

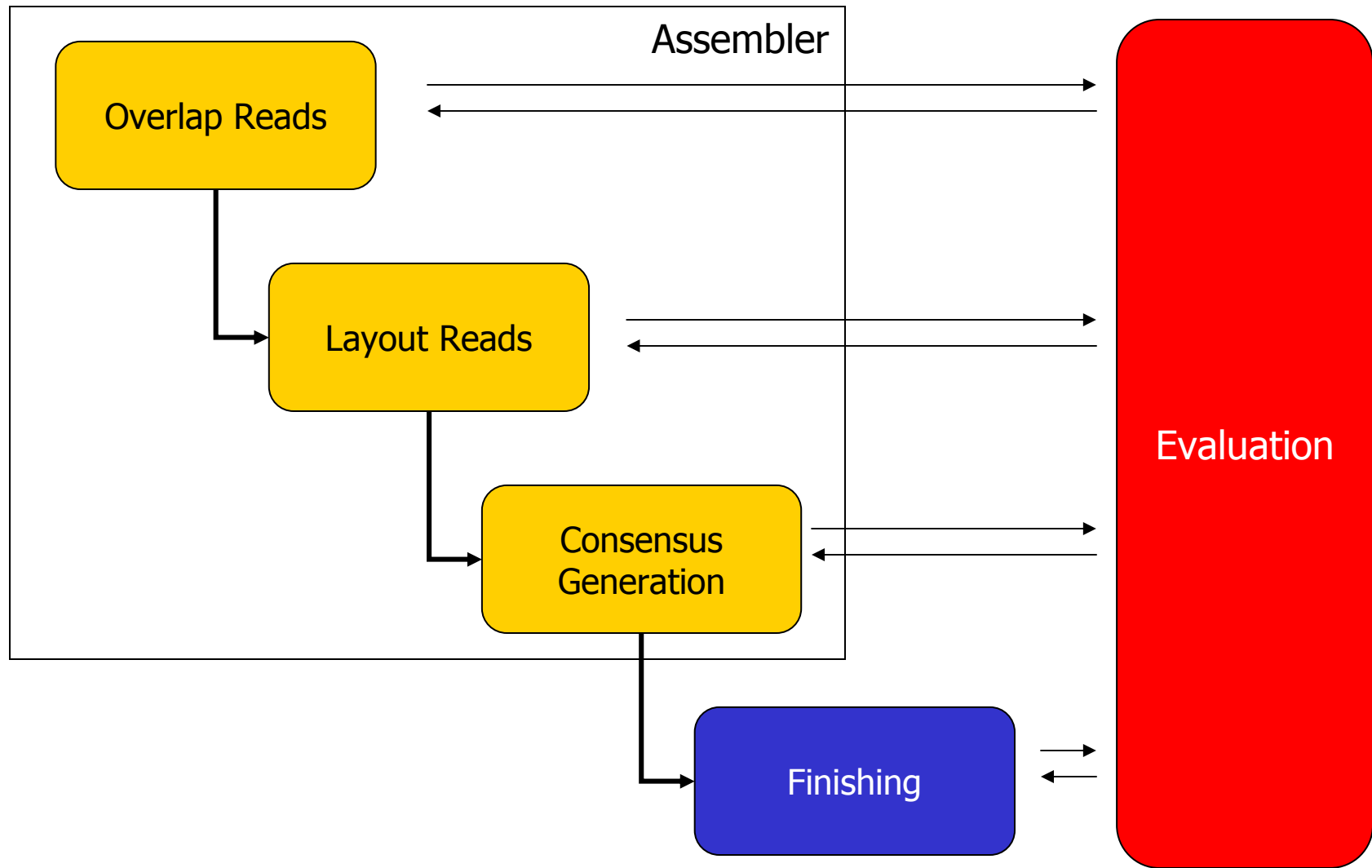
Interactive visual analytic tools for genome assemblies

Michael Schatz

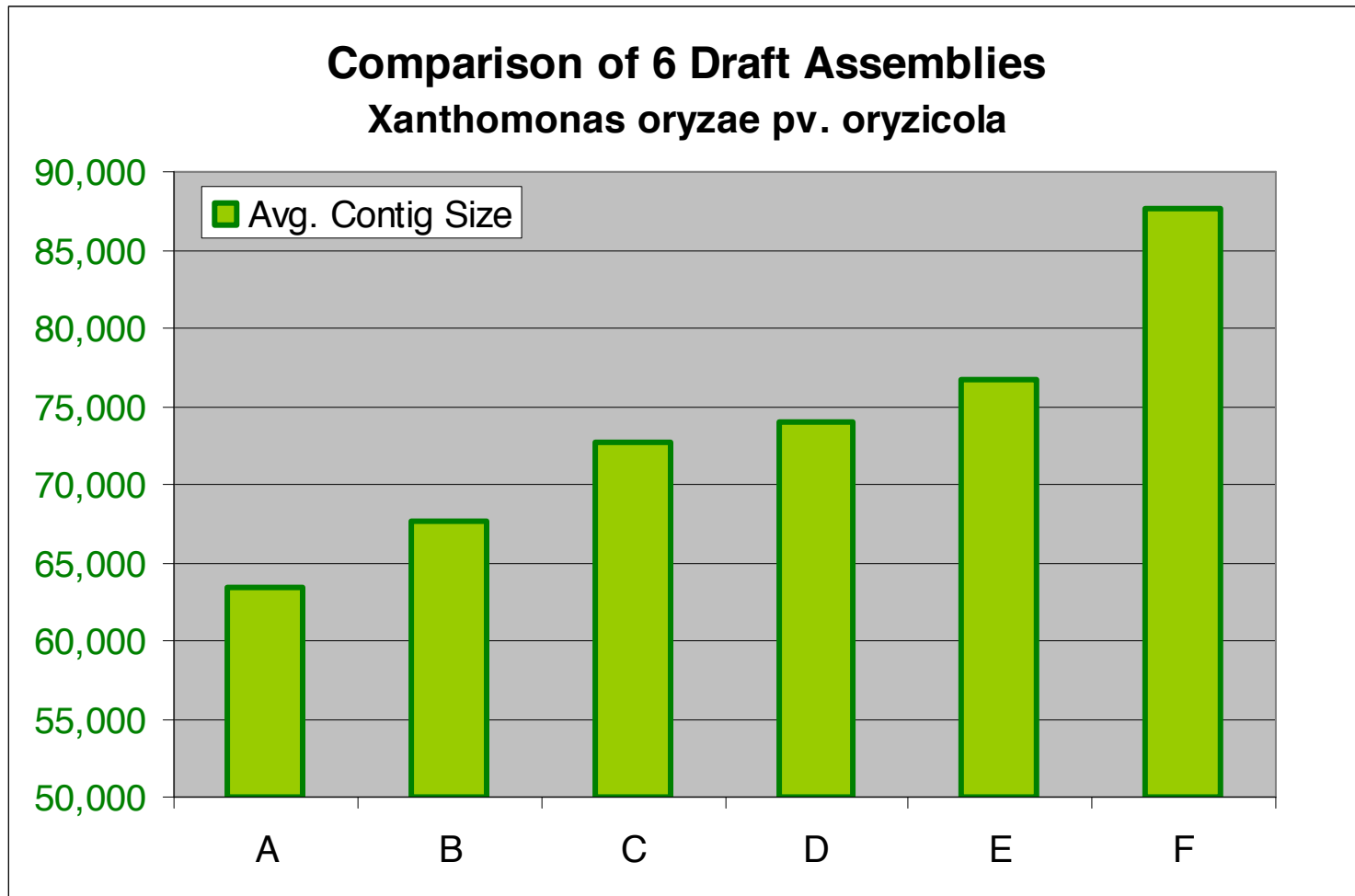
Center for Bioinformatics and Computational Biology
University of Maryland

