Comparative genomics tools for biological discovery

Inna Dubchak, Ph.D.
Staff Scientist

Lawrence Berkeley National Laboratory, NERSC
Ildubchak@lbl.gov
(510)495-2419

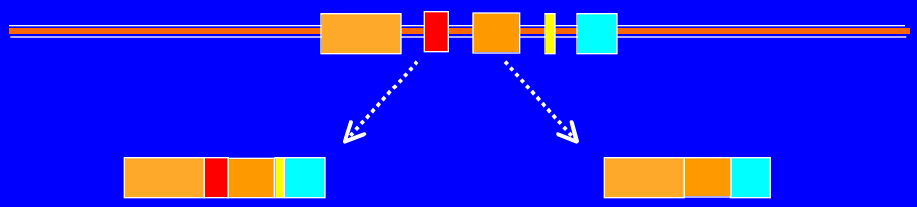
Why comparative genomics?

- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

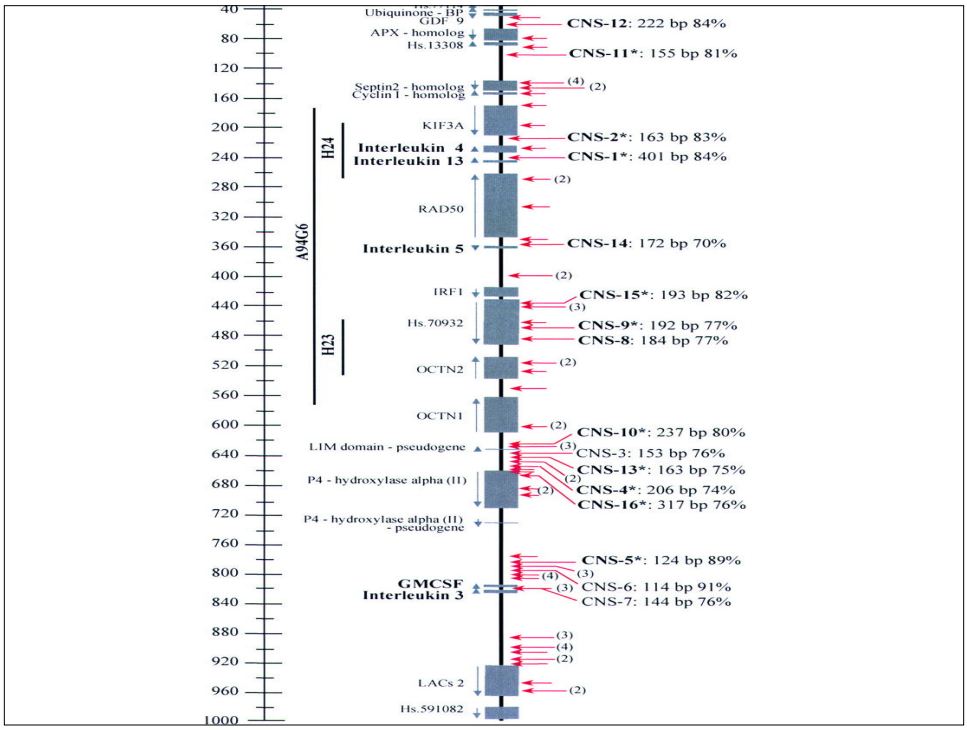
- Biological importance of conserved non-coding elements
- Discovering them in DNA sequence
- Tools for sequence visualization and analysis

~ 5% coding
~ 95% non-coding

Gene A



Protein A **Protein A'**



90 Elements in 1 Megabase

Are most conserved noncoding sequences "functional" or are they a product of passive evolution?

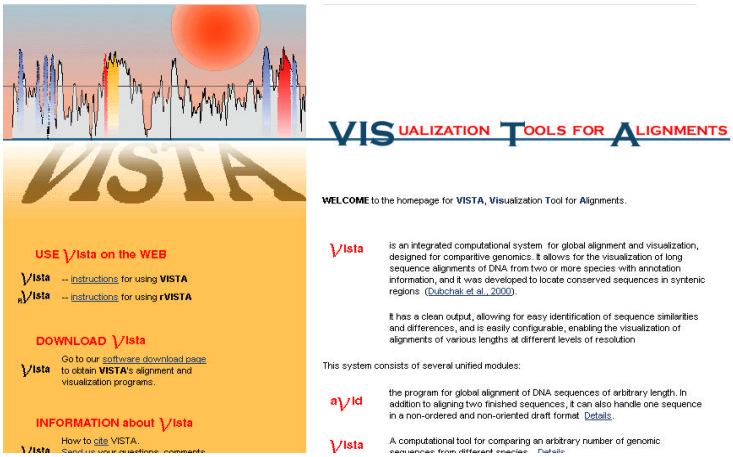
Evolutionarily Conserved Non-Coding Sequences

