Mouse-Human Genomic Sequence Comparisons

Webb Miller
Penn State

Outline

1. A “pip” (percent identity plot) graphically summarizes a set of local alignments between two sequences.
2. Human-mouse alignments sometimes help find genes and regulatory elements.
3. PipMaker (**http://bio.cse.psu.edu/**) compares two genomic sequences.
4. Sequences from any two close species can be compared.
5. A detailed example.

Visualizing a Long Alignment

1. Divide it into segments between successive gaps.

2. Represent each segment by a line showing:

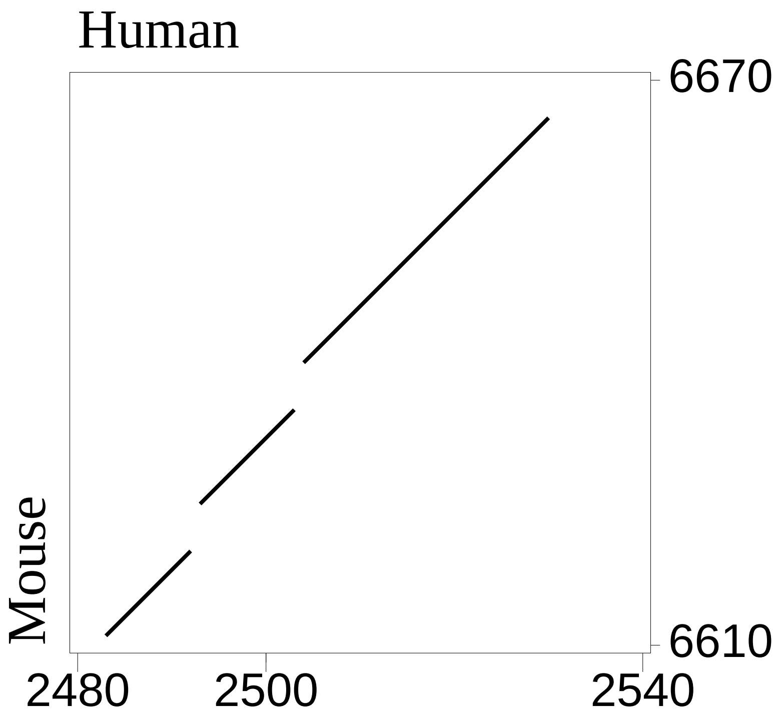
   (a) position in the first sequence vs. position in the second sequence (*dotplot*) or

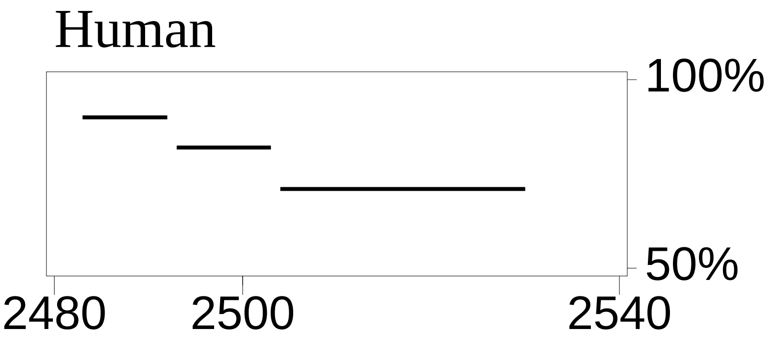
   (b) position in the first sequence vs. percent identity (*pip)*

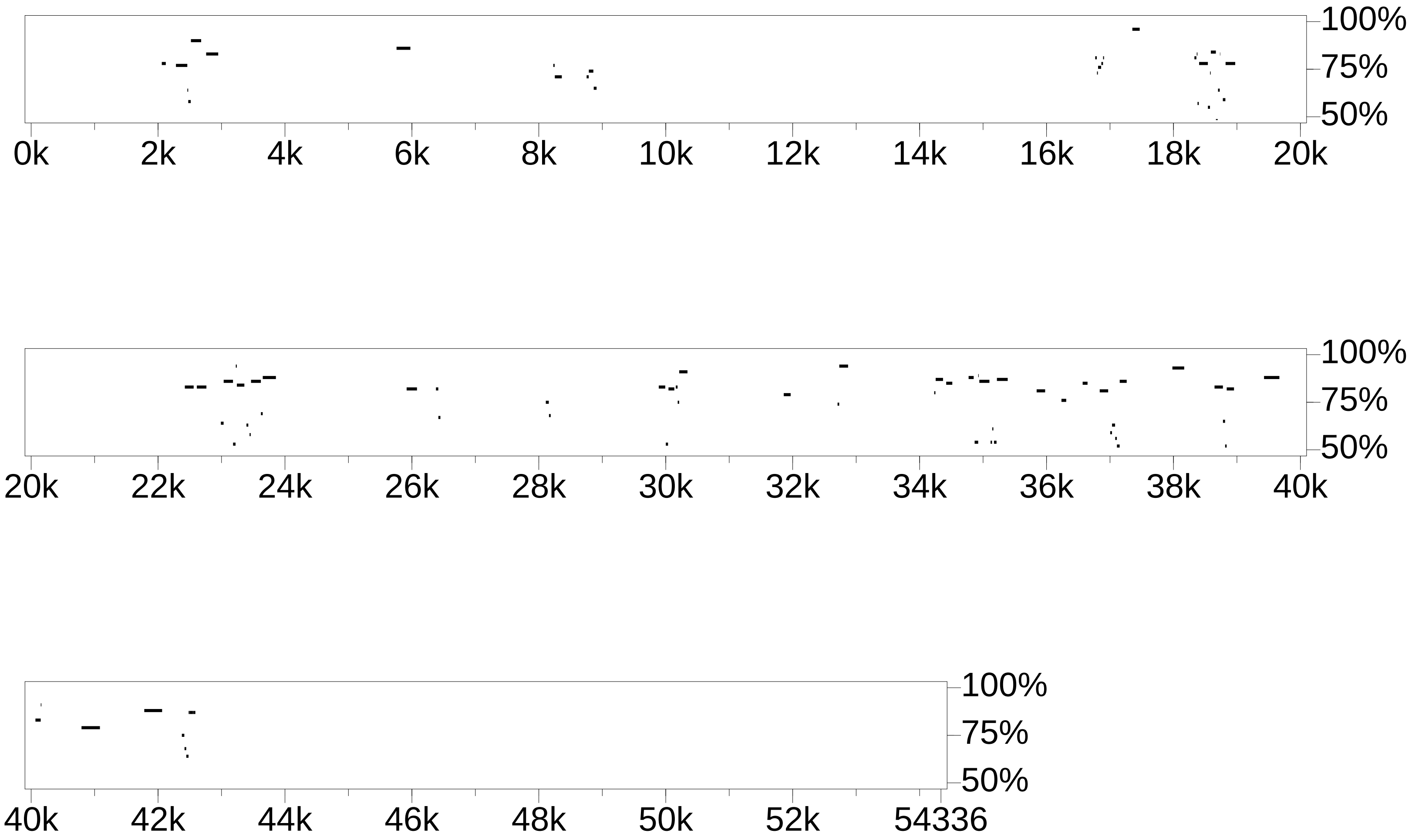
```
2483     2492   2493   2503    2504                   2530
   |--------|   |---------|    |-------------------------|
   CCAGGGCAAT   ATTTTCAAAGA    CAAATCTGATTCGTTCTCCCCATCCCA
   ||||||| ||---|||||||:|:|----|  |||||||||:|| :|: ||::|||
   CCAGGGCTATCCCTATTTTCAGAAATTGCTTATCTGATTCATTGCCTACACTCCA
   |--------|   |---------|    |-------------------------|
6611     6620   6625   6635    6640                   6666
```

