Consed 19.0, Crossmatch Adapted for Solexa and 454 Reads

David Gordon and Phil Green

HHMI at University of Washington

Major Changes in Consed 19.0, Cross_match

- New Features Aimed at Mutational Profiling/Resequencing more than Finishing
- Handles Millions of Solexa/454 Reads
- Cross_match so fast (faster than maq) that it now is used as a database search engine
- Phaster than cross_match

Speed: Cross_match vs MAQ

- One million solexa reads searched against human genome
- MAQ time: 130 minutes
- Cross_match time: 85 minutes (only 18 minutes if filtered repeats)
- Cross_match settings such that it was slightly more sensitive than MAQ

Phaster is faster

- Takes 35 seconds to 1min:38 seconds for a million exact match simulated reads against the human genome
- Not based on Burrows-Wheeler Transform
- Will do gapped alignments
- Only 2 weeks old

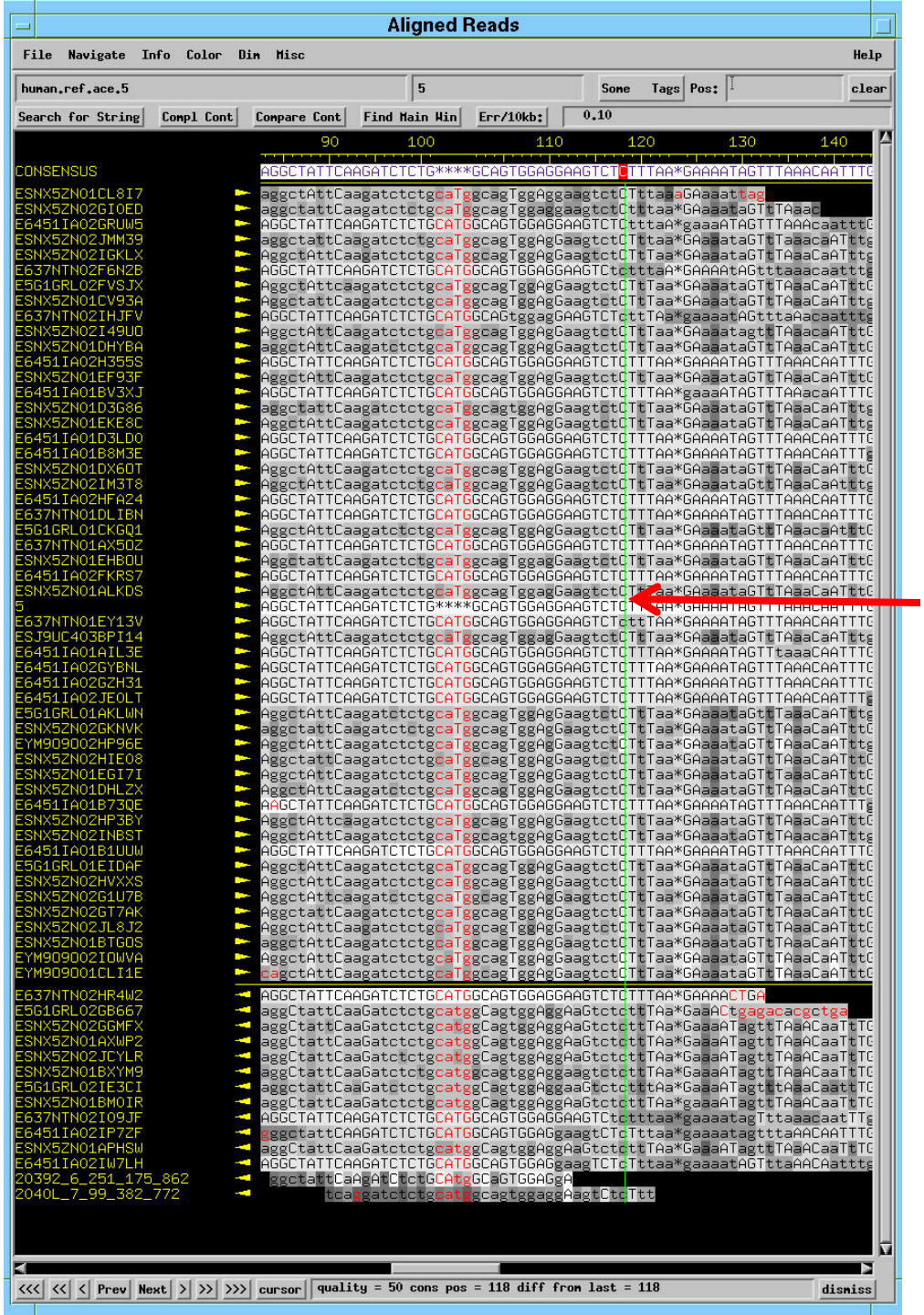
Cross_match to Consed

- Prepares a consed-ready assembly in a fraction more time
- 3 to 4 min to start Consed
- Peak Memory Usage: 4.6 Gb

New Consed Features

- Finding variants
- Handles Millions of Solexa/454 Reads
- Handling depth of coverage of millions of reads
- Assembly View shows consistent/inconsistent 454 and Solexa reads

