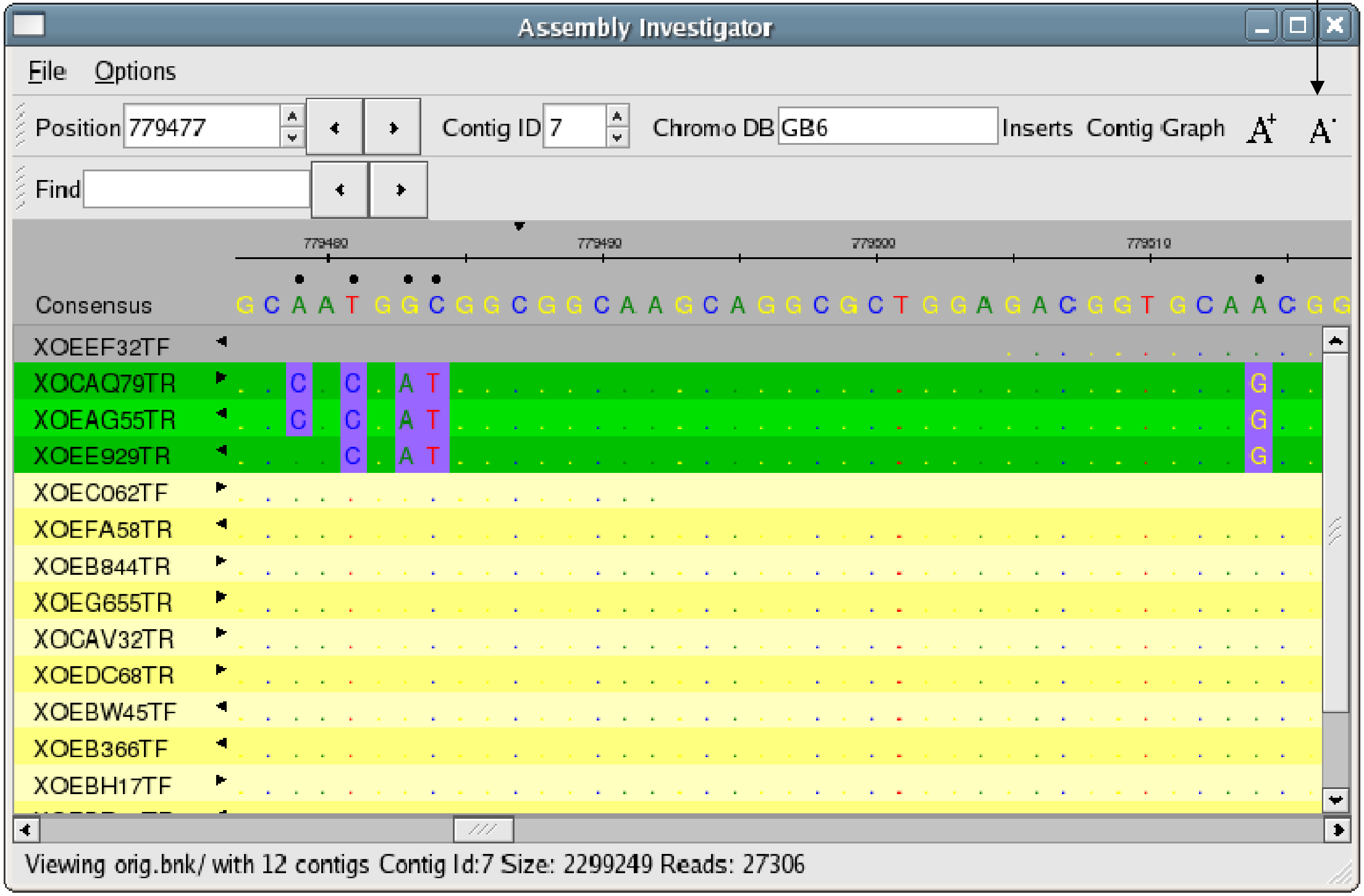
Viewing orig.bnk/ with 12 contigs Contig Id:7 Size: 2299249 Reads: 27306