October 29, 2006
9th Annual Computational Genomics Conference

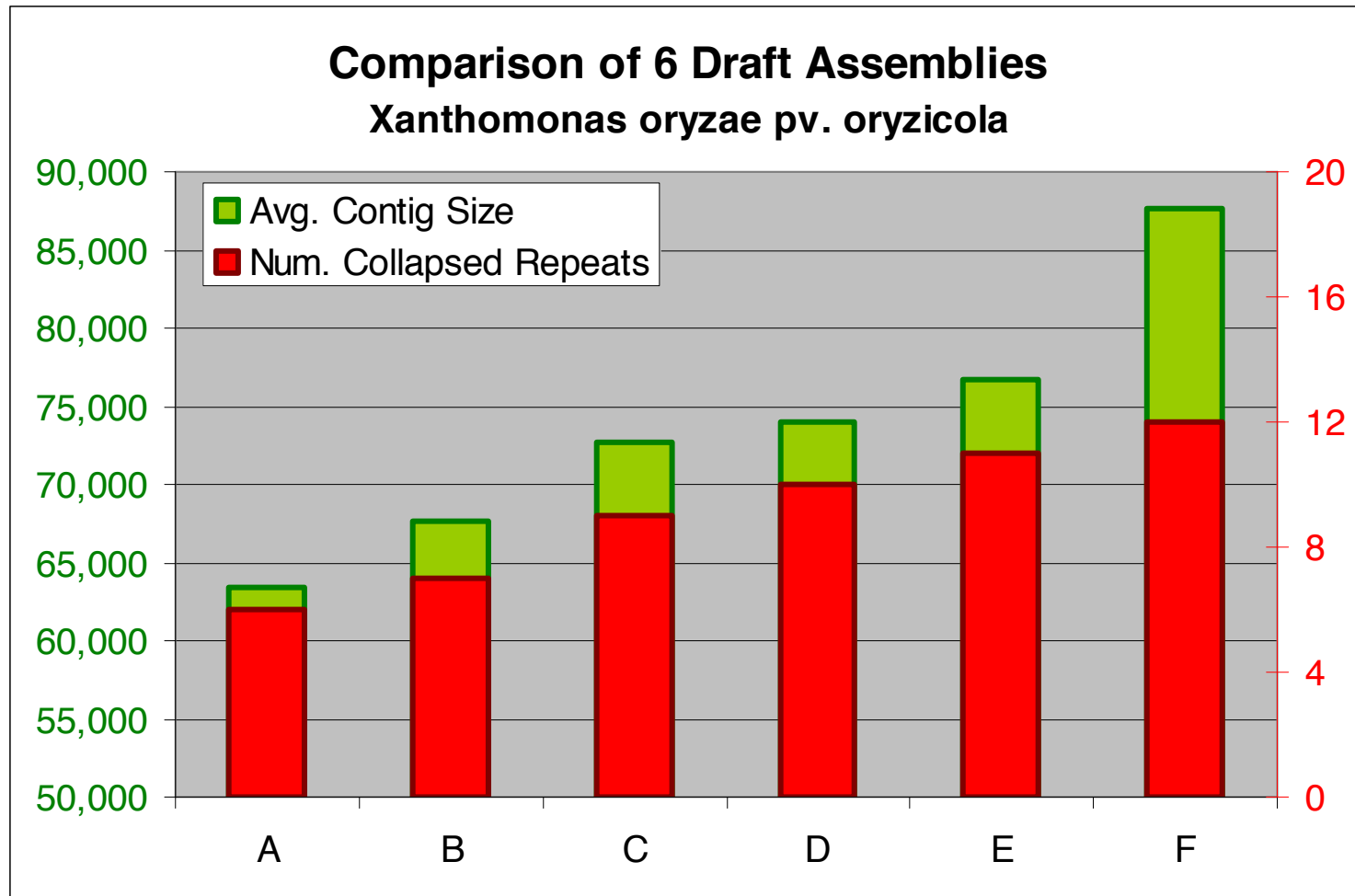
Genome Assembly



Assembly Evaluation

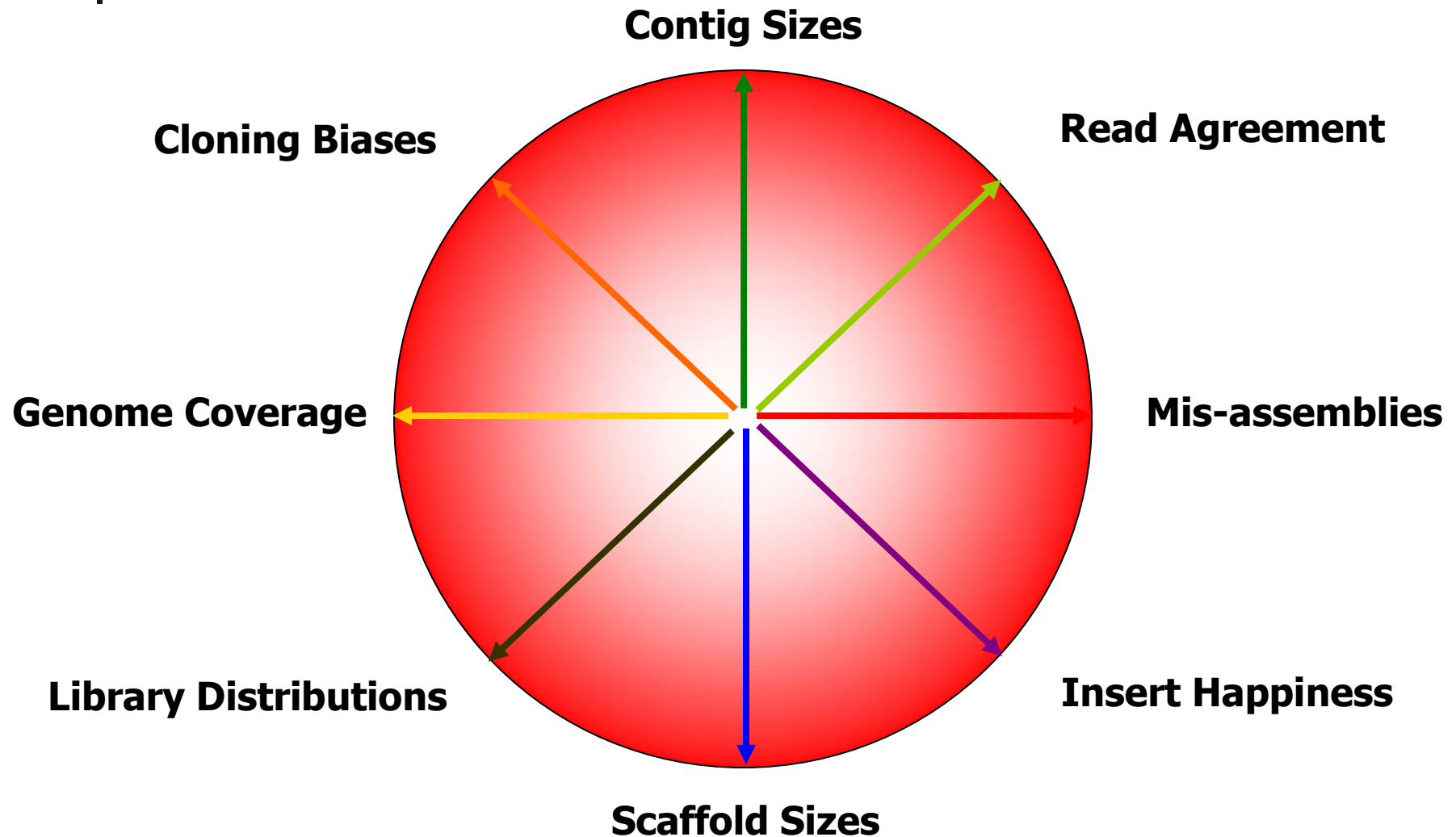


Assembly Evaluation



* Bigger is not always better

Dimensions of Assembly Quality



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

Hawkeye Goals

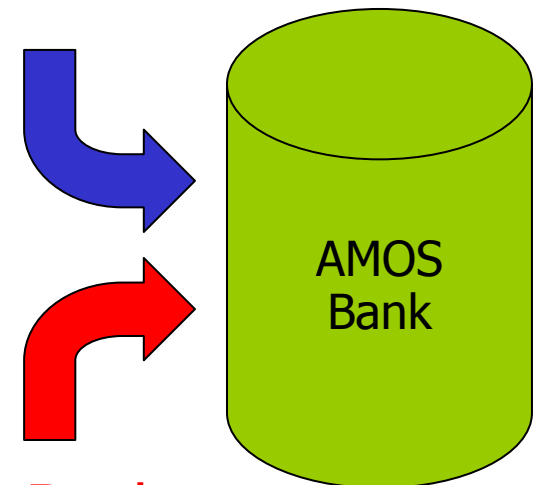
Interactively explore and analyze