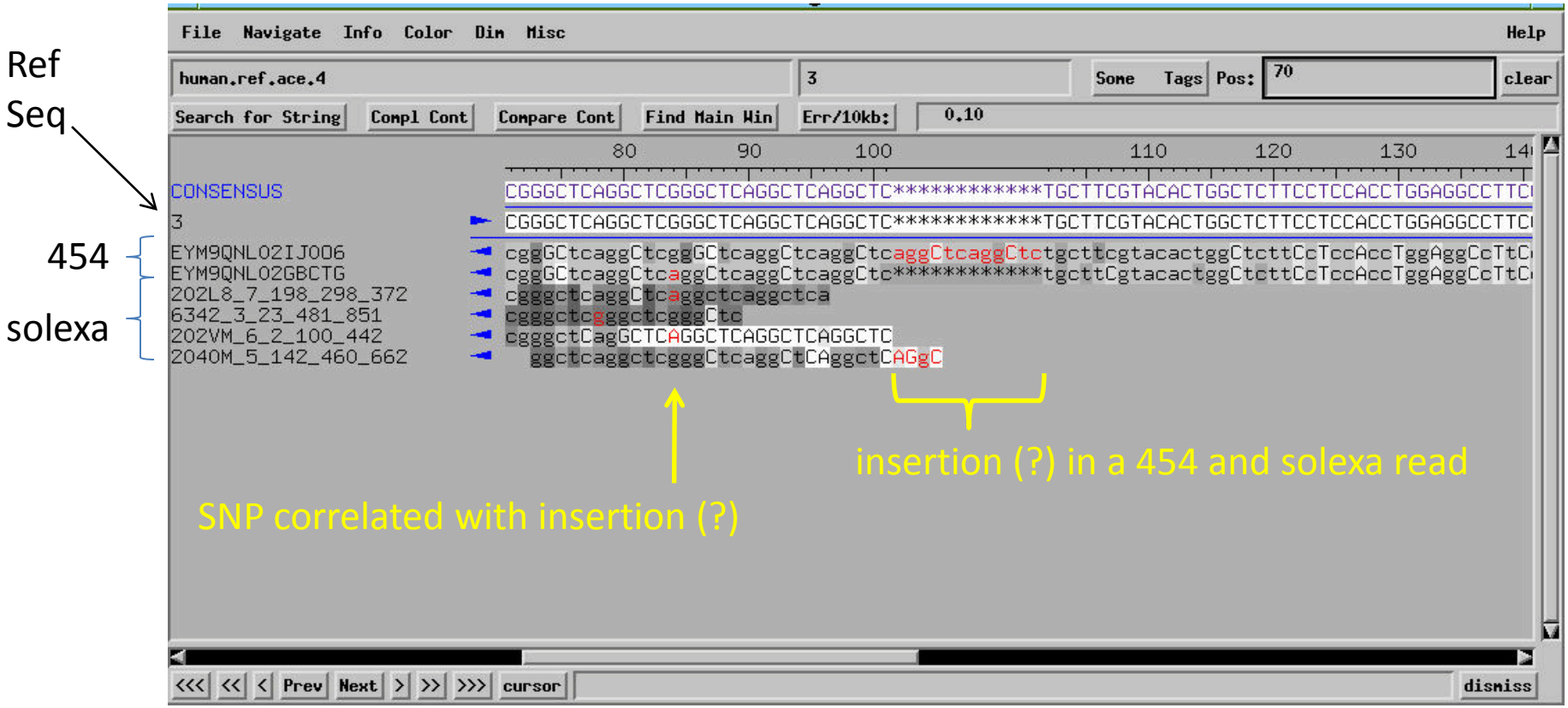
Can Mix Different Read Types



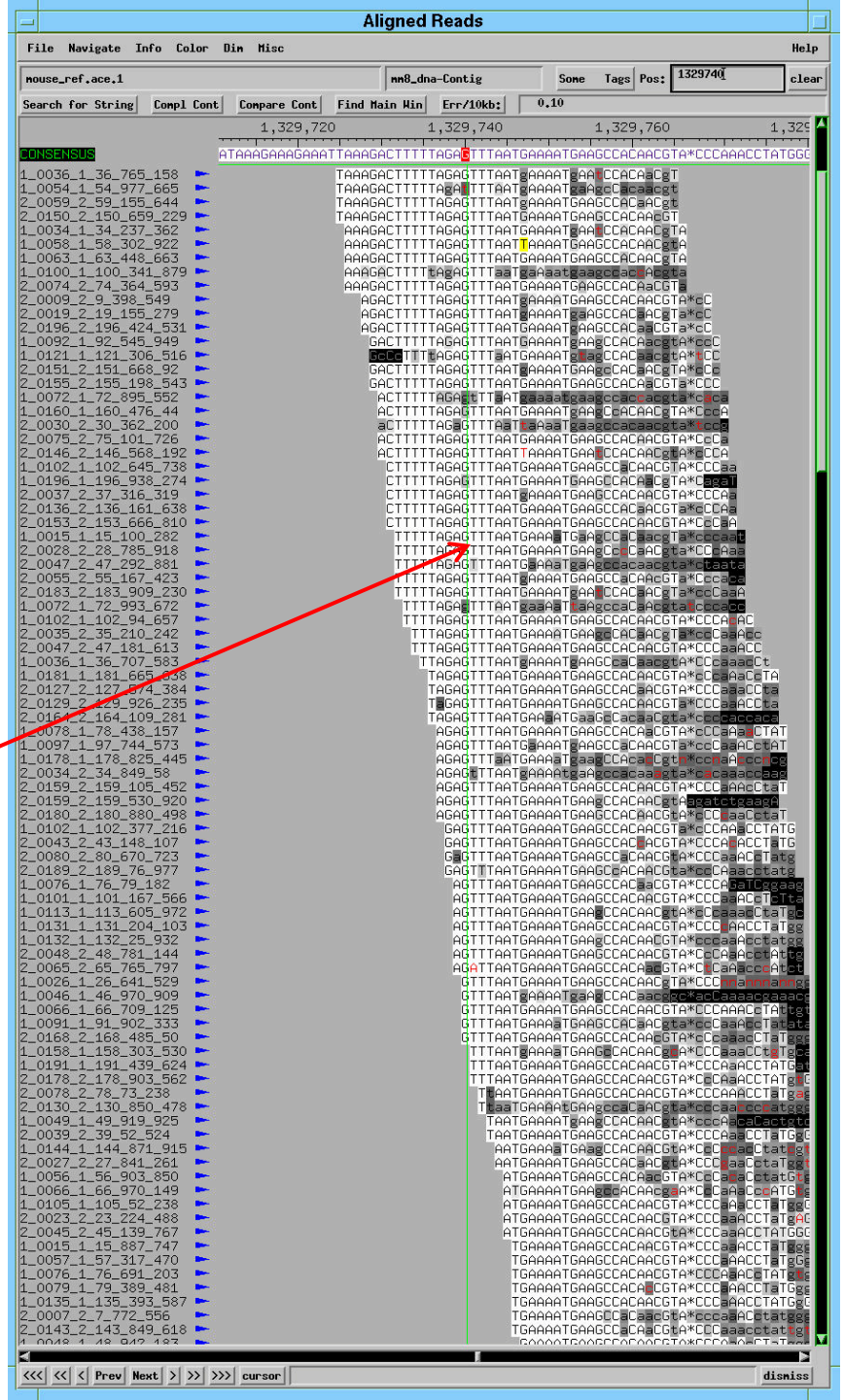
454 reads

solexa reads

green vertical cursor aids eye

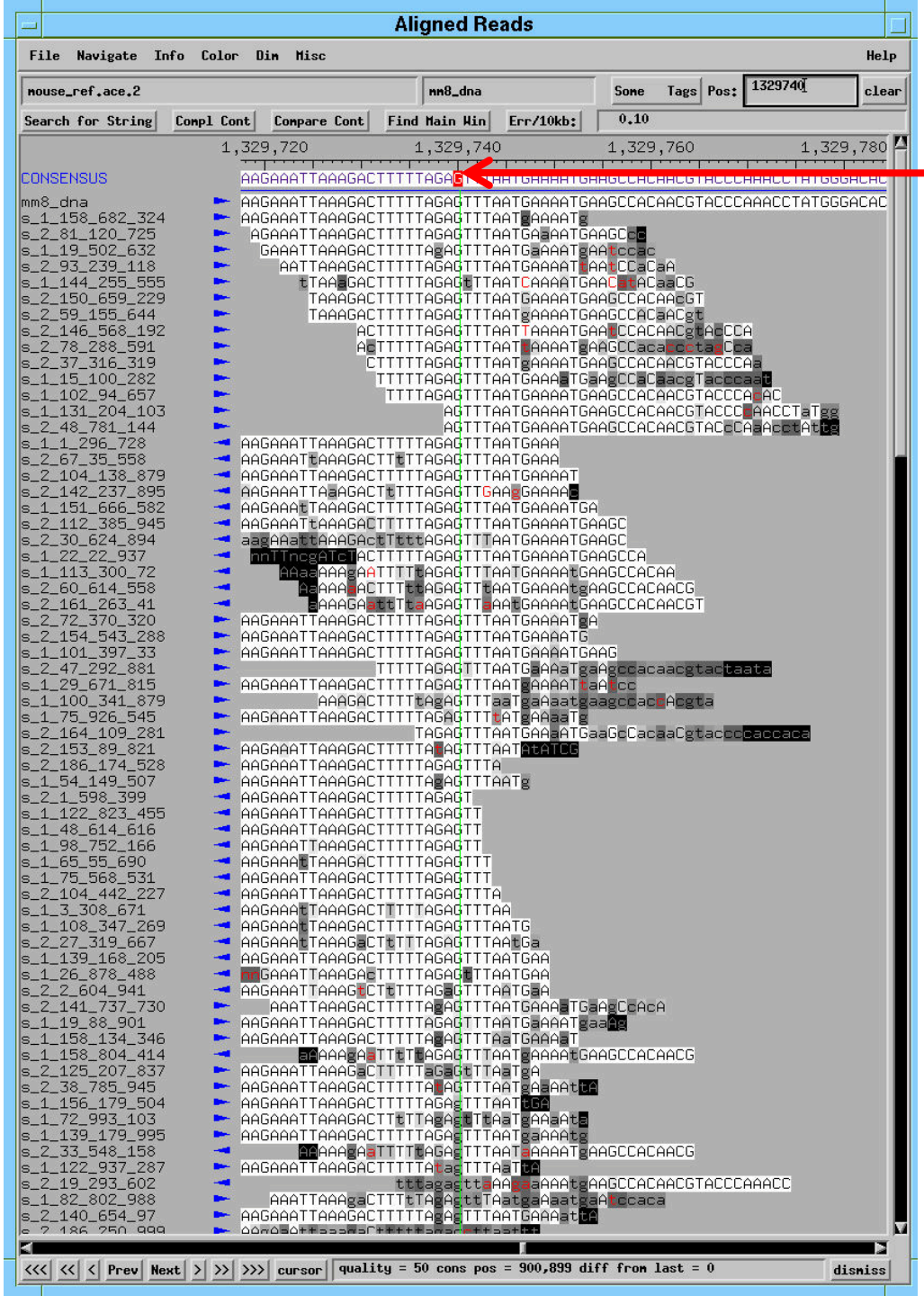

Ref Seq
454
solexa
File Navigate Info Color Dim Misc
Help
human.ref.ace.4
3
Some Tags Pos: 70
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb: 0.10
80 90 100 110 120 130
CONSENSUS
CGGGCTCAGGCTCGGGCTCAGGCTCAGGCTC**********TGCTTCGTACACTGGCTCTTCCTCCACCTGGAGGCCTTC
3
CGGGCTCAGGCTCGGGCTCAGGCTCAGGCTC**********TGCTTCGTACACTGGCTCTTCCTCCACCTGGAGGCCTTC
EYM9QNL02IJ006
cggGCtcaggCtcggGCtcaggCtcaggCtcaggCtcaggCtctgcttcgtacactggCtcttCcTccAccTggAggCcTtC
EYM9QNL02GBCTG
cggGCtcaggCtcaggCtcaggCtcaggCtc**********tgcttCgtacactggCtcttCcTccAccTggAggCcTtC
202L8_7_198_298_372
cggggctcaggCtcaggctcaggctca
634Z_3_23_481_851
cggggctcgggctcgggCtc
202VM_6_2_100_442
cggggctCagGCTCAGGCTCAGGCTCAGGCTC
2040M_5_142_460_662
ggctcaggctcgggCtcaggCtCAggctCAGgC
insertion (?) in a 454 and solexa read
SNP correlated with insertion (?)
<<< << < Prev Next > >> >>> cursor
dismiss

