



A  
M  
o  
s

A large, stylized, black-outlined letter "A" is positioned above the letters "M", "o", and "s". The letters "M", "o", and "s" are stacked vertically, with "M" on top, "o" in the middle, and "s" on the bottom. All three letters have a thick, black, brushstroke-like outline.

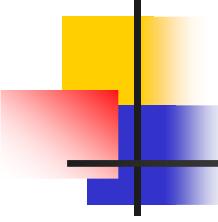
# AMOS Assembly Validation and Visualization

---

Michael Schatz

Center for Bioinformatics and Computational Biology  
University of Maryland

August 13, 2006  
University of Hawaii



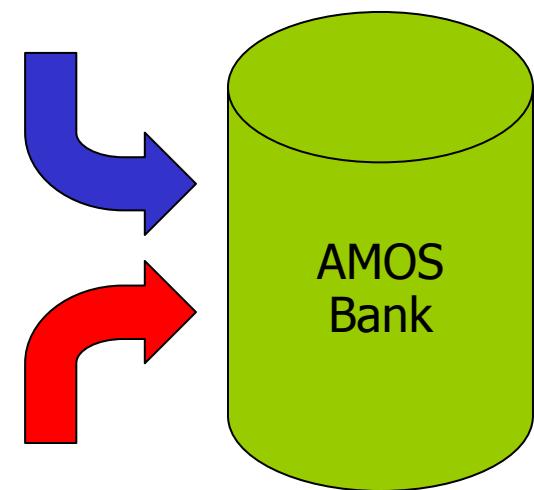
# Outline

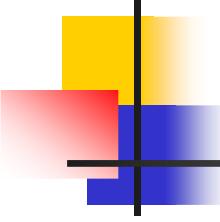
---

- 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

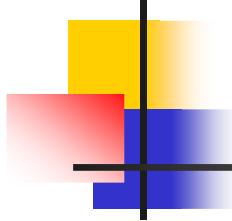
- Automatically scan an assembly to locate misassembly signatures for further analysis and correction
- **cav validate** prefix (.frg, .asm)
  1. Load CA Assembly Data into Bank
  2. Evaluate Mate Pairs & Libraries
  3. Evaluate Read Alignments
  4. Evaluate Read Breakpoints
  5. Analyze Depth of Coverage
  6. List Surrogates
  7. Load Misassembly Signatures into Bank
- **amosvalidate** prefix (.afg)
  - Same as **cav validate**, except skips surrogates





# Mate-Happiness: asmQC

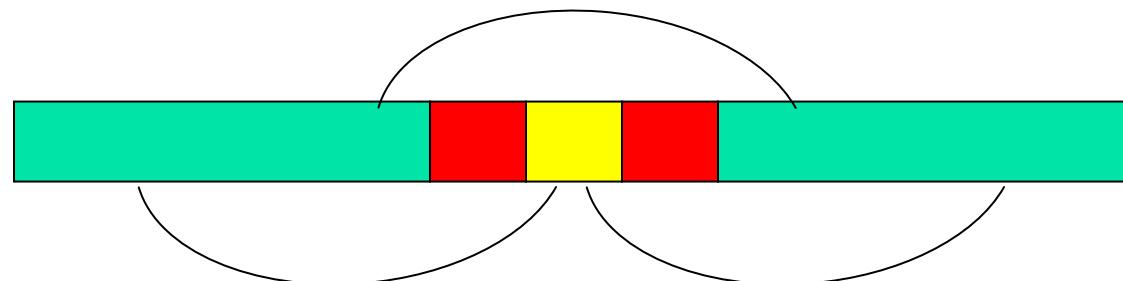
- Evaluate mate “happiness” across assembly
  - Happy = Correct orientation and distance
- Finds regions with multiple:
  - Compressed Mates
  - Expanded Mates
  - Invalid same orientation ( $\rightarrow \rightarrow$ )
  - Invalid outie orientation ( $\leftarrow \rightarrow$ )
  - Missing Mates
    -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



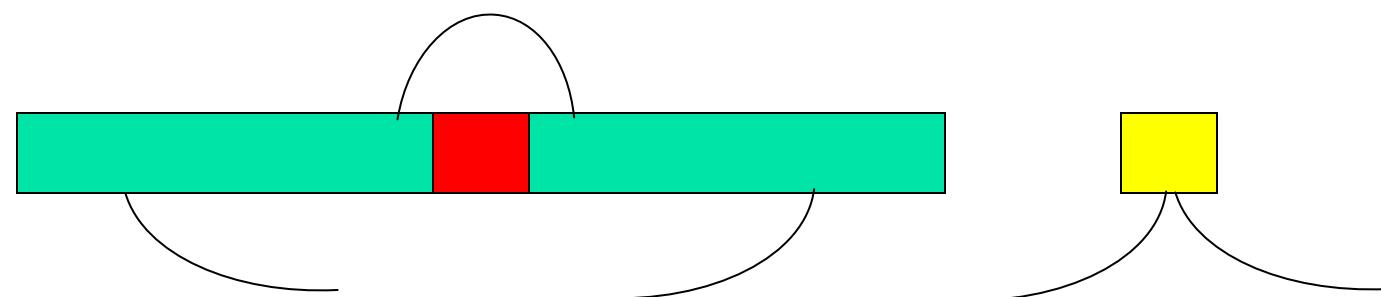
# Mate-Happiness: asmQC

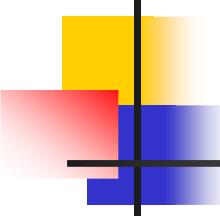
- Excision: Skip reads between flanking repeats

- Truth



- Misassembly: Compressed Mates, Missing Mates

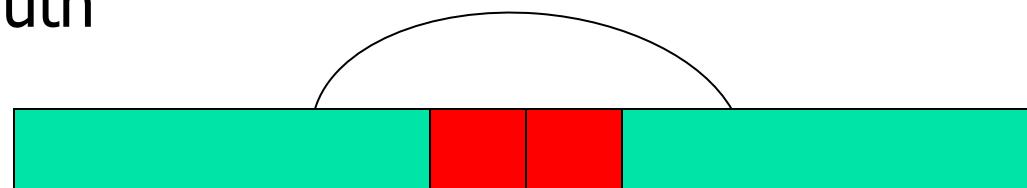




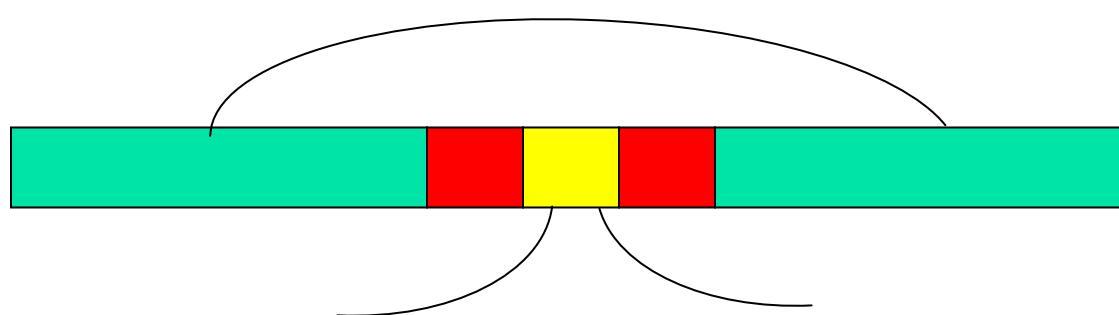
# Mate-Happiness: asmQC

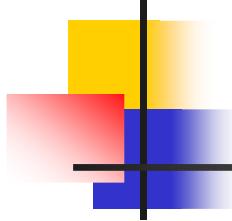
- Insertion: Additional reads between flanking repeats