Human	*Mouse*	*Identity*
2483-2492	6611-6620	90%
2493-2503	6625-6635	82%
2504-2530	6640-6666	67%

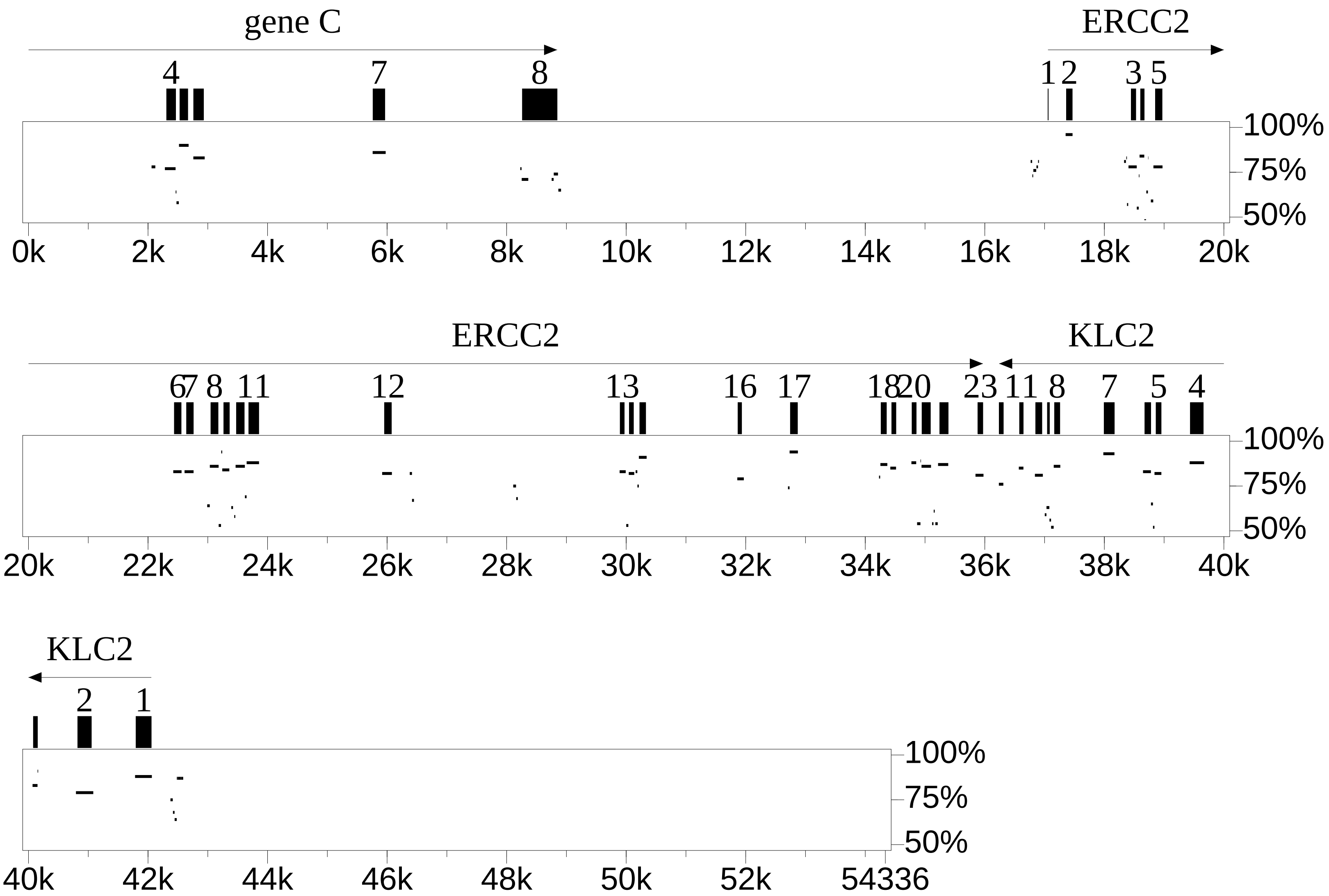
Human	*Mouse*	*Identity*
2483-2492	6611-6620	90%
2493-2503	6625-6635	82%
2504-2530	6640-6666	67%