SNP View

Zoom Out

SNP Sorted Reads

Polymorphism View



Position 779477

Contig ID 7

Chromo DB GB6

Find

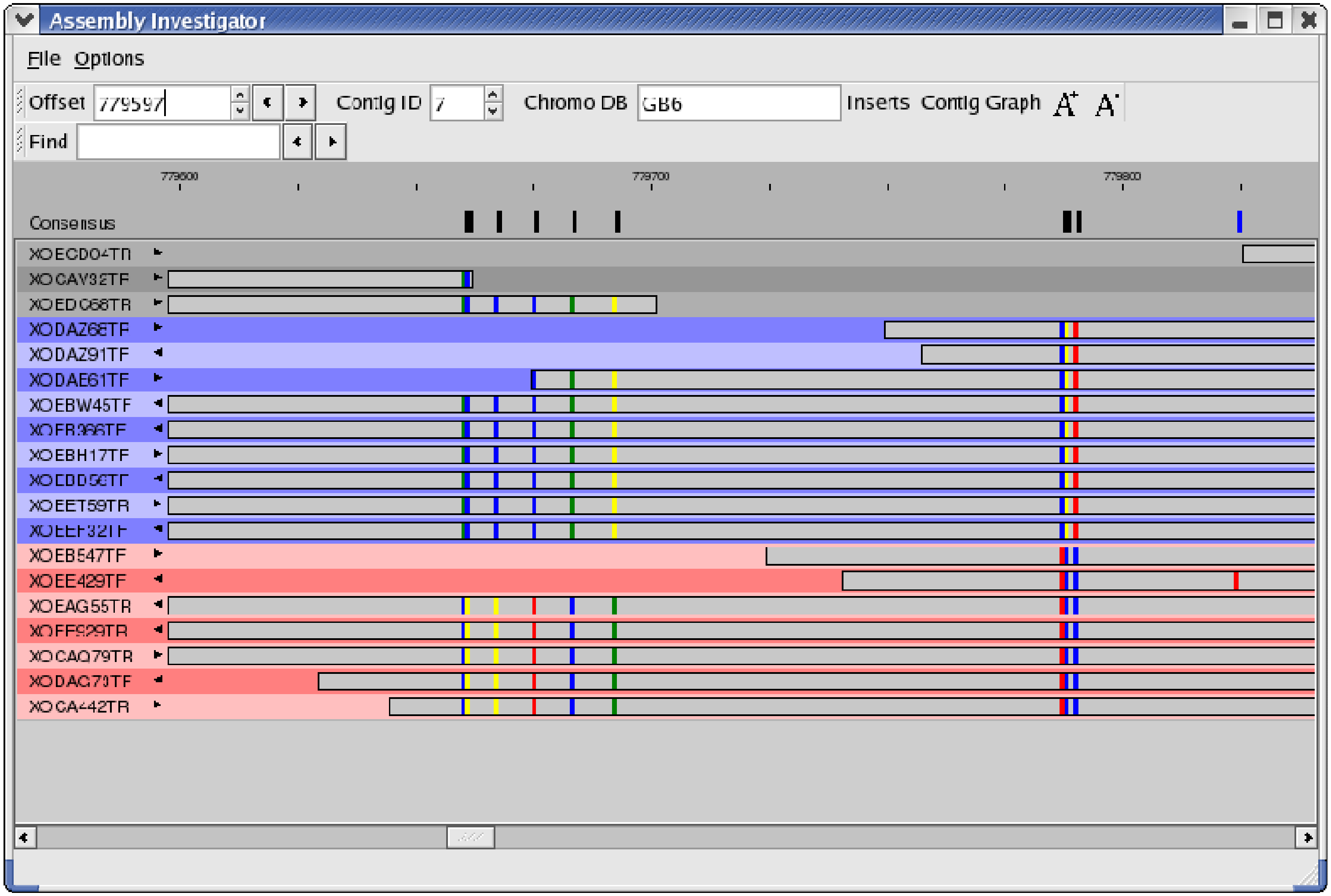
Consensus G C A A T G G C G G C G G C A A G C A G G C G C T G G A G A C G G T G C A A C G G

XOEEF32TF
XOCAQ79TR C C A T G
XOEAG55TR C C A T G
XOEE929TR C A T G
XOEC062TF
XOEFA58TR
XOEB844TR
XOEG655TR
XOCAV32TR
XOEDC68TR
XOEBW45TF
XOEB366TF
XOEBH17TF

