Improving Genome Assemblies without Sequencing

Michael Schatz

April 25, 2005 TIGR Bioinformatics Seminar

Assembly Pipeline Overview

- Sequence shotgun reads 1.
- Call Bases phred/TraceTuner/KB Base Caller 2.
- Trim Reads 3.
- Assemble 4
- **Electronic Finishing** 5.
 - Second generation base-caller
 - Automatic Gap Closure
 - Research Techniques

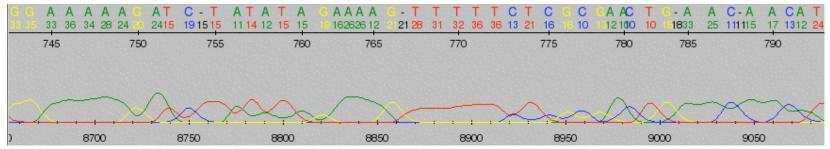
lucy

CA/TA/Arachne

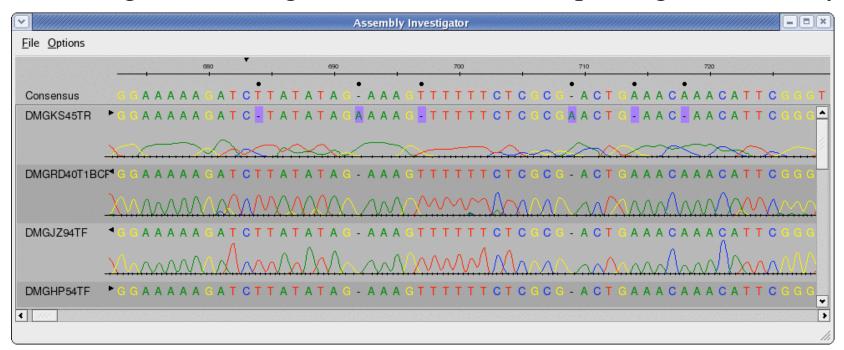
AutoEditor **AutoJoiner**

AutoEditor

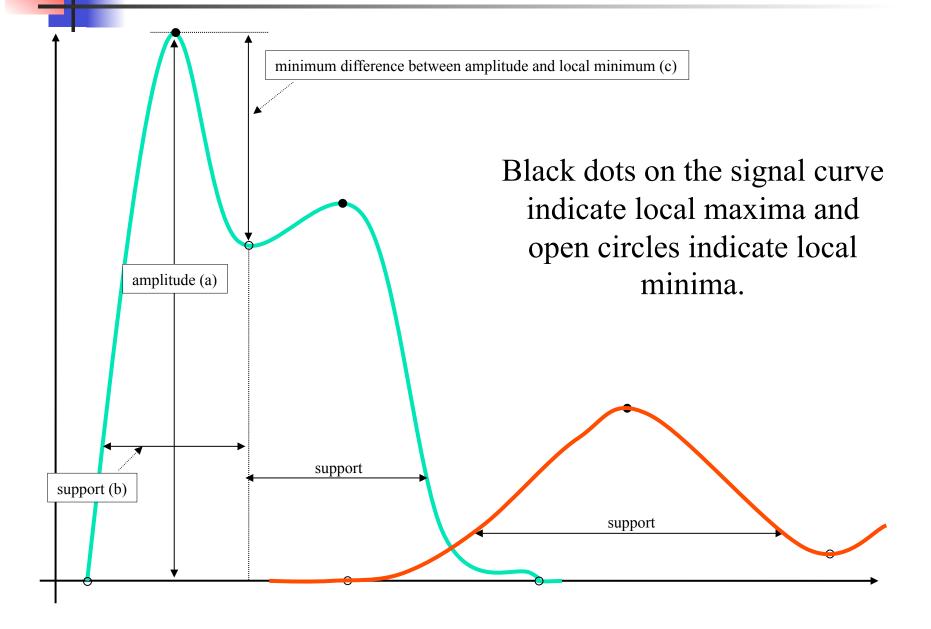
Base-calling in the context of single chromatogram is hard...



but finding base-calling "mistakes" in a multiple alignment is easy.



Signal Parameters

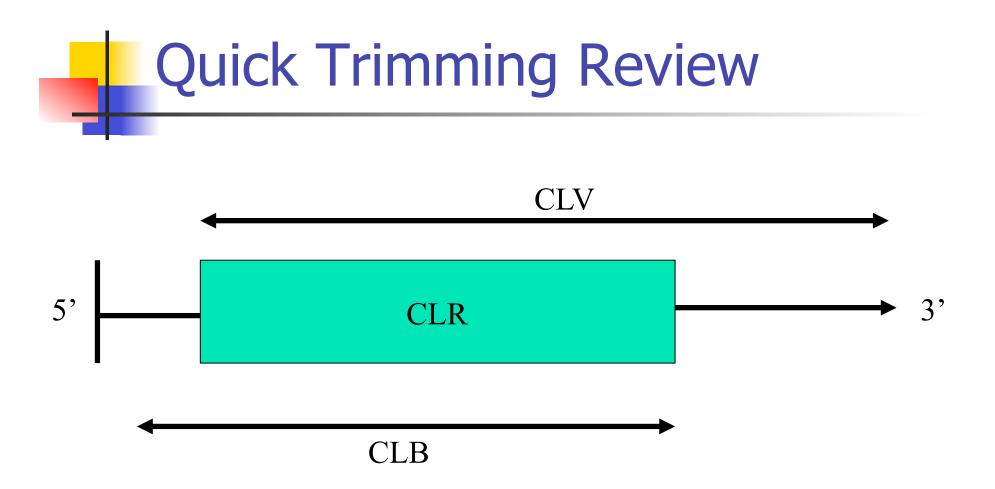


AutoEditor Results

- Corrects 80% of all discrepant base-calls with an error rate better than 1/8800.
- Increase consensus quality, decrease finishing costs
- Remaining discrepancies highlight assembly problem regions or interesting biological events.