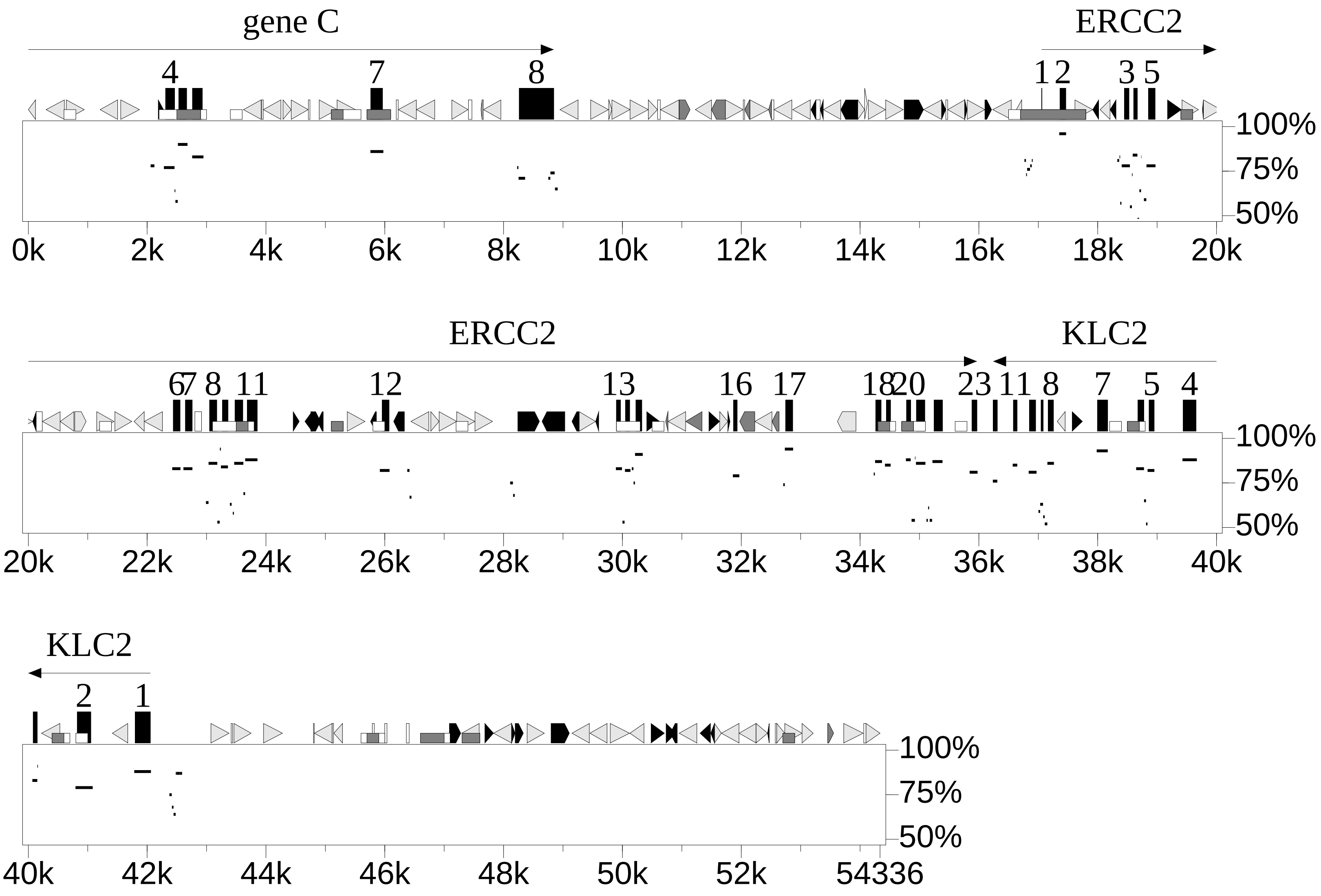



100%
75%
50%
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
100%
75%
50%
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
100%
75%
50%
40k
42k
44k
46k
48k
50k
52k
54336

Human ERCC2 region:



Human ERCC2 region:

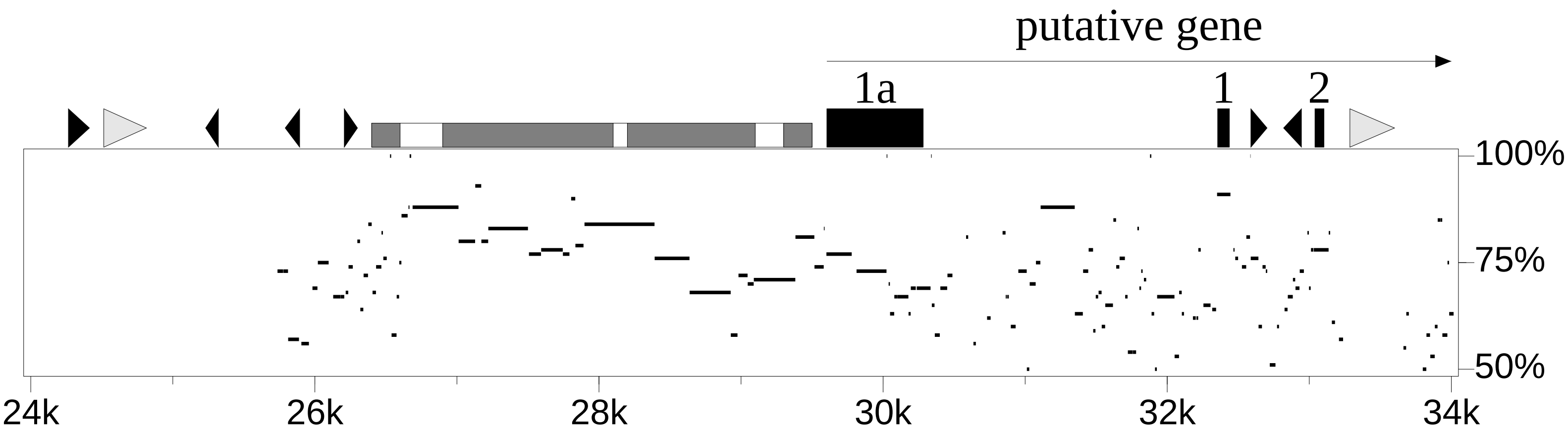


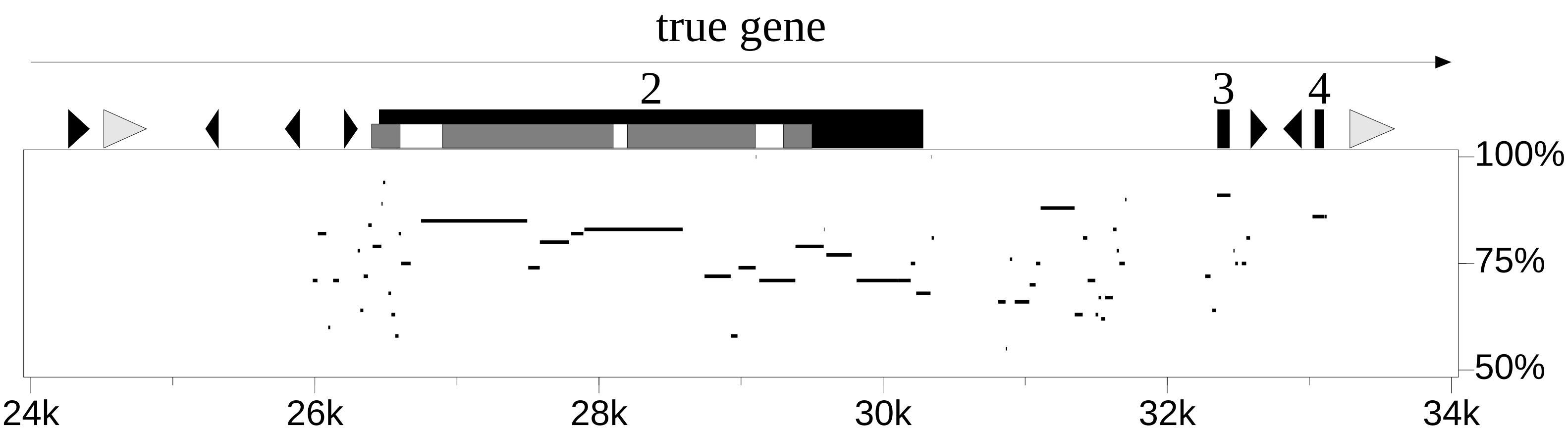
Computational Exon-Finding

- *Ab initio* gene-finders (GenScan, Grail) can be confused by alternative splicing and nested genes.