Viewing orig.bnk/ with 12 contigs Contig Id:7 Size: 2299249 Reads: 27306

SNP Barcode

SNP Sorted Reads



Offset 779597

Contig ID 7

Chromo DB GB6

Find

Colored Rectangle indicate the positions and composition of the SNPs

Scaffold View

Coverage

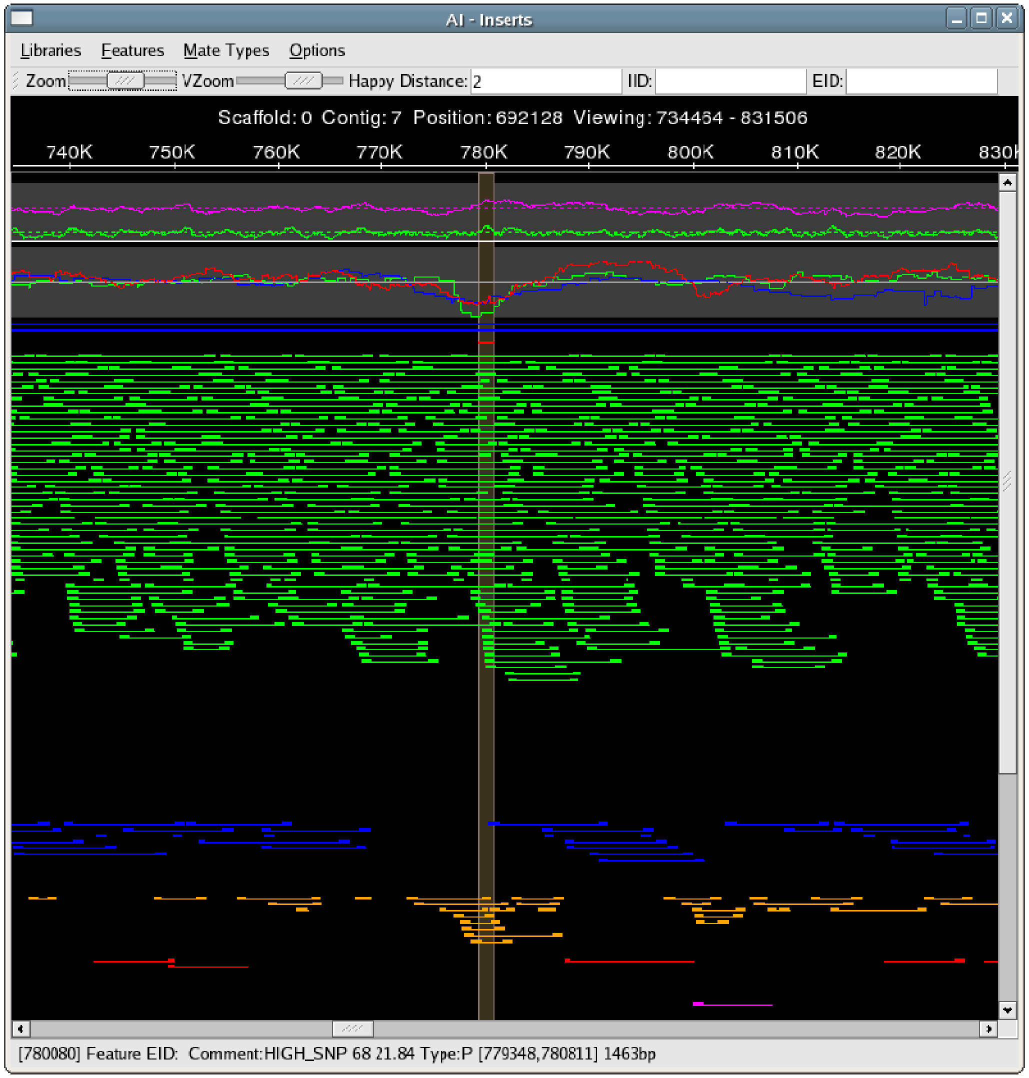
CE Statistic

Happy

Stretched

Compressed

Misoriented



Zoom VZoom Happy Distance: 2 IID: EID:

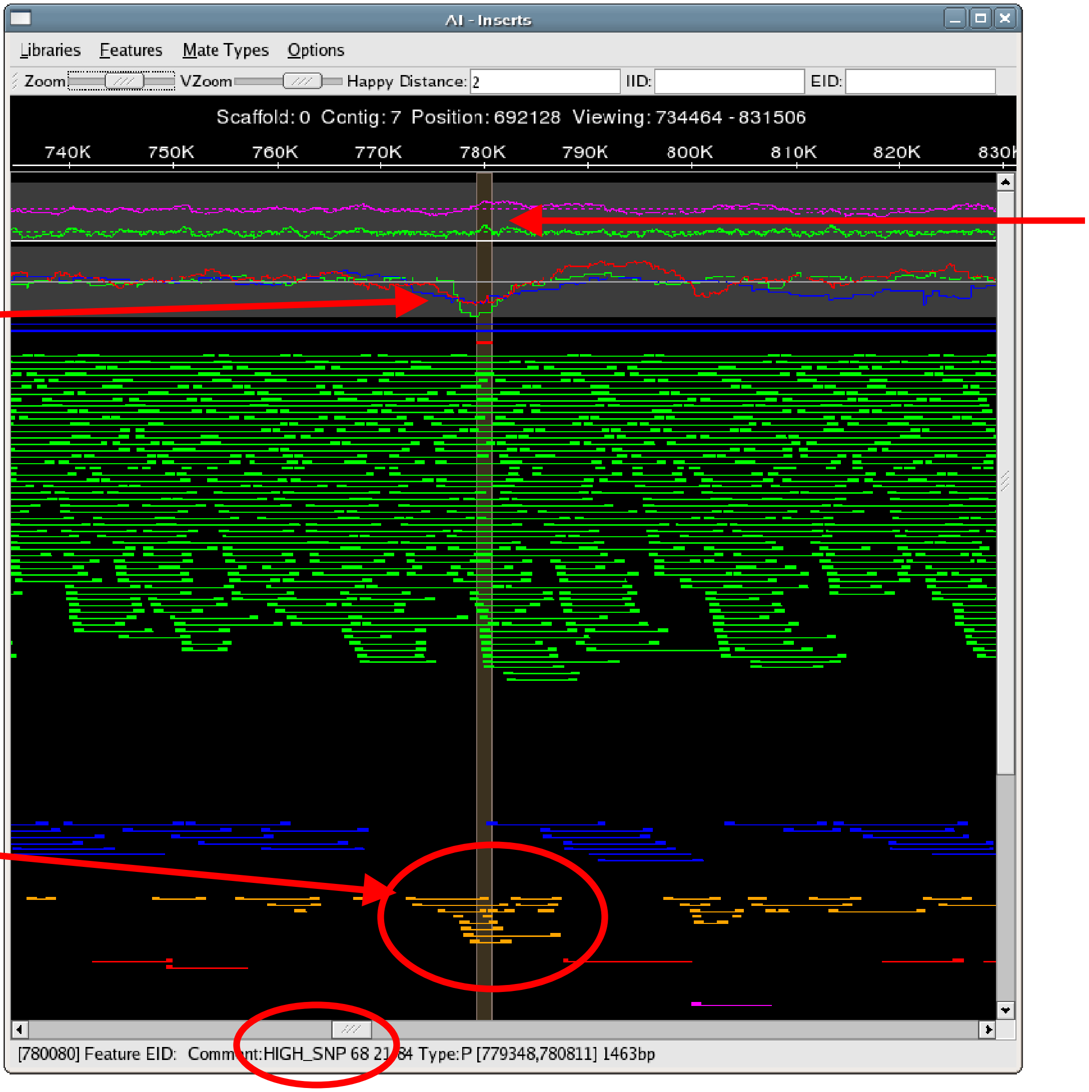
Scaffold: 0 Contig: 7 Position: 692128 Viewing: 734464 - 831506

[780080] Feature EID: Comment:HIGH_SNP 68 21.84 Type:P [779348,780811] 1463bp

SNP Feature

Linking

Collapsed Repeat



Zoom VZoom Happy Distance: 2 IID: EID:

Scaffold: 0 Contig: 7 Position: 692128 Viewing: 734464 - 831506

[780080] Feature EID: Comm nt:HIGH_SNP 68 21 84 Type:P [779348,780811] 1463bp

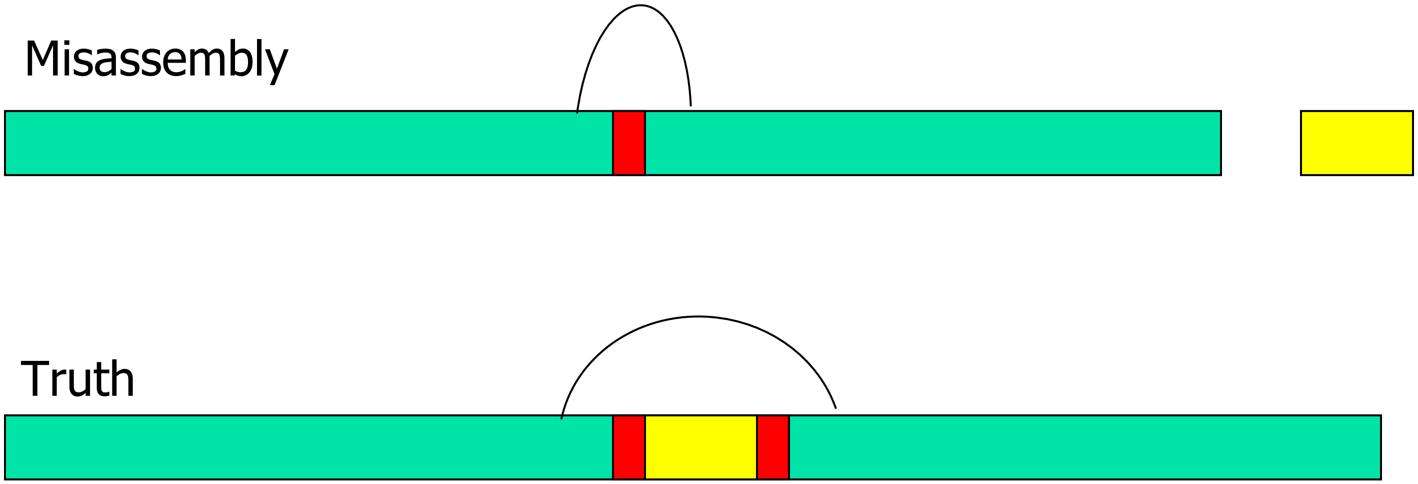
Read Coverage Spike

-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

Fixing collapsed repeats with AMOS

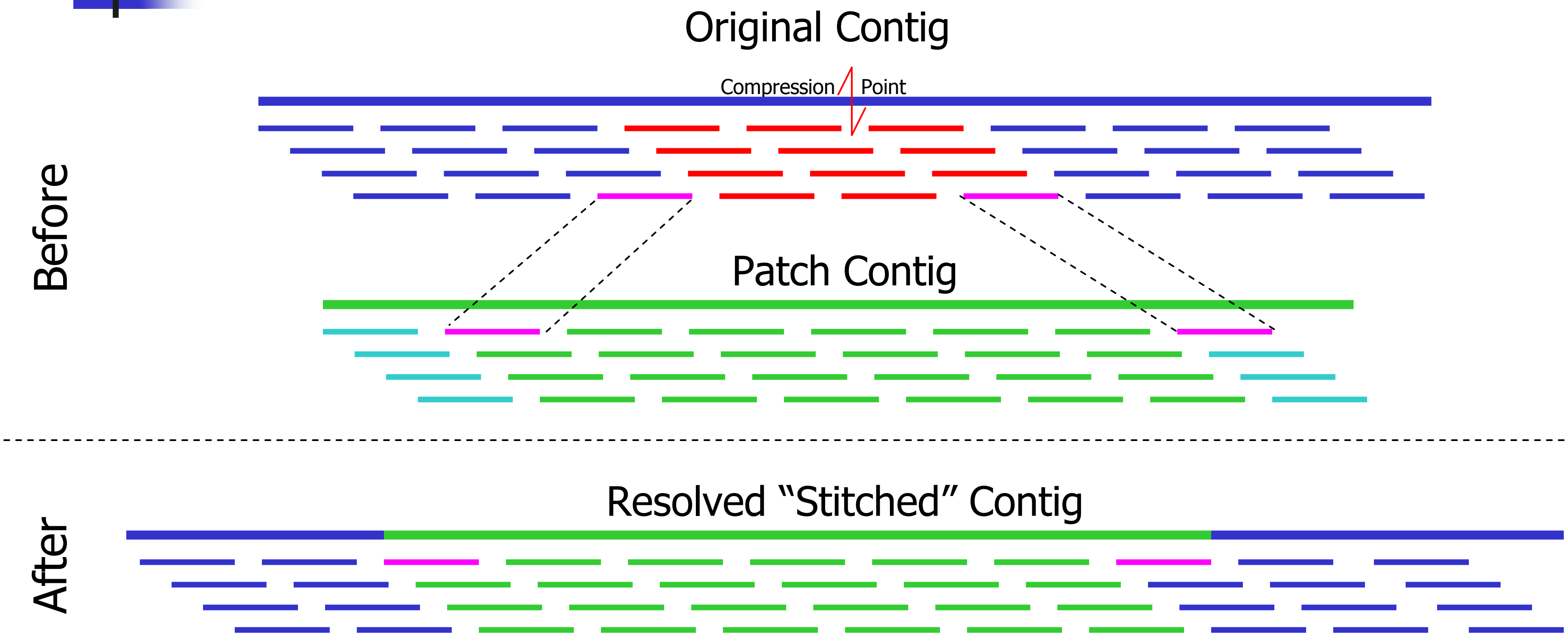
1. Select reads and mates in region of collapse.
   - AMOS: findMissingMates, select-reads

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