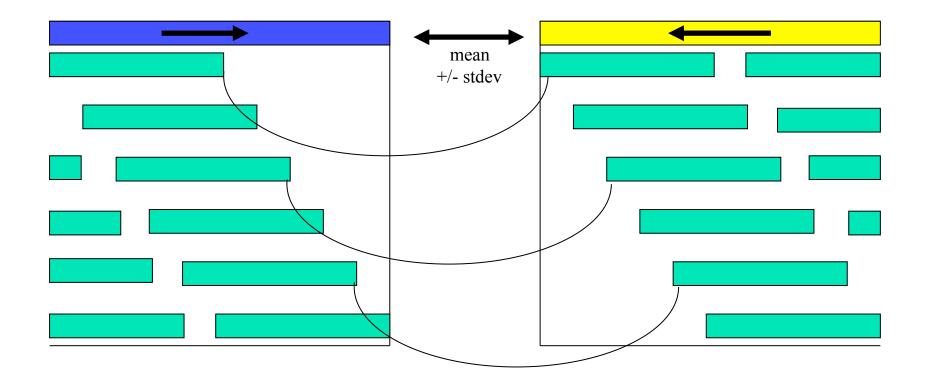
Organism	Read length	Corrections	AE errors
Listeria monocytogenes	37 420 828	145 274	4
Wolbachia sp.	11 446 011	51 163	0
Burkholderia mallei	47 407 080	99 711	28
Brucella suis	26 629 877	112 359	2
Streptococcus agalactiae	23 485 615	105 878	3
Coxiella burnetii	29 135 115	117 232	30
Campylobacter jejuni	15 013 845	792 37	11
Chlamydophila caviae	10 286 694	36 972	6
Dehalococcoides ethenogenes	10 724 521	46 416	12
<i>Neorickettsia sennetsu</i> Miyayama	8 805 232	37 425	0
Fibrobacter succinogenes	46 463 268	196 150	4
Mycoplasma capricolum	9 353 819	15 444	0
Prevotella intermedia	20 084 365	94 162	3
Pseudomonas syringae	50 369 232	177 897	46
Total	346 625 502	1 315 320	149

Ask Pawel for more information!



Trimming identifies the regions of good quality for the assembler to use (CLR), as the intersection of the region free of vector (CLV) and the region free of bad quality (CLB).

Quick Assembly Review



The individual reads (green) have been assembled into 2 contigs (blue & yellow). The mate relationship between the reads allows for the contigs to be oriented and the gap size to be estimated.

Read Coverage

	Insert	S				// = = ×
Display Types Options						
Zoom	IID: 46708	EID: XOEFB35	TF			
615K 616K 617K 618K	619K 620K	621K 622K	623K	624K	625K	626K
- man	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~				، بەي بەي بەي مىر بەر مەركىمى	~~~~
		•				
Image: A state of the state						•
Insert *[e:XOEFB35TF i:46708 c:16] <=> [e:XOEFB	335TR i:46709 c:15] [H]	Actual: 8187 Expected: 55	42 - 9918			

Sequencing Gap

✓				Inserts						
<u>D</u> isplay Types <u>O</u> p	tions									
§ Zoom 🦳	Happy Distanc	e: 2	I	ID: 46708		EID: XOEFB35	FF			
615K 616K	617K	618K	619K	620K	621K	622K	623K	624K	625K	626K
	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~		~					an the second		·
			~~							
	······						~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~	~~~~~~	
								_		
			_							
										_
				_						
										-
•				700 40157			12 0012			>
Insert *[e:XOEFB3	5TR i:46709 c:15]	<=> [e:XOEF	-B35TF i:46	708 c:16] [H]	Actual: 81	187 Expected: 55	42 - 9918			

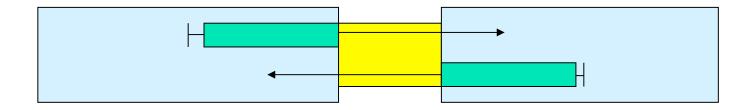
### AutoJoiner Architecture

- 1. All-vs-All Alignment
- 2. Analyze Alignments
- 3. Extend and

Join Contigs

- 4. Contig Fattening
- 5. AutoEdit Result

autoJoin
nucmer
aj_evaluateOverlaps
aj_joinContigs
fattenSlice
zipclap
fattenSlice
autoEditor



### All-vs-All Alignment

🔲 read-alig	nment.txt	: + (/local/	asmg/so	ratch/l	ignment-	104	72838	4743	4-1047	28384	743	5) - GVIM4		×
<u>F</u> ile <u>E</u> dit	<u>T</u> ools <u>S</u> y	/ntax <u>B</u> u	ffers <u>W</u>	/indow <u>I</u>	<u>H</u> elp								19	
		@ 1 % I	b 🗗 i	۵ 🖗	۵ 🖗	63	3	16		010		1		
<pre>% show-coor 35001 53419 1012 769 % show-alio ====================================</pre>	35998 53479 1133 980 gns out. ======= nts betw	1023   1023   285   1023 delta 104 ====================================	14 963 405 813 4728384 4728384 2838474	34 and 1	1010 61 121 211 EFB35TF EFB35TF (0EFB357	1   1   1   ====	95.1 100.0 90.2 99.0	0   4	53479 53479 1163 980	1023 1023 1023 1023		1047283847434 1047283847434 XODA905TF XOEC861TR	XOEFB35T XOEFB35T XOEFB35T XOEFB35T	F
53419 1023		cgtctggad cgtctggad												
53468 974	ctgagcc ctgagcc													
END a:	lignment	[ +1 534	419 - 5	3479	1 1023	- 9	63]	== <b>2</b>						•
												25,	,60	Top

The first AutoJoiner!