- Truth



- Misassembly: Expanded Mates, Missing Mates

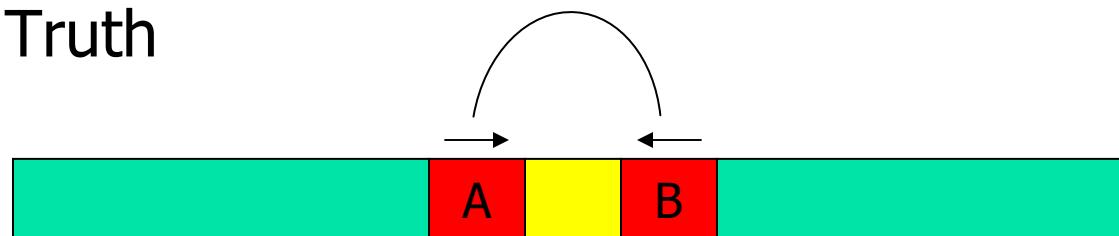




# Mate-Happiness: asmQC

- Rearrangement: Reordering of reads

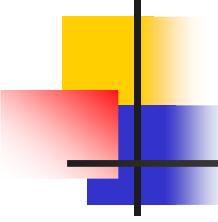
- Truth



- Misassembly: Misoriented Mates



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

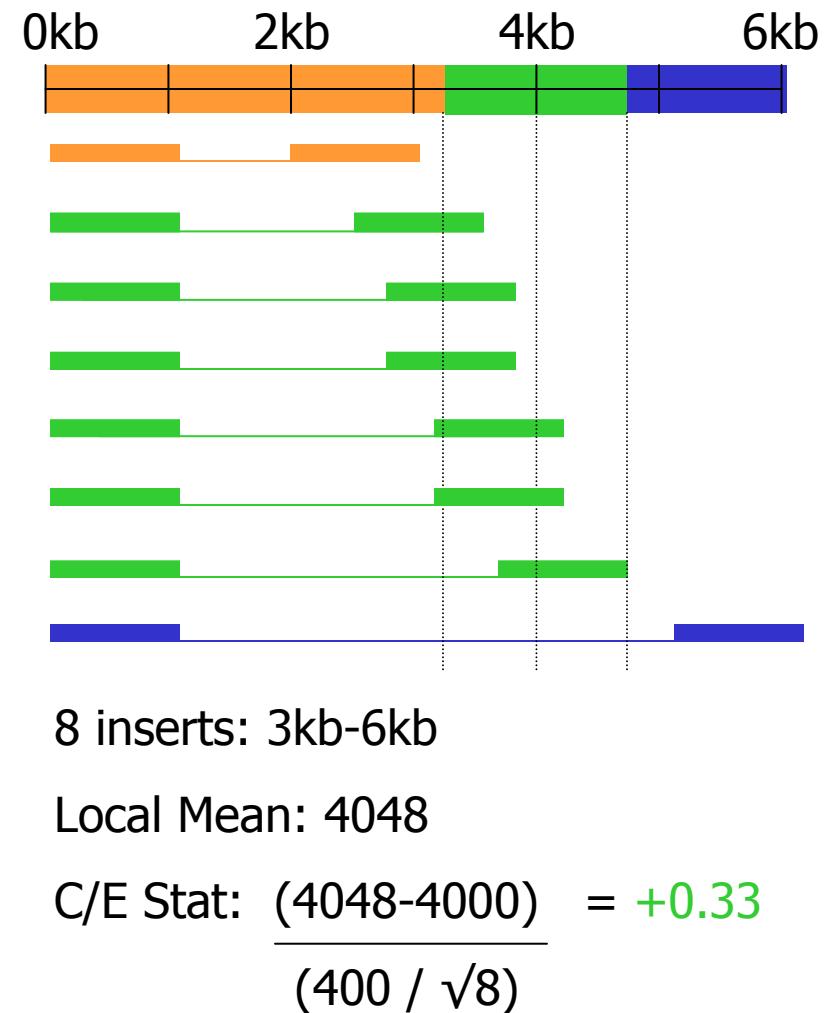
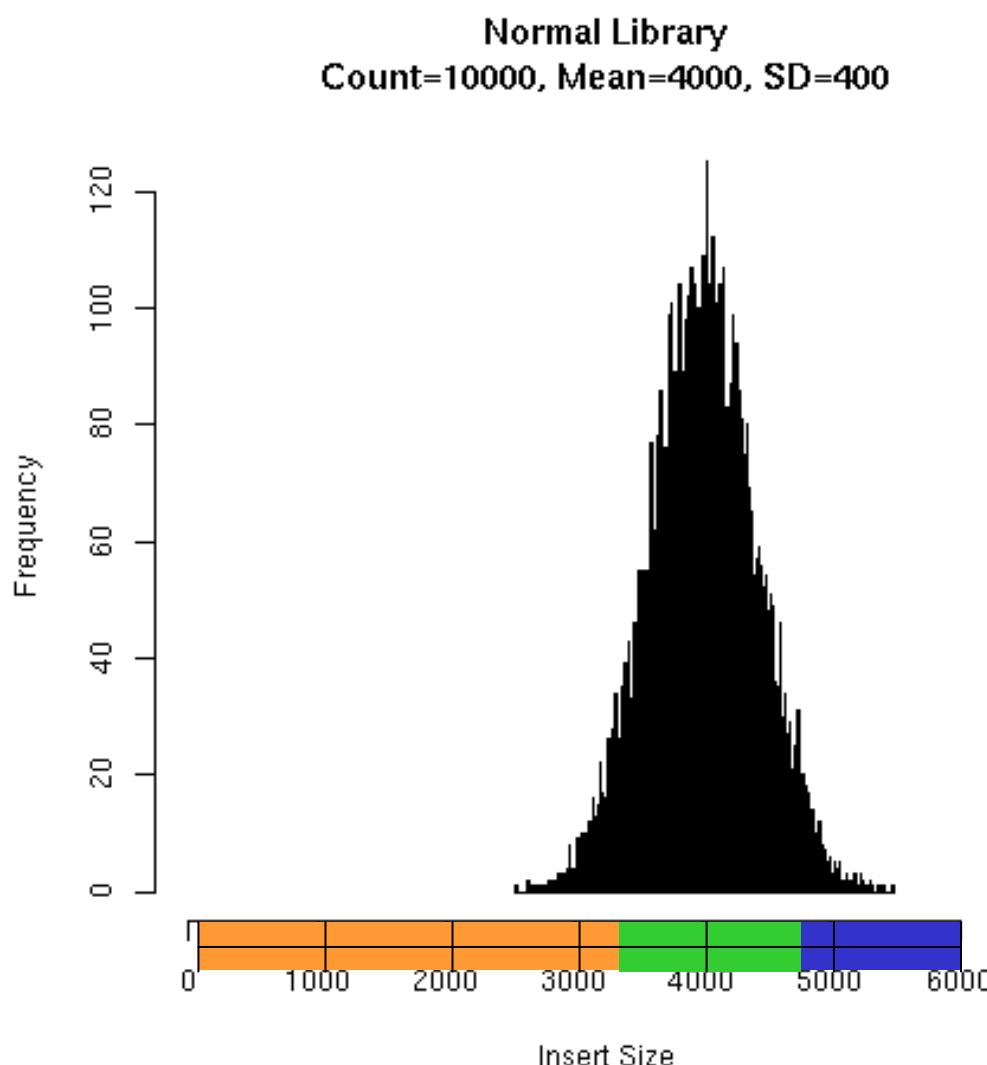


# C/E Statistic

---

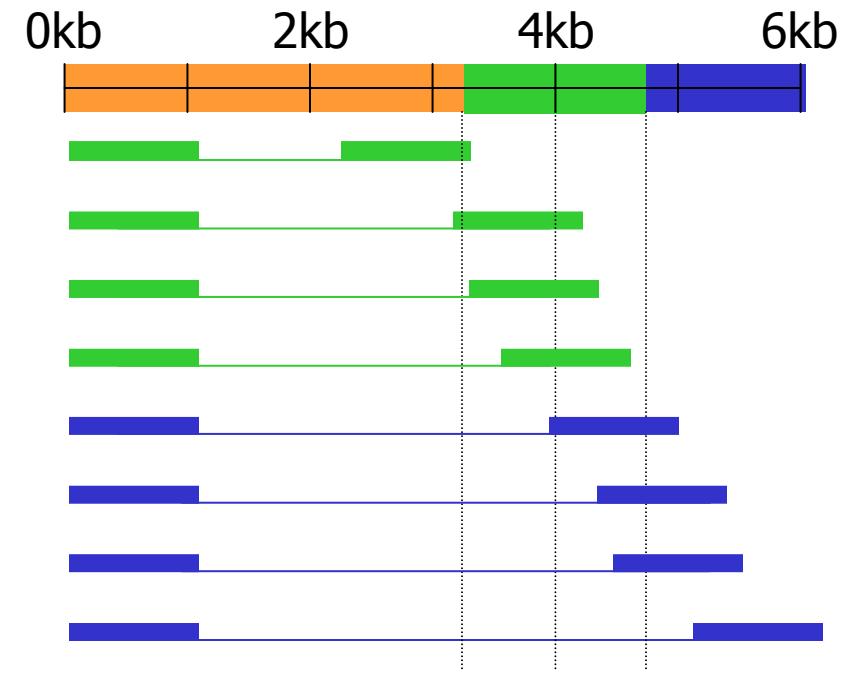
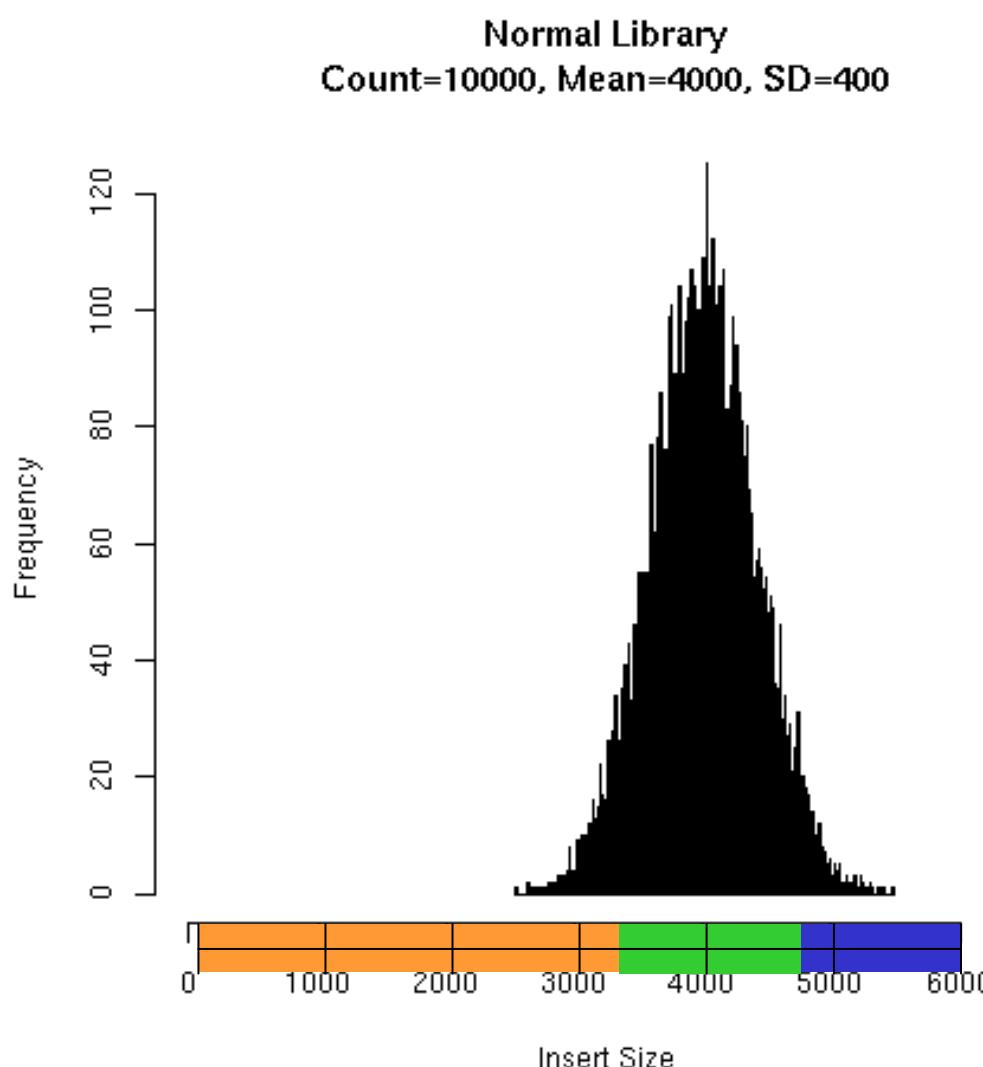
- The presence of individual compressed or expanded mates is rare but expected.
- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential misassemblies
  - Even if each individual mate is “happy”
- Compute the statistic at all positions
 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