- Database searches may miss low-expressing genes and genes expressed in few tissues or developmental stages. Also, some exons may be missed (esp. with ESTs).

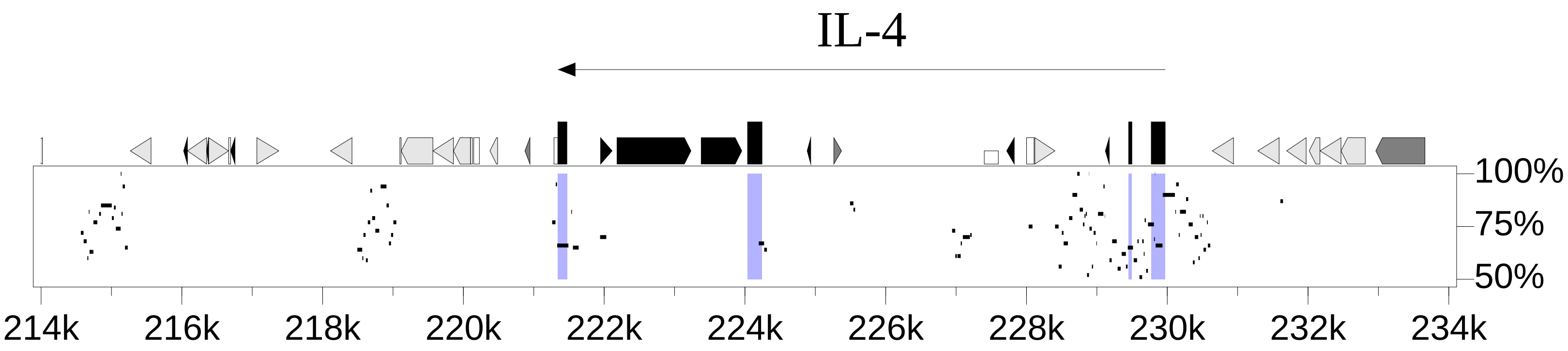
- Human-mouse comparisons complement these approaches quite effectively.

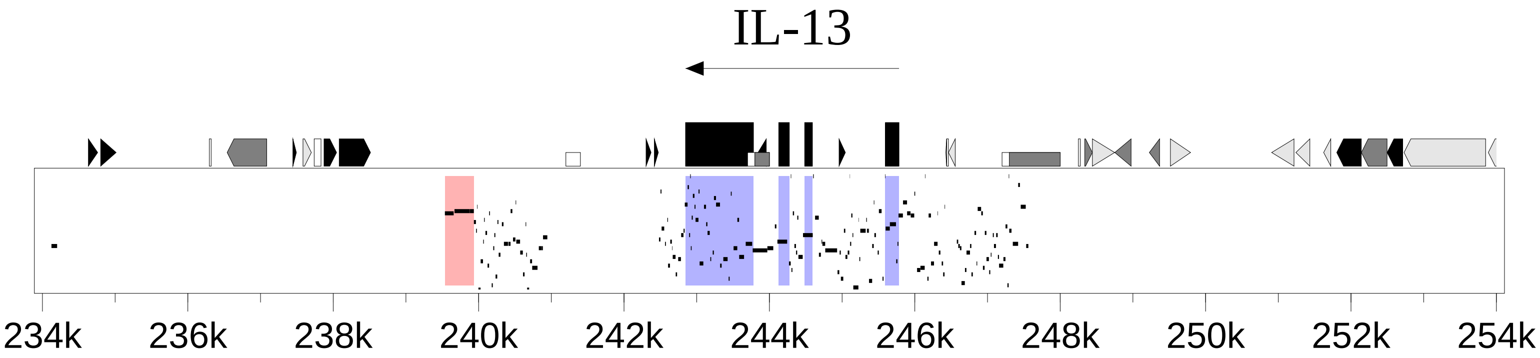

putative gene
1a
1
2
100%
75%
50%
24k
26k
28k
30k
32k
34k


true gene
2
3
4
100%
75%
50%
24k
26k
28k
30k
32k
34k

Bad News; Good News

- In some cases, because of low gene conservation and/or high background of conservation, coding regions don’t stand out in the pip.
- The pip frequently highlights non-coding functional regions, for which we have essentially no computational alternative.