### Alignment Reports

overlap.report (/local/asmg/scratch/ment-1047283847434-1047283847435) - GVIM3	- 8 ×
<u>F</u> ile <u>E</u> dit <u>T</u> ools <u>S</u> yntax <u>B</u> uffers <u>W</u> indow <u>H</u> elp	
<pre> 1047283847434(BE) &lt;-&gt; 1047283847435(BE) 6.547 [171.273]  21: Gap: -2 Id*Dist: 61.00 (100.00%) Overhang: 0 (x-2+,x-1+,-1,-1) 28: 1047283847434 &lt;-&gt; XOEFB35TF 100.00% 61bp </pre>	<b>^</b>
##1047283847434(0) [] 53479. <1 53479> {1 53479} (53419 53479) 53479  53419 53479  #XOEFB35TF(134) [RC] 797. <135 931> {830 34} (1023 963) 965  -58 2	
[ <=(B) { Eval: id len dir t18(0 -58) gap(0,-2,-2) 3'() 5'(q) ex(q) success.	
@=1047283847434 0 1047283847435 0 qex XOEFB35TF 193 @ @	-
1,1	A11

Why did AutoJoiner make this join?

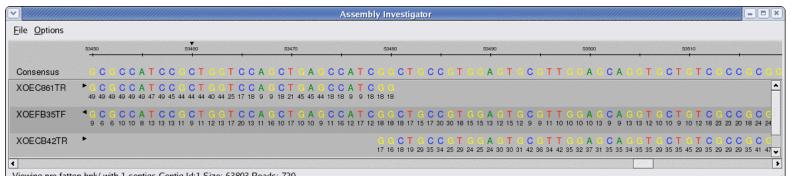
### **Contig Extension**

Assembly Invest	igator 🙀				- <b>- X</b>
<u>F</u> ile <u>O</u> ptions					
Position 110	€ • ► Contig ID 16	Chromo DB Xoc	Inserts Contig	Graph A⁺ A' ∐Find	• •
	110	120	130 +	140	150
Consensus	CTCAAGC	A C G C C T A C G A	CCTGTCCG	A T G A A G C G G T G	T G C G A A C G T 1
XOECB42TR	CTCAAGC	ACGCCTACGA	CCTGTCCG	ATGAAGCGGT	G T G C G A A C G T ⁻
	40 44 44 44 44 40 41	41 47 41 41 47 44 36 35 38 29	29 31 44 49 49 49 49 41 47	47 47 47 47 47 41 47 47 41 47 4	1 47 47 44 38 4C 44 31 44 35 4
XOEFB35TF	CTCAAGC	A C G C C T A C G A	CCTGTCCG	A T G A A G C G G T C	GTGCGAACGT [•]
	35 35 35 28 19 10 10	8 10 10 10 12 18 25 40 38 34	44 44 44 18 18 10 9 <mark>19</mark>	18 19 22 9 9 18 18 44 47 47 4	4 26 26 20 21 28 33 47 29 31 2
<b>⊥</b> ]					Ŀ

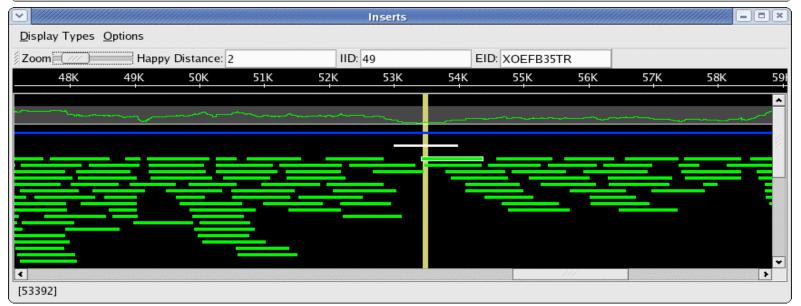
**Extension Procedure:** 

- If necessary, extended selected read by aligning trimmed bases to existing consensus.
- Untrim to desired base, promote untrimmed bases to consenus, shift offsets.

### **Contig Joining**

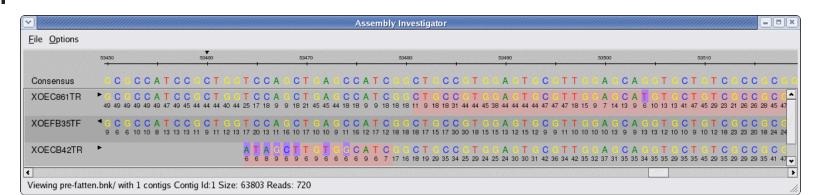


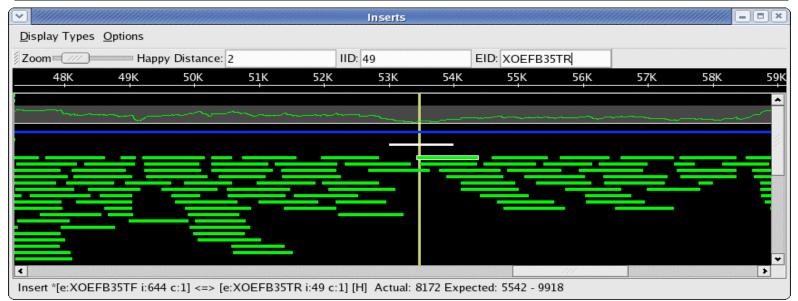
Viewing pre-fatten.bnk/ with 1 contigs Contig Id:1 Size: 63803 Reads: 720



"Zip" together contigs by pairwise alignment between consensi.

