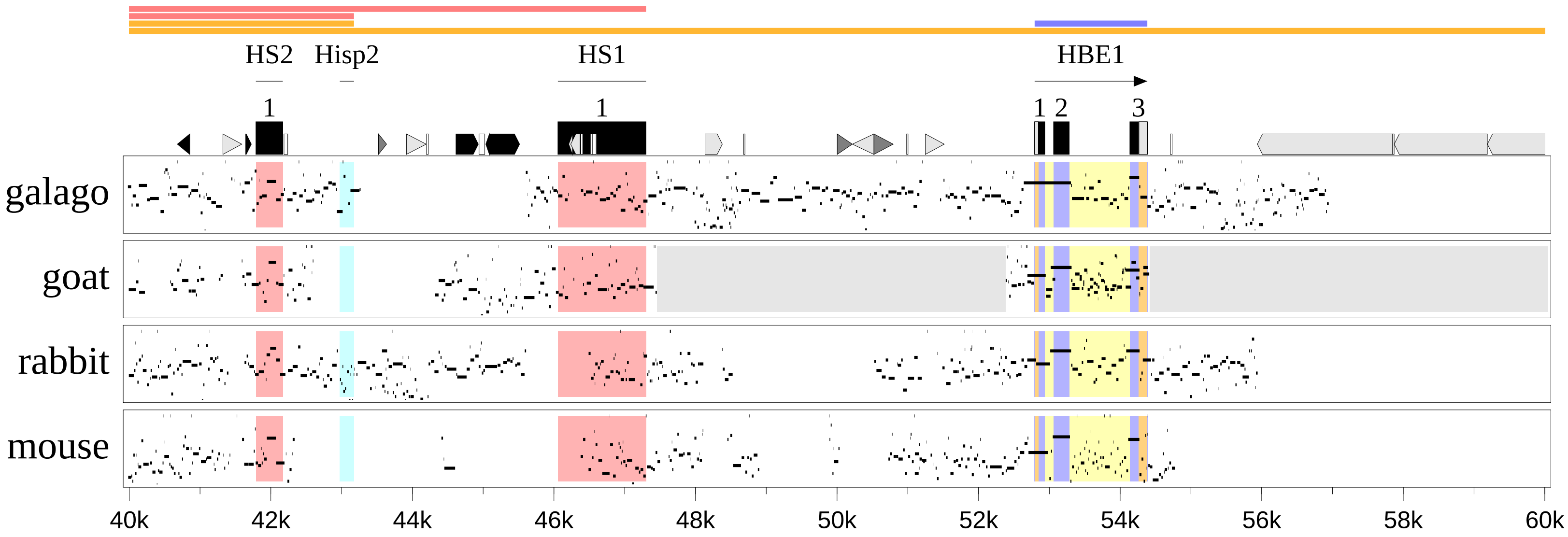
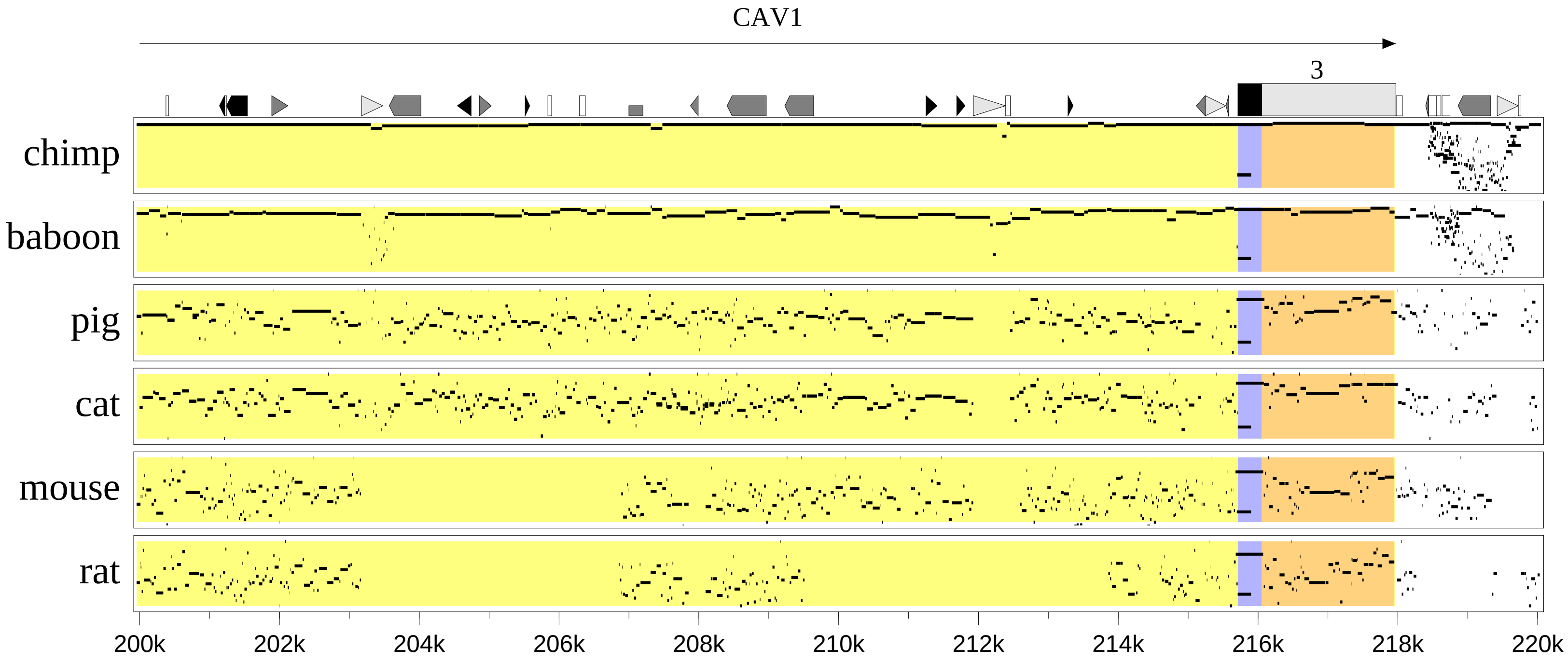
Recent Additions to PipMaker