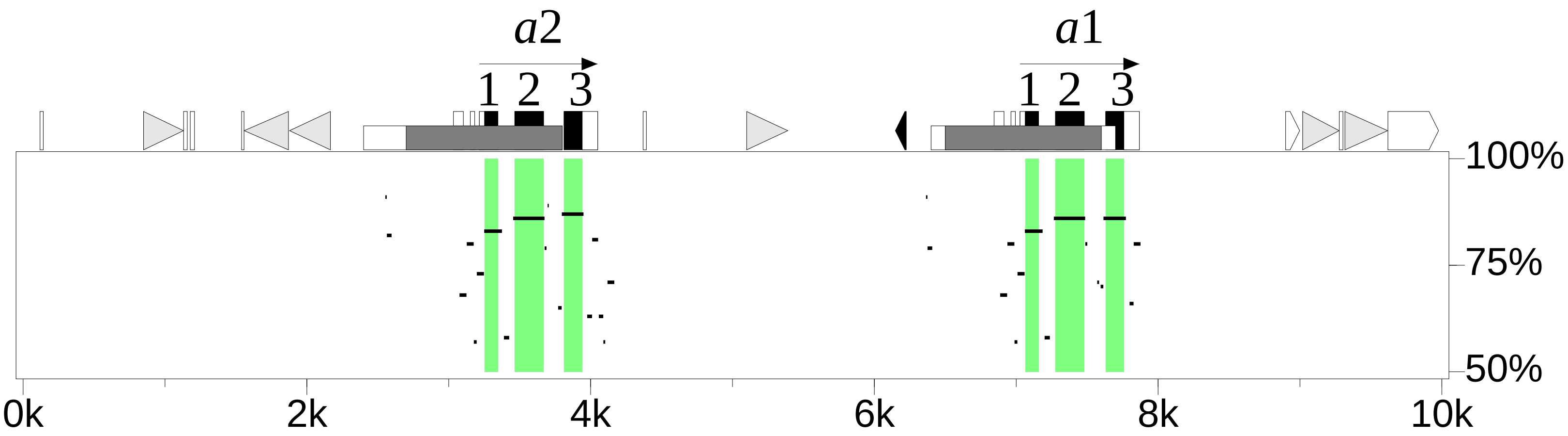

IL-4
100%
75%
50%
214k
216k
218k
220k
222k
224k
226k
228k
230k
232k
234k

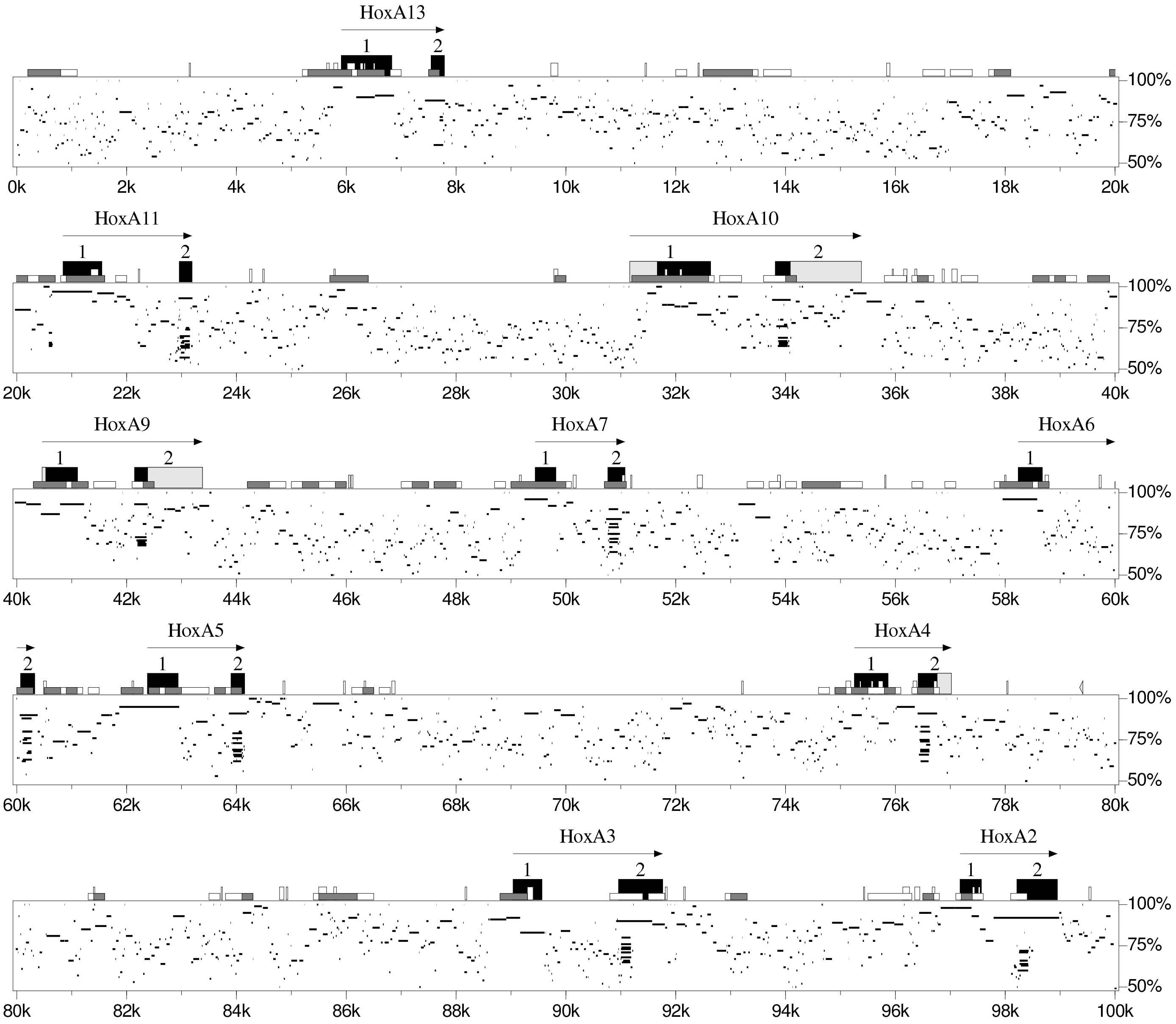

IL-13
234k
236k
238k
240k
242k
244k
246k
248k
250k
252k
254k

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

region	aligns	high	G+C	bp	masked	reference
HOXA	99.3	21.3	50.7	93211	15.2	unpublished
TCR	77.8	7.0	44.0	77115	21.0	Koop and Hood 1994
FHIT	58.1	7.6	37.1	331123	42.1	Shiraishi et al. 2001
CFTR	53.2	4.9	34.9	247331	41.3	Ellsworth et al. 2000
BTK	49.6	4.9	41.1	43504	41.0	Oeltjen et al. 1997
SNCA	44.4	1.0	34.6	84504	29.8	Touchman et al. 2001
DIST1	40.9	0.8	55.3	64841	45.7	Flint et al. 2001
MECP2	39.7	5.9	47.8	59670	56.9	Reichwald et al. 2000
CD4	35.6	3.3	51.9	106531	50.8	Ansari-Lari et al. 1998
CECR	21.3	1.8	45.9	368778	52.5	Footz et al. 2001
MYO15	15.4	3.7	56.9	46035	47.7	Liang et al. 1999
ERCC2	11.0	0.0	58.5	15721	61.7	Lamerdin et al. 1996

Network Resources for Genomic Alignments

Name	Http Address	Type	Reference
Alfresco	www.sanger.ac.uk/Software/Alfresco	P	Jareborg and Durbin 2000
CGAT	inertia.bs.jhmi.edu/roger/CGAT/CGAT.html	P	Lund *et al.* 2000
EnteriX	globin.cse.psu.edu/enterix	A	Florea *et al.* 2000a
GLASS	plover.lcs.mit.edu	S	Batzoglou *et al.* 2000
Gibbs	www.wadsworth.org/res&res/bioinfo	P,S	Wasserman *et al.* 2000
Intronerator	www.cse.ucsc.edu/∼kent/intronerator	S	Kent and Zahler 2000
LAJ	bio.cse.psu.edu	A,P	Wilson *et al.* 2001
LAJ	web.uvic.ca/∼bioweb/laj.html	A	Wilson *et al.* 2001
MUMmer	www.tigr.org/softlab	P	Delcher *et al.* 1999
PipMaker	bio.cse.psu.edu	S	Schwartz *et al.* 2000
SynPlot	www.sanger.ac.uk/Users/jgrg/SynPlot	P	Göttgens *et al.* 2000
VISTA	www-gsd.lbl.gov/vista	S	Dubchak *et al.* 2000
WABA	www.cse.ucsc.edu/∼kent/xenoAli/index.html	P,S	Kent and Zahler 2000

A = archived alignments; P = programs; S = server

PipMaker Input

- one completed genomic sequence
- a second sequence, perhaps in pieces
- optional positions of interspersed repeats
- optional positions of genes and exons
- optional positions and colors of stripes
- optional hyperlinks to network sites

Schwartz *et al.*, *Genome Research* **10** (2000), 577-586.