- 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 ($\rightarrow \rightarrow$)
 - 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

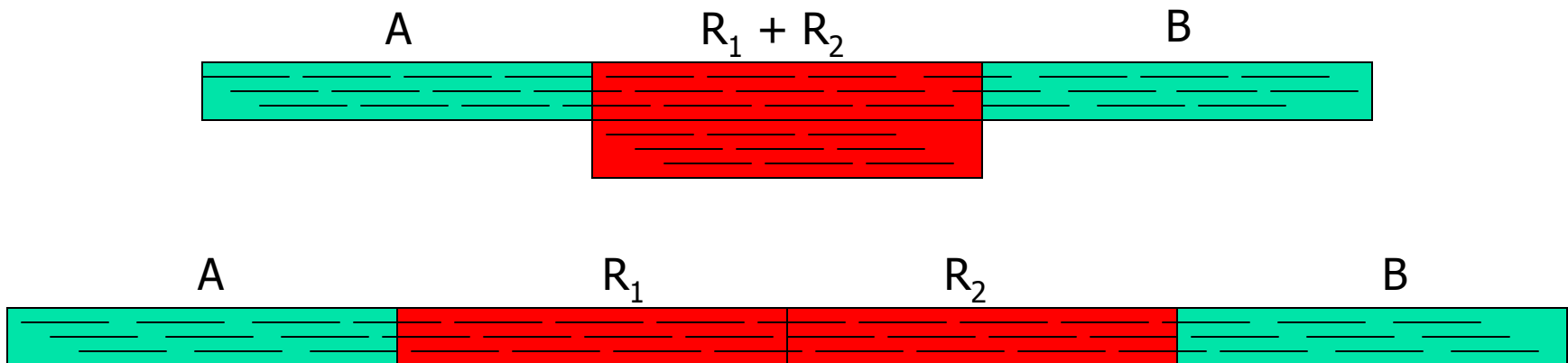


C/E Statistic

- 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
 - $(\text{Local Mean} - \text{Global Mean}) / \text{Scaling Factor}$
 - $> +3$ 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)

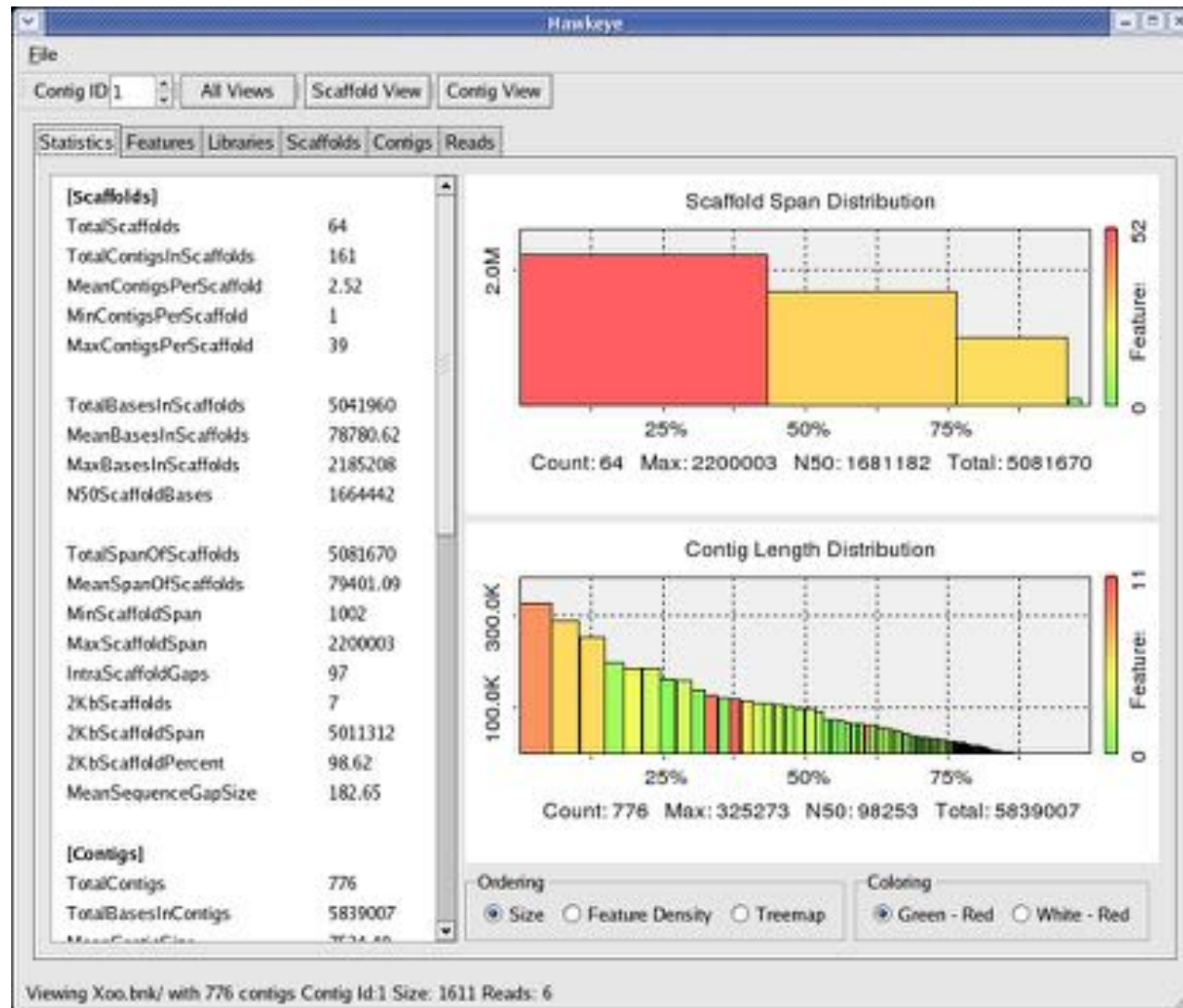
A	G	C
A	G	C
A	G	C
A	G	C
A	G	C
A	G	C
C	T	A
C	T	A
C	T	A
C	T	A
C	T	A