### **Contig Fattening**





"Fatten" addition reads in the join region to increase coverage.

### Join Report

blasted.joinreport (/local/asmgc/Blasted/autoJoin/results) - GVIM	100		_ = ×
<u>F</u> ile <u>E</u> dit <u>T</u> ools <u>S</u> yntax <u>B</u> uffers <u>W</u> indow <u>H</u> elp			
	<b>%</b> 🗆	010	0
<pre>[]autoJoin] NumSequenceGaps=39 NumJoined=6 NumDoubleExtend=0 NumSingleExtend=5 NumNoExtend=1 PercJoin=0.154 ExpectedGap.Range=-20.000,15.034 ExpectedGap.Mean=-4.39 ExpectedGap.Mean=-4.39 ExpectedGap.SD=14.25 ExpectedGap.Median=-3.95 ExpectedGap.Median=-3.95 ExpectedGap.Mad=19.67 ActualGap.Range=-82,67 ActualGap.Mean=19.83 ActualGap.Mean=19.83 ActualGap.Median=42.75 ActualGap.Median=42.75 ActualGap.Mad=23.72</pre>			
- "blasted.joinreport" 18L, 338C	1,1		• A11

What did AutoJoiner do?

### **AutoJoiner Validation**

Project	Gaps	Joined	%	Invalid	Degenerate	Gap Size	Mean	Join ID	Mean
blm	106	14	13.20%	0	0	-200.5:156	21.18	98.39:100.00	99.26
dmg	52	13	25.00%	0	3	-666.5:36.5	-140.73	98.99:100.00	99.64
gb6	32	10	31.20%	0	0	-13:146	42.3	98.50:99.87	99.31
gba	110	38	34.50%	0	0	-452:229	29.18	97.18:100.00	99.36
gbm	43	6	14.00%	0	0	-17:22	-4.67	99.24:100.00	99.78
gbr	32	11	34.40%	0	0	-62.5:103.5	15.36	98.89:99.92	99.44
gbs	31	5	16.10%	0	0	-5:32.5	10.9	98.99:99.65	99.33
gcb	10	2	20.00%	0	0	-37.5:-4.5	-21	99.38:99.81	99.59
gcj	22	11	50.00%	0	0	-53:139	27.45	98.71:99.81	99.45
gcp	25	8	32.00%	0	4	-555:184.5	-75.31	99.03:99.85	99.42
gde	82	17	20.70%	0	1	-113:203.5	22.29	97.04:99.93	99.14
ges	27	17	63.00%	0	0	-779:-302	-586.71	100.00:100.00	100
gfs	131	33	25.20%	0	6	-182.5:212	21.79	98.81:100.00	99.51
gmcap	10	2	20.00%	0	0	-11.5:171	79.75	98.67:99.84	99.25
gpi	150	52	34.70%	0	0	-231.5:181	20.3	97.96:99.93	99.4
gps	162	43	26.50%	0	0	-1069.5:213.5	-36.13	98.76:100.00	99.51
gsa	262	32	12.20%	0	0	-618:136	-43.44	94.90:100.00	99.25
crypt_1	20	8	40.00%	0	0	-36:186.5	63.62	98.74:99.88	99.43
crypt_2	7	5	71.40%	1	0	-39:148	27.8	98.63:99.56	99.18
crypt_3	21	8	38.10%	0	0	-93:67.5	-3.06	97.83:100.00	99.31
crypt_4	25	7	28.00%	0	0	-90:159	45.21	98.94:100.00	99.52
crypt_5	23	8	34.80%	0	0	-111.5:249	35.12	98.98:99.92	99.43
crypt_6	14	7	50.00%	0	0	-14:192	37.21	98.41:99.93	99.58
crypt_7	17	6	35.30%	0	0	-3.5:230.5	66.67	99.09:100.00	99.62
crypt_8	15	6	40.00%	0	0	-19:57.5	15	99.20:100.00	99.74
crypt_9	12	6	50.00%	0	0	-423:34	-120.5	99.16:100.00	99.82
crypt_10	14	7	50.00%	1	0	-777:124	-91.21	95.23:100.00	99.04
crypt_11	12	2	16.70%	0	0	-6:69.5	31.75	99.63:99.69	99.66
crypt_12	10	4	40.00%	0	0	-340:77.5	-69.88	99.77:100.00	99.86
crypt_13	13	7	53.80%	1	0	-213.5:144	19.07	99.38:100.00	99.7
Composite	1490	395	26.51%	3	14	-1069.5:249	-25.89	94.90:100.00	99.45

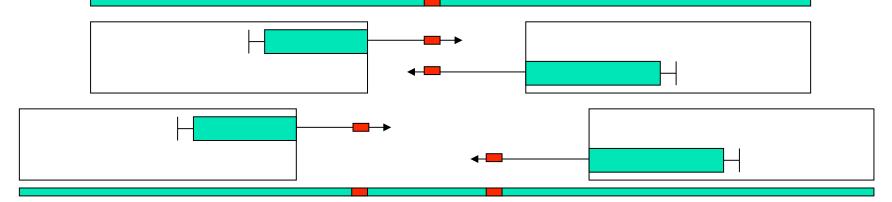
25%+ of all sequencing gaps closed with 3 mistakes.