PipMaker Example (see handout)

GenScan predicted 6 genes; Blast search categorized them:

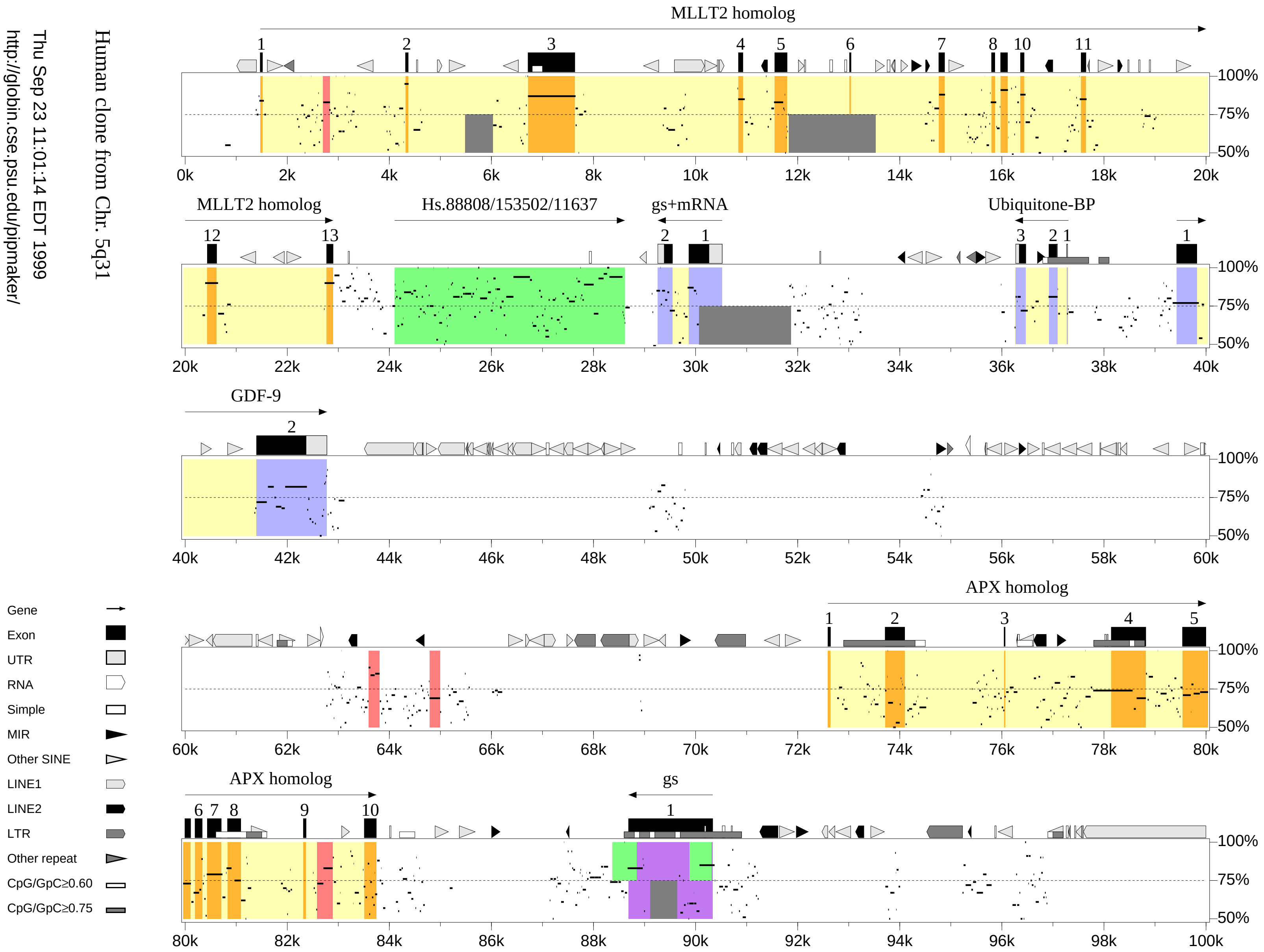
- 3 had characterized mRNAs (blue)
- 2 had weak protein hits (orange)
- 1 had only EST hits (purple)

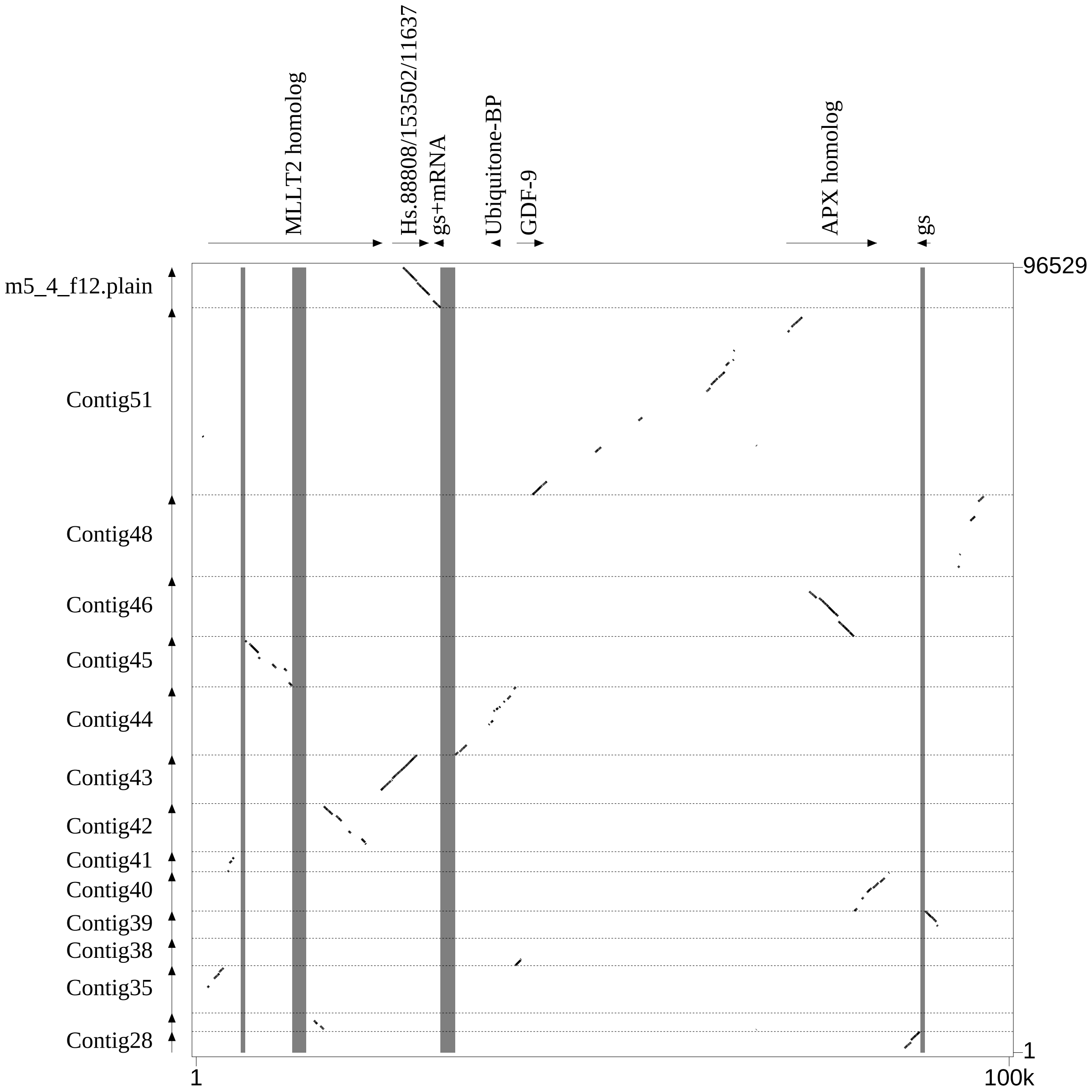
In addition, there was a cluster of EST hits (green).