Identification

Verification

Analysis

http://www-gsd.lbl.gov/vista



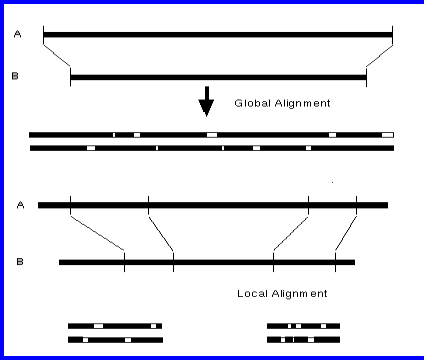
Well tested - ~2500 queries processed on-line

Modules of VISTA:

- Novel program for global alignment of DNA fragments of any length
- Visualization of alignment and various sequence features
- Evaluation and retrieval of all regions with predefined levels of conservation in two or more species

Local alignment algorithms are designed to search for highly similar regions in two sequences that may not be highly similar in their entirety. The algorithm works by first finding very short common segments between the input sequence and database sequences, and then expanding out the matching regions as far as possible.

For cross-species comparison one needs to accurately align two complete sequences. It is insufficient to find common similar regions in the two sequences, rather, what is needed is a global map specifying how the two sequences fit together, much like understanding how the pieces in a puzzle connect up with each other. This problem is called **global alignment**



Aligning large genomic regions

- Long sequences lead to memory problems
- Speed becomes an issue
- Long alignments are very sensitive to parameters
- Draft sequences present a nontrivial problem
- Accuracy is difficult to measure and to achieve

References for some existing programs:

Glass:
Domino Tiling, Gene Recognition, and Mice.
Pachter, L. *Ph.D. Thesis, MIT* (1999)
Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Batzoglou, S., Pachter, L., Mesirov, J., Berger, B., Lander, E. *Genome Research* (2000).
MUMmer
Delcher, A.L., Kasif S., Fleischmann, R.D., Peterson J., White, O. and Salzberg, S.L.
Alignment of whole genomes. *Nucleic Acids Research* (1999)
PipMaker
PipMaker: A Web Server for Aligning Two Genomic DNA Sequences.
Scott Schwartz, Zheng Zhang, Kelly A. Frazer, Arian Smit, Cathy Riemer, John Bouck, Richard Gibbs, Ross Hardison, and Webb Miller. *Genome Research* (2000)
Scan2
Dbscan/Scan2: Fast alignment of mega-sequences.
Seledtsov I.A., Solovyev V.V. To Appear. Web site http://softberry.com/

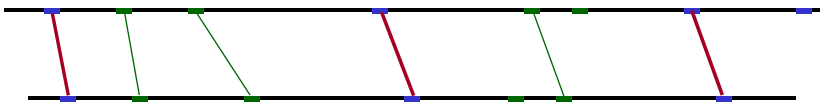
AVID
Aligner for VISTA including Draft

Main features

- Fast global alignment of long sequences irrespective of similarity (applications: comparison of assemblies, human – chimp comparison, human – mouse comparison, megabase alignments)
- Orientation and ordering of draft contigs by comparison to a related finished sequence (applications: ordering and orientation of draft contigs for better assembly, error detection in assembled BACs)

Algorithm of AVID

(outline)



- ORDER and ORIENT
- FIND all the (maximal) matches
- ALIGN the maximal matches
- FIX matches with good local homology
- REFINE with shorter matches

Ordering and Orientation

- Build matrix of overlaps by hashing and then using Fourier transforms
- Merge and remove overlapping contigs
- Find matches between super-contigs and the finished sequence
- Use matching information to order and orient

Finding Matches

- Build suffix tree with back pointers using McCreight's algorithm
- Finding maximal matches in linear time by building a tree from k sequences
- Sort the matches by length and local homology for the purpose of finding anchors

Results and Features

- Very fast alignment – times depend on sequence similarity and length, and vary from seconds to minutes.
- ~100Mb RAM / megabase of sequence
- Detailed information about oriented and ordered contigs, and accurate placement in the finished sequence
- Full integration with repeat masking to avoid confusing matches
- Robust performance on sequence pairs that range from human/chimpanzee to human/fugu
- VISTA ready output