### **Complicating Issues**

- Poly-monomer tails
  - Use dust to filter low complexity sequence



- Undetected repeats
  - Require strict agreement with scaffold

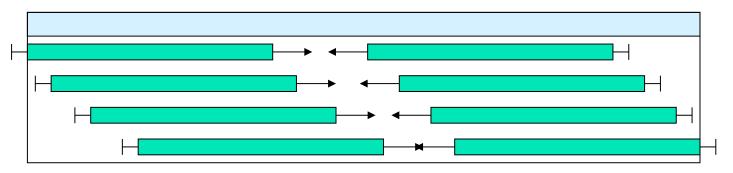


- Chimeric reads / Hard Stops
  - Good: Require high alignment similarity.
  - Better: Recognize hard stops by coverage gradients, other clues.
  - Best: Recognize unreliable sequence at chromatogram level.

### **Pre-Production Techniques**

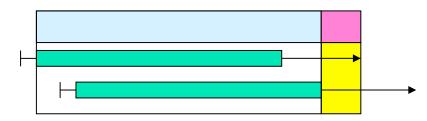
### Contig Fattening

• TVG coverage increased from 5.83X to 6.10X (mean extension: 80.5bp)



### Contig Growing

Extended 6144 edges in TVG (mean extension: 59.0bp)



### Measuring Assembly Quality

<ul> <li>Gross Status</li> <li>scaffold &amp; contig sizes</li> </ul>	QC file
<ul> <li>Connectedness Status</li> </ul>	Cloe, AssemblyViewer,
read & clone coverage	getCoverage, cvgChop, asmQC
Insert Status	
mate happiness, Q	C file, Assembly Viewer, asmQC
library randomness	
Consensus Status	Cloe, getCoverage, getqc
Quality Class	
Consensus Quality Value	e
Read Status	
Correlated SNPs	findTcovSnps

### Finding Suspicious Regions

#### % **ls**

blasted.frg blasted.asm

#### % /local/asmg/Linux/bin/cavalidate blasted

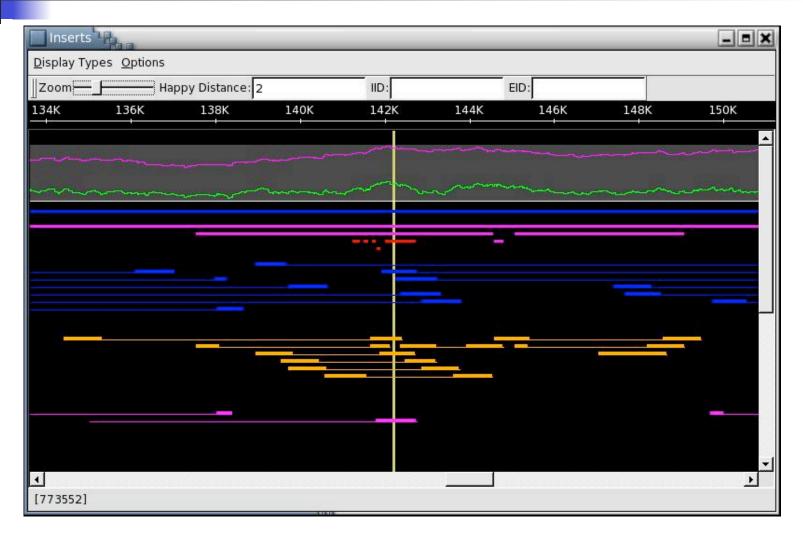
- Doing step 10: toAmos
- Doing step 20: bank-transact
- Doing step 30: asmQC
- Doing step 40: bank2contig
- Doing step 50: getCoverage
- Doing step 60: findTcovSnps
- Doing step 70: ClusterSnps
- Doing step 80: Load SNP Features
- Doing step 90: Find Surrogates
- Doing step 100: Load Surrogates
- Doing step 1000: Dump Features
- Doing step 1010: Get Suspicious Features
- Doing step 1020: Create Suspicious Regions
- % sort -nrk 6 blasted.snp.feat | head
- 1047283847436 P HIGH_SNP 141991 142712 46 15.67