Human clone from Chr. 5q31

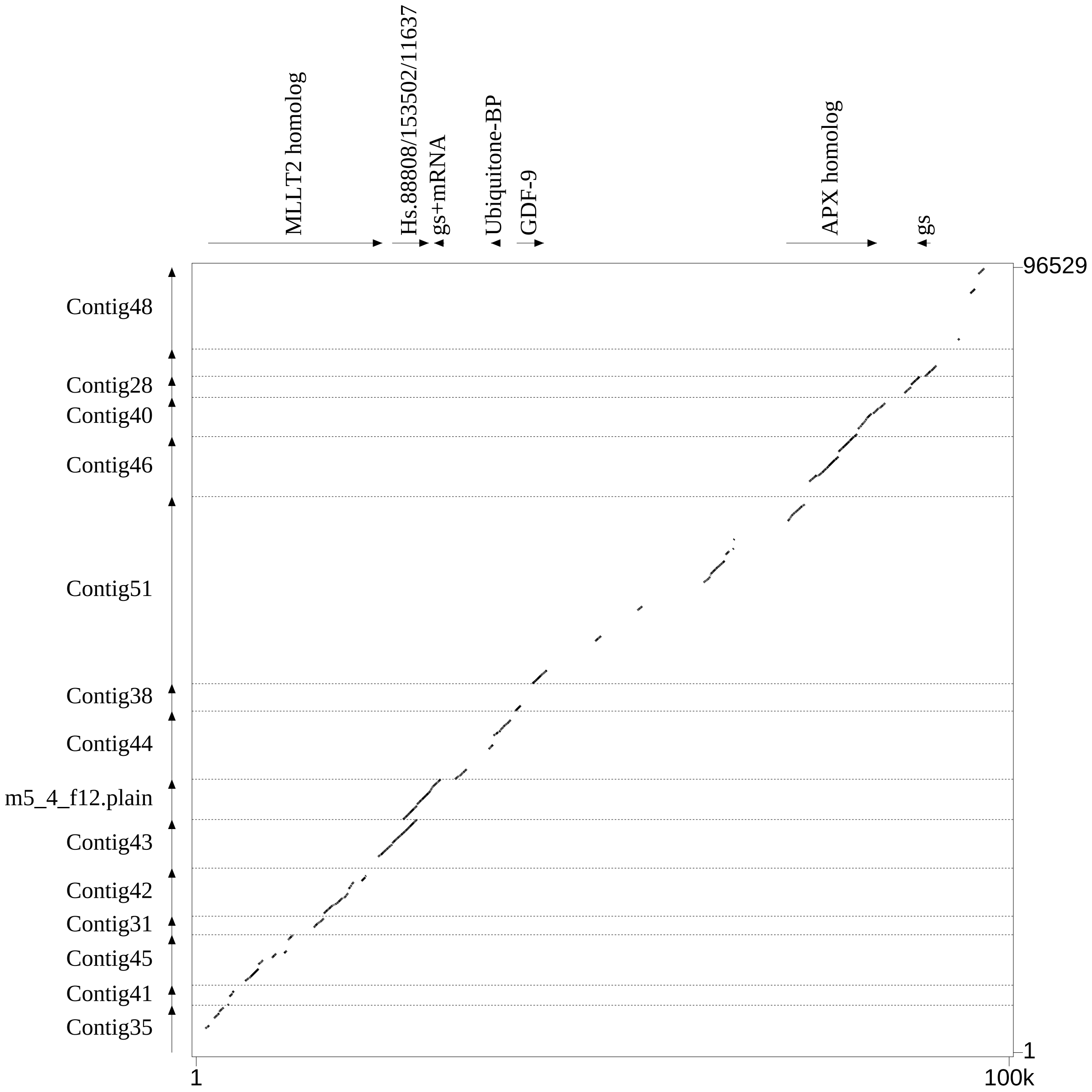
Thu Sep 23 11:01:14 EDT 1999

http://globin.cse.psu.edu/pipmaker/





MLLT2 homolog
Hs.88808/153502/11637
gs+mRNA
Ubiquitone-BP
GDF-9
APX homolog
gs
96529
1
m5_4_f12.plain
Contig51
Contig48
Contig46
Contig45
Contig44
Contig43
Contig42
Contig41
Contig40
Contig39
Contig38
Contig35
Contig28
1
100k