# Sampling the Genome



Near 0 indicates overall happiness

# C/E-Statistic: Expansion

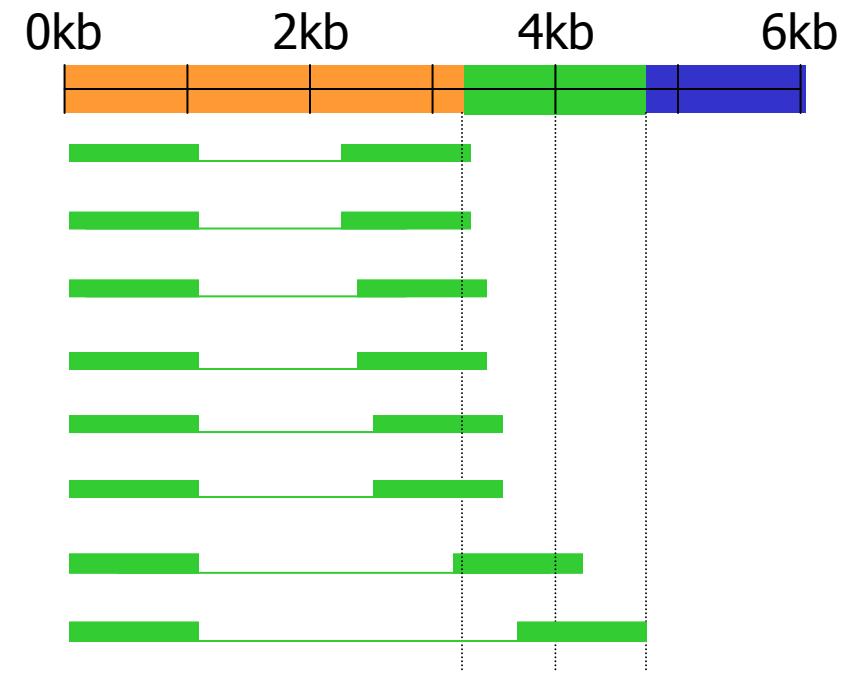
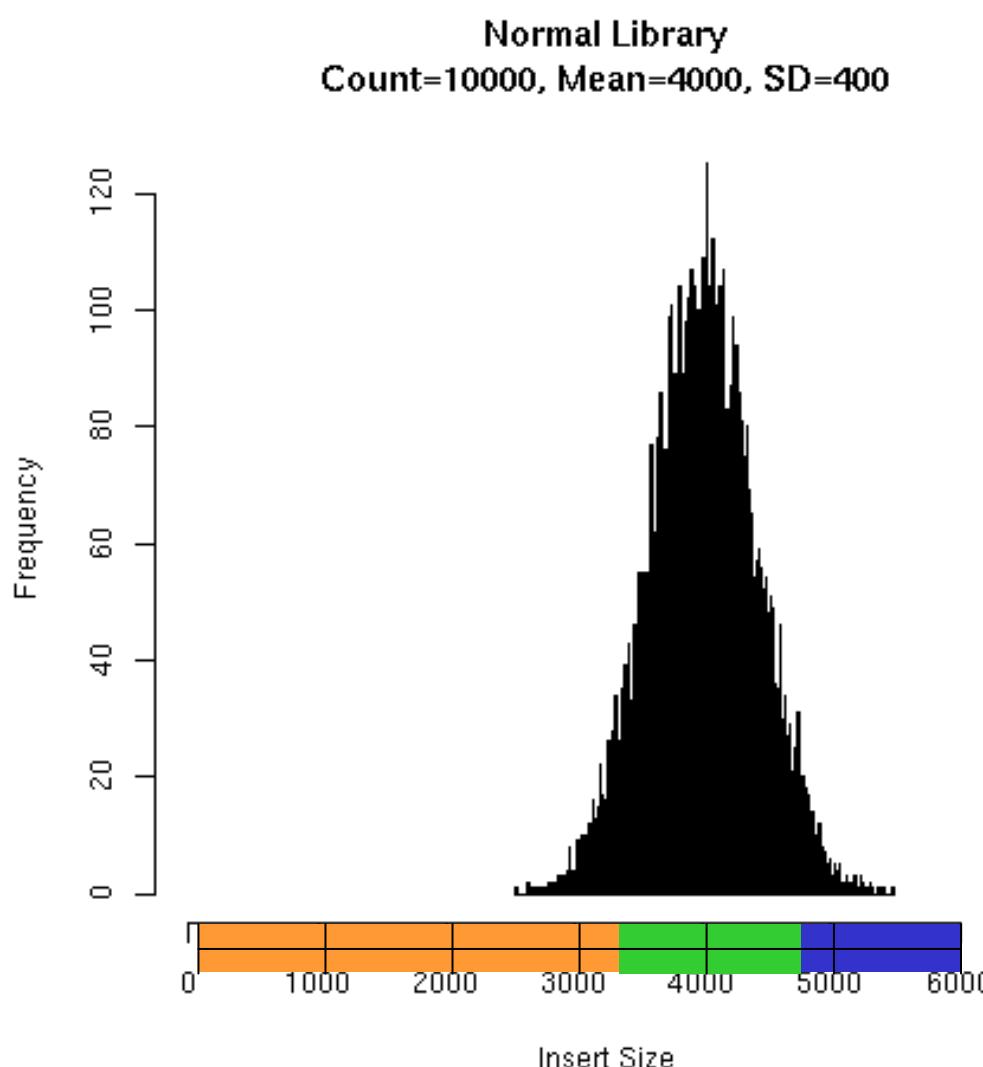


Local Mean: 4461

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

C/E Stat  $\geq 3.0$  indicates Expansion

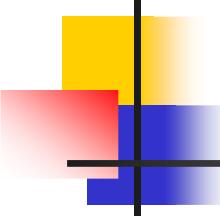
# C/E-Statistic: Compression



Local Mean: 3488

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

C/E Stat  $\leq -3.0$  indicates Compression



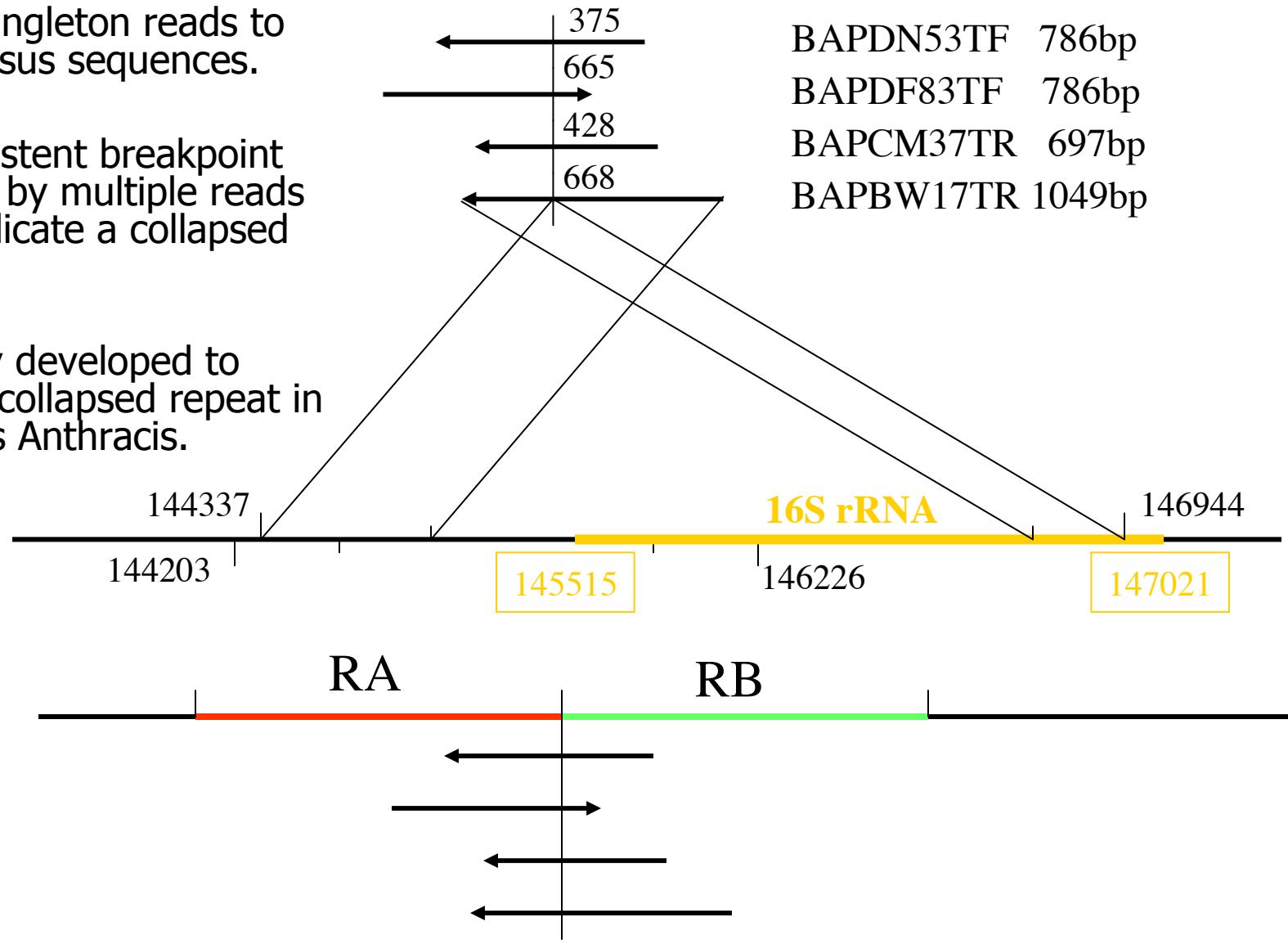
# Read Alignment

- 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  $\geq 40$  (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

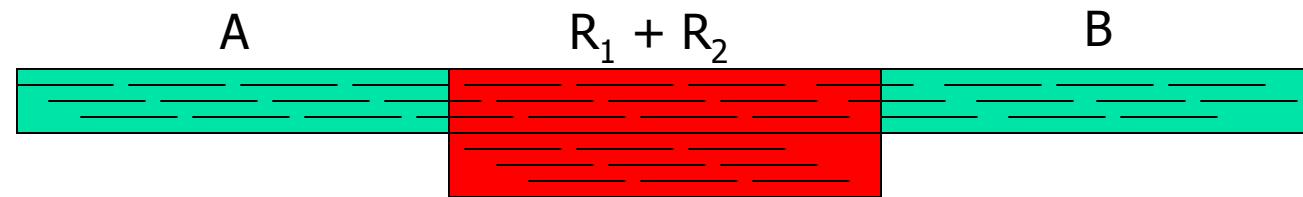
# Read Breakpoints

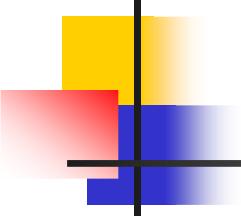
- Align singleton reads to consensus sequences.
- A consistent breakpoint shared by multiple reads can indicate a collapsed repeat.
- Initially developed to detect collapsed repeat in *Bacillus Anthracis*.