Hawkeye

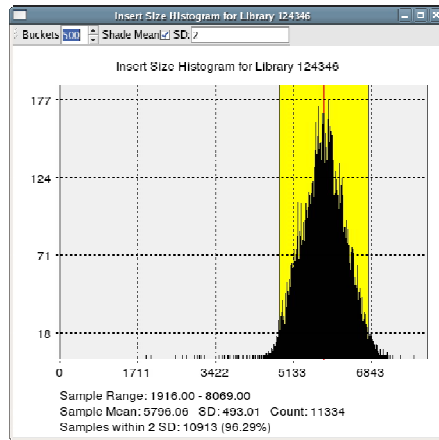


Launch Pad

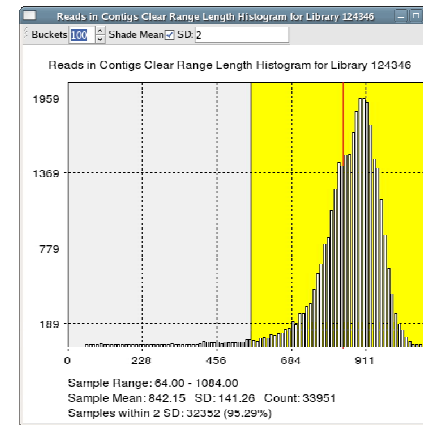


Histograms & Statistics

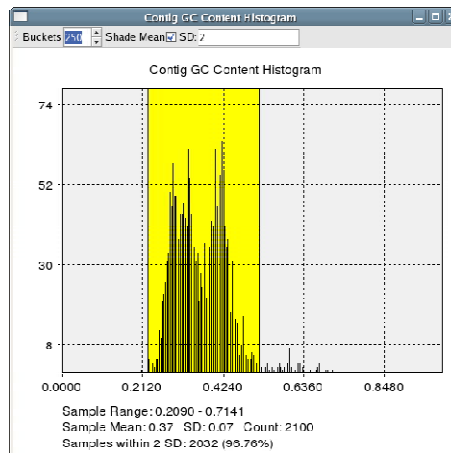
Insert Size



Read Length



GC Content



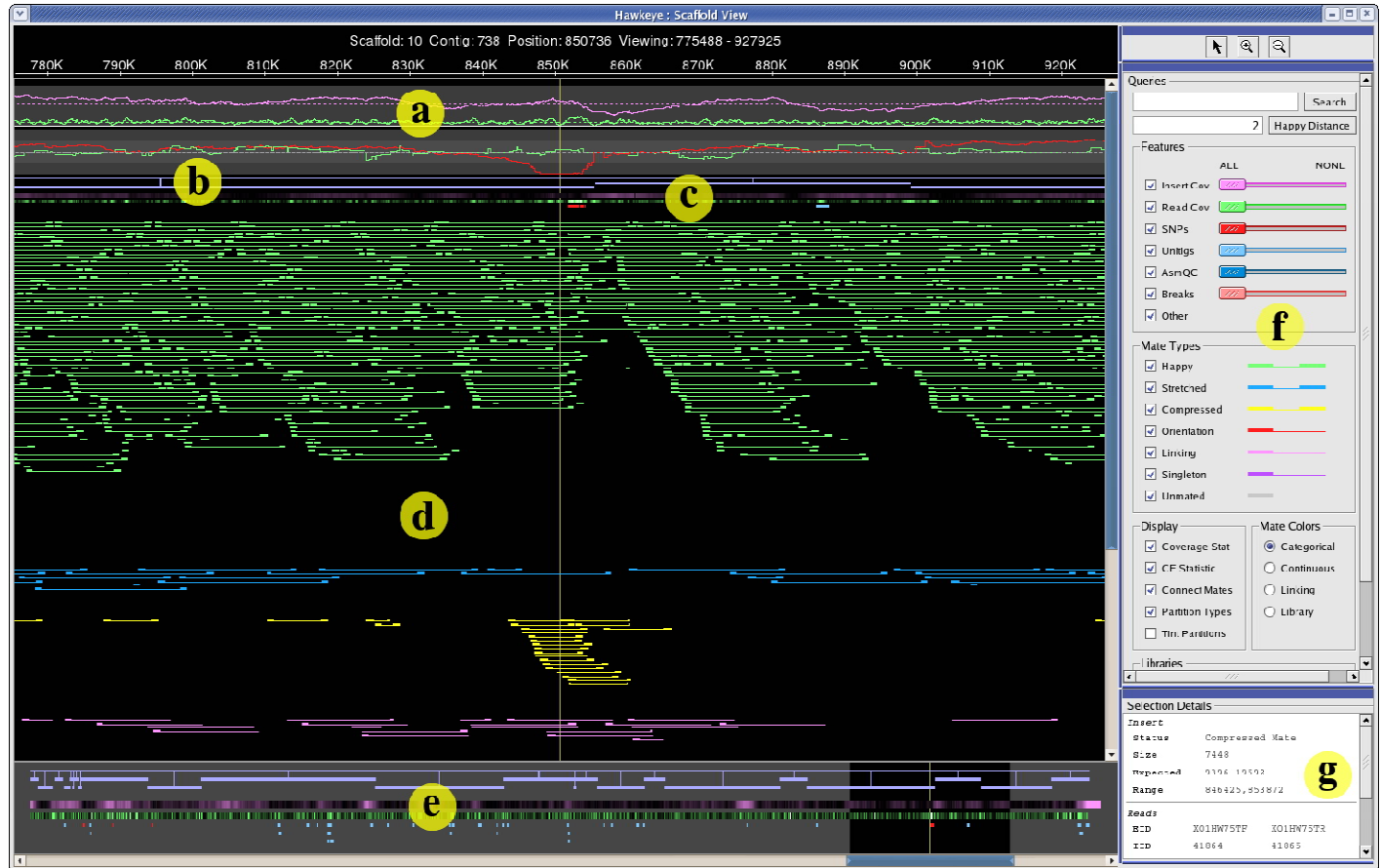
Overall Statistics

Field	Value
[Scaffolds]	
TotalScaffolds	1076
TotalContigsInScaffolds	1396
MeanContigsPerScaffold	1.30
MinContigsPerScaffold	1
MaxContigsPerScaffold	15
TotalBasesInScaffolds	7511900
MeanBasesInScaffolds	6981.12
MaxBasesInScaffolds	279040
N50ScaffoldBases	75935
TotalSpanOfScaffolds	780540
MeanSpanOfScaffolds	7253.24
MinScaffoldSpan	1007
MaxScaffoldSpan	285205
IntraScaffoldGaps	320
2KbScaffolds	200
2KbScaffoldSpan	644092
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



Insert Happiness

Both mates present



Happy

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



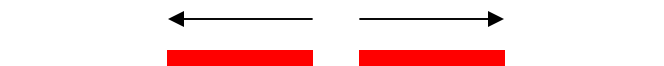
Stretched

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



Compressed

- Oriented Correctly &&
- $\text{Insert Size} < \text{Library.mean} - \text{Happy-Distance} * \text{Library.sd}$



Misoriented

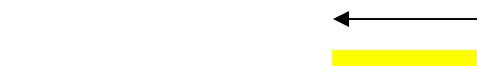
- Same or Outies

