New tools for new tech data at the WTSI

Darren Grafham

The Wellcome Trust Sanger Institute

Outline

» Gap5

» 12 pooled bacs on 454

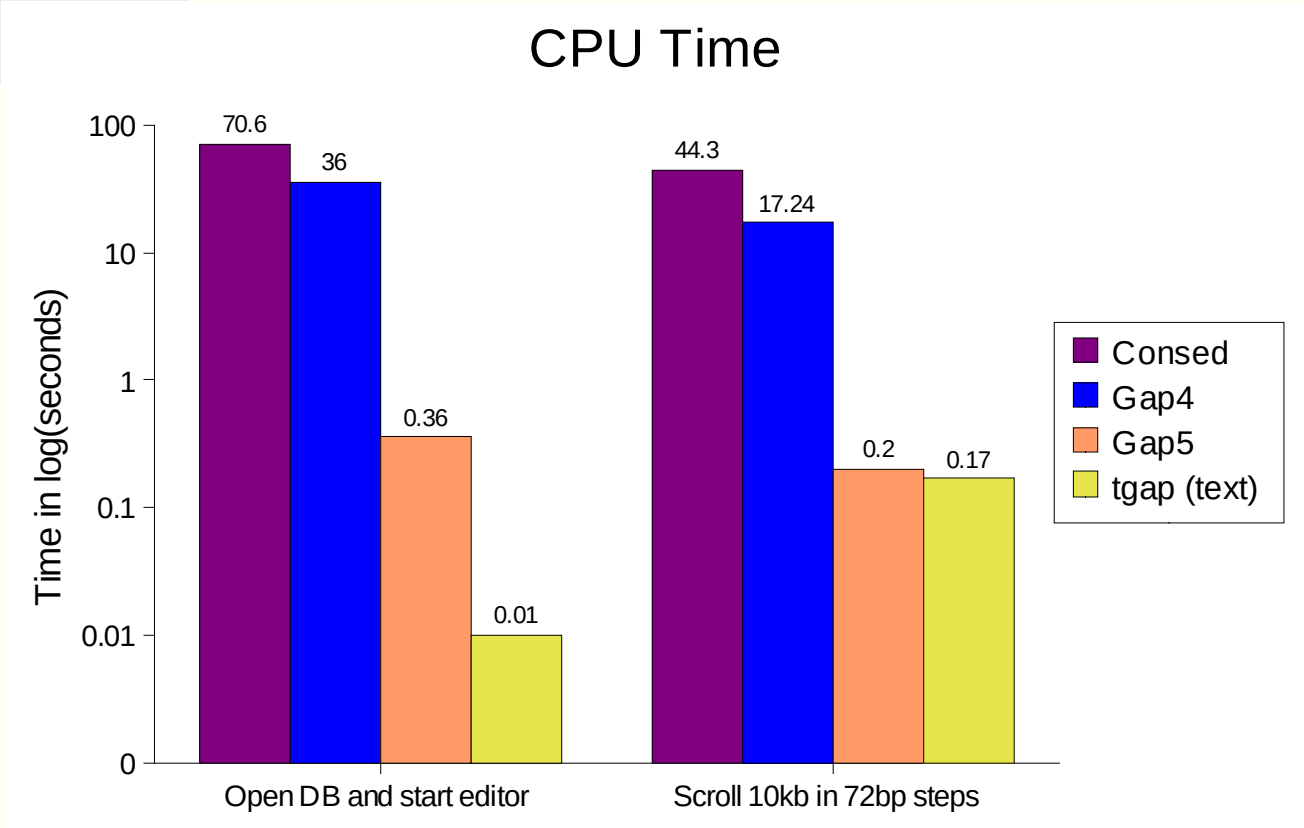
» 55 pooled bacs on Illumina

» Most recent trials and future plans

Running Gap5

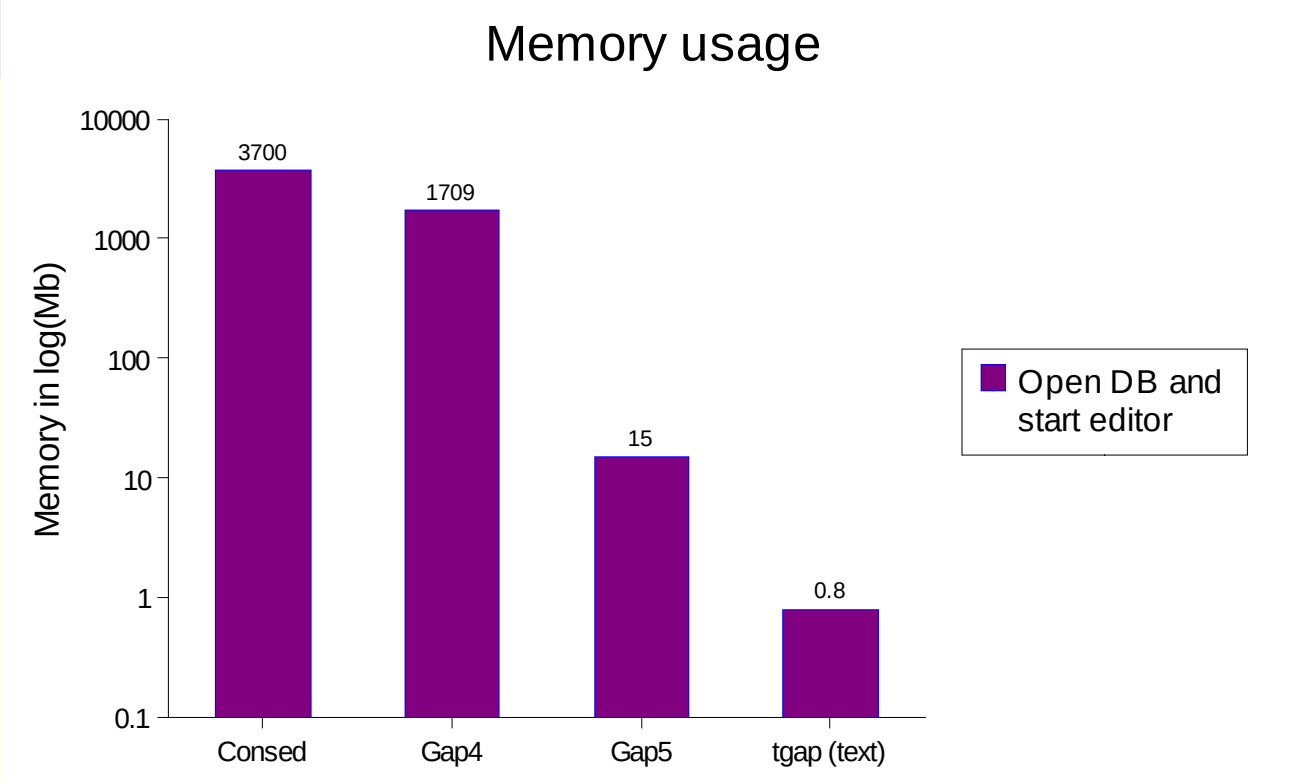
- tg_index-1.1 -o *g5db.0* -p -m *input.map*
  - Builds Gap5 databases from various input formats.
  - Supports short and long format MAQ output; BAM files (but not SAM); plain text ".aln"; BAF; ACE (sort of).
  - Processes ~20,000 sequences/sec (from BAM).
- tg_view-1.1 *g5db.0*
  - Text based viewer (aka tgap).
  - Very similar to "samtools tview foo.bam".
- gap5-1.1 [-ro] *g5db.0*
  - Graphical viewer and/or editor.
  - http://sourceforge.net/projects/staden/

Benchmarks – cpu time



» Tested on S.Suis: 1.7million fragments, 2Mb genome.

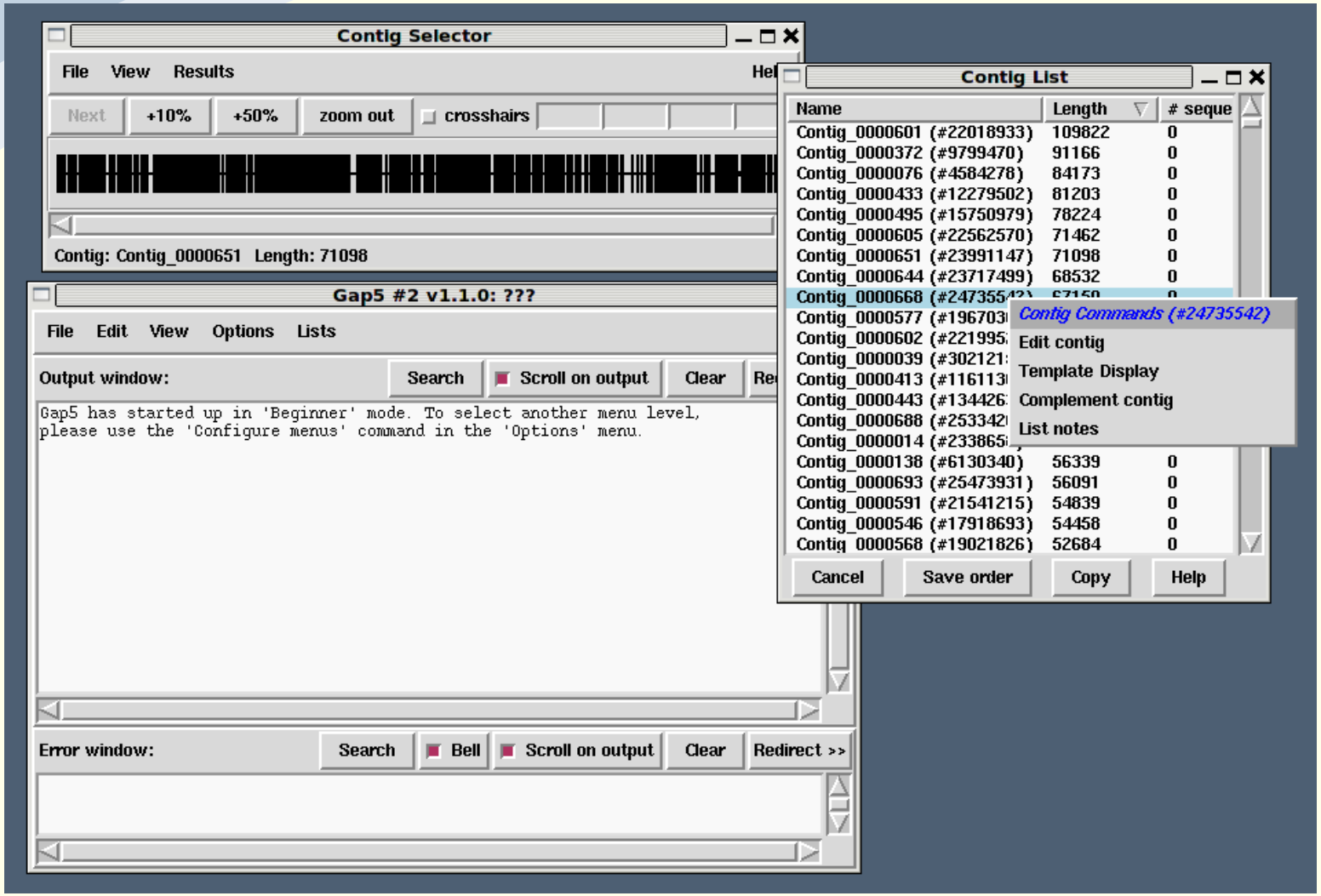
Benchmarks – memory



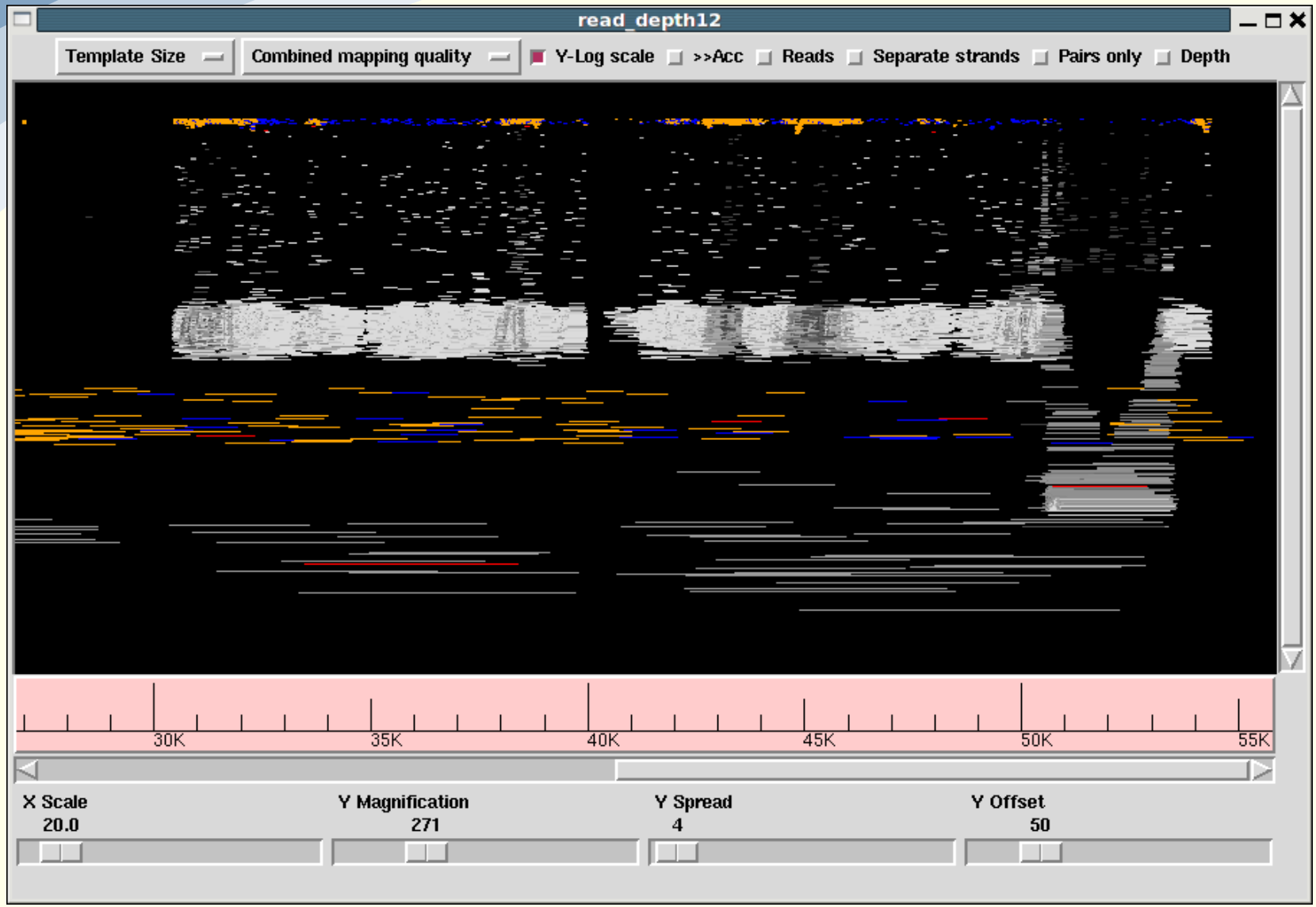
» Gap5: proportional to number of contigs

» Gap4: proportional to number of

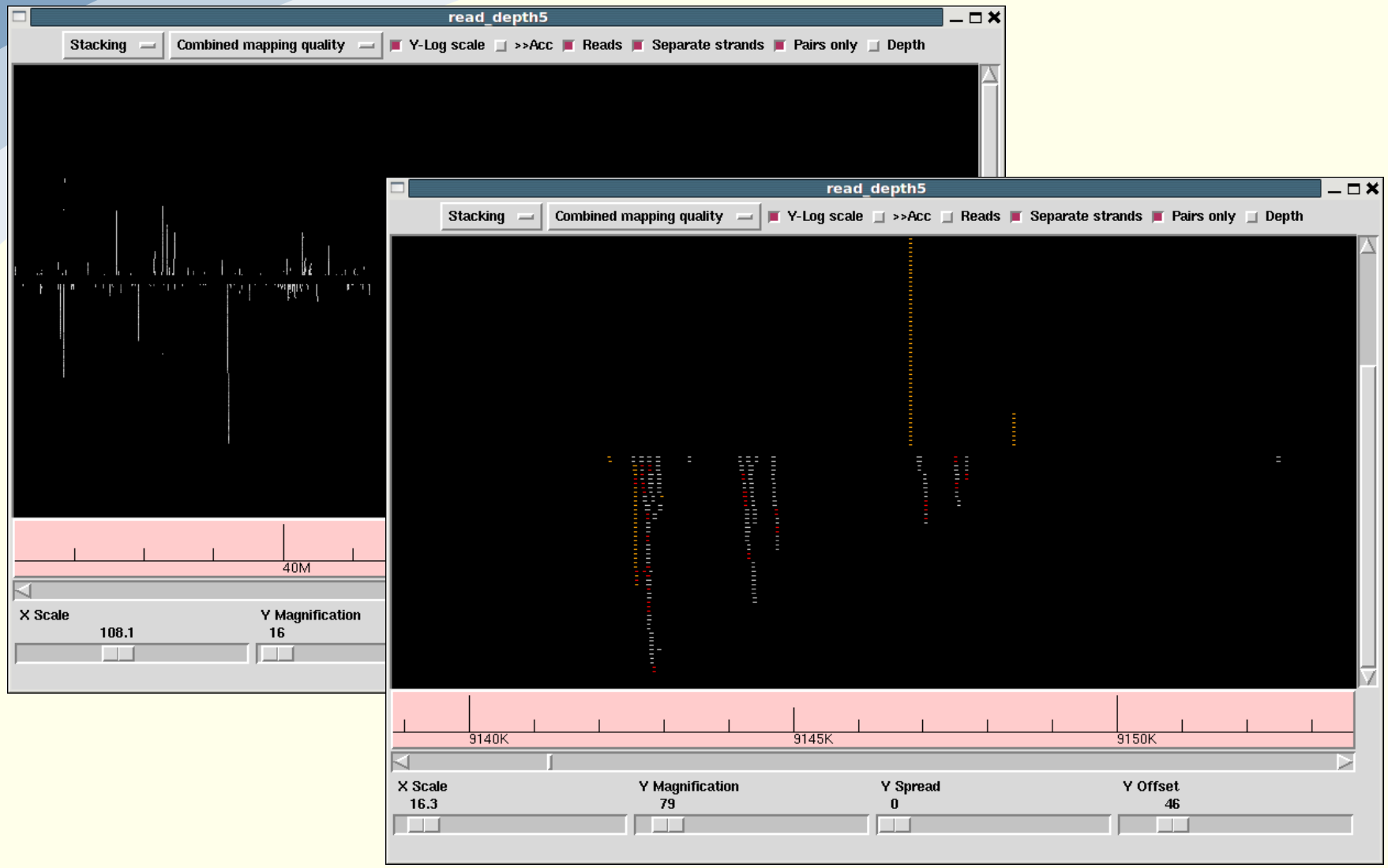
Gap5 start-up



Template display- by size



Template display- by strand



wellcome trust sanger institute

12 clone pool

BAC	Number of gaps in *de novo* capillary data	Number of gaps in *de novo* 454 data	Number of gaps in 454 assembly covered by WGS data	Number of gaps remaining in 454 data after addition of WGS data
zH117H1	13	23	15	8
zH141B18	13	22	13	9
zH151M17	28	26	18	8
zH117E7	10	26	18	8
zH137D22	14	25	20	5
zH97A24	22	20	n/a	20
zH146D21	23	36	20	16
zH147D24	7	16	12	4
bE2F11	20	21	n/a	n/a
bE156J20	13	46	n/a	n/a
bE240L11	36	68	n/a	n/a

454 to gap

- http://genome.imb-jena.de/software/roche454ace2caf/Poster_UserMeeting_GS20_Munich_070328.pdf

- In house modification to produce a gap4 assembly from a 454 assembly. Allows traditional visualisation of sequence reads
- Looking to automate the PCR gap closure process via ABACUS
- More details can be found on poster 82 by Sarah Pelan

Second Set with 50 Zfish Clones

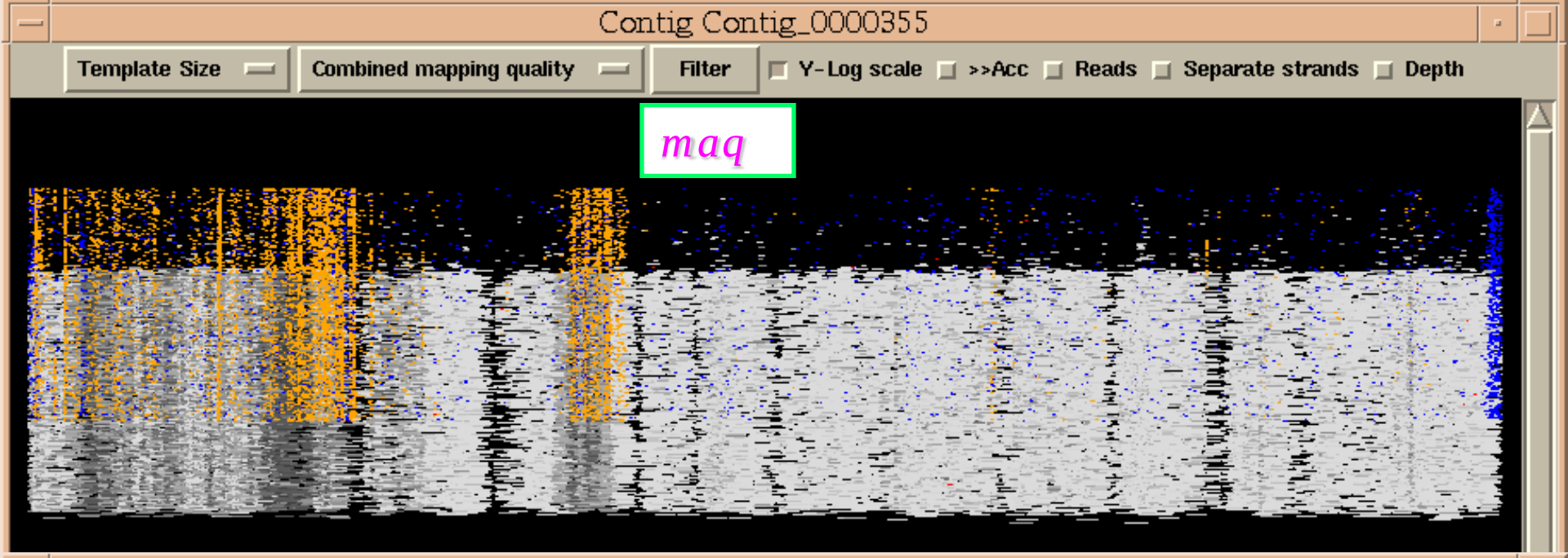
Solexa reads:

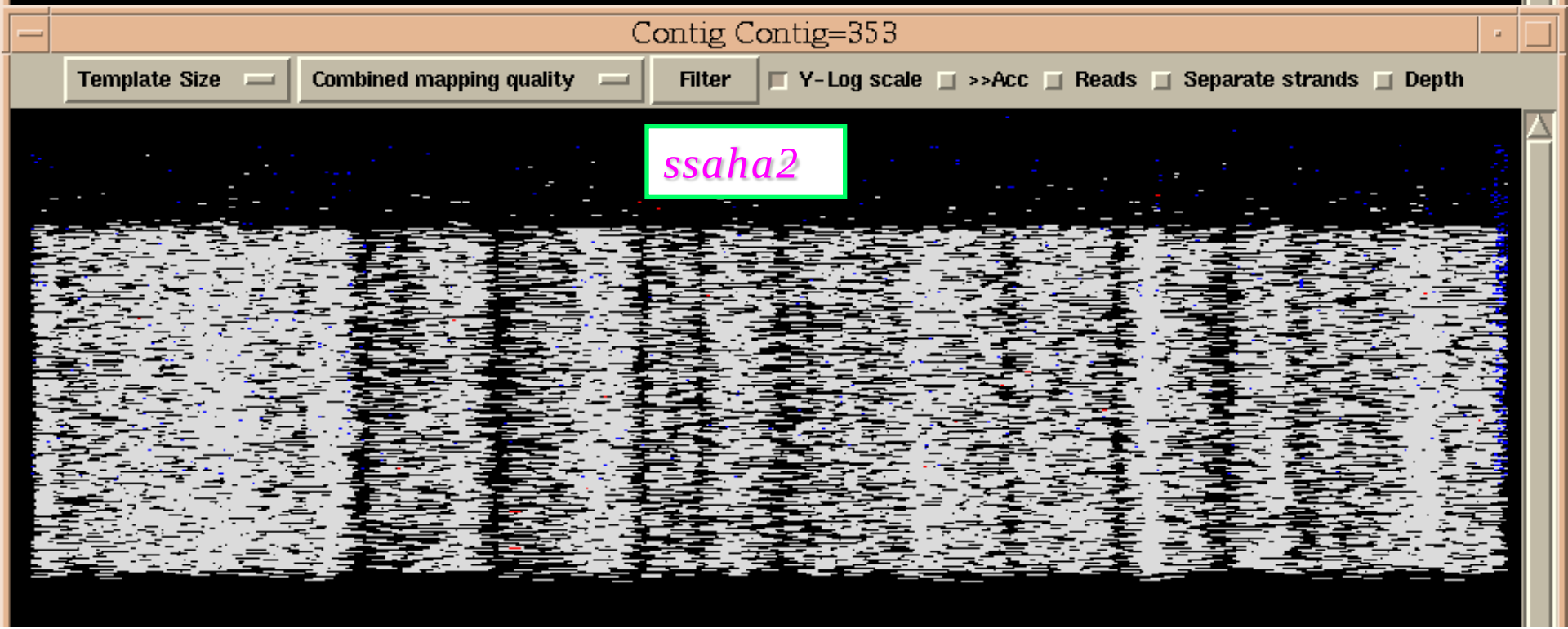
Number of reads:	17.5 million;
Estimated size of covered region :	~9.0 Mbp;
Read length:	2x54bp;
Estimated read coverage:	~190X;
Insert size:	260/50-400 bp;

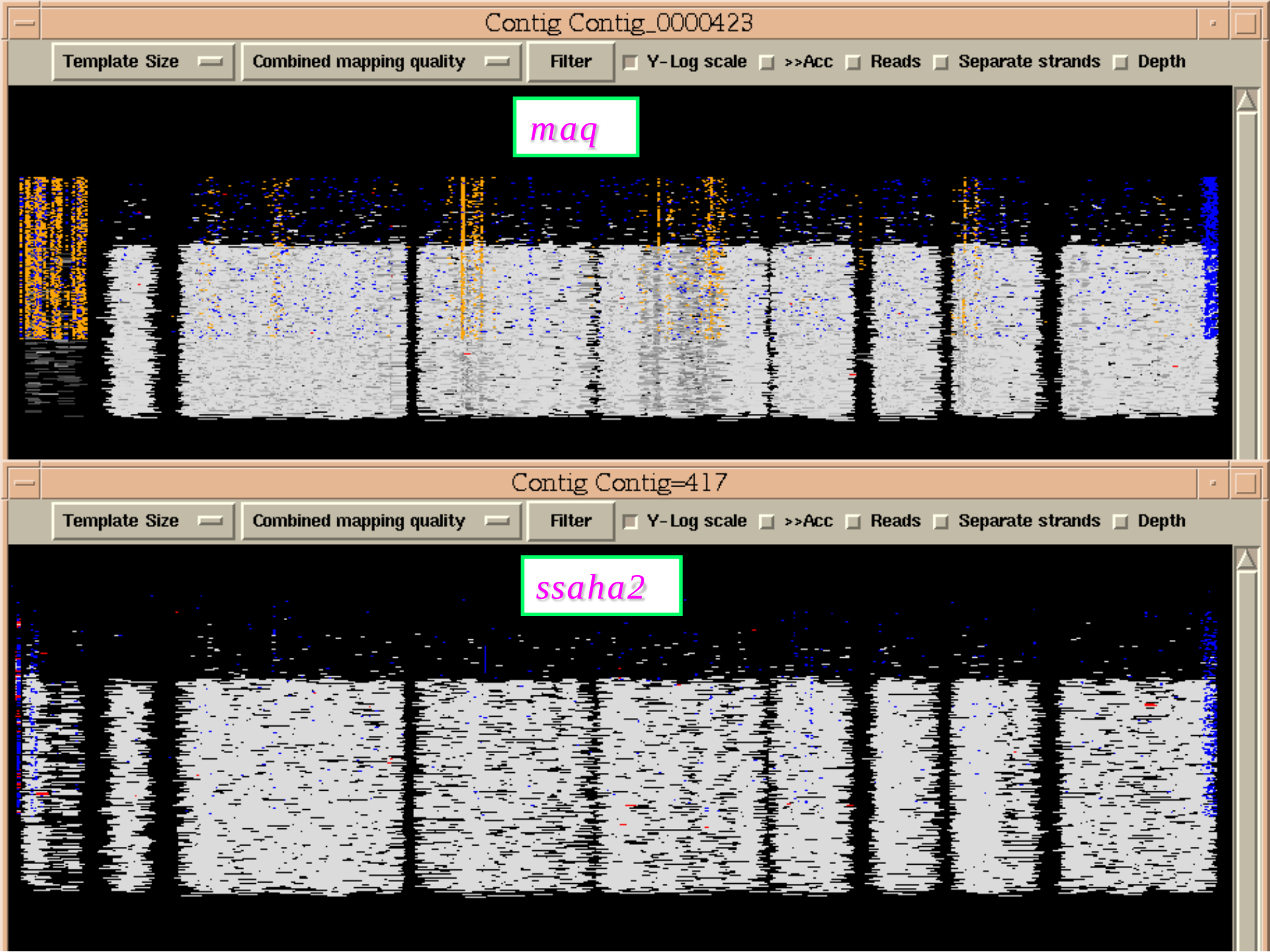
Zfish DH capillary reads: 112,583

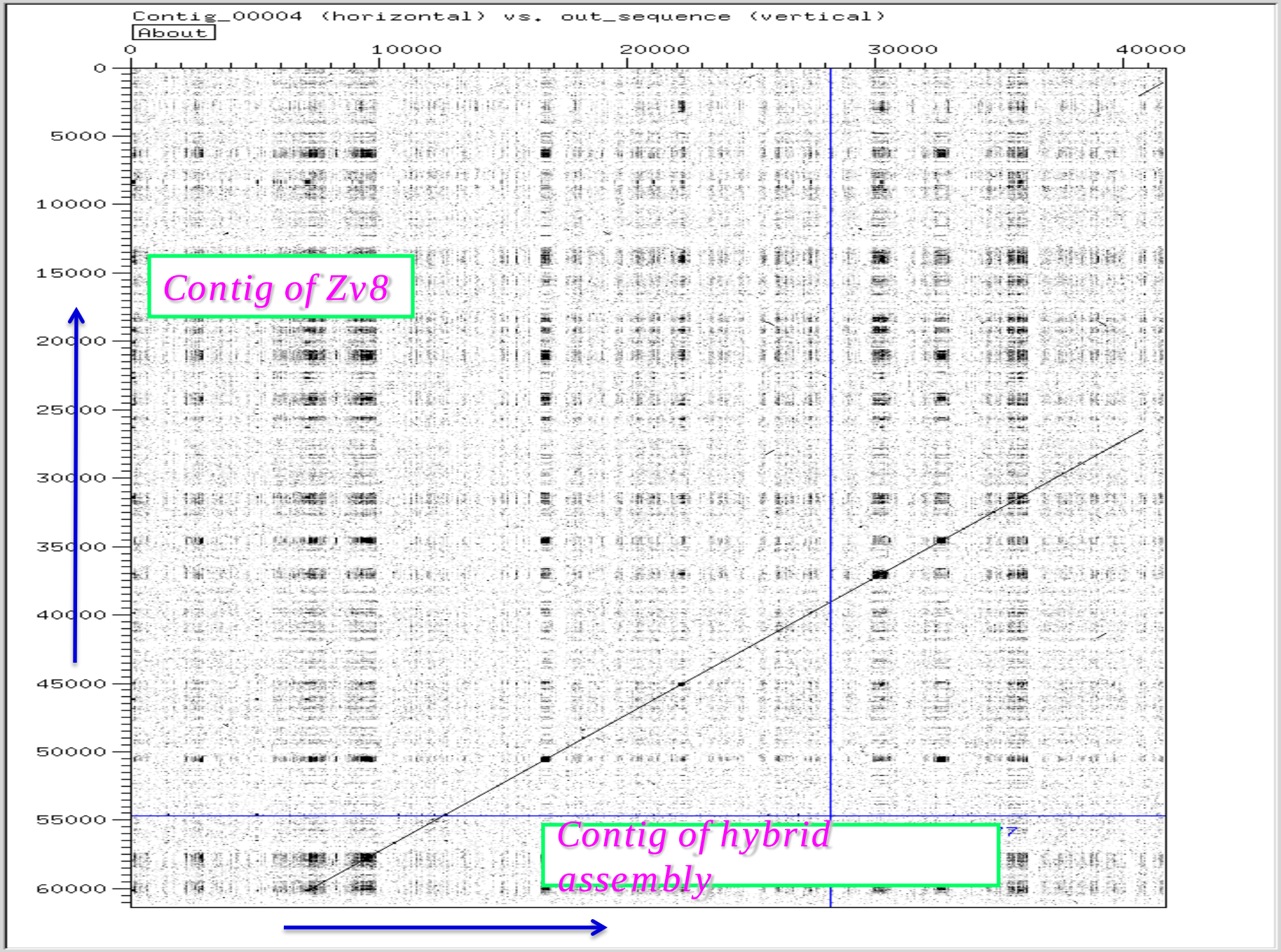
Assembly features: - contig stats

	Solexa	Hybrid_Ctg	Hybrid_Super
N contigs:	3,143	688	359
Bases:	4.01 Mbp	8.39 Mbp	8.43 Mbp
N50 size:	3,189	24,448	70,703
Largest	23,018	108,090	274,224
Averaged:	1,275	12,194	23,493
Coverage:	~50%	~93%	~94%
Errors:	?	?	?


Contig Contig_0000355
Template Size
Combined mapping quality
Filter
Y-Log scale
>>Acc
Reads
Separate strands
Depth
maq


Contig Contig=353
Template Size
Combined mapping quality
Filter
Y-Log scale
>>Acc
Reads
Separate strands
Depth
ssaha2


Contig Contig_0000423
Template Size
Combined mapping quality
Filter
Y-Log scale
>>Acc
Reads
Separate strands
Depth
maq
Contig Contig=417
Template Size
Combined mapping quality
Filter
Y-Log scale
>>Acc
Reads
Separate strands
Depth
ssaha2


Dotter Contig_00004 vs. out_sequence
Contig_00004 (horizontal) vs. out_sequence (vertical)
About
0
10000
20000
30000
40000
5000
15000
25000
35000
45000
50000
55000
60000
Contig of Zv8
Contig of hybrid assembly

Reference: TATTACAAAAAATGTACATATATAAAGTAATATATATAGTATATATTACAGTATGCACTATATATATATACTATATAT ATATATATATATATATATATATTTTTTTTTTTTTTTTTCATTA

» SSAHA CTG417: TATTACAAAAAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNTATTTTTTTTTTTTTTTTTTCATTA

»

» Reference: TATTTTTTTTTTTTTTTTTTCATTATGCCAGTAATAAAATAAAGTAATTGCTGATTTTCTTAATTTATTTTCTTTATTTA TTTATTTATT

» SSAHA CTG417: TATTTTTTTTTTTTTTTTTTCATTANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATTTAGTTTCTTT ATTTATTTATTTATT

»

»

» Reference: CAAATTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATT TTATTTTATTTTATTTTATTCAGCC

» SSAHA CTG417: CAAATTTATTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTATTTTATTTTATTTTATTTTATTTTATT TTATTTTATTTTATTTTATTTTATNCAGCC

» MAQ CTG: CAAATTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTA TTTTATTTT

»

» Reference: CAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTAATTTTATTTAAT TTTATTTAATTTTATTTAATTTTATTTAATTTAATTTAACTAGG

» SSAHA CTG417: CAATTTAATTTAATTTAATTTAATTTAATTNNNNNNNNNNNNAATTTAATTTAATTTAATTTAATTTAATTTTATTTAAT TTTATTTAATTTTATTTAATTTTATTTAATTTAATTTAACTAGG

»

» Reference: AATAATAAAATAATAATAATAATAATAATAATAATAATAATAATTGTAATAATAATAATAATAATAATAATAATAATAATAAAAATC ATCATCATAACAATATTAATAATTATAATAATAAAATAATAATAATAATAATAATAATAATAATAAAATAATA

» SSAHA CTG417:

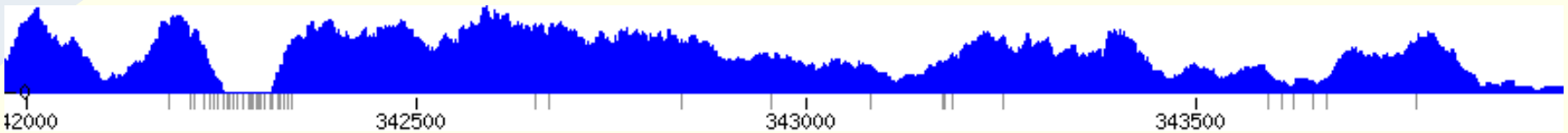
» AATAATAAAATAATAATAATAATAATAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATAATAATAAA ATAATA

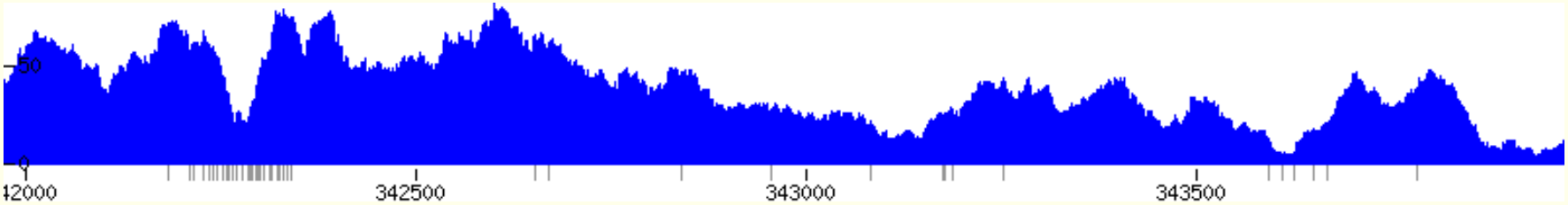
»

Where is the missing sequence?

- In the 50 clone trial the sequence was missing from the sequencing library. Development of no PCR library generation for subsequent trial removes the sequencing issue and pushes the problem to be an assembly issue.
- 55 clones is too many to handle currently not in terms of data generation but splitting back without other information and assigning to each clone reliably for manual intervention
- Zemin's advice was to pool 12 bacs on 1 lane of solexa.

Is this seen in malaria ?





» Snpomatic will only place a read with an exact match, and uses prior data on possible SNP locations (tick marks below the line) when mapping reads. Maq will allow base mismatches during mapping, but will also map reads randomly if they have an equal mapping score. Both approaches show the same distribution of read depth coverage in a sample.

New tech clone trial 3

- Trial number III is now underway and is in 2 parts:
-
- To investigate if all problematic, repetitive zebrafish clones can be sequenced and finished to Phase III using the 454 Titanium platform in pools of 12.
-
- To investigate if all other zebrafish clones can be sequenced and finished to Phase III using the Illumina platform in pools of 12 clones in combination with previously capillary sequenced Whole Genome Shotgun where possible

New tech clone trials

- 454 data is now available for a set of 12 known problem BACs.
- Initial results gave an assembly with many contigs (1440) and took 4 days
- Discussions with Jim Knight have given us the new assembler which does a better job and faster (391 contigs) and 15 minutes to run.
- Still a lot of contigs for 12 clones and work is ongoing

New tech clone trials

- 6 pools of 12 unknown BACs are currently being sequenced in 6 separate lanes using Illumina.
- The 6 pools will be assembled by Zemin in combination with the WGS capillary data before being available as 6 separate databases containing a pool of 12 BACs each
- This data set will be the first time we have had no capillary clone data to refer to, or to help with ordering and orientating the contigs. Therefore it will be our first true “Live” test. The initial plan is to make use of all the data we do have available to us to help get the correct contigs together such as BAC ends sequences, associated supercontigs and any available overlap data.
- We hope to then make use of ABACAS to order primers for PCRs using the supercontigs as a reference.

Conclusions

- New tools are arriving such as 454togap and gap5 to help visualisation of new tech data.
- New programs like ABACAS (Algorithm Based Automatic Contiguation of Assembled Shotgun sequences) and CARMA (Correction And Reference Morphing Algorithm) (not covered today) are expediating the process but if you require reference level sequence then manual intervention/review will be required.
- Better algorithims to assemble with are required along with co-assembly of data

Conclusions

- » Clones can be finished via new technology but there are constraints
- » High AT organisms benefit from NO PCR library step when using solexa
- » Assembly alogorithms need to be able to cope with SSR's

Acknowledgments

» Gap5 – James Bonfield

» Ssaha and Maq assembly – Zemin Ning, Yong Gu, Tony Cox

» Jim Knight-454 assembly

» Clone finishing – Sarah Pelan, Siobhan Whitehead, Karen Holt, Helen Beasley