# 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

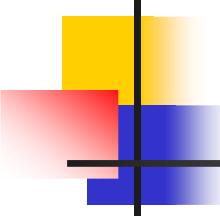




Hawkeye

---





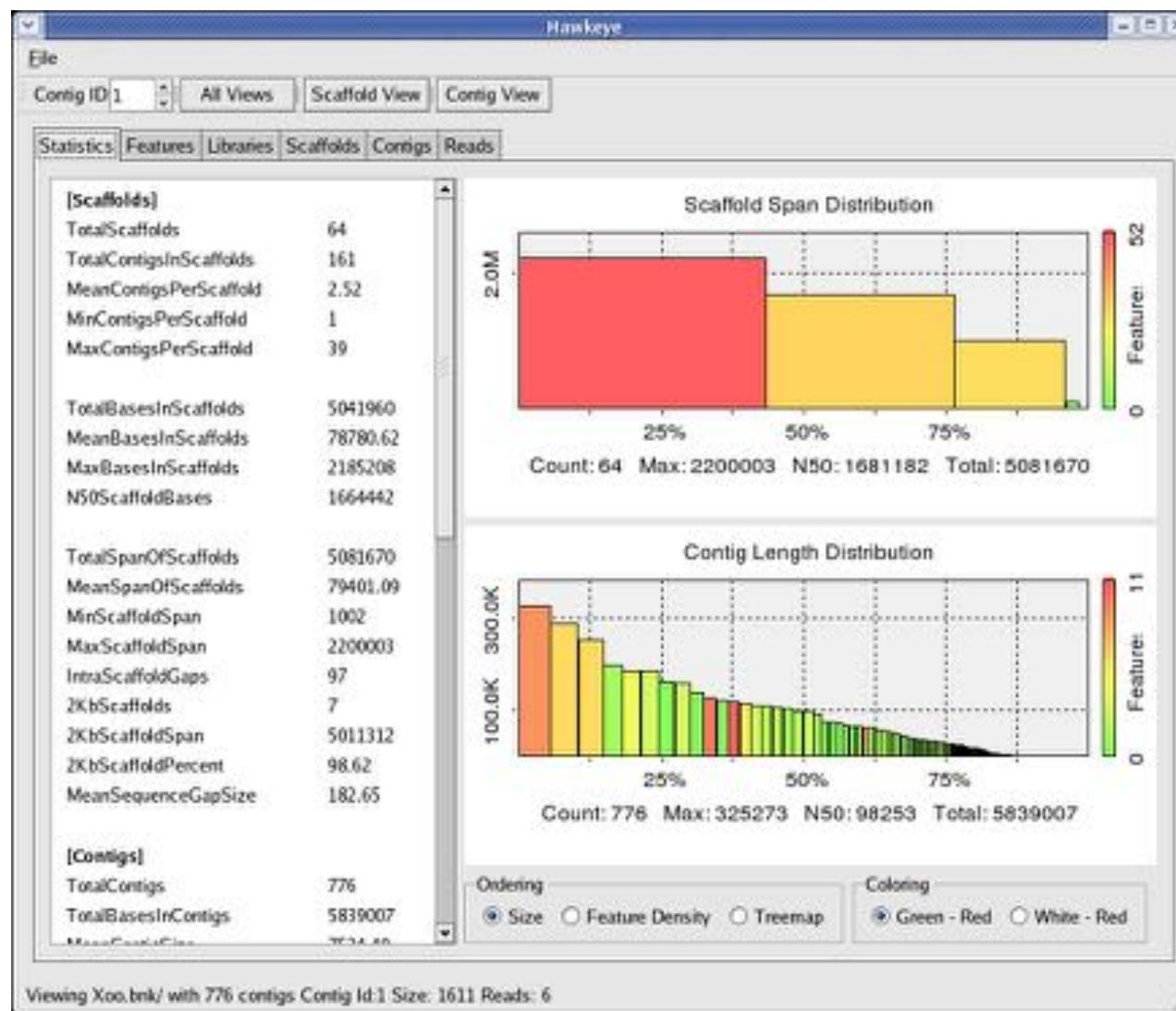
# Hawkeye Goals

---

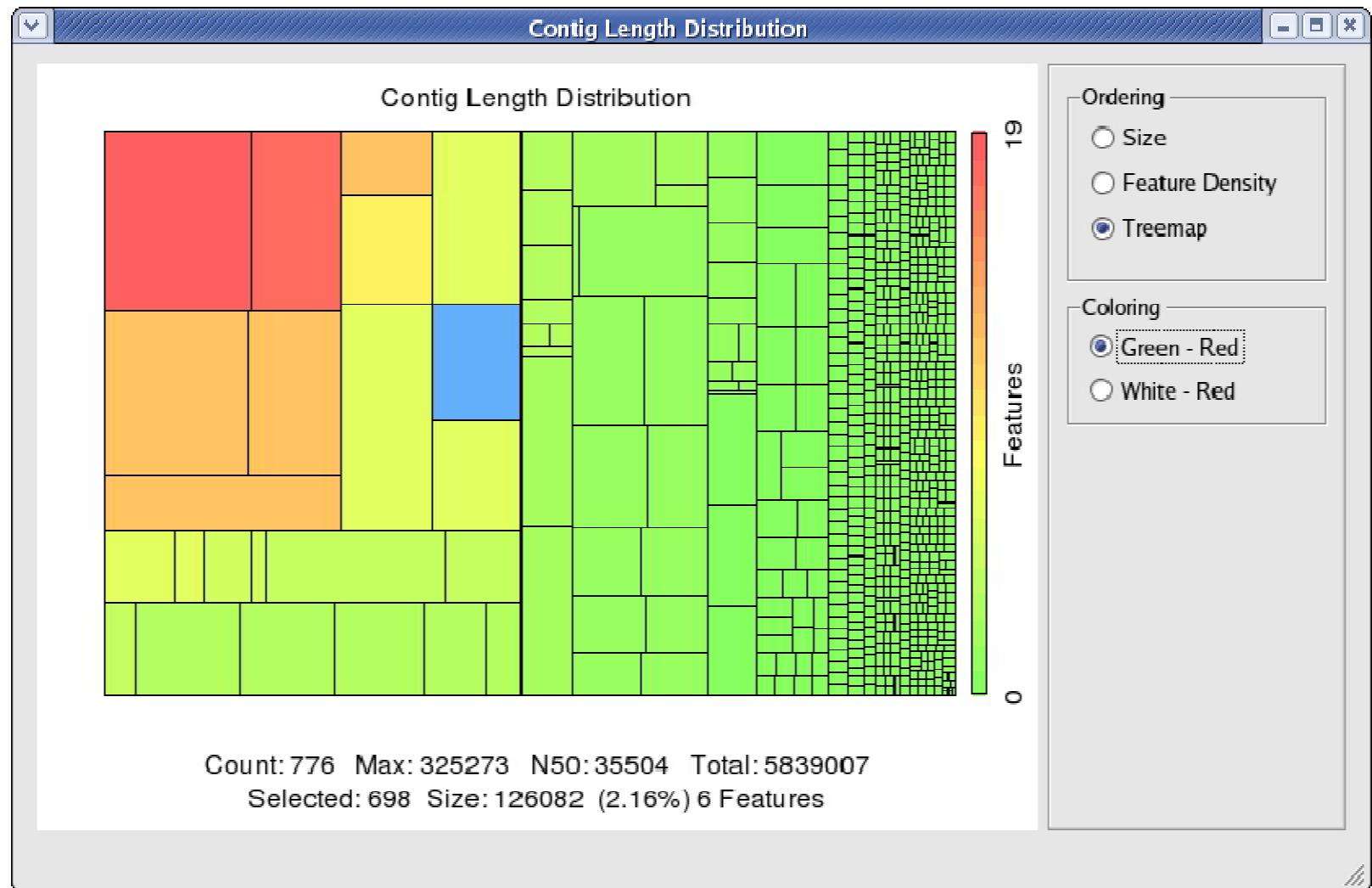
Interactively explore and analyze

- 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

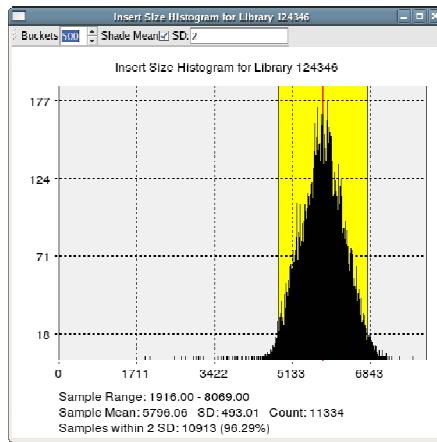


# Contig Length Distribution

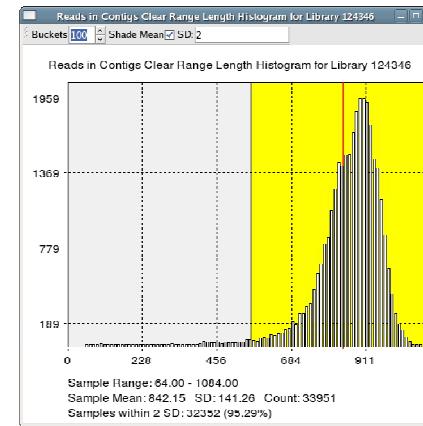


# Histograms & Statistics

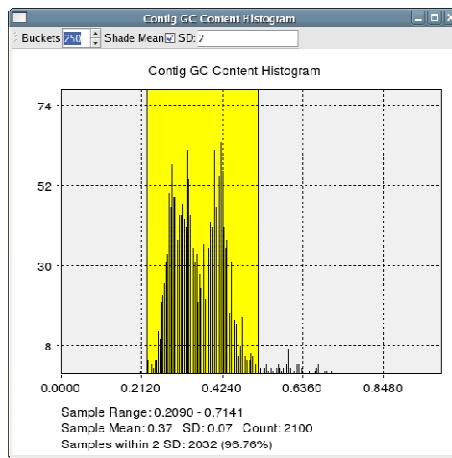
Insert  
Size



Read  
Length



GC  
Content



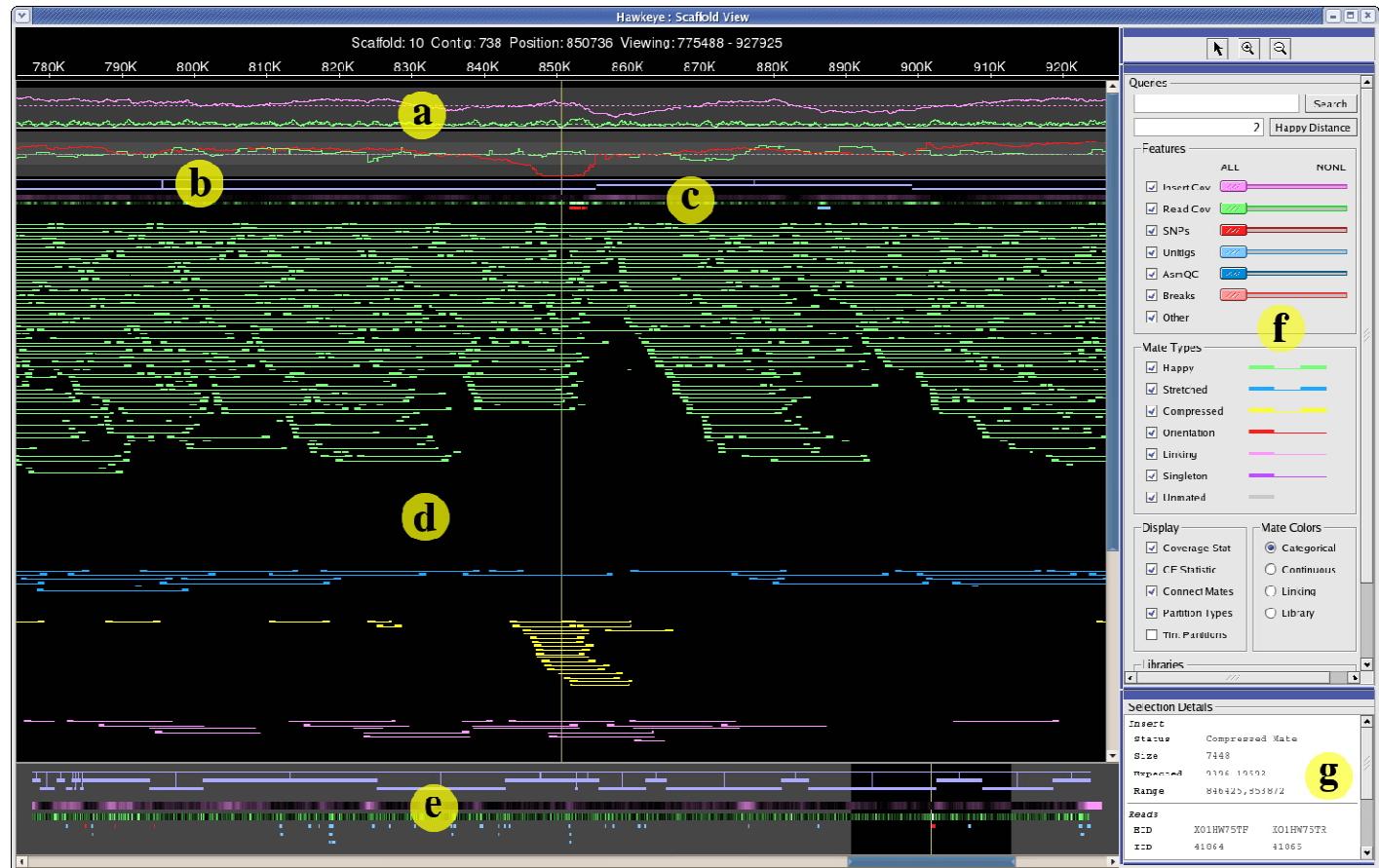
Overall  
Statistics

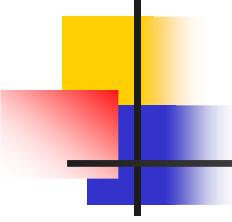
Field	Value
[Scaffolds]	
TotalScaffolds	1076
TotalContigsInScaffolds	1396
MeanContigsPerScaffold	1.30
MinContigsPerScaffold	1
MaxContigsPerScaffold	15
TotalBasesInScaffolds	7511900
MeanBasesPerScaffold	6981.12
MaxBasesInScaffolds	279040
N50ScaffoldBases	75935
TotalSpanOfScaffolds	780540
MeanSpanOfScaffolds	7253.29
MinTotalSpan	1007
MaxScaffoldSpan	285705
IntraScaffoldGaps	320
2KbScaffolds	200
2KbScaffoldSpan	6264092
2KbScaffoldPercent	32.82
MeanSequenceGapSize	355.37
[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