Only 1 read present



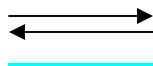
Linking

- Read's mate is in some other scaffold



Singleton

- Read's mate is a singleton



Unmated

- No mate was provided for read

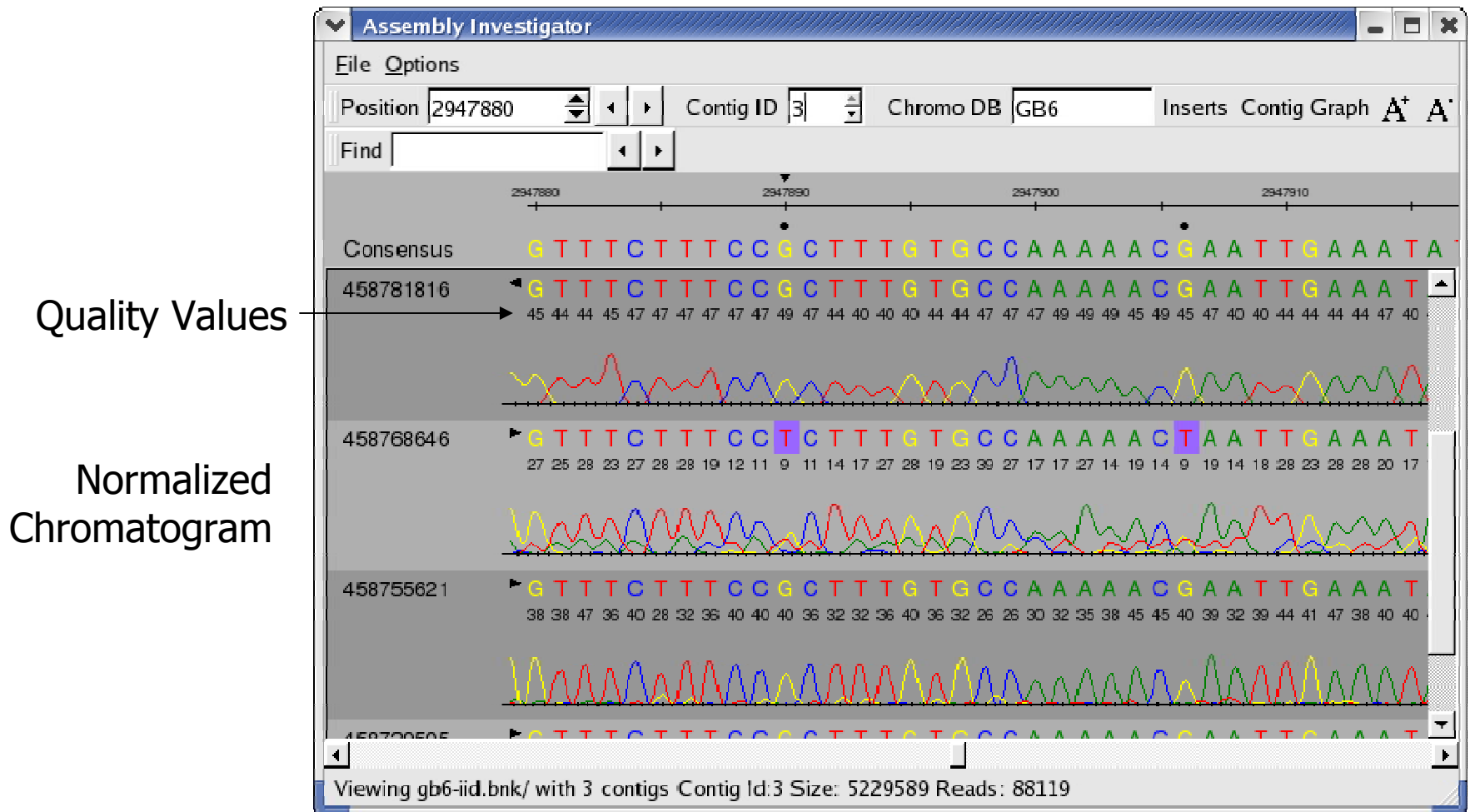
Contig View

The screenshot displays the 'Assembly Investigator' window. At the top, there are navigation and search controls: 'Discrepancy Navigation' (left), 'Contig Quick Select' (center), and 'Discrepancy' (right). Below these are input fields for 'Position' (116659), 'Contig ID' (738), and 'Chromo DB' (GB6). A 'Find' field is also present. The main area shows a 'Consensus' sequence: 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. Below the consensus is a 'Scrollable Read Tiling' section with multiple rows of reads, each starting with a right-pointing arrow. Some letters in the reads are highlighted in purple, indicating discrepancies. At the bottom, a 'Summary' bar shows 'Viewing Xoo.bnk/ with 776 contigs Contig Id:738 Size: 119783 Reads: 1114'. A 'Read Orientation' arrow points to the right-pointing arrows in the reads, and a 'Discrepancy Highlight' arrow points to the purple highlights.