### **Overcollapsed Repeat?**

Position 14207	7	€_	•	•	Co	nti	g ID	17	a la la		Ch	ron	no	DB	X	oc	8					Ins	sert	s C	Cont	tig	Gra	ph	A ⁺	A	· ]]I	Find	1							•	•
	-	1	12080	)				*		142	090				T.			81	421	00			P			14	2110	6			-			142	120						
Consensus	т	G G	т	G G	i C	C.	ΑT	с	G	сс	C A	G	с	Å	A	τ		4 C	G	G	сс	5 G	С	AA	A G	c	A	G G	с	G	ст	G	G	AC	ΞA	с	• G	G -	ΓG	с	А
XOEE429TF	۹Ţ	GG	Т	G G	i C	С	A T	C	G	C (	A	G	С	А	A	Т	A	4 C	G	G	c c	G	С	A A	A G	C	A (	G T	С	G											
XOEGI79TR	. т	GG	T	GG	C	C	AT	С	G	c c	A	G	С	A	A	Τ,	A	4 0	G	G	cc	G	с	A	A G	C	A (	G G	С	G	ст	G	G	AC	5 A	С	G	G .	r G	c	А
XOEDS21TF	• 1	GG	Т	GG	i C	С	A T	С	G	c c	A	G	с	А	A	T,	A	4 C	G	G	c c	G	С	A	A G	С	A (	G G	C	G	ст	G	G	A	G A	C	G	G T	T G	С	A
XOEE929TR	▲ T	GG	Т	GG	i C	C	AT	С	G	co	A	G	С	А	A	T ,	A	A C	G	G	CC	G	C	A	A G	C	A	G G	C	G	СТ	G	G	AC	5 A	C	G	G	T G	С	A
XOCAQ79TR	۰T	GG	т	G G	i C	C	AT	C	G	c c	A	G	С	А	A	T ,	A	A C	G	G	c c	G	С	A	A G	C	A	- G	С	G	ст	G	G	AC	G A	c	4	G	T G	С	А
XODAG73TF	• T	GG	T	GG	C	C	AT	С	G	c c	A	G	С	A	A	T,	A	4 0	G	G	cc	G	С	A	A G	C	A	G G	C	G	C T	G	G	AC	5 A	C	G	G .	T G	С	A
XOCA442TR	۳T	GG	Т	GG	i C	С	A T	С	G	c c	A	G	С	А	A	T,	A	4 C	G	G	c c	G	С	A A	A G	C	A (	G G	С	G	ст	G	G	AC	G A	C	G	G '	T G	С	A
XOEB547TF	.⊁ Т	GG	Т	GG	i C	C.	A T	С	G	C C	A	G	C	А	A	T ,	A	A C	G	G	C C	G	C	A	A G	C	A	G G	C	G	CT	G	G	AC	5 A	C	G	G '	T G	C	A
XOEG358TF	• 1	GG	Т	G G	C	C	AT	C	G	c (	A	G	С	А	A	T ,	A	A C	G	G	c c	G	C	A	A G	C	A (	G G	С	G	ст	G	G	AC	G A	C	G	G	T G	С	А
XOCAW58TF	▲ T	GG	T	GG	C	C	AT	С	G	c c	A	G	С	A	A	T,	A	4 C	G	G	CC	G	С	A	A G	C	A (	G G	С	G	C T	G	G	AC	G A	C	G	G .	T G	С	А
XOEDG13TR	• т	GG	Т	G G	i C	С	A T	С	G	c c	C A	G	С	А	A	T ,	A	4 C	G	G	C C	G G	С	A	A G	C	A (	G G	C	G	ст	G	G	A (	G A	C	G	G T	T G	С	А
XOEBD56TF	<b>▲</b> T	CG	Т	GG	i C	C.	A T	С	G	C C	C A	G	С	С	A	Т	G	G C	G	G	C C	G	С	A	A																
XOEET59TR	► T	C G	Т	GG	i C	C	A T	С	G	СС	C A	G	С	С	A	Т	G	G C	G	G	C C	G	С	AA	A G	С	А	G	С	G	с т	G	G	AC	G A	8					
XOEEF32TF	• T	C G	Т	GG	i C	C	A T	С	G	c c	C A	G	С	С	A	Т	G	G C	G	G	C C	G	С	A	A G	С	A	G G	С	G	с т	G	G	AC	G A	С	G	G T	T G	С	А
XODAE61TF	۴T	C G	Т	G G	i C	C	A T	С	G	c c	A	G	С	С	A	т	G	G C	G	G	c c	G	С	AA	A G	С	A	G G	С	G	с т	G	G	AC	G A	C	G	G	T G	С	А
XODAZ68TR	► T	C G	Т	GG	i C	C.	A T	С	G	C C	C A	G	С	С	A	Т	G	G C	G	G	C C	G	С	A	A G	C	A	G G	С	G	с т	G	G	AC	G A	С	G	G T	T G	С	A
XODAZ91TF	<b>▲</b> T	C G	Т	GG	i C	C.	A T	С	G	c c	C A	G	С	С	A	Т	G	G C	G	G	C C	G	С	AA	A G	С	A	G G	С	G	с т	G	G	AC	GΑ	C	G	G T	T G	С	Α
XOECD04TR	► T	C G	Т	GG	i C	C	A T	С	G	c c	C A	G	С	С	A	Т	G	G C	G	G	c c	G G	С	A	A G	C	A	G G	С	G	с т	G	G	AC	G A	С	G	G T	T G	С	Α
XOEB475TR	<b>▲</b> T	C G	Т	GG	i C	C	A T	С	G	C C	C A	G	С	С	A	Т	G	G C	G	G	CC	G	С	AA	A G	С	A	G G	С	G	с т	G	G	AC	G A	С	G	G	T G	С	Α
XODAJ66TF	► T	CG	Т	GG	i C	C	A T	С	G	C C	C A	G	С	С	A	Т	G	GC	G	G	CC	G	С	AA	A G	C	A	G G	С	G	с т	G	G	AC	G A	С	G	G T	T G	С	Α
XODB891TF	۴T	CG	Т	GG	i C	C	A T	С	G	CC	A	G	С	C	A	Т	G	G C	G	G	CC	G	С	AA	A G	C	A	GG	С	G	СТ	G	G	AC	G A	C	G	G T	T G	С	A

For a bacterial sample, correlated discrepancies strongly suggest a repeat has been collapsed.

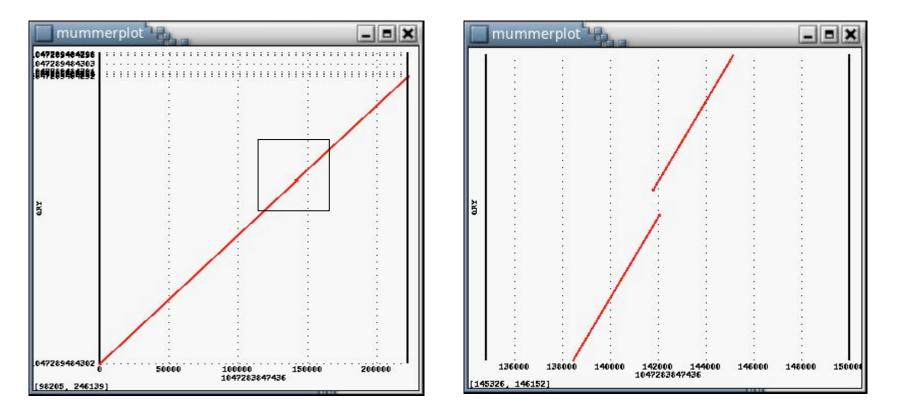
### Mate View



The shrunken mates (orange) suggest the assembly has a deletion from the true sequence.