High Depth of Coverage Solexa Reads



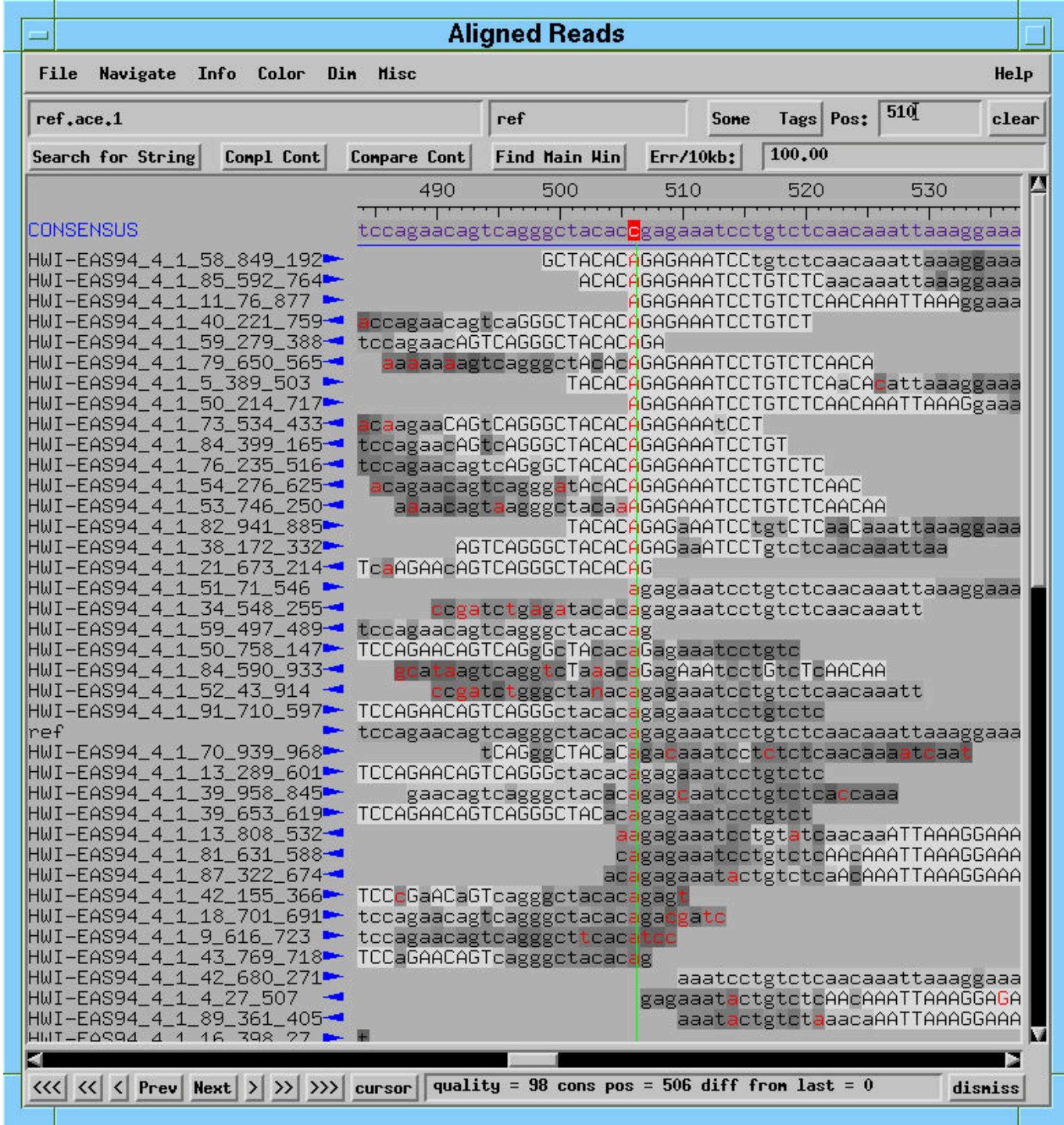
green vertical cursor aids eye

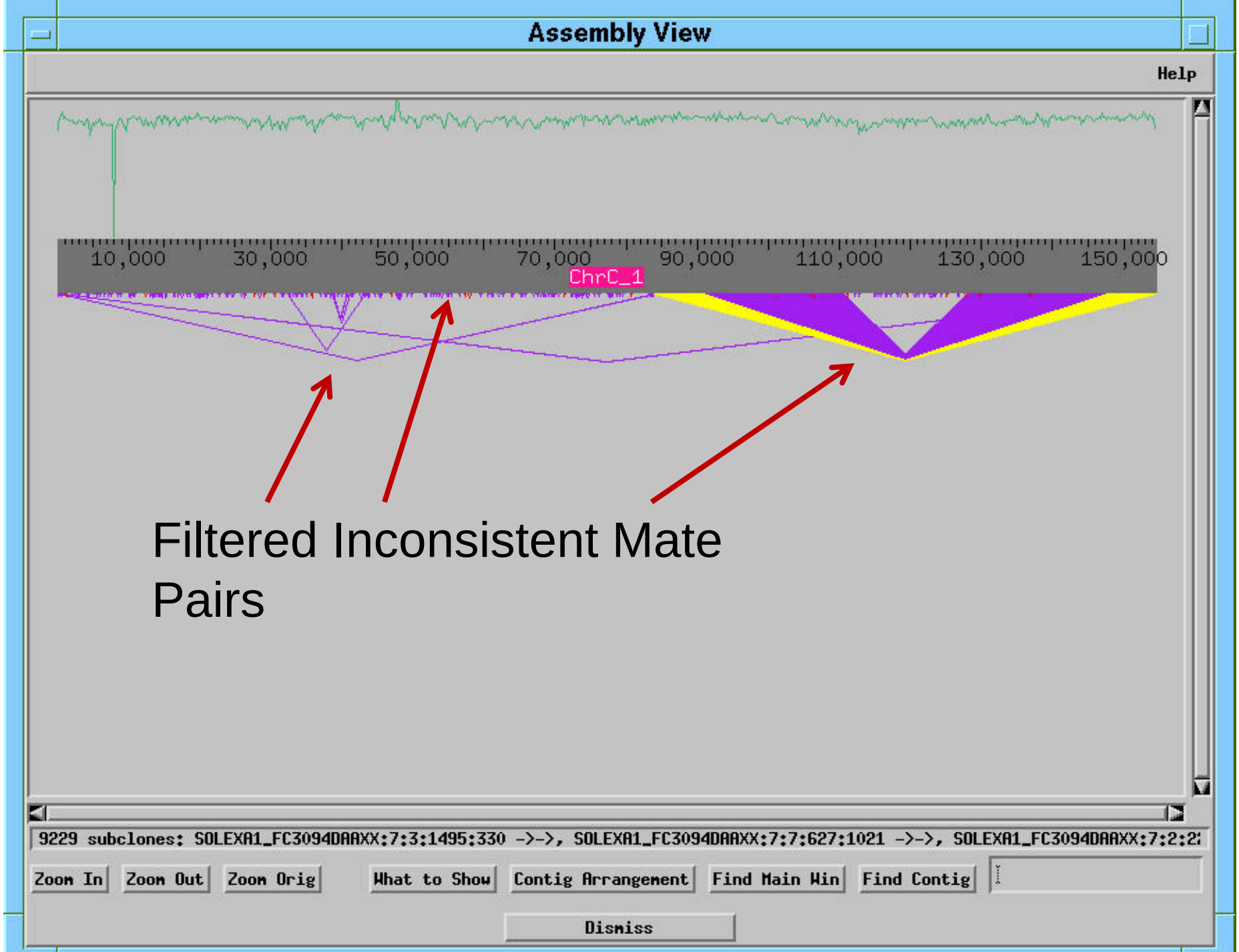
Same Reads Sorted by Quality At Cursor Column