Annotations on the left side of the image:

- Regular Expression Consensus Search
- Consensus & Position
- Scrollable Read Tiling
- Summary
- Read Orientation
- Discrepancy Highlight

Contig View Expanded



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

Assembly Reports

Misassembly Walkthrough: Correlated SNPs

Contigs

Id	IID	EID	Status	Length	Reads	GC Content
144	144	1047283847442	P	519090	6280	0.6390
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	1	END_BREAK: 175763
B	C		145	F	1563	1	END_BREAK: 22996
B	C		156	F	197501	1	END_BREAK: 3284
B	C		130	F	5853	5854	END_BREAK: 60701
B	C		144	F	512056	512057	END_BREAK: 6420
B	C		159	F	87187	87188	END_BREAK: 690
D	C		23	F	2055	3454	HIGH_READ_COVERAGE 32
D	C		84	F	899	2463	HIGH_READ_COVERAGE 32
D	C		41	F	634	1675	HIGH_READ_COVERAGE 35
D	C		28	F	4463	5735	HIGH_READ_COVERAGE 36
P	C		2	F	299	1393	HIGH_SNP 10 121.67
P	C		23	F	1561	3317	HIGH_SNP 10 195.22
P	C		164	F	29745	30597	HIGH_SNP 10 94.78
P	C		153	F	21586	22457	HIGH_SNP 10 96.89
P	C		37	F	772	2506	HIGH_SNP 12 157.73
P	C		124	F	268	1196	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	71	342	1308	967	F	28	994	86919	0.5890
8396	XODA243TF	71	720	1686	967	F	985	20	86918	0.5896
40100	XOEA207R	71	795	1711	917	R	933	16	86919	0.5911
8007	XODAG50TF	71	748	1710	963	F	20	982	86918	0.5946
121	XOCA035TFB	71	344	1198	855	F	23	877	86920	0.6030
36894	XOEDC38TR	71	291	1206	916	F	19	934	86919	0.6055
42027	XOEDT12TF	71	284	1056	773	F	74	847	86919	0.6080
17934	XOEA62TR	71	135	1140	1006	R	1035	40	86919	0.6151
52159	XOEF11TF	71	169	1106	938	R	963	27	86919	0.6154
43894	XOEF980TR	71	199	1140	942	R	976	36	86919	0.6170
24879	XOECN79TR	71	232	1040	809	R	830	22	86919	0.6225
18209	XOEA132TR	71	86	1082	997	R	1015	22	86919	0.6234
28667	XOEBN27TF	71	163	1050	888	F	21	907	86919	0.6253
4238	XOCAN73TF	71	82	970	879	F	29	906	86920	0.6271

Select from 23 reads

Scaffolds

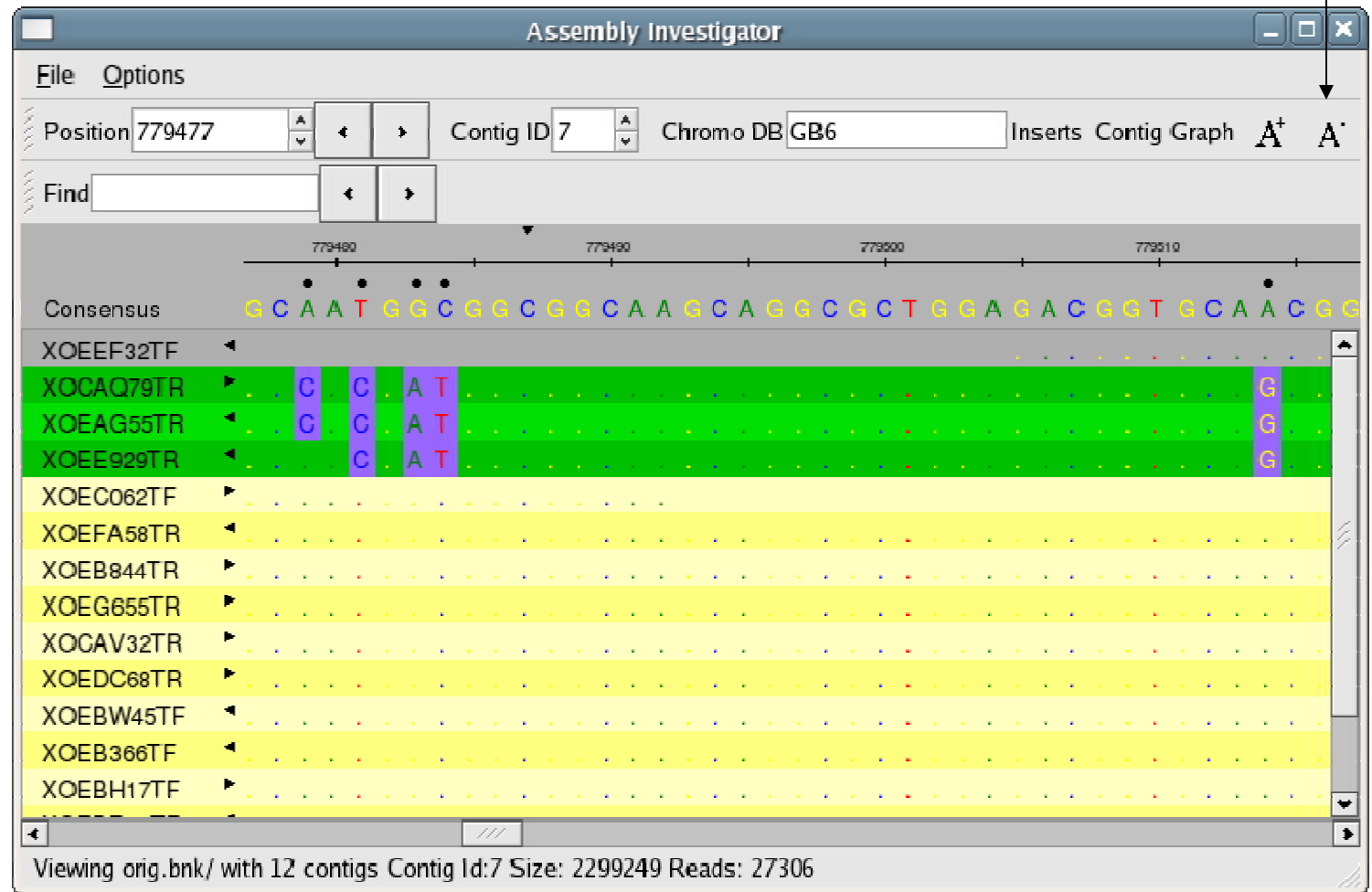
Id	IID	EID	Offset	Span	Contigs
1	173	1047283847471		2559	1
2	174	1047283847472		2725904	25
3	175	1047283847473		2111083	24
152	152	1047283847450	0	259589	BE
153	153	1047283847451	259820	61666	BE
154	154	1047283847452	321466	24156	BE
155	155	1047283847453	345602	73623	BE
156	156	1047283847454	419250	200997	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