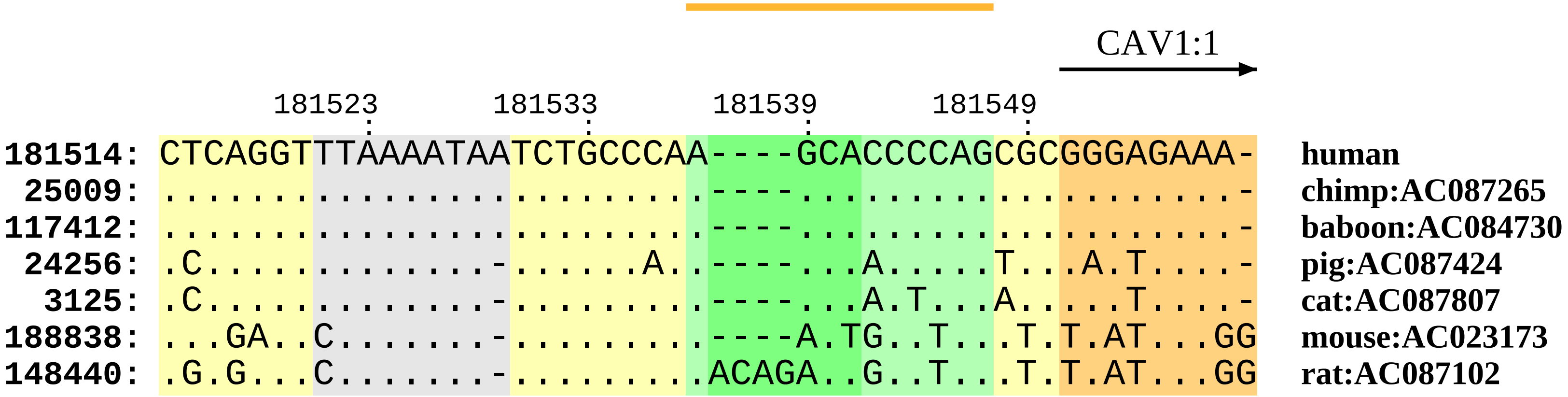
Webb Miller
Penn State

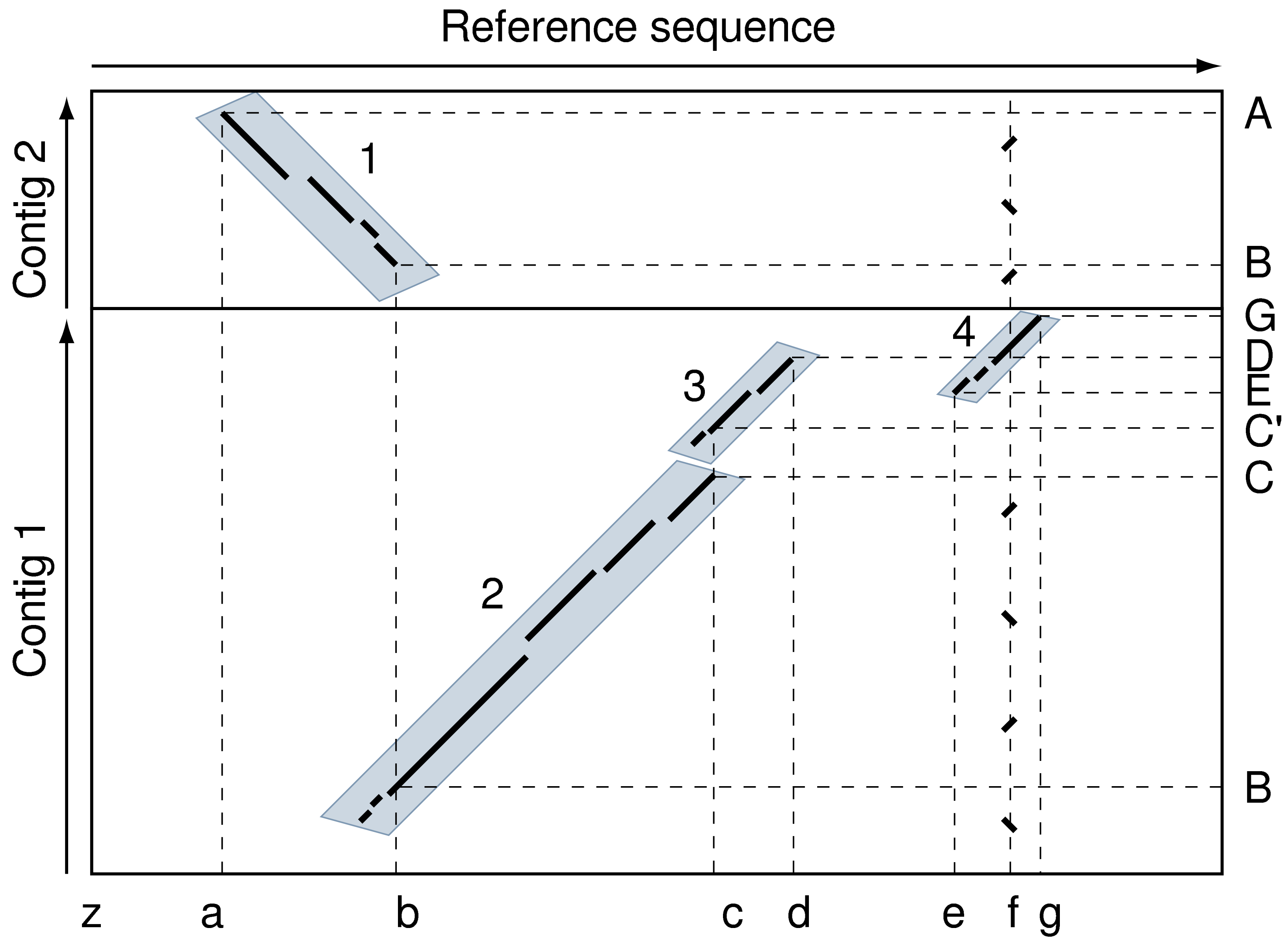
Outline

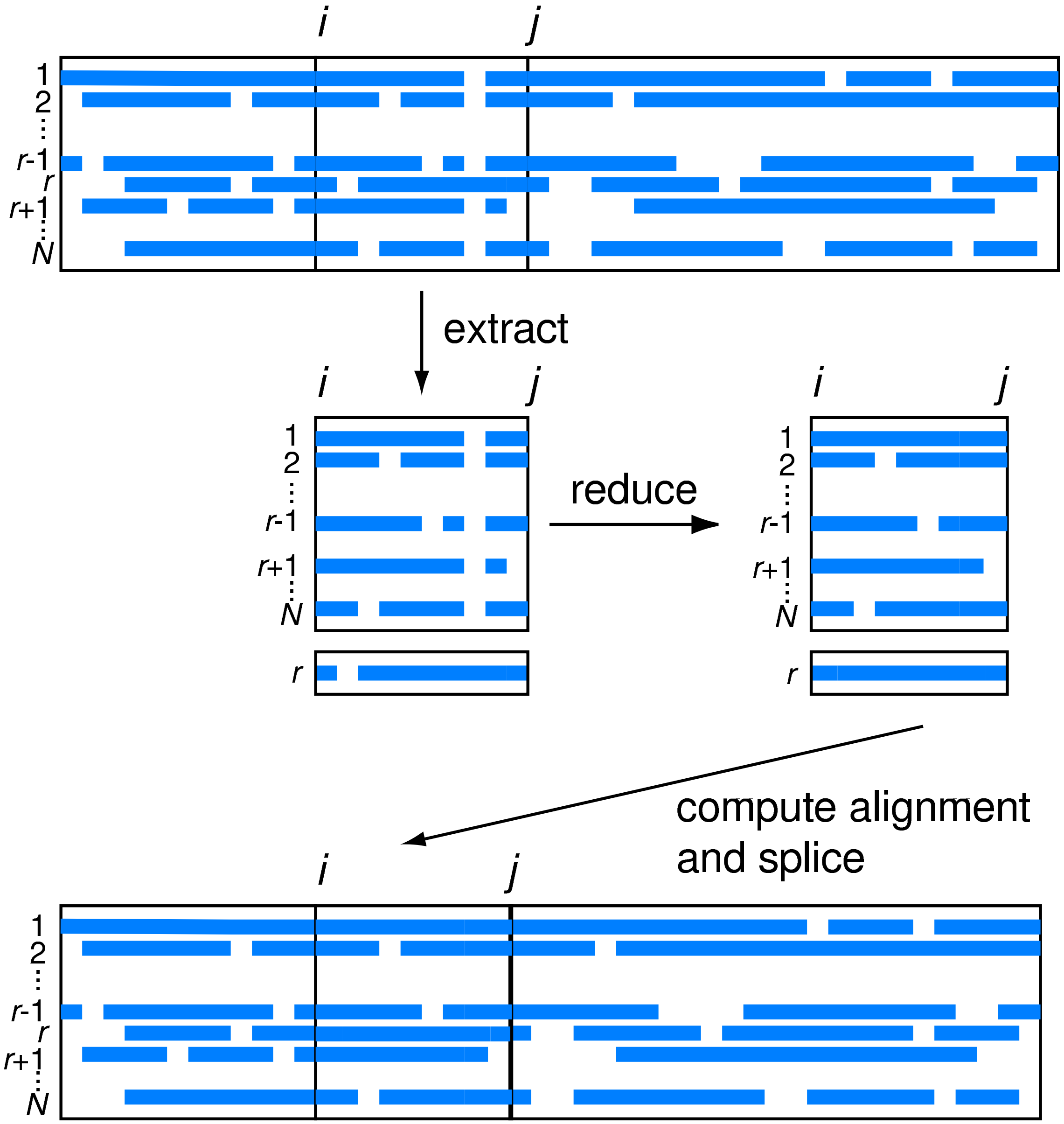
1. MultiPipMaker — simultaneous analysis of more than two sequences.
2. LAJ — locally run program for interactive viewing of PipMaker alignments
3. PipTools — locally run tools to facilitate use of PipMaker
4. Enterix — archived alignments of enteric bacterial genomes
5. PipDispenser — archived alignments human and mouse genomes


HS2
Hisp2
HS1
HBE1
1
1
1
2
3
galago
goat
rabbit
mouse
40k
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k


CAV1
3
chimp
baboon
pig
cat
mouse
rat
200k
202k
204k
206k
208k
210k
212k
214k
216k
218k
220k