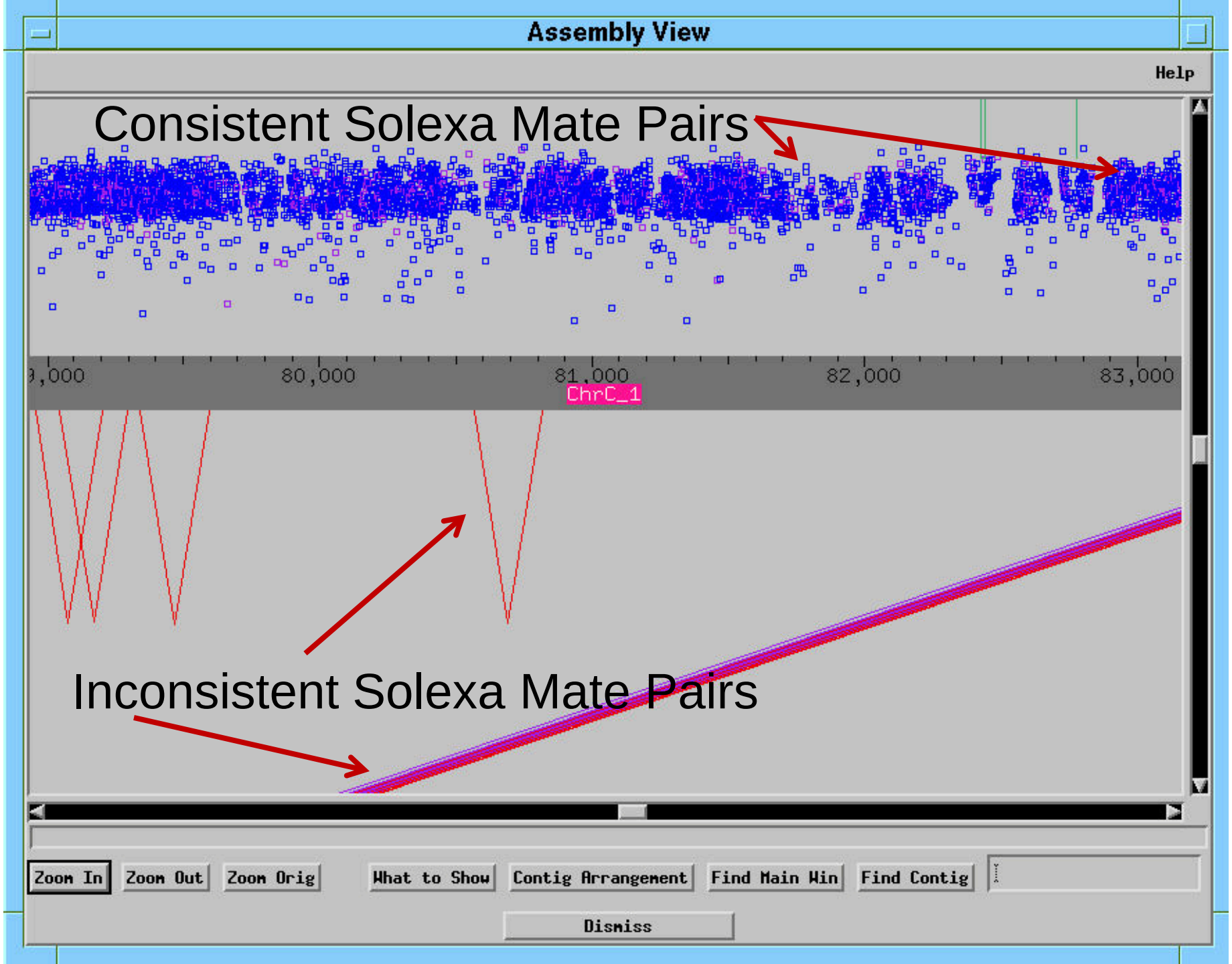


Cursor

Overstrike Consensus

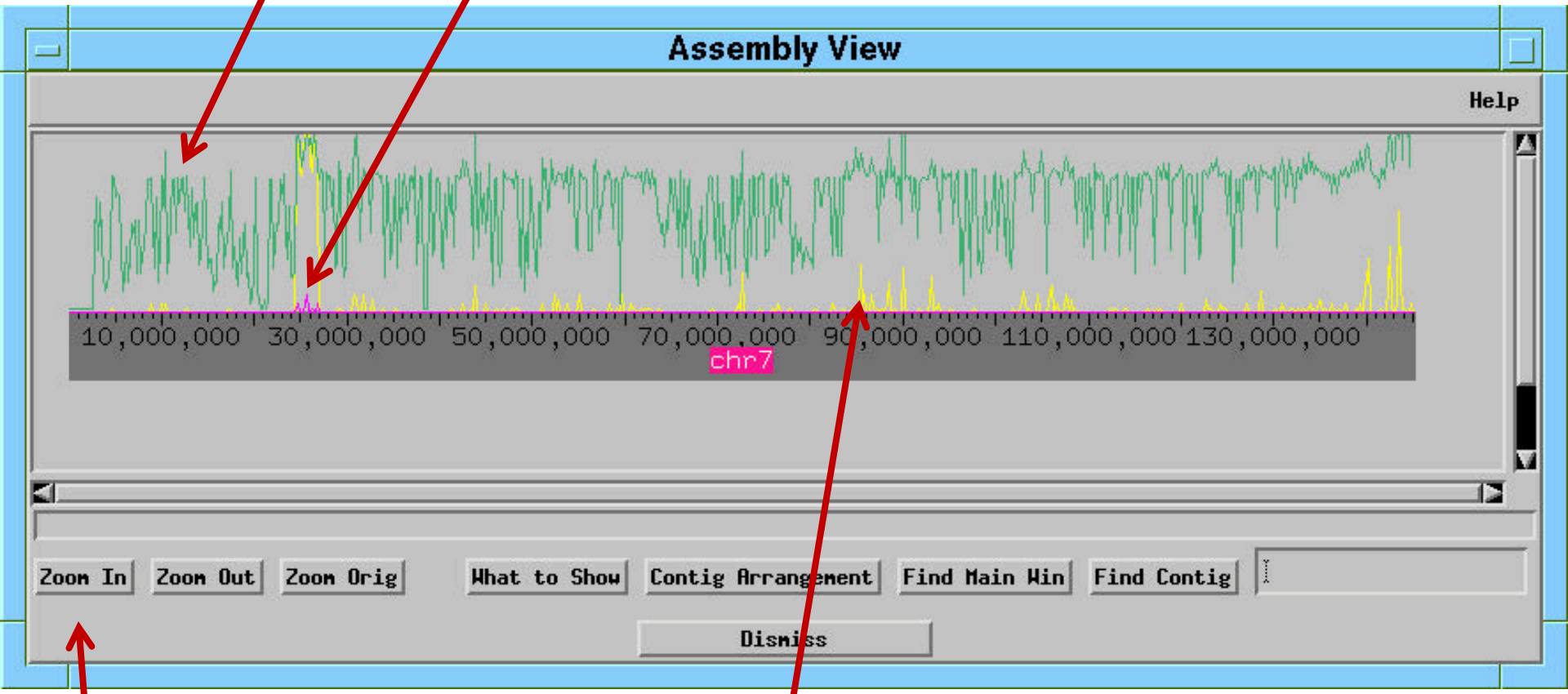



Assembly View
Help
10,000
30,000
50,000
70,000
90,000
110,000
130,000
150,000
ChrC_1
Filtered Inconsistent Mate Pairs
9229 subclones: SOLEXA1_FC3094DAAXX:7:3:1495:330 ->->, SOLEXA1_FC3094DAAXX:7:7:627:1021 ->->, SOLEXA1_FC3094DAAXX:7:2:2
Zoom In
Zoom Out
Zoom Orig
What to Show
Contig Arrangement
Find Main Win
Find Contig
Dismiss


Assembly View
Help
Consistent Solexa Mate Pairs
80,000
81,000
82,000
83,000
ChrC_1
Inconsistent Solexa Mate Pairs
Zoom In
Zoom Out
Zoom Orig
What to Show
Contig Arrangement
Find Main Win
Find Contig
Dismiss

Read Depth of Coverage (green)