Zoom Out

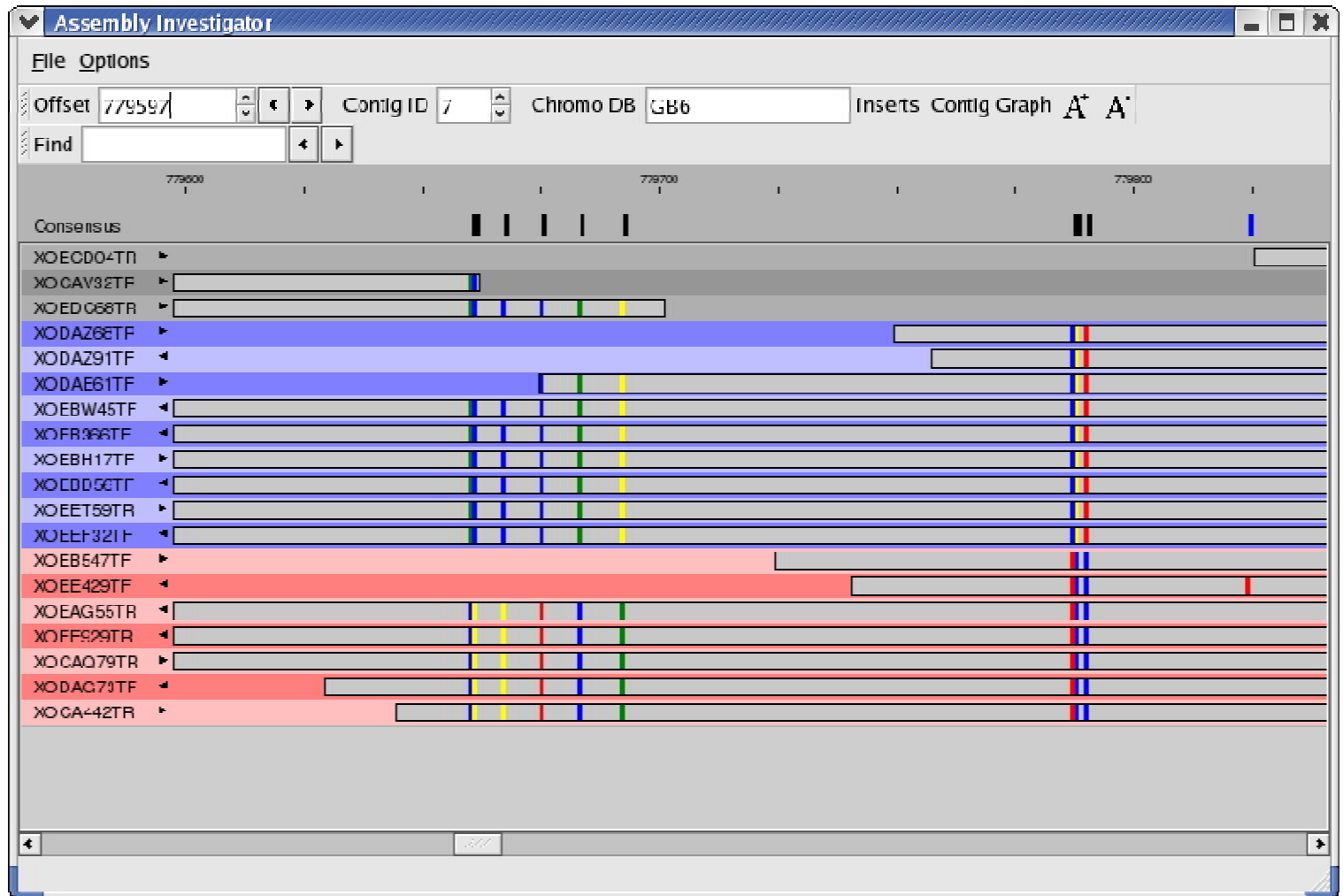


SNP Sorted Reads

Polymorphism View

SNP Barcode

SNP Sorted Reads



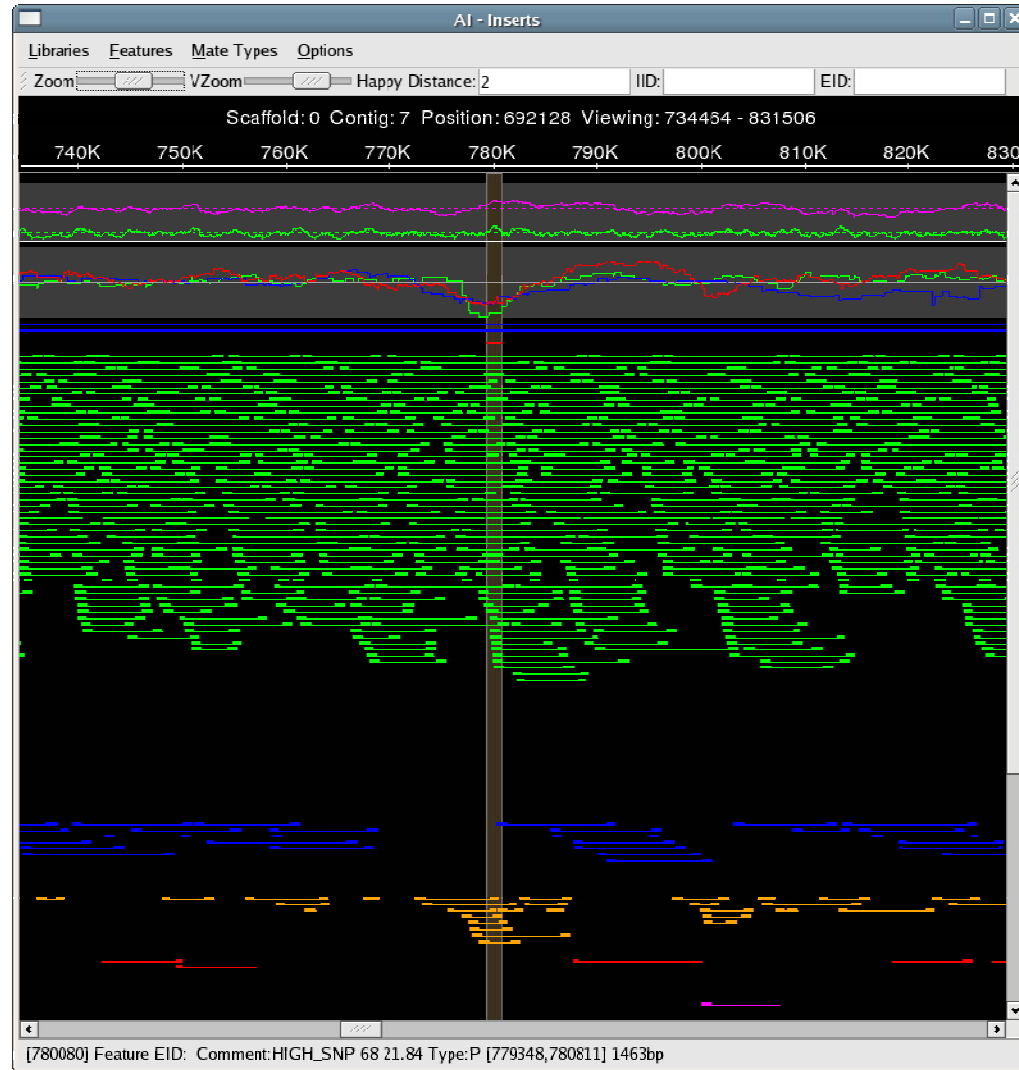
Colored Rectangle indicate the positions and composition of the SNPs