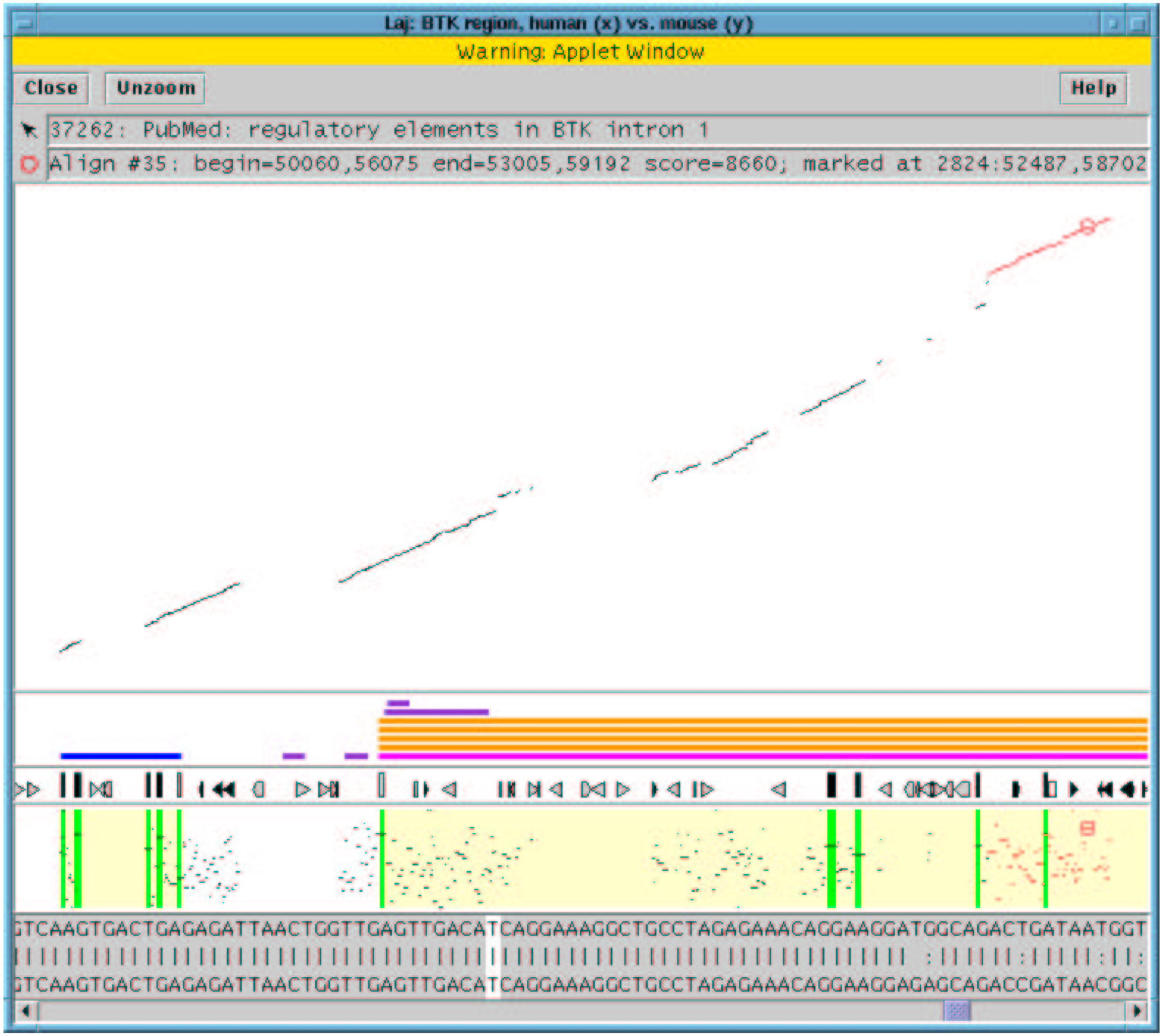

CAV1:1
181523
181533
181539
181549
181514: CTCAGGTTTAAAATAATCTGCCCAA----GCACCCCAGCGCGGGAGAAA- human
25009: ....----....................- chimp:AC087265
117412: ....----....................- baboon:AC084730
24256: .C..............-......A..----...A.....T...A.T....- pig:AC087424
3125: .C..............-.........----...A.T...A.....T....- cat:AC087807
188838: ...GA..C.......-.........----A.TG..T...T.T.AT...GG mouse:AC023173
148440: .G.G...C.......-.........ACAGA..G..T...T.T.AT...GG rat:AC087102


Reference sequence
Contig 2
Contig 1
1
2
3
4
A
B
G
D
E
C'
C
B
z
a
b
c
d
e
f
g


i
j
1
2
r-1
r
r+1
N
extract
reduce
compute alignment
and splice

LAJ

LAJ (”Local Alignments in Java”) is an interactive viewer for alignments generated by Blastz (PipMaker’s pairwise alignment program). Both dotplot and PIP views of the alignments are given. The user can zoom in, click to see a nucleotide-level view, click on hyperlinks.



```
37262: PubMed: regulatory elements in BTK intron 1
Align #35: begin=50060,56075 end=53005,59192 score=8660; marked at 2824:52487,58702
```

```
GTCAAGTGACTGAGAGATTAACTGGTTGAGTTGACATCAGGAAAGGCTGCCTAGAGAAACAGGAAGGATGGCAGACTGATAATGGT
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  :||||||:|||||:||:
GTCAAGTGACTGAGAGATTAACTGGTTGAGTTGACATCAGGAAAGGCTGCCTAGAGAAACAGGAAGGAGAGCAGACCGATAACGGC
```

LAJ for an “Electronic Supplement”