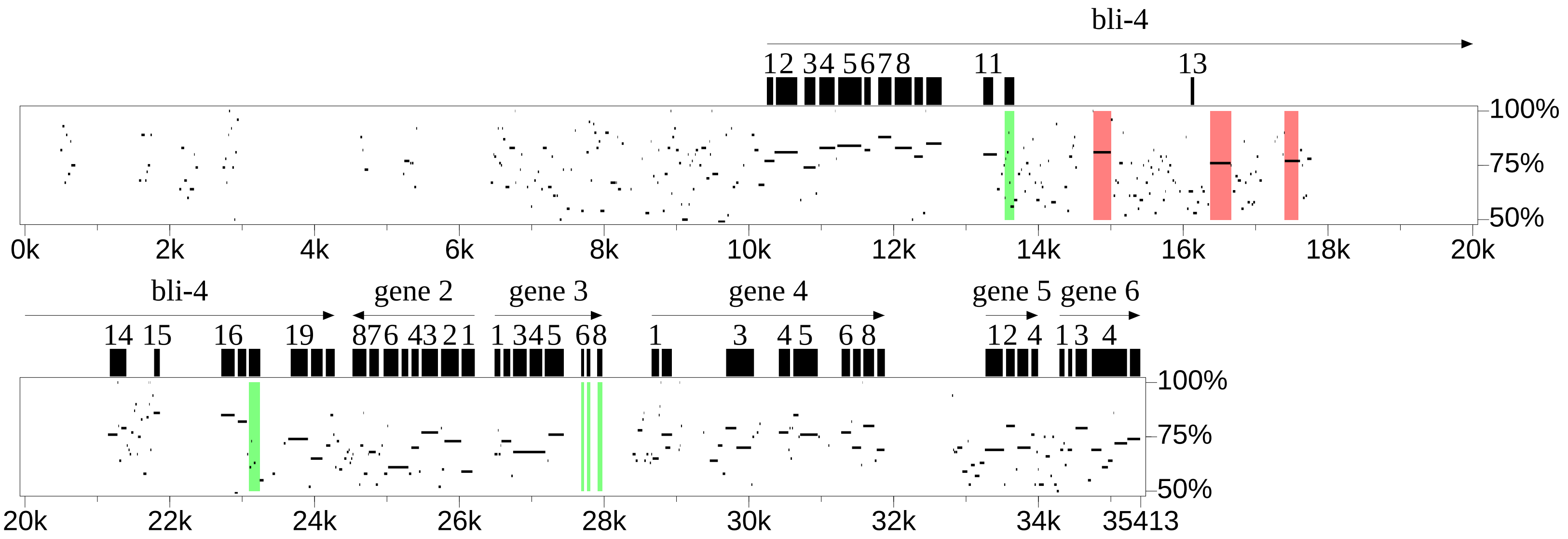

MLLT2 homolog
Hs.88808/153502/11637
gs+mRNA
Ubiquitone-BP
GDF-9
APX homolog
gs
96529
Contig48
Contig28
Contig40
Contig46
Contig51
Contig38
Contig44
m5_4_f12.plain
Contig43
Contig42
Contig31
Contig45
Contig41
Contig35
1
1
100k

When is PipMaker Appropriate?

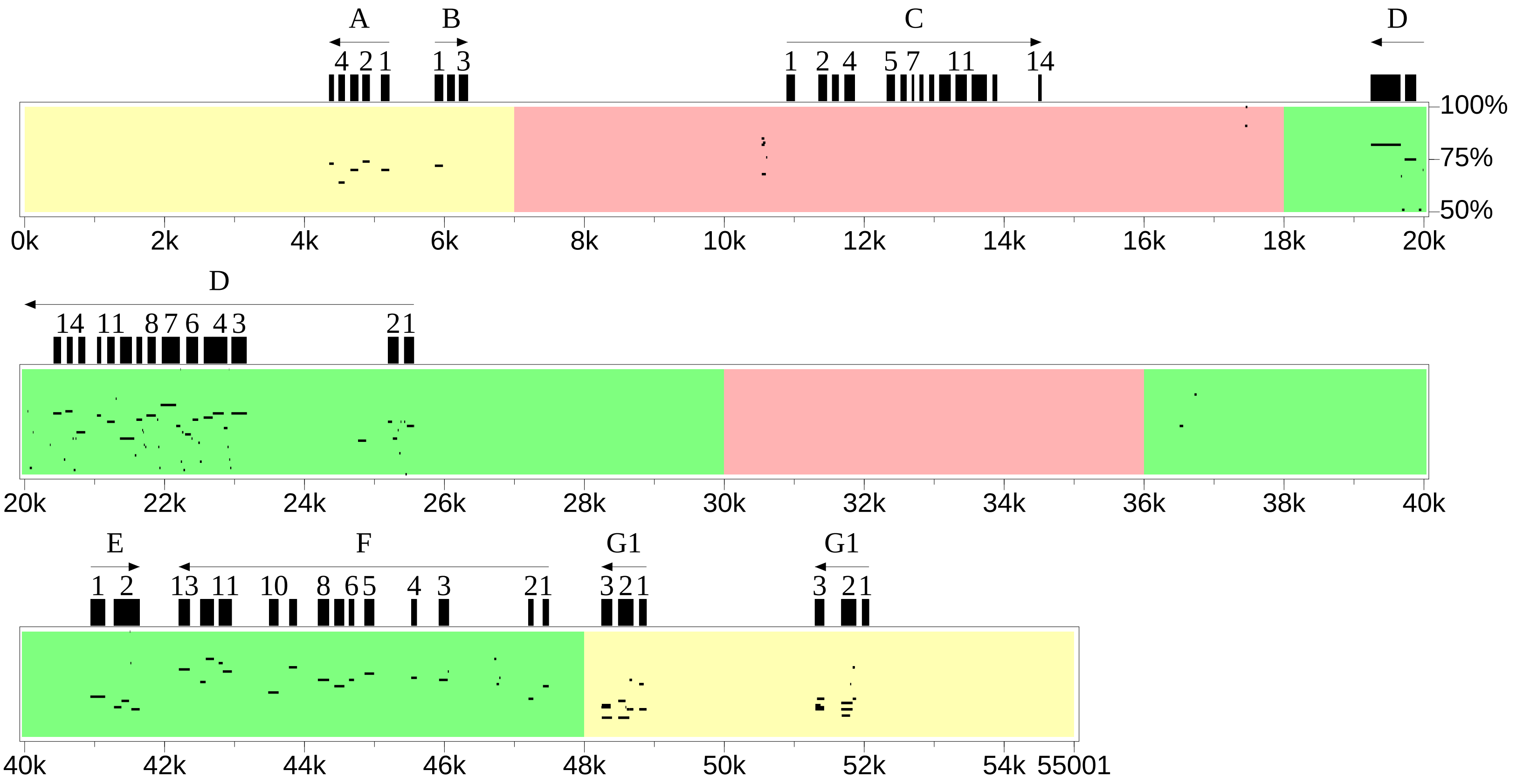
Aligning long DNA sequences (say, between 10 Kb and 10 Mb) from species that diverged, say 30-400 MYA.

Examples: *C. elegans* vs. *C. briggsae* and *E. coli* vs. *Salmonella*.

C. elegans (GenBank annotations from AF039719)



Fugu


A
4 2 1
B
1 3
C
1 2 4 5 7 11 14
D
100%
75%
50%
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
D
14 11 8 7 6 4 3 2 1
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
E
1 2
F
13 11 10 8 6 5 4 3 2 1
G1
3 2 1
G1
3 2 1
40k
42k
44k
46k
48k
50k
52k
54k
55001

Major Points

- Mouse-human comparisons help find genes, regulatory sites, and other functional regions.
- Rates of evolution vary among genomic locations, affecting what can be seen.
- The approach can be used for sequence pairs at an appropriate evolutionary
- Authors can make a pip available on their Web sites as an “electronic supplement” to a sequence-analysis publication. The pip can be hyperlinked to other network resources.