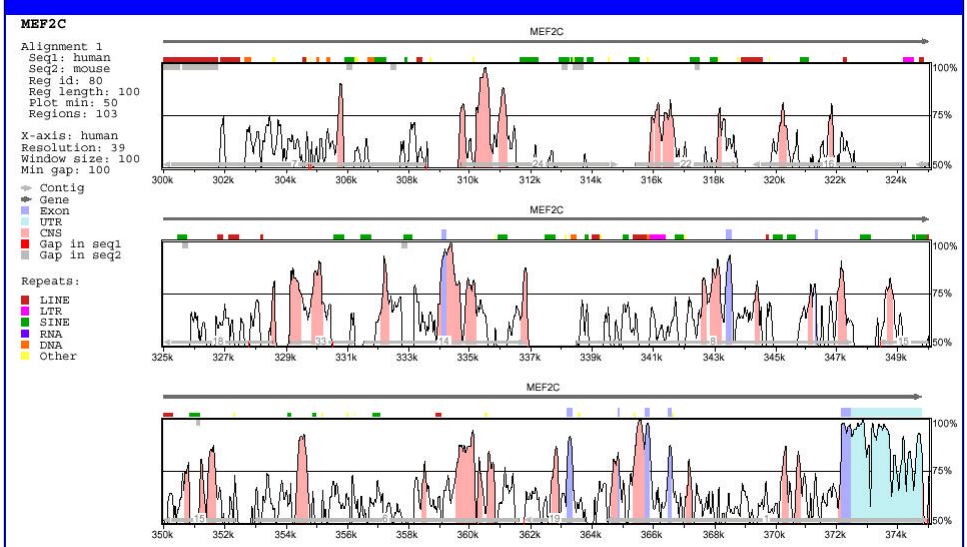
Visualization

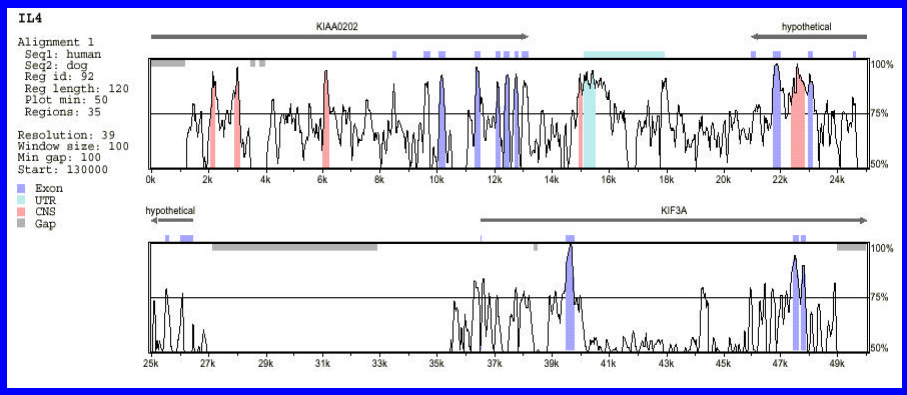
```
tggtaacattcaaattatg-----ttctcaaagtgagcatgaca-actttttccatgg
|| | ||||  |  |  ||     || | | |    |||||| |  ||   |   | ||
tgatgacatctatttgctgtttcctttttagaaactgcatgagagcctggctagtaggg
```

Window of length L is centered at a particular nucleotide in the base sequence

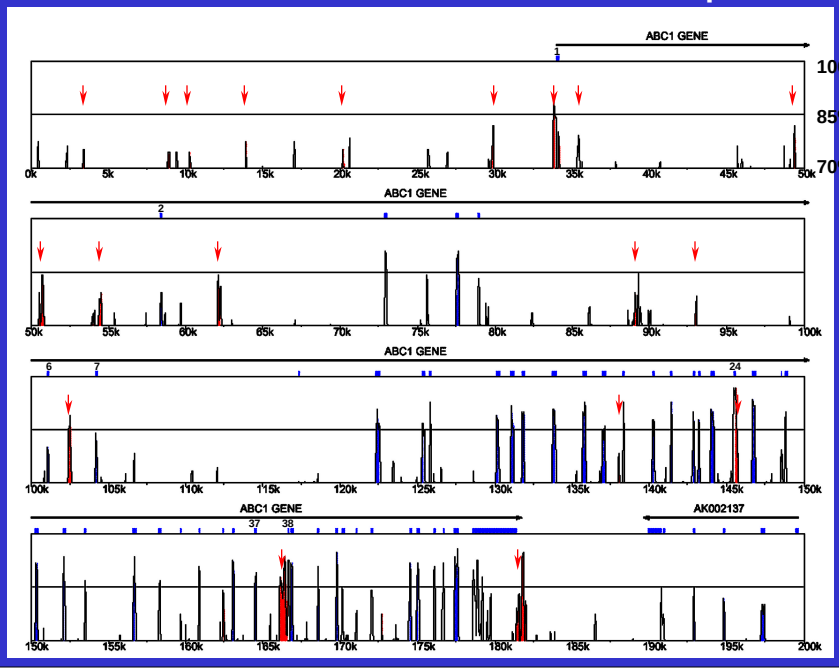
Percent of identical nucleotides in L positions of the alignment is calculated and plotted

Move to the next nucleotide

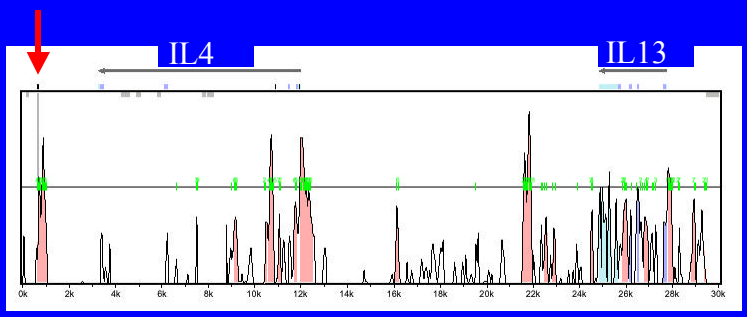