### Local Assembly

- % run_CA -local -noedit -noupload local.frg -dir ca-0.003 -e 0.003
- % nucmer 1047283847436.fasta ca-0.003/local.fasta
- % /local/asmg/Linux/bin/mummerplot out.delta -R 1047283847436.fasta -Q ca-0.003/ local.fasta -layout -filter



### **Resolved Repeat**

Zoom	Hap	py Distance: 2		IID: 544	Ell	D: XODA533TR		
134K	136K	138K	140K	142K	144K	146K	148K	150K
					~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~~~~		
~~~	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	~	Sur and		man m		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	m
		-		_		-		
-		-					_	_

Unfortunately, size violated mates are only a clue. Ask Mihai for current research techniques.

### **Final Results**

	Original Assembly	Reassembled
TotalScaffolds	21	5
TotalSpanOfScaffolds	4770228	4819528
IntraScaffoldGaps	51	33
MeanSequenceGapSize	410.18	96.94
[Top5Scaffolds]		
1	25:2156009:2164583:357.25	34:4810208:4813407:96.94
2	13:826284:830667:365.25	1:2558:2558:0.00
3	3:437076:439209:1066.50	1:1473:1473:0.00
4	3:333768:334012:122.00	1:1056:1056:0.00
5	5:310971:311756:196.25	1:1034:1034:0.00
TotalContigsInScaffolds	72	38
N50ContigBases	151430	253084
TotalDegenContigs	125	1
DegenContigLength	129182	959
MeanDegenContigSize	1033.46	959
[Top5Contigs]		
1	5939:536280	6291:516881
2	2751:259499	6203:436501
3	3005:238048	3805:385244
4	2199:220720	4292:364829
5	2509:196450	4025:355502

#### Expected Genome Size: 4.85Mb - 5.04Mb

### **Research Directions**

- AutoEditor 2.0: Better results, better engineering
- Context Based trimming
  - Partial Overlaps
  - Reference sequence
- Advanced CA Techniques
  - Contained Stones bug fixes
  - Blasting Degenerate and Surrogate Unitigs
  - Assembling in the gap
- Arachne & Other Assemblers
- Assembler Reconciliation
- AMOS Framework
- Assembly Forensics
- Assembly Visualization / Navigation

### Conclusions

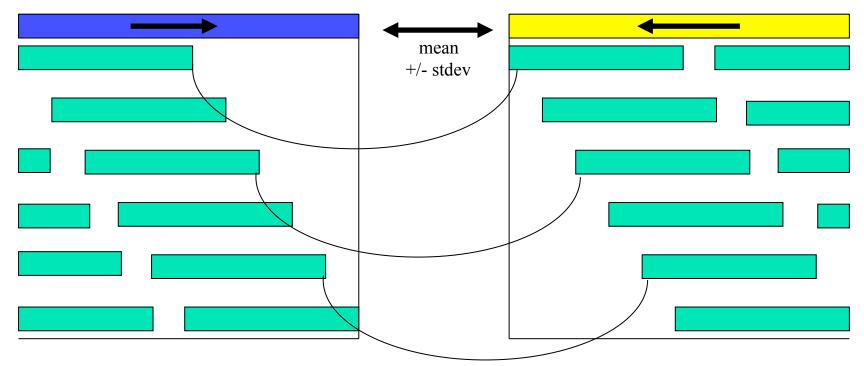