Scaffold View

Coverage
CE Statistic

Happy

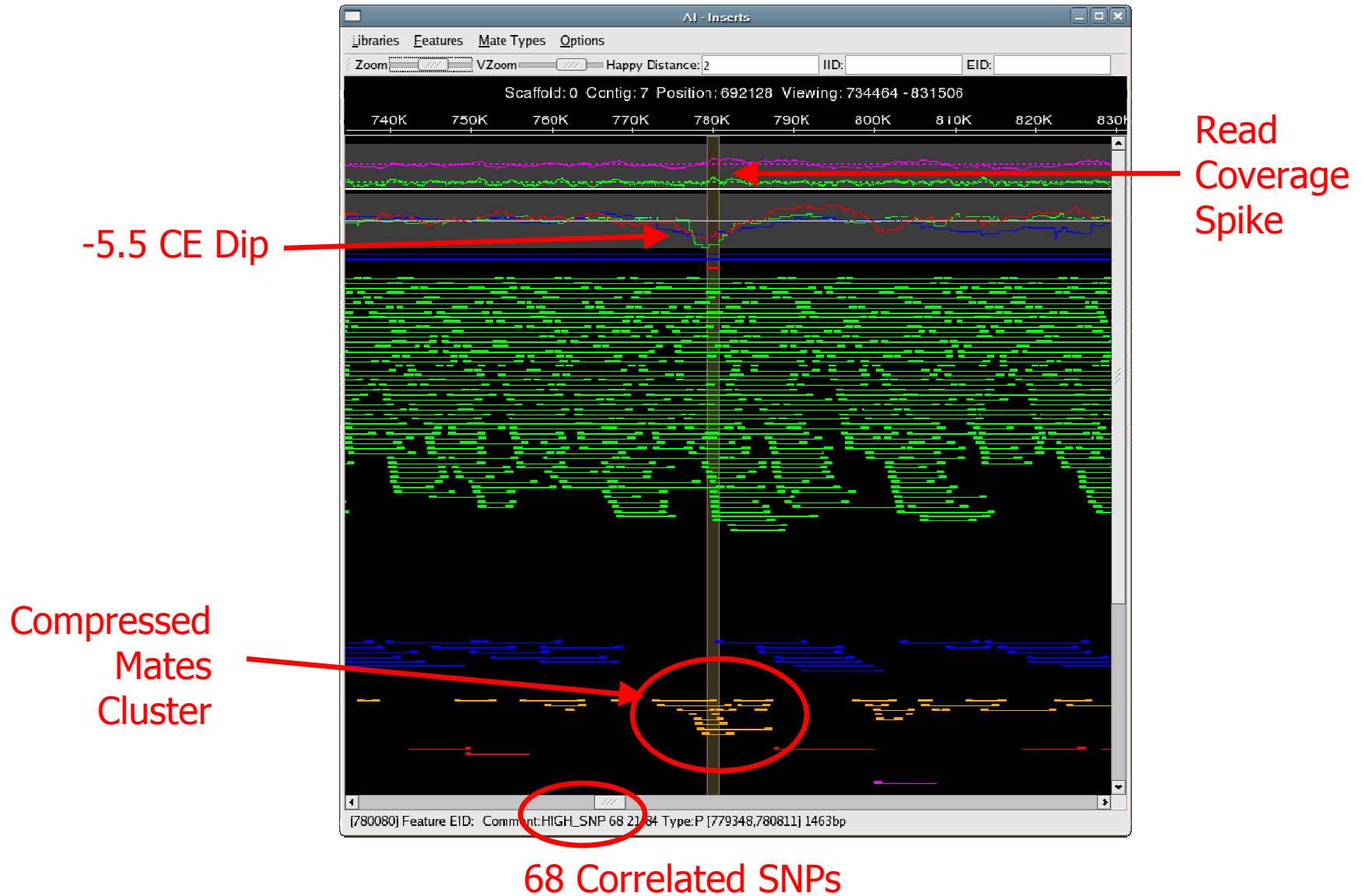
Stretched
Compressed
Misoriented



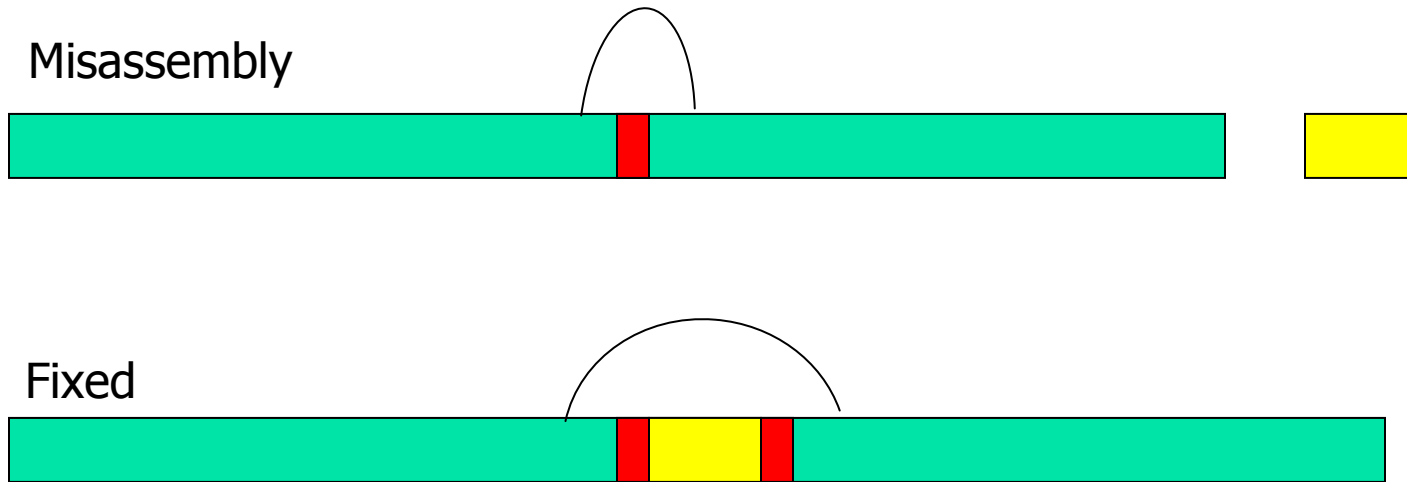
SNP Feature

Linking

Collapsed Repeat



Confirmed Misassembly



Collapsed repeat

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

More Information

- Hawkeye Webpage:
 - <http://amos.sourceforge.net/hawkeye>

A

M

O

S

- Contact AMOS
 - [amos-help \[at \] lists.sourceforge.net](mailto:amos-help@lists.sourceforge.net)

- Acknowledgements



Adam Phillippy



Ben Shneiderman



Steven Salzberg



Mihai Pop



Art Delcher