A biologist can establish a Web site as an “electronic supplement” to a sequence analysis project, where LAJ is provided as an applet that can be loaded by any Java-compliant browser. That way, other biologists can browse the electronic supplement using the full power of LAV. For instance, see:

http://linus.ceh.uvic.ca/ mdwilson/laj.html

PipTools for Preparing Annotations
(i.e., repeats, exons and underlay files)

Program	**From**	**To**
exons2underlays	exons file	underlay file
genbank2exons	GenBank	exons file
genbank2repeats	GenBank	repeats file
genscan2exons	Genscan	exons file
genscan2underlays	Genscan	underlay file
rmask2repeats	RepeatMasker	repeats file
sim4	cDNA sequence	exons file

PipTools for Modifying Annotations
(e.g. if the reference sequence changes)

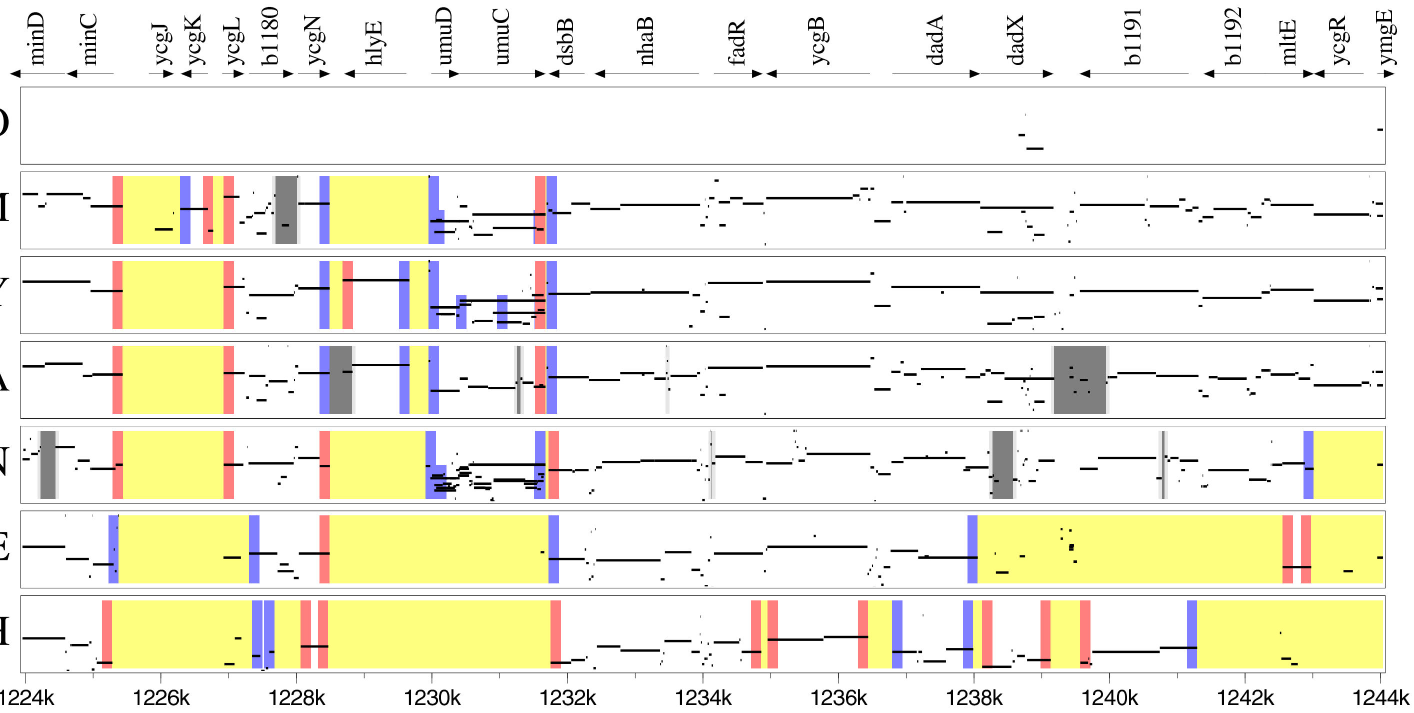
Program	**Function**
exons2mrna	extract putative cDNA sequence
shift-pos	shift positions in annotations
transform-pos	transfer positions to other sequence

PipTools for Analyzing Alignments

Program	**Function**
strong-hits	find strong hits in a pairwise alignment
strong-hits2underlays	color strong hits
infocon	find strong hits in a multiple alignment
slice	extract part of a multiple alignment
multi-pat	match patterns in a multiple alignment