# Standard Feature Types

## [B] Breakpoint

Alignment ends at this position

Loading Features:

```
$ loadFeatures bankname featfile
```

## [C] Coverage

Location of unusual mate coverage (asmQC)

Featfile format:

```
Contigid type end5 end3 comment
```

## [S] SNPs

Location of Correlated SNPs

## [U] Unitig

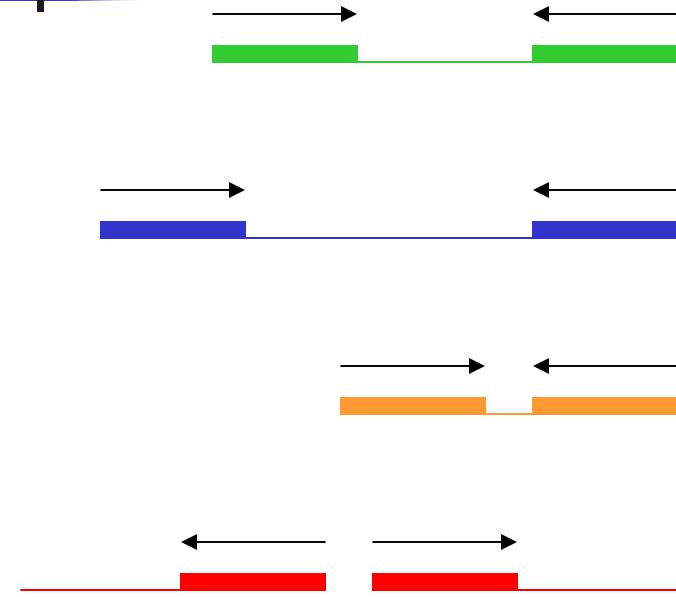
Used to report location of surrogate unitigs in CA assemblies

## [X] Other

All other Features

# Insert Happiness

Both mates present



## Happy

- Oriented Correctly &&
- $|Insert\ Size - Library.mean| \leq Happy-Distance * Library.sd$

## Stretched

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

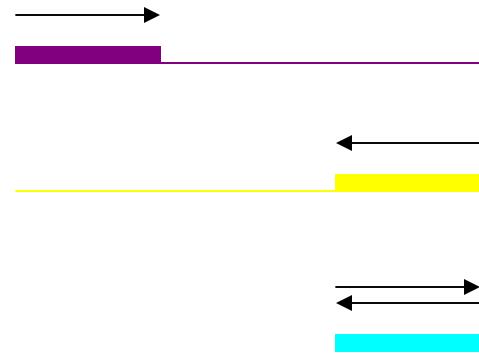
## Compressed

- 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

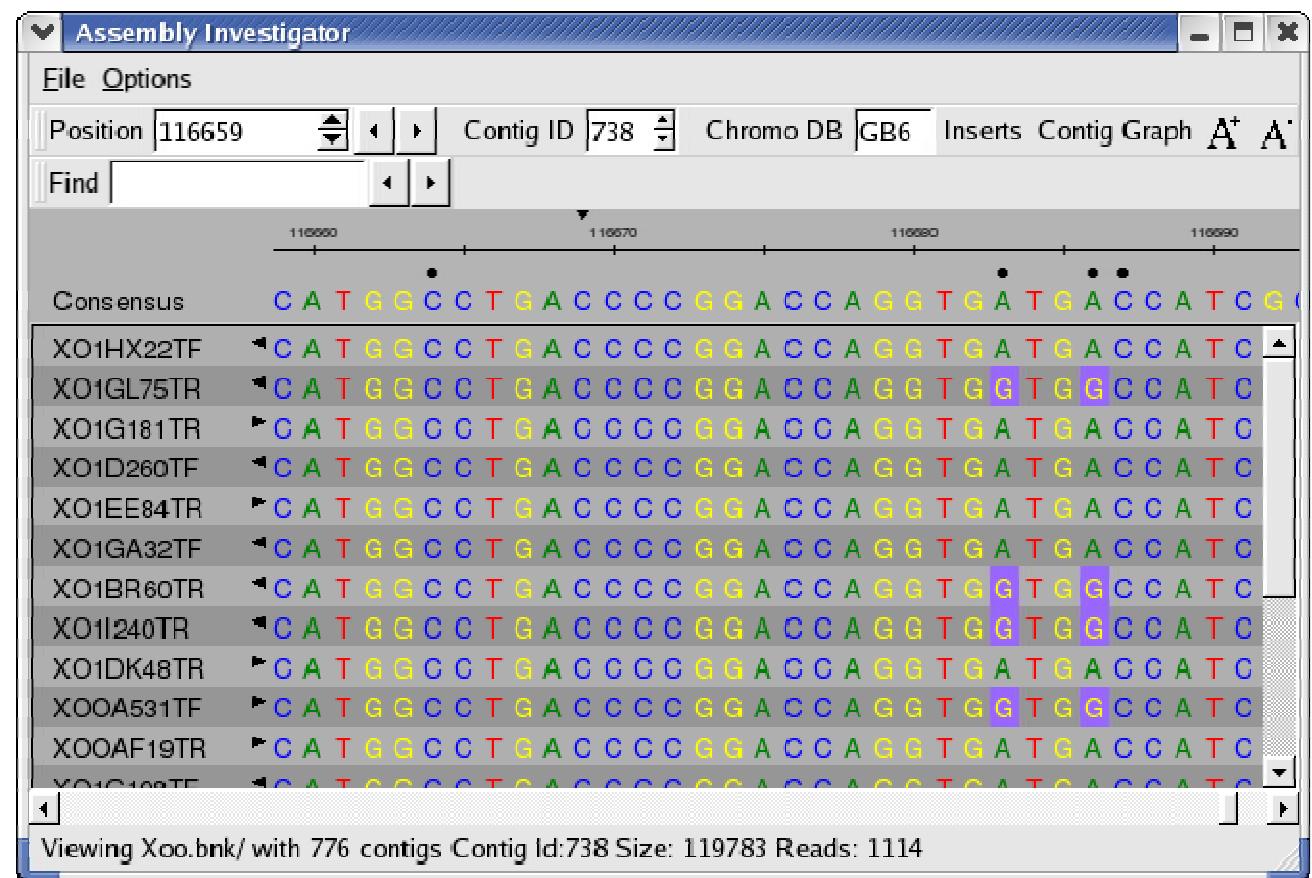
## Singleton

- Read's mate is a singleton

## Unmated

- No mate was provided for read

# Contig View



# Contig View

Discrepancy Navigation Contig Quick Select Discrepancy

Regular Expression Consensus Search

Consensus & Position

Scorable Read Tiling

Summary

Read Orientation

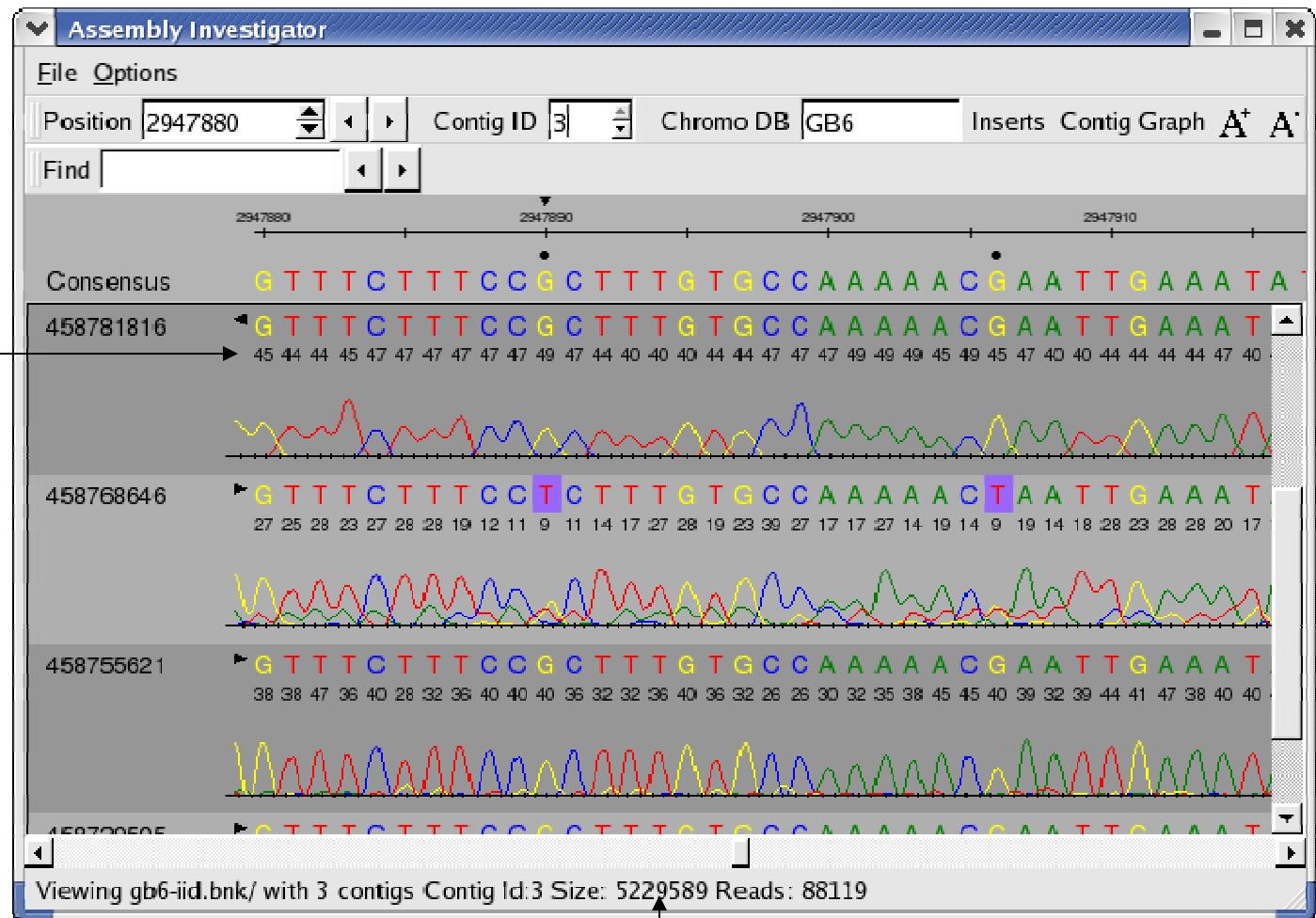
Discrepancy Highlight

The screenshot shows the Assembly Investigator software interface. At the top, there's a menu bar with 'File' and 'Options'. Below the menu is a toolbar with buttons for 'Position' (set to 116659), 'Contig ID' (set to 738), 'Chromo DB' (set to GB6), 'Inserts', 'Contig Graph' (with two A+ icons), and 'Find' (with left and right arrows). The main window displays a consensus sequence at the top, followed by a list of reads. Each read is shown as a colored sequence of letters (A, T, C, G) with a small arrowhead pointing to its start position on the consensus. The reads are grouped into several horizontal blocks, indicating they overlap or are part of the same tiling. The bottom of the window shows a status bar with the text: 'Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114'.