- Overriding strategy: Start conservatively, and iteratively build as more information becomes available.
- 95.5% 99.2% of genome in a single scaffold not typical yet, but it could be.
  - Be aware of potential size/quality tradeoffs, though.
- Assembly is complicated by genome structure, repeat characteristics, data quality, data management- one size does not fit all, ask for help.
  - Use Data Support!

### Acknowledgements

- Steven Salzberg
- Martin Shumway
- Jason Miller
- Pawel Gajer
- Art Delcher
- Mihai Pop
- Adam Phillippy

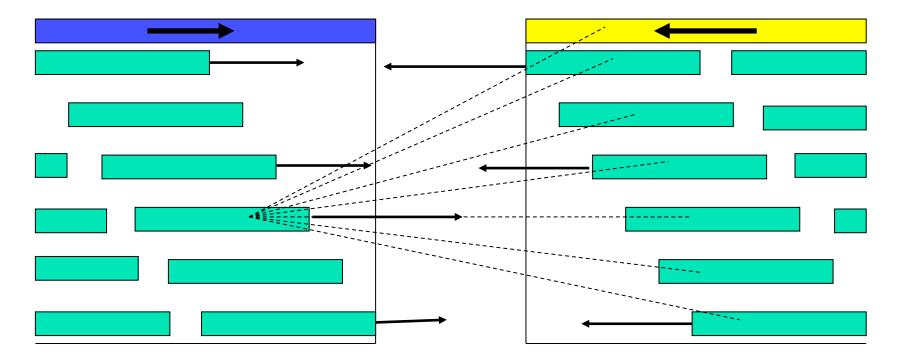
- WGA
- SE
- Data Support
- Jane Carlton
- Vish Nene

# Sequencing Gap



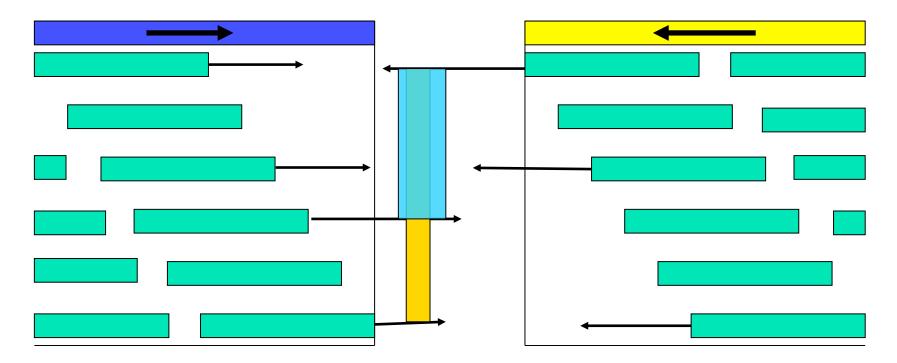
The individual reads (green) have been assembled into 2 contigs (blue & yellow). The mate relationship between the reads allows for the contigs to be oriented and the gap size to be estimated.

## All-vs-all Alignment



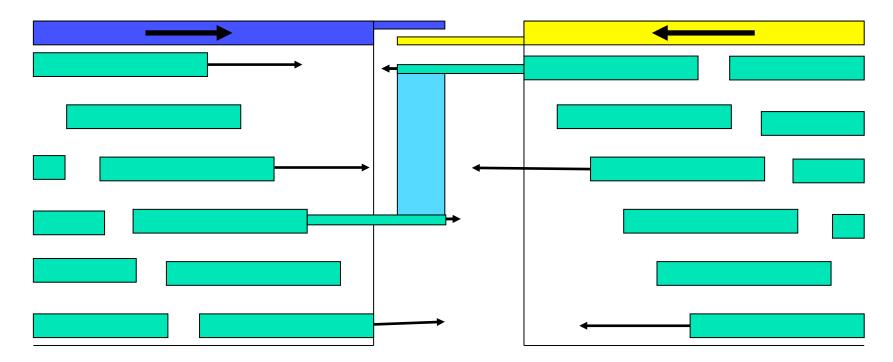
1. An all-vs-all pairwise alignment between the full range sequences from the flanking contigs is computed.

# Alignment Analysis

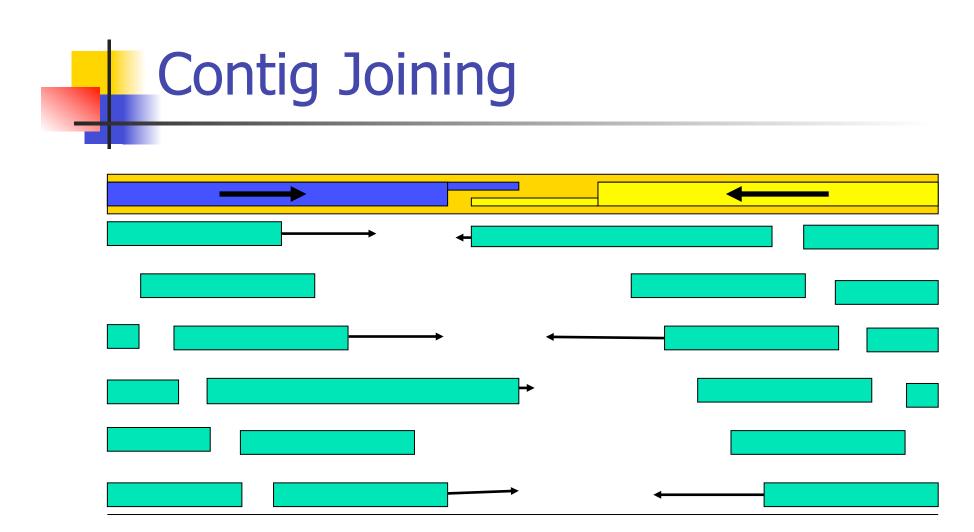


2. The alignments are tested for consistency with the scaffold and for being of sufficient quality. If any alignments satisfy the requirements, the best alignment (blue) is selected for joining the contigs.



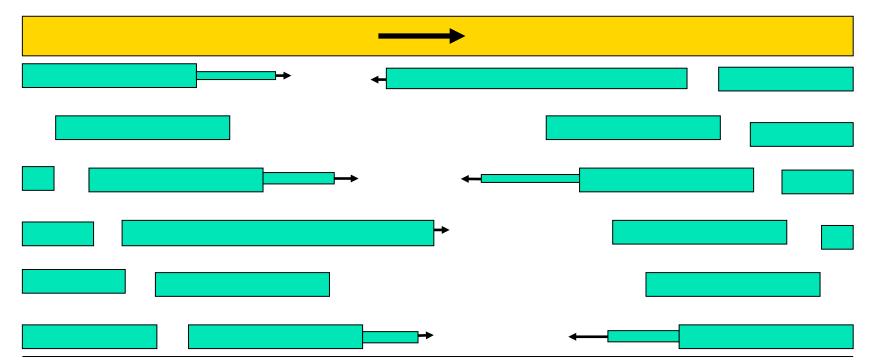


3. The contigs are extended by extending the selected reads beyond their original clear range to the desired position. If necessary, the reads are first aligned to the existing consensus.



4. The contigs are joined by aligning the newly extended consensi. Alignment gaps inserted into the conseni are promoted into the appropriate positions in the underlying multiple alignment. The joined contig (orange) replaces the original two in the scaffold.





5. The join region is fattened to increase the depth of coverage and enhance the consensus quality.