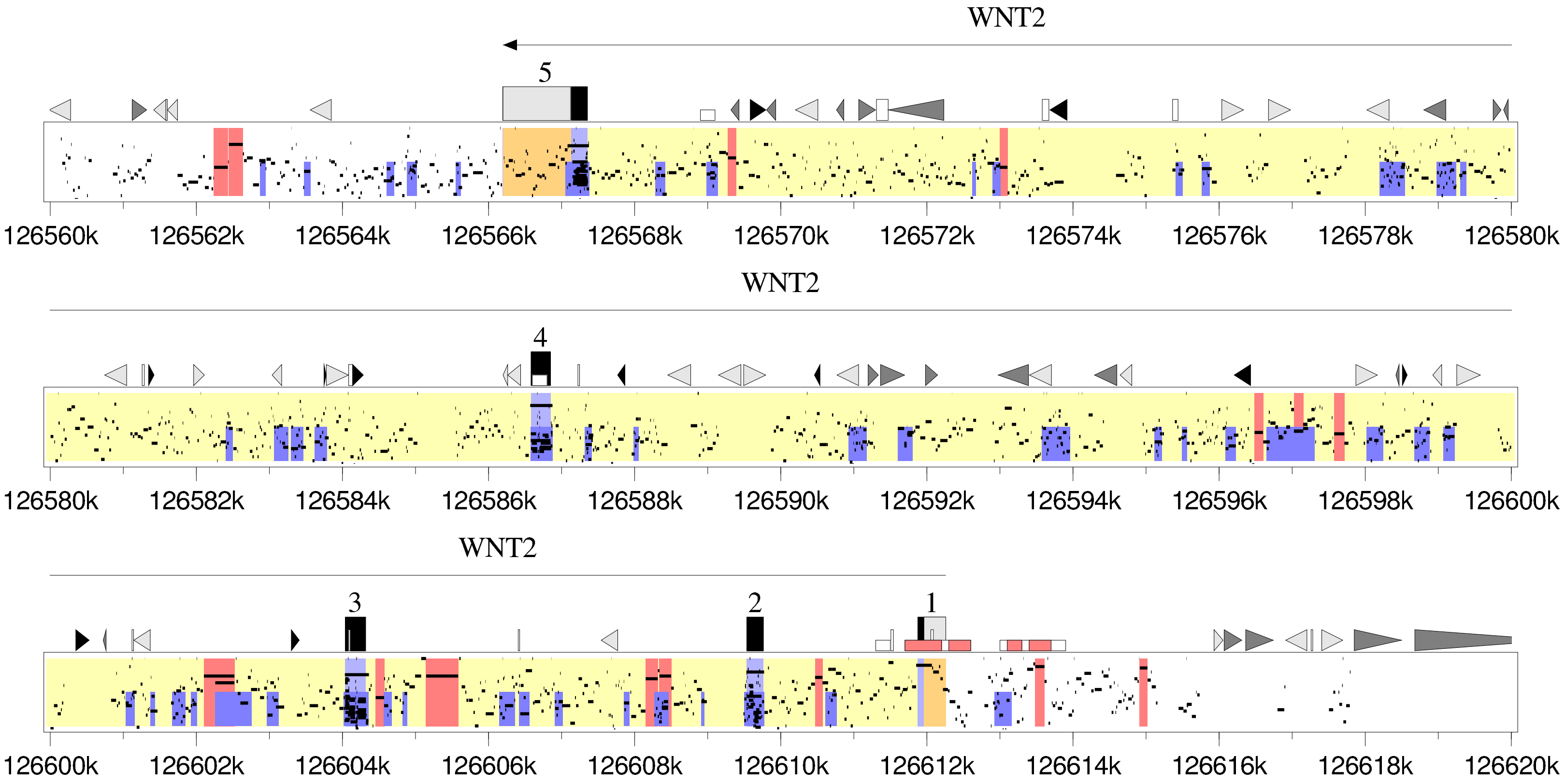
http://globin.cse.psu.edu/cgi-bin/enteric?organism=ecoli;address=1234567 Thu Apr 6 11:49:20 EDT 2000

yellow = E. coli sequence not found in the other species
red = sequence in the other species whose immediate neighbor has a homolog elsewhere in E. coli
blue = sequence in the other species whose immediate neighbor has no detectable homolog in E. coli
gray = apparently not sequenced in the other species
purple = overlapping colors, such as red and blue



PipDispenser

At our website you can request a pip of any desired gene or region in the human genome, aligned to the mouse. We intend to add the rat genome sequence in the near future.


WNT2
5
126560k
126562k
126564k
126566k
126568k
126570k
126572k
126574k
126576k
126578k
126580k
WNT2
4
126580k
126582k
126584k
126586k
126588k
126590k
126592k
126594k
126596k
126598k
126600k
WNT2
3
2
1
126600k
126602k
126604k
126606k
126608k
126610k
126612k
126614k
126616k
126618k
126620k

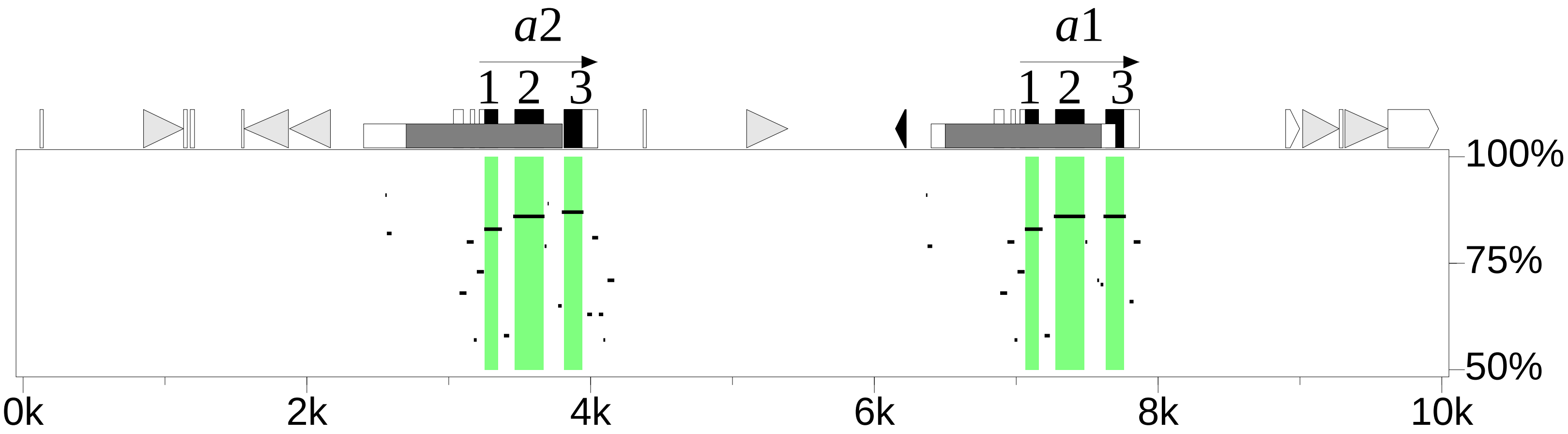
Aligning Whole Genomes

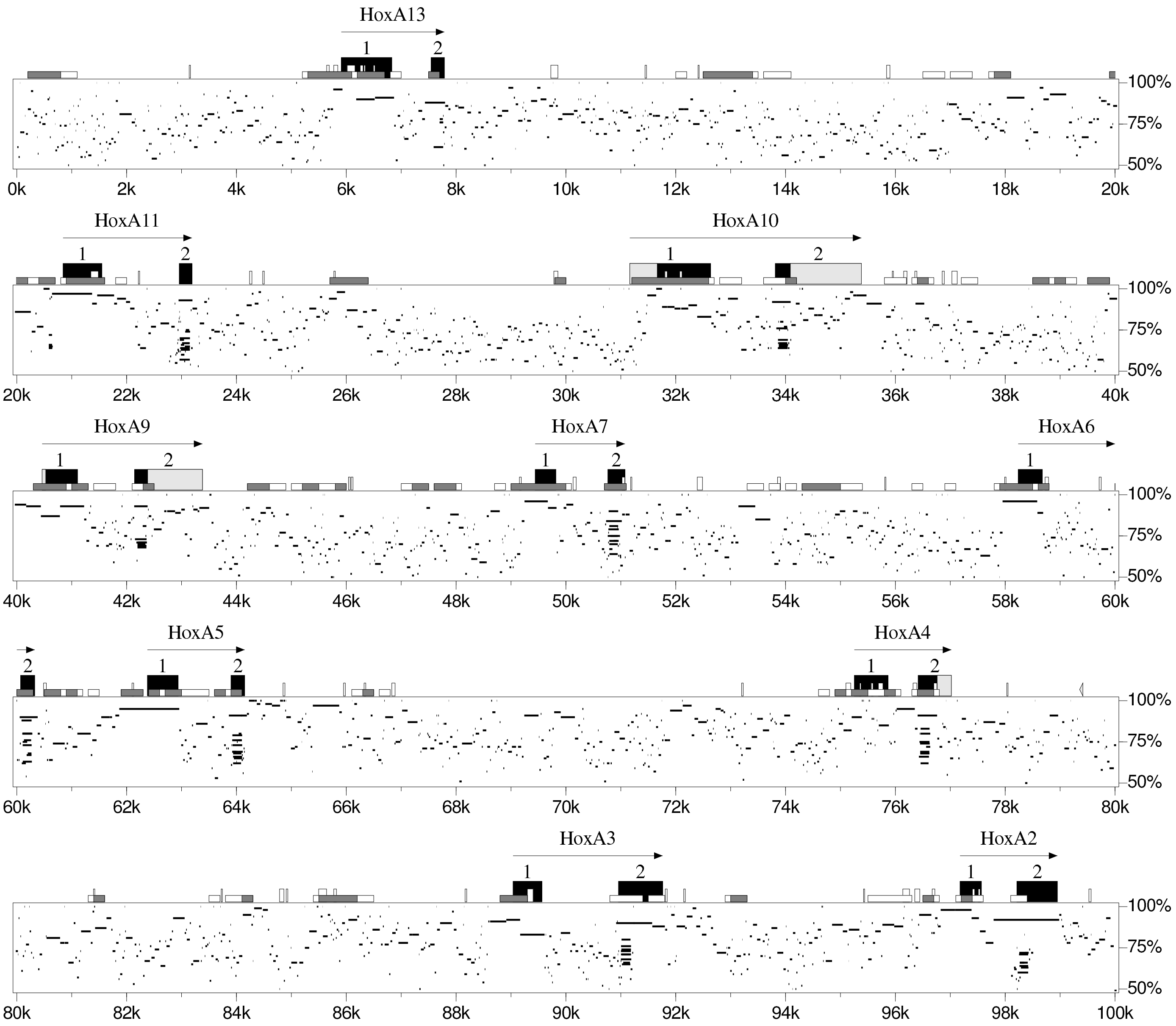
Alignments for PipDispenser are computed on a 1000-CPU cluster belonging to David Haussler of the University of California at Santa Cruz. The computation takes half a day.

Varying Rate of Conservation

The rate of human-mouse conservation varies widely among different genomic loci. At some, only the protein-coding regions can be reliably aligned. At others, most or all of the non-coding DNA aligns.

Alpha-globin gene cluster




HoxA13
1
2
100%
75%
50%
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
HoxA11
HoxA10
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
HoxA9
HoxA7
HoxA6
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
HoxA5
HoxA4
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k
HoxA3
HoxA2
82k
84k
86k
88k
90k
92k
94k
96k
98k
100k

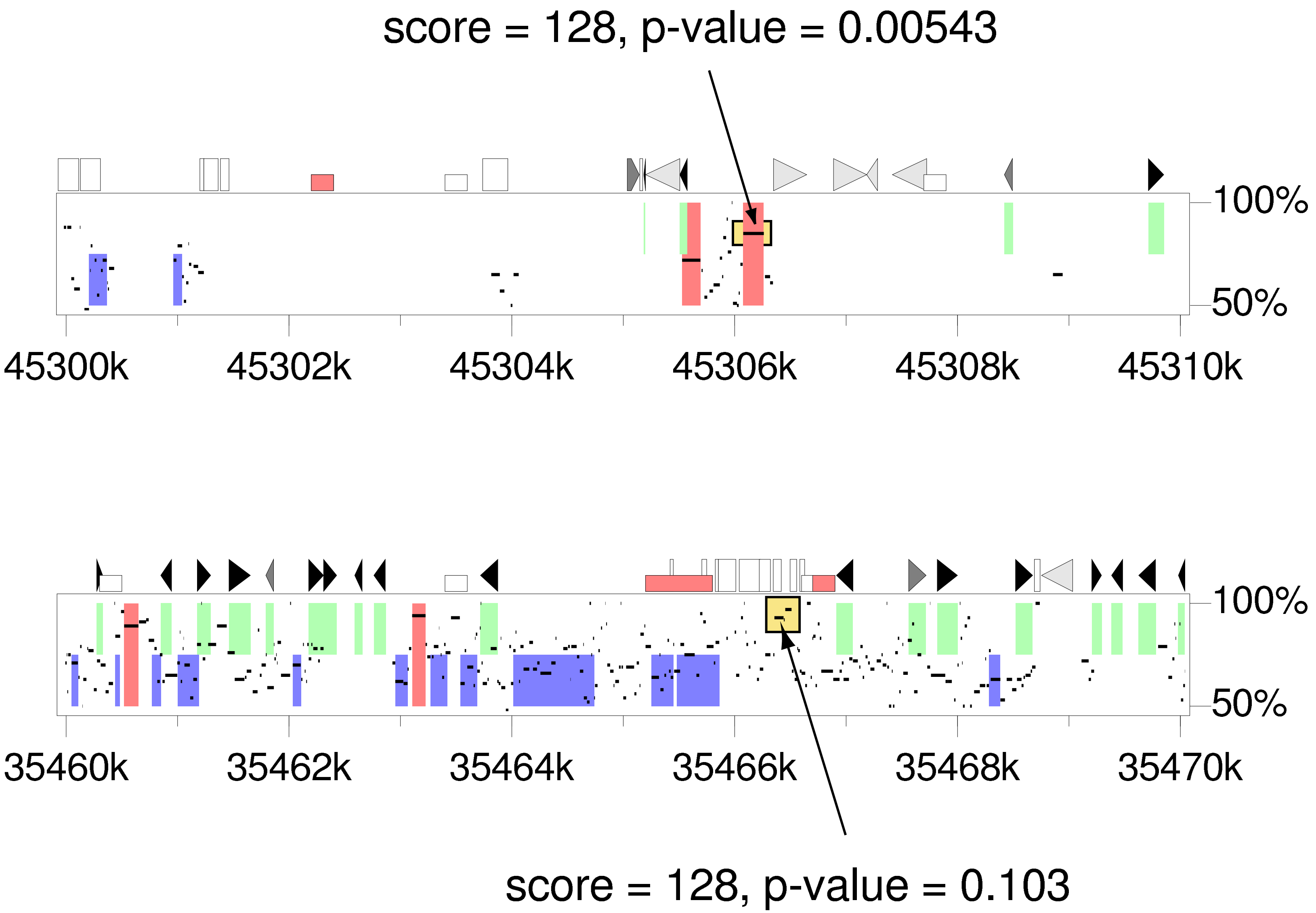
region	aligns	high	%G+C	%rept	%ident	ts/tv	ave seg	ave gap
HOXA	99.3	21.3	50.7	3.4	78.9	1.53	42.0	5.3
TCR	77.8	7.0	44.0	14.3	70.4	1.55	31.0	7.4
FHIT	58.1	7.6	37.1	42.0	68.9	1.34	30.7	7.1
CFTR	53.2	4.9	34.9	38.9	69.9	1.37	28.1	7.3
BTK	49.6	4.9	41.1	41.2	72.8	1.41	32.3	8.8
SNCA	44.4	1.0	34.6	31.8	66.7	1.28	26.0	7.7
DIST1	40.9	0.8	55.3	38.0	69.8	1.51	26.5	7.7
MECP2	39.7	5.9	47.8	47.5	74.2	1.66	34.2	8.1
CD4	35.6	3.3	51.9	36.9	73.0	1.44	30.0	7.3
CECR	21.3	1.8	45.9	47.8	70.0	1.34	27.3	6.7
ERCC2	11.0	0.0	58.5	53.9	73.4	1.34	28.5	8.4

Statistical Significance of Matches

Working with Jia Li of Penn State's Statistics Department, we have developed a method for assigning statistical significance to strongly matching regions within a long genomic region.

1. Segment the region according to extent of divergence using a Hidden Markov Model.

2. Using statistical theory developed by Dembo and Karlin (which generalizes that used for Blast p-values), assign p-values to strongly matching regions according to their local degree of background divergence.



score = 128, p-value = 0.00543

score = 128, p-value = 0.103

Acknowledgements

- almost everything — Ross Hardison
- PipMaker — Scott Schwartz
- MultiPipMaker — Eric Green (ZooSeq), Scott Schwartz
- LAJ, PipTools — Cathy Riemer, Laura Elnitski
- PipDispenser — Scott Schwartz; David Haussler and Jim Kent (U.C. Santa Cruz)
- Enterix — Liliana Florea (now at Celera), Scott Schwartz, Cathy Riemer
- p-values for conserved regions — Jia Li

Web Sites

- http://bio.cse.psu.edu — (Multi)PipMaker, Pip Dispenser, LAJ, PipTools, Enterix
- http://genome-test.cse.ucsc.edu — Santa Cruz Genome Browser test site, including human-mouse alignments
- http://pipeline.lbl.gov — Vista alignment generator, human-mouse alignments