# Contig View Expanded

Quality Values

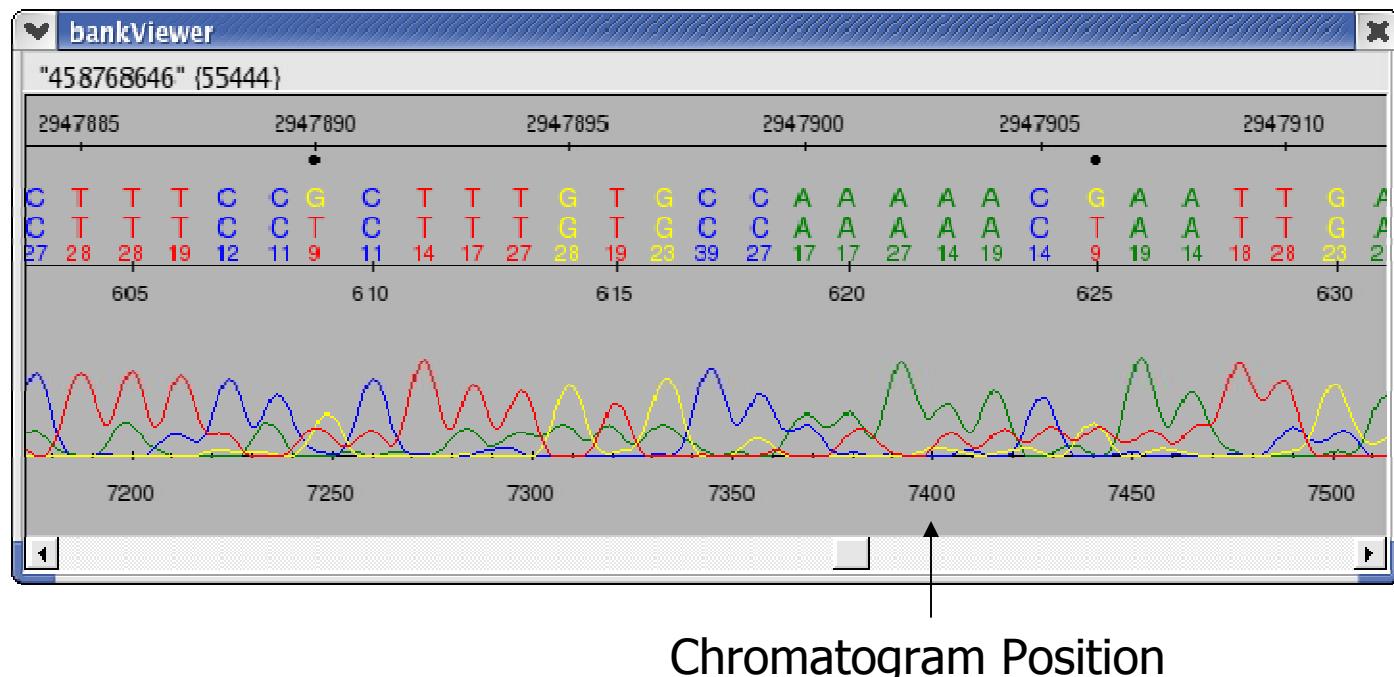
Normalized Chromatogram



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

Contig Chooser							
Display		Options					
IID		EID:					
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	
-151	151	1047283847440	P	166666	0	0.6275	

Select from 172 contigs in xoc4.bnk

Features

Feature Browser									
EID	Type	Source Type	Source IID	Dir	Start	End	Length	Comment	
8	C	164	F	3259	3260	1		END_BREAK: 175763	
8	C	145	F	1563	1564	1		END_BREAK: 22996	
8	C	156	F	197501	197502	1		END_BREAK: 3244	
8	C	130	F	5853	5854	1		END_BREAK: 60701	
8	C	144	F	512056	512057	1		END_BREAK: 6420	
8	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

Read Chooser										
Display		Options								
IID		EID:								
IID	EID	MateType	Offset	End Offset	Length	Dir	CLR Begin	CLR End	Lib ID	GC Content
3885	XOEDAK61TF	71	342	1308	967	F	28	994	86019	0.5890
8196	XODA24JTH	71	720	1686	967	R	985	20	86018	0.5896
40106	XODA24JTR	71	795	1711	917	R	933	16	86019	0.5911
8007	XODAQ50TF	71	748	1710	963	R	20	982	86018	0.5946
121	XOCAN51TFB	71	344	1198	855	F	23	877	86020	0.6059
35894	XOEDC38TR	71	293	1206	916	F	19	934	86019	0.6055
40027	XOEDC38TR	71	354	1056	773	F	22	817	86019	0.6059
17934	XOEA6K2TR	71	135	1140	1006	R	1025	40	86019	0.6151
52159	XOEFP21TR	71	169	1186	938	R	823	27	86019	0.6154
43894	XOEF98TR	71	199	1140	942	R	876	36	86019	0.6170
24879	XOECN97TR	71	232	1040	809	R	830	22	86019	0.6225
18209	XOELA32TR	71	86	1082	997	R	1015	22	86019	0.6234
28667	XOEBN27TR	71	163	1050	888	F	21	907	86019	0.6253
4238	XOCAN73TF	71	92	970	879	F	29	906	86020	0.6271

Select from 29 reads

Scaffolds

Scaffold Information									
Display		Options							
IID		EID:							
IID	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	86627	BE			

Select from 10 scaffolds in xoc4.bnk

- Full Integration: “Double click takes you there”

# Assembly Reports

## Misassembly Walkthrough: Correlated SNPs

Contigs

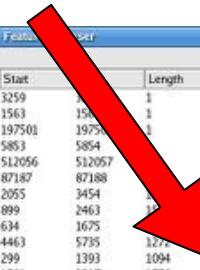
Contig Chooser							
Display		Options					
IID	EID			Status	Length	Reads	GC Content
-144	144	1047283847442	P	519090	6280	0.6399	
-141	141	1047283847439	P	326218	3784	0.6391	
-160	160	1047283847458	P	315606	3611	0.6372	
-152	152	1047283847450	P	259589	3402	0.6422	
-171	171	1047283847469	P	254579	2555	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	
-151	151	1047283847440	P	166666	0.6235		

Select from 172 contigs in xoc4.bnk

Features

Feature Chooser								
EID	Type	Source Type	Source IID	Dir	Stat	Length	Comment	
8	C	164	F	3259	1	1	END_BREAK: 175763	
8	C	145	F	1563	1563	1	END_BREAK: 22996	
8	C	156	F	197501	197501	1	END_BREAK: 3244	
8	C	130	F	5853	5854	1	END_BREAK: 60701	
8	C	144	F	512056	512057	1	END_BREAK: 6420	
8	C	159	F	87187	87188	1	END_BREAK: 690	
D	C	23	F	2055	3454	1	HIGH_READ_COVERAGE 32	
D	C	84	F	899	2463	1	HIGH_READ_COVERAGE 32	
D	C	41	F	634	1675	1	HIGH_READ_COVERAGE 35	
D	C	28	F	4463	5735	1	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

Read Chooser										
Display		Options								
IID	EID	LaneType	Offset	End Offset	Length	Dir	CLR Begin	CLR End	Lib ID	GC Content
3885	XOEDAK61TF	71	342	1308	967	F	28	994	86019	0.5890
8196	XODA24JTH	71	720	1686	967	R	985	20	86018	0.5896
40106	XODA24JTR	71	795	1711	917	R	933	16	86019	0.5911
8007	XODAQ50TF	71	748	1710	963	R	20	982	86018	0.5946
121	XOCAN51TFB	71	344	1198	855	F	23	877	86020	0.6055
35894	XOEDC38TR	71	293	1206	916	F	19	934	86019	0.6055
40027	XOEDC38TR	71	354	1056	773	F	22	817	86019	0.6059
17934	XOEA6K2TR	71	135	1140	1006	R	1025	40	86019	0.6151
52159	XOEFP21TR	71	169	1186	938	R	823	27	86019	0.6154
43894	XOEF98TR	71	199	1140	942	R	976	36	86019	0.6170
24879	XOECN97TR	71	232	1040	809	R	830	22	86019	0.6225
18299	XOELA32TR	71	86	1082	997	R	1015	22	86019	0.6234
28667	XOEBN27TR	71	163	1050	888	F	21	907	86019	0.6253
4238	XOCAN73TF	71	92	970	879	F	29	906	86020	0.6271

Select from 29 reads

Scaffolds

Scaffold Information						
Display		Options				
IID	EID		Offset	Span	Contigs	
1	173	1047283847471		2559	1	
2	174	1047283847472		2725904	25	
3	175	1047283847473	0	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		677786	315606	BE
161	161	1047283847459		1063385	86827	BE

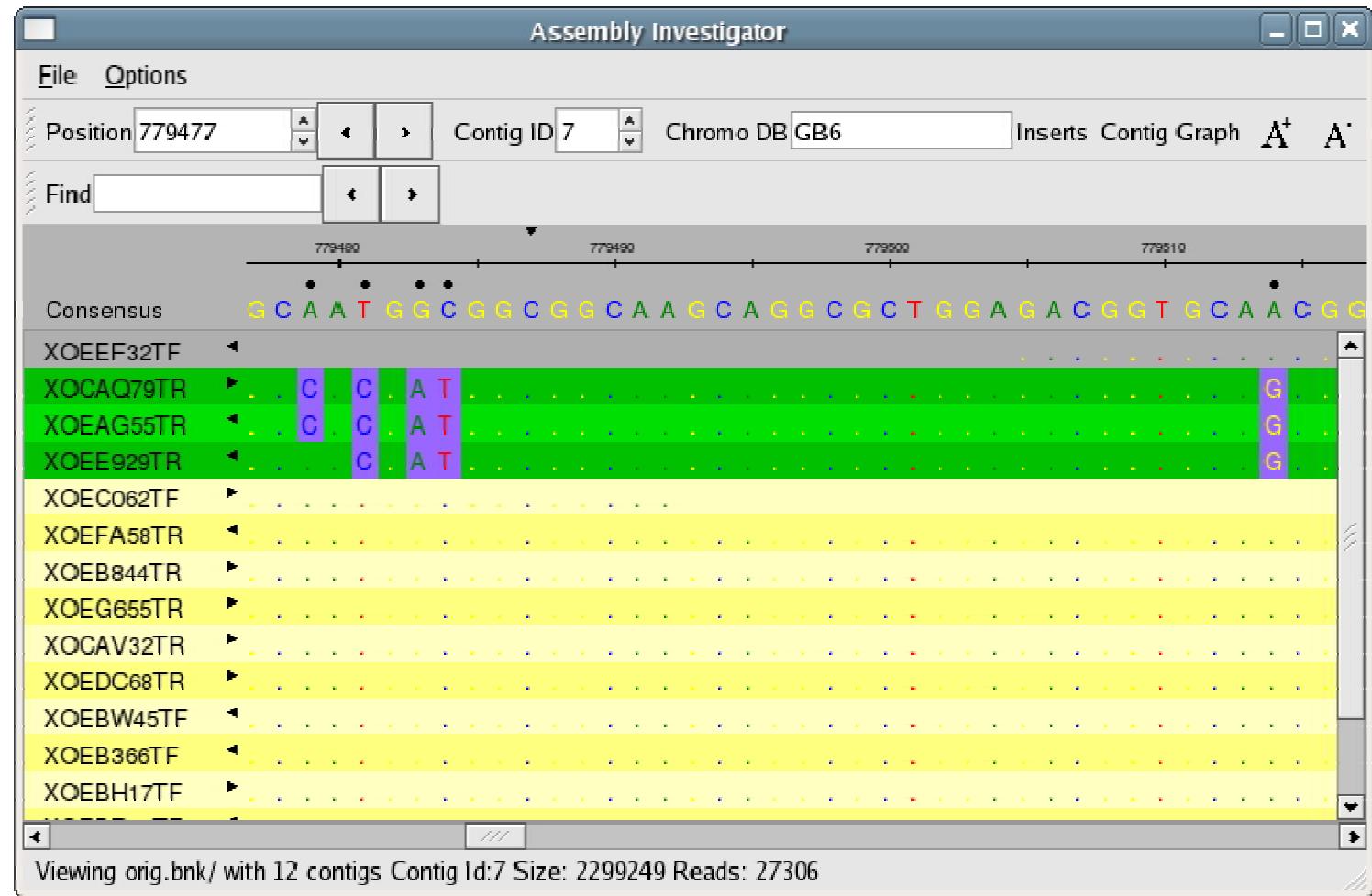
Select from 10 scaffolds in xoc4.bnk

- Full Integration: “Double click takes you there”

# SNP View

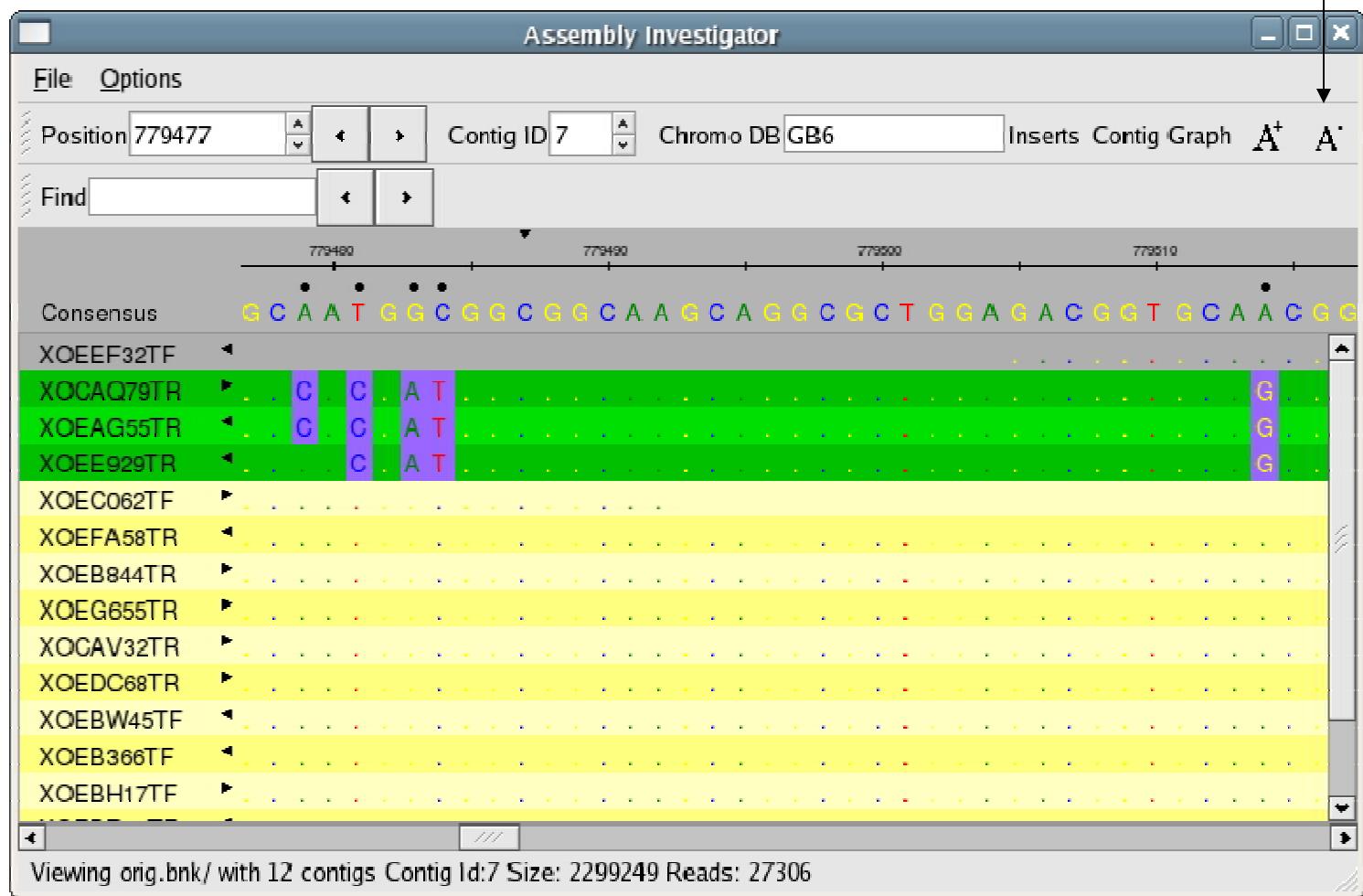
SNP Sorted  
Reads

Polymorphism  
View



# SNP View

Zoom Out

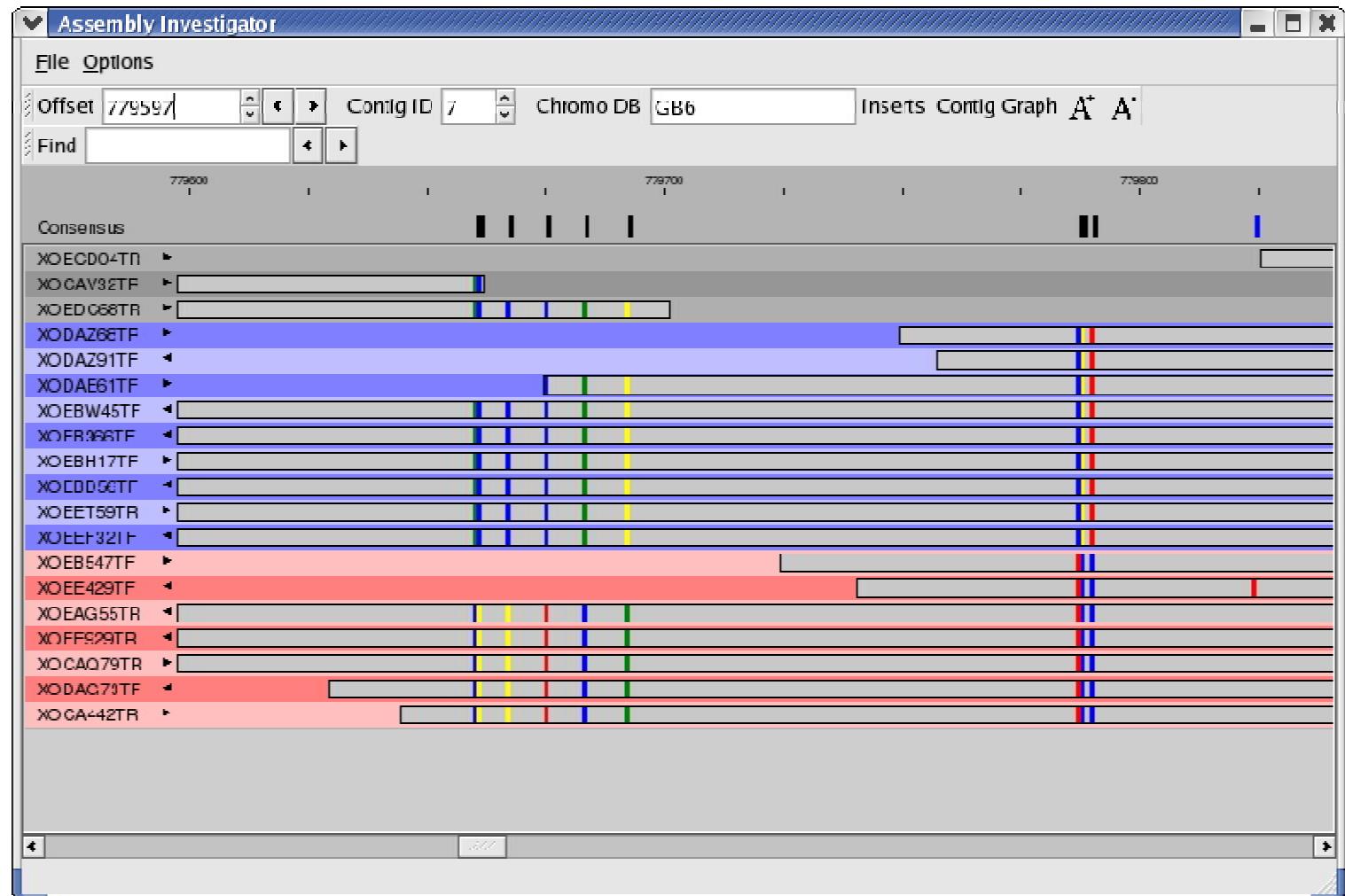


SNP Sorted  
Reads

Polymorphism  
View

# SNP Barcode

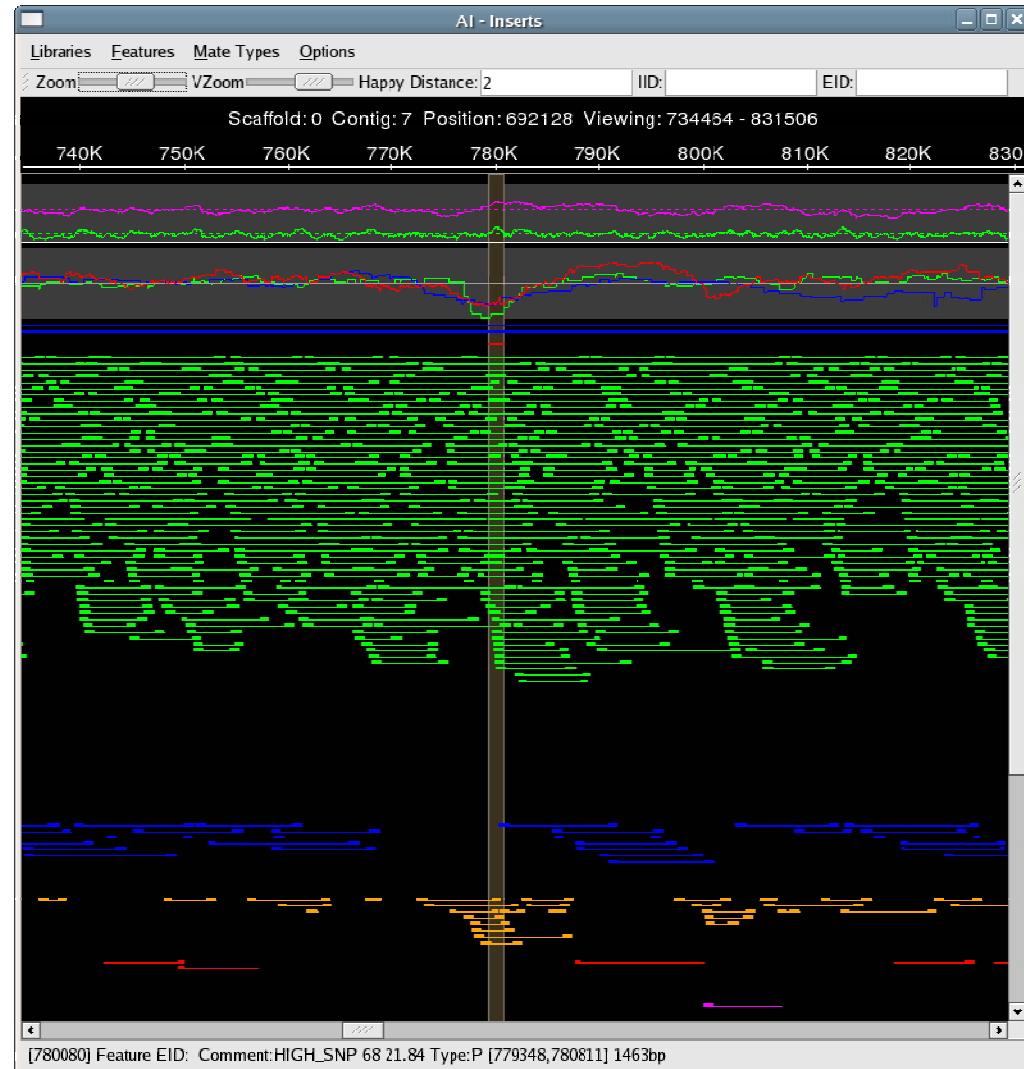
SNP Sorted  
Reads



Colored Rectangle indicate the positions and composition of the SNPs

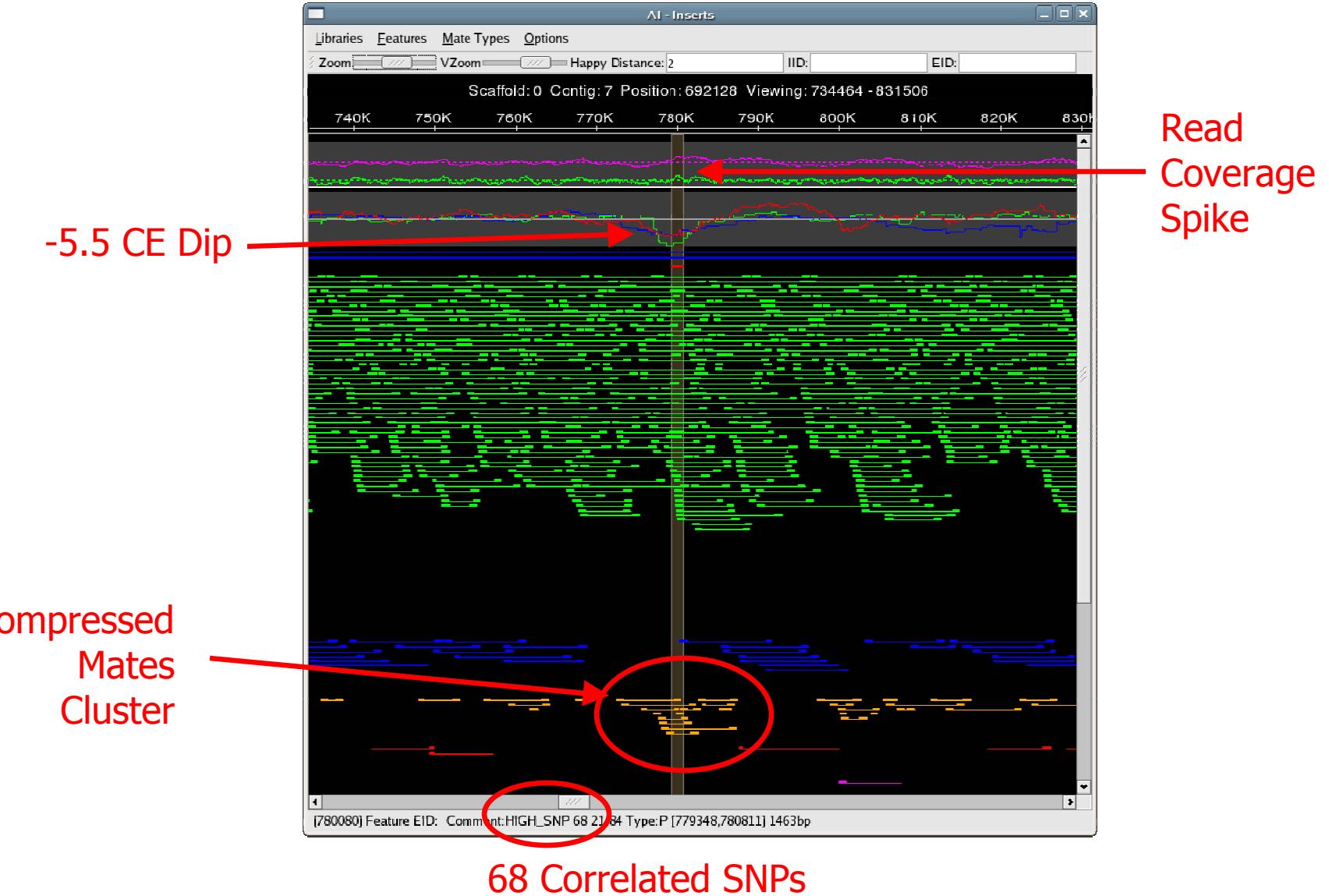
# Scaffold View

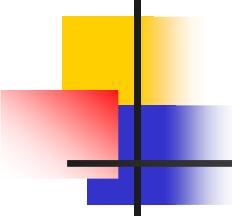
Coverage  
CE Statistic  
  
Happy  
  
Stretched  
Compressed  
Misoriented



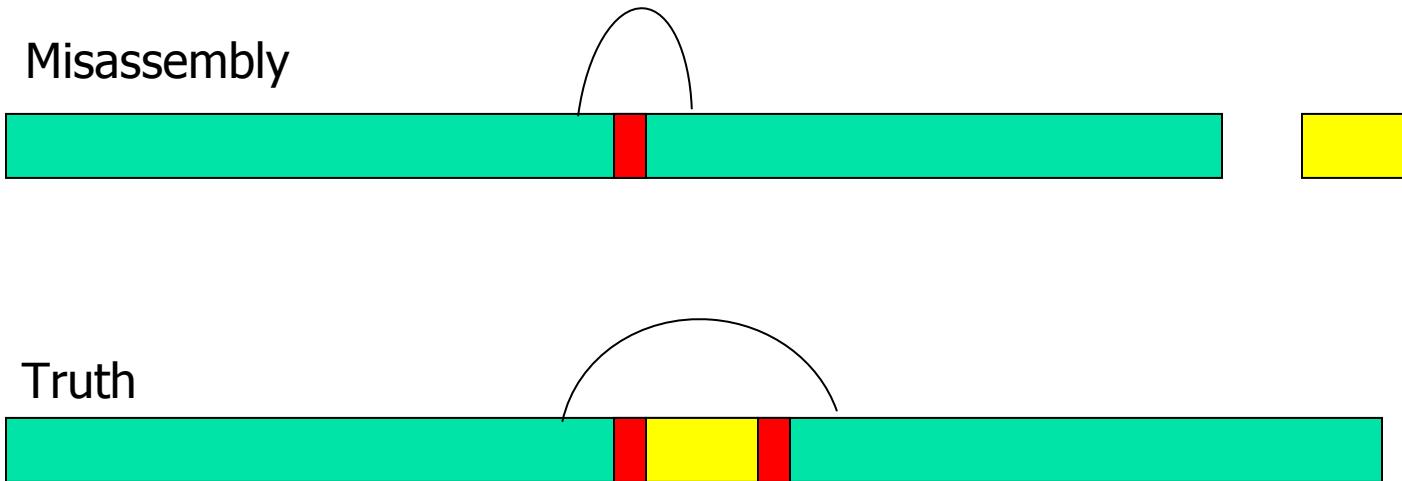
SNP Feature  
  
Linking

# Collapsed Repeat



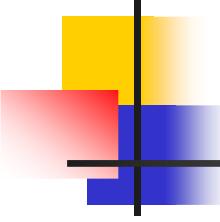


# Confirmed Misassembly



## Collapsed repeat

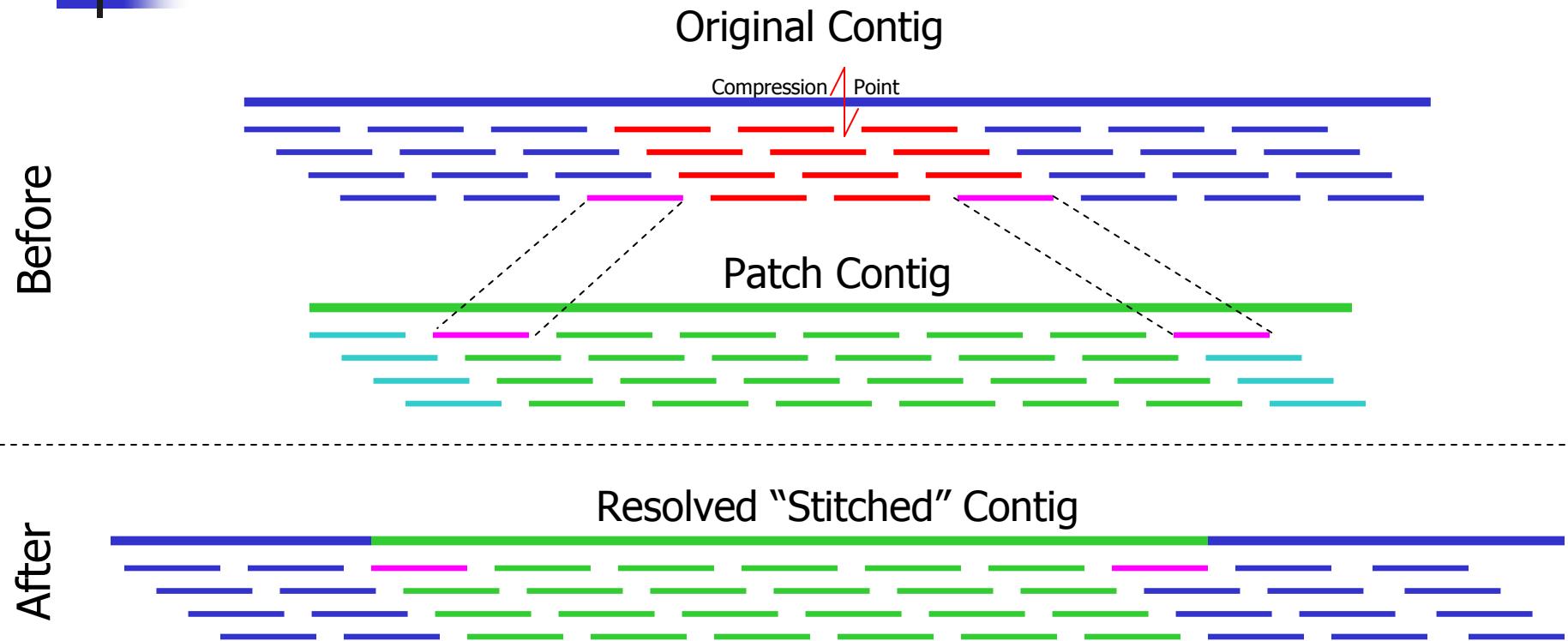
- Compressed mates (-5.5 CE Stat)
- Correlated SNPs (68 Positions within 1400bp)
- Spike in Read Coverage



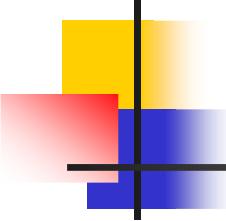
# Fixing collapsed repeats with AMOS

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



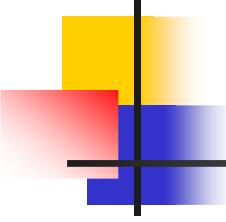
- Replace the reads between the stitch reads in the original contig with corresponding region in the patch contig.
- Can also close gaps or fix contig ends



# 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

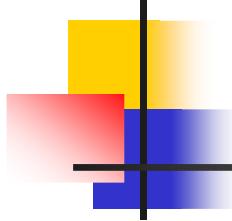




# Current Research

- 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](mailto:amos-help@lists.sourceforge.net)
- A
- M
- O
- S
- Hawkeye Webpage:
  - <http://amos.sourceforge.net/hawkeye>
- Acknowledgements
  - Adam Phillippy
  - Ben Shneiderman
  - Steven Salzberg
  - Mihai Pop