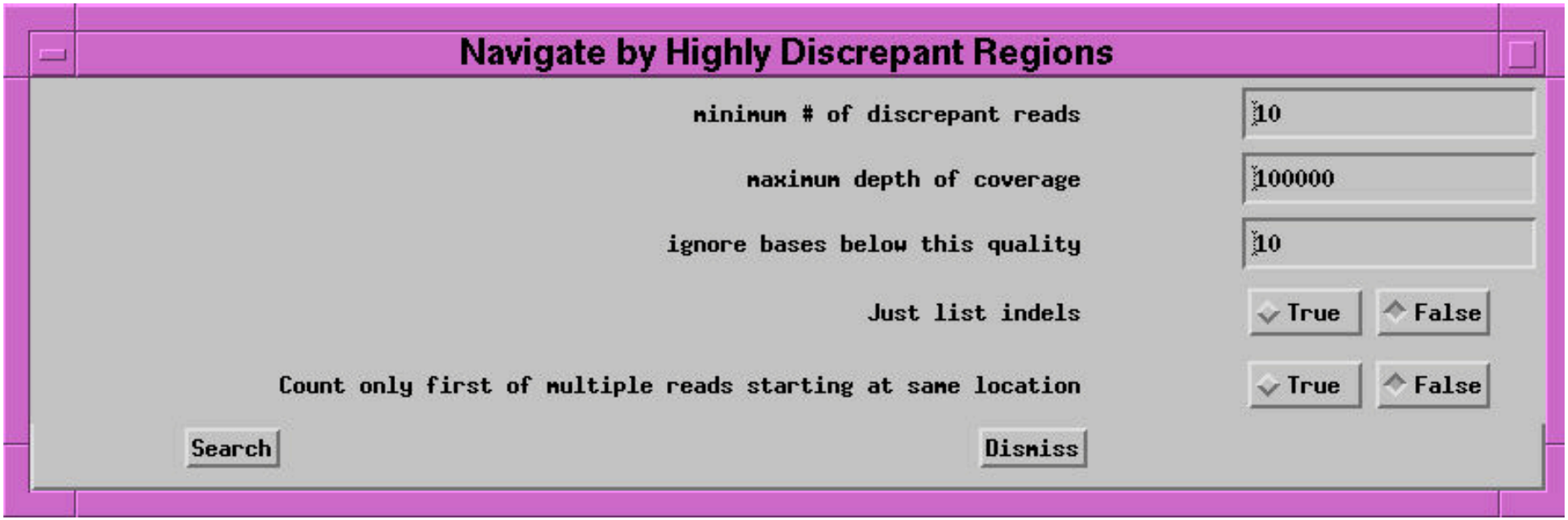
Indel discrepancies (magenta)



Non-indel discrepancies (yellow)

Look!

Looking for variants



Report of Variants

Highly Discrepant Positions

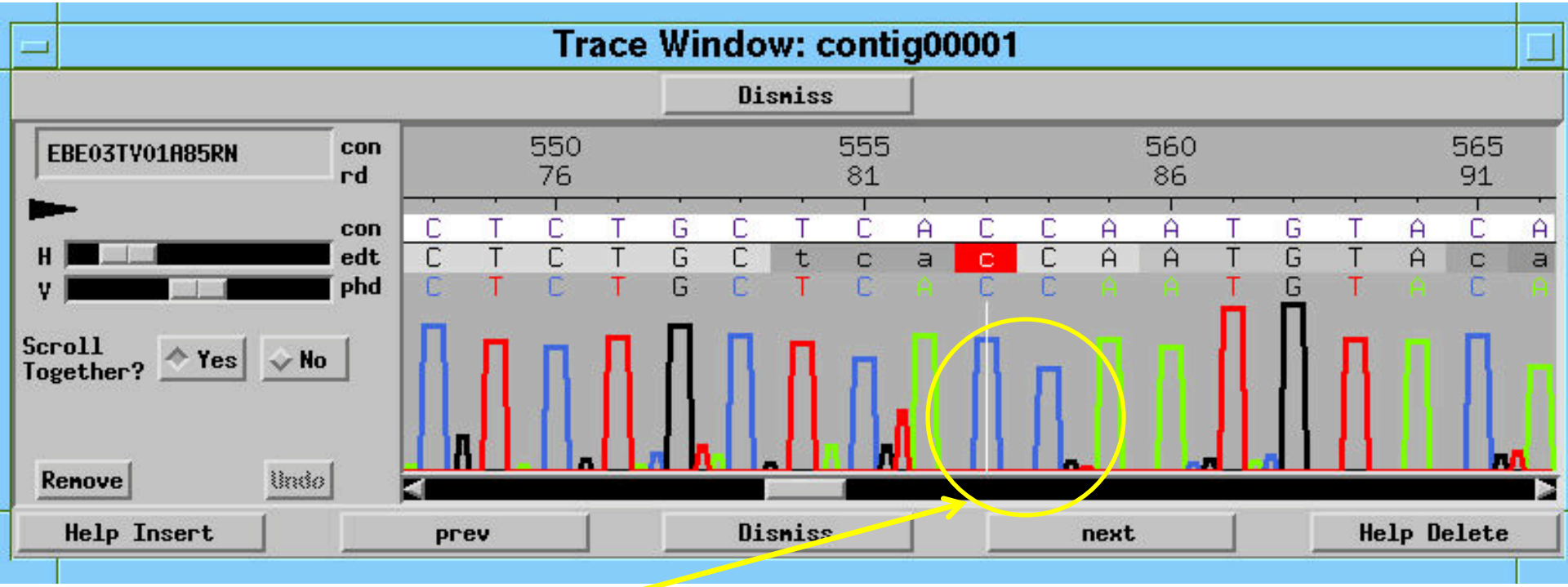
min # of discrepant reads: 10, min quality: 20, "r": base of reference seq
max depth of coverage: 100000 and ignoring reference seq

A		C		G		T		*		pos	contig
0	0.0%	0	0.0%	10	40.0%r	15	60.0%	0	0.0%	9,096,328	chr7
0	0.0%	2	7.4%	4	14.8%r	21	77.8%	0	0.0%	9,096,331	chr7
0	0.0%	161	100.0%	0	0.0%	0	0.0%r	0	0.0%	10,656,718	chr7
0	0.0%	178	100.0%	0	0.0%r	0	0.0%	0	0.0%	10,656,739	chr7
0	0.0%	12	92.3%	0	0.0%	1	7.7%r	0	0.0%	17,563,643	chr7
13	8.4%	141	91.6%r	0	0.0%	0	0.0%	0	0.0%	24,746,149	chr7
13	8.1%	148	91.9%r	0	0.0%	0	0.0%	0	0.0%	24,746,164	chr7
12	8.6%	126	90.6%r	0	0.0%	1	0.7%	0	0.0%	24,747,122	chr7
17	9.4%	163	90.6%r	0	0.0%	0	0.0%	0	0.0%	24,747,907	chr7
0	0.0%	0	0.0%	183	91.5%r	17	8.5%	0	0.0%	24,747,945	chr7
0	0.0%	0	0.0%	157	94.0%r	10	6.0%	0	0.0%	24,748,343	chr7
0	0.0%	0	0.0%	143	89.9%r	16	10.1%	0	0.0%	24,748,352	chr7
0	0.0%	0	0.0%	131	87.9%r	18	12.1%	0	0.0%	24,748,353	chr7
10	8.5%	0	0.0%	107	91.5%r	0	0.0%	0	0.0%	24,748,629	chr7
12	6.2%	180	93.8%r	0	0.0%	0	0.0%	0	0.0%	24,752,205	chr7

Go | Prev | Next | Save | Dismiss

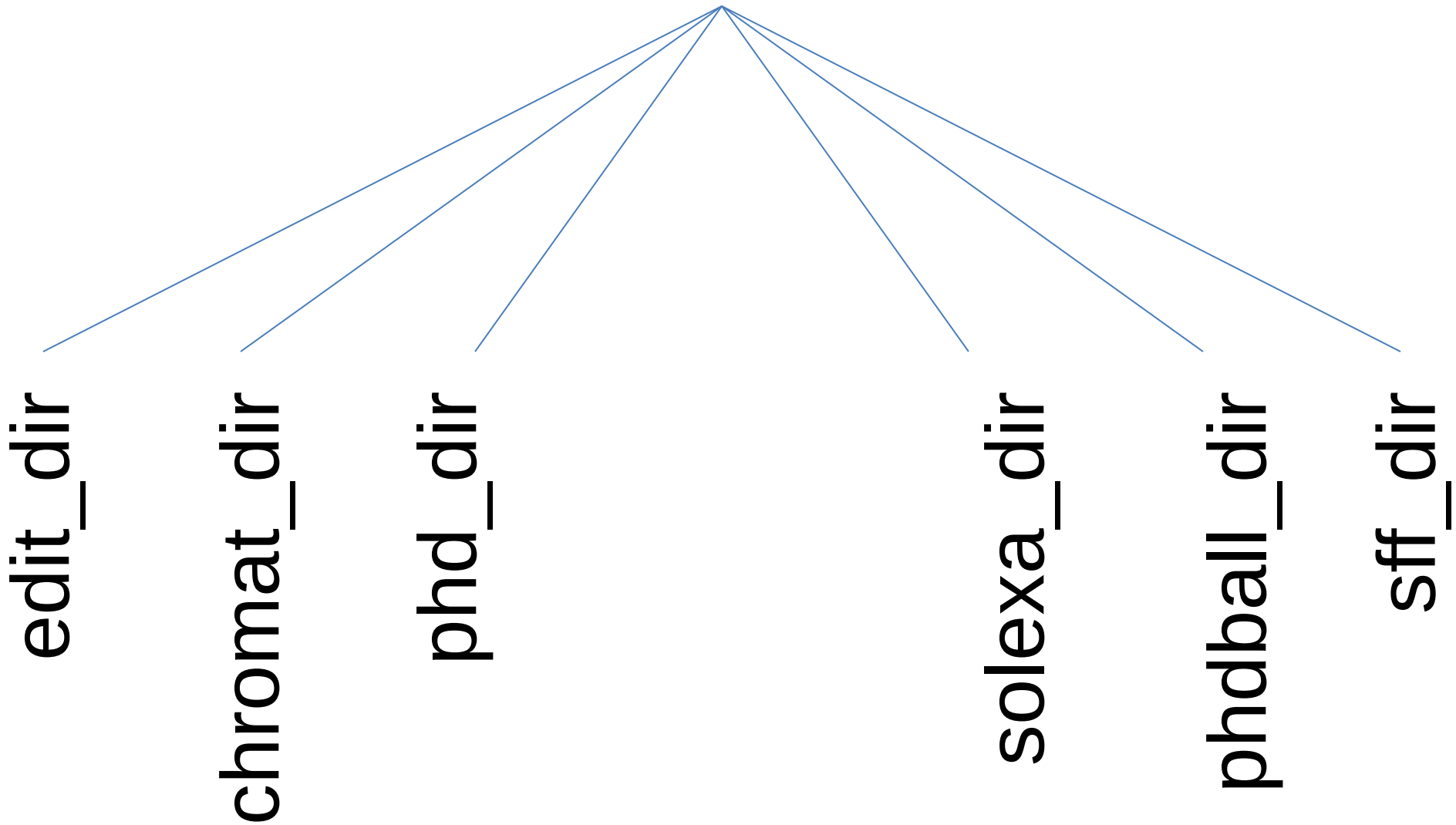
This is also available as a report which can be automatically generated and saved to a file.

454 Read Signal



Disparate signal heights alerts the user to uncertainty in number of repeats of base C

New Consed Directories



Newbler Assembler 454's and Consed

454's Newbler assembler's "–consed" option produces consed-ready files

Cross_match Options

- -minscore 30

  Must be decreased for short reads. 25 is sufficient for most applications. 20 is more sensitive, but makes run take longer.

- -gap1_only

  Limit alignment to 2 1-base gaps (1 in each direction from word-match). Greatly speeds up search. Use it.

- -output_nonmatching_queries

  useful for successively more sensitive runs

flexible reporting of hits: -minmargin and -masklevel

	dist A	dist B	dist C
score	# of hits	# of hits	# of hits
25	5	1	1
24	100	100	0
23	2	2	0
22	0	0	0
21	1	1	100

- -score_hist

  print a histogram of the scores

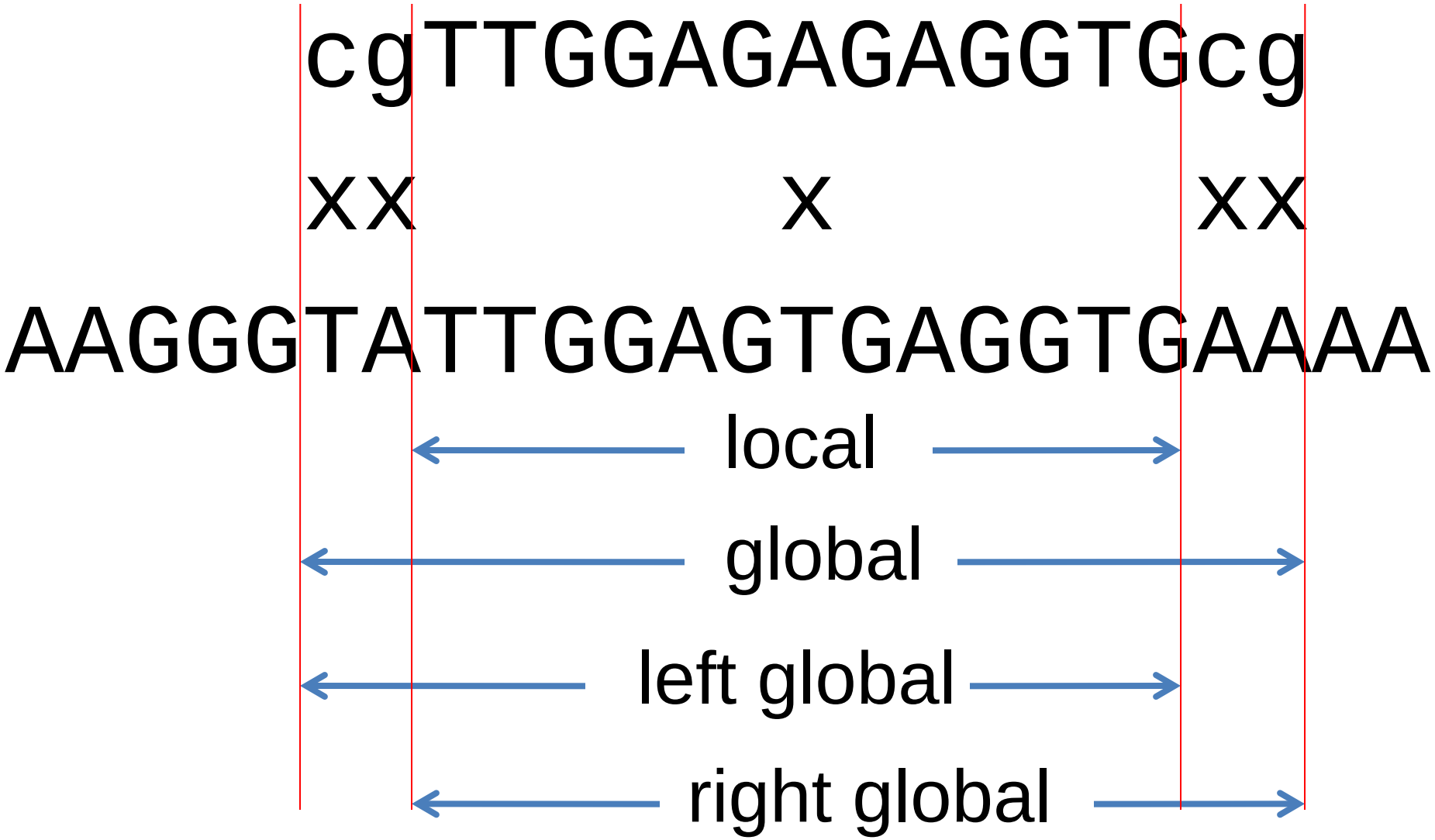
- print mapping scores (and filter by mapping score)

- -output_bcdsites

  table giving, for each stretch of reference sequence bases, how many reads agree, show deletion, show insertion, or show substitution with respect to the reference sequence

Types of Alignments

(-globality)



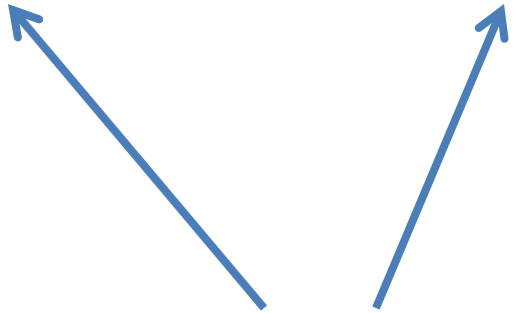
Spliced Alignments

transcript:

GGTATT**...**CGGTGAAA

genome:

AAGGGTATTGT...AGCGGTGAAA



U2 splice sites

What is Coming: Billion-Read Consed

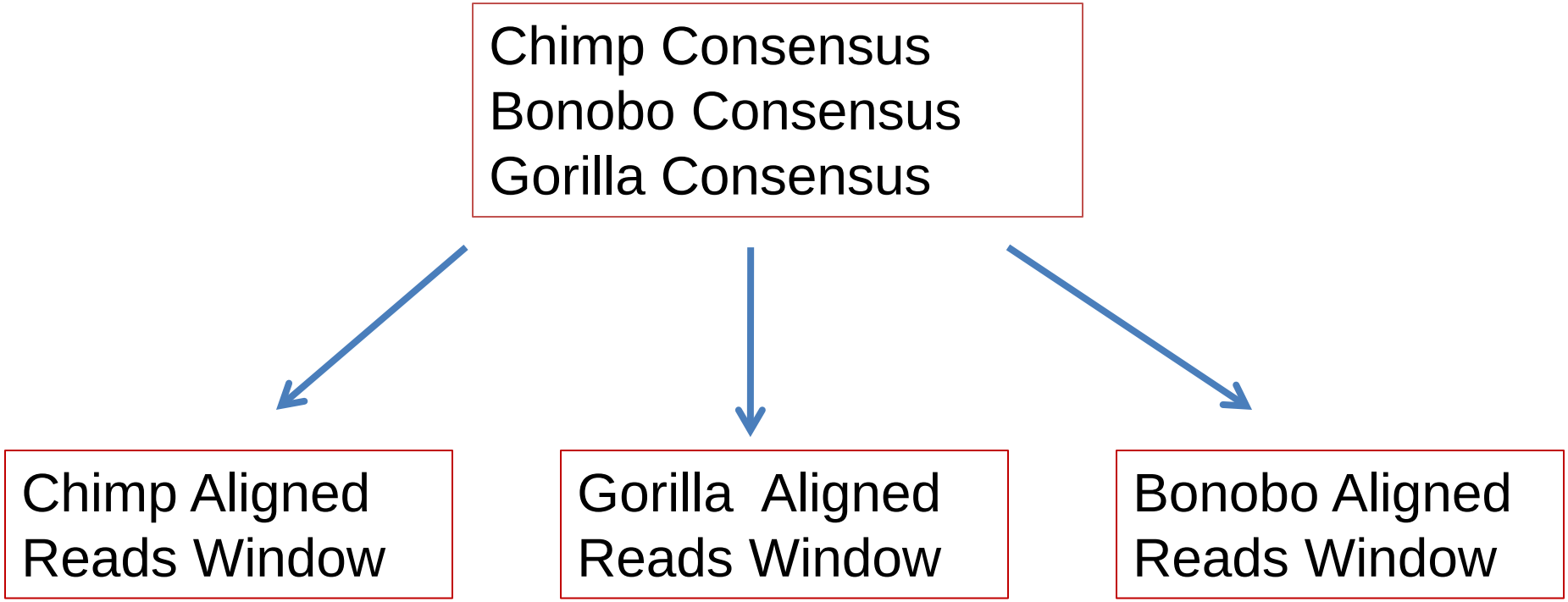
Requires new file format:

CALF: Compact Alignment Format

not all reads in memory at once and rapidly find necessary reads

Alignment of Different Assemblies

Align consensus sequences of different species:



Consed: Other Assemblers

Directly read Velvet output

More as requested

What is Coming—More Automation

- Ability to change consensus in batch
- Automatically fixing consensus
- Batching joining/tearing
- Automatically join/extend when add new reads

What is Coming--variation

- determine SNPs using a probabilistic model
- detecting structural variants

Features for high depth of coverage regions

- enable user to select a read to stay visible as scroll up and down
- show a less overwhelming view of data (optionally hide reads that only differ in low quality bases)

Acknowledgements

Curt Jamison, Pat Minx, Bob Fulton, Chris Munk, Shunsheng (Cliff) Han, Hajnalka Kiss, Joel Martin, Aye Wollam, Dave Dooling, Sven Klages, Dena Leshkowitz, Yaron Butterfield, Johanna Swanson, …

dgordon@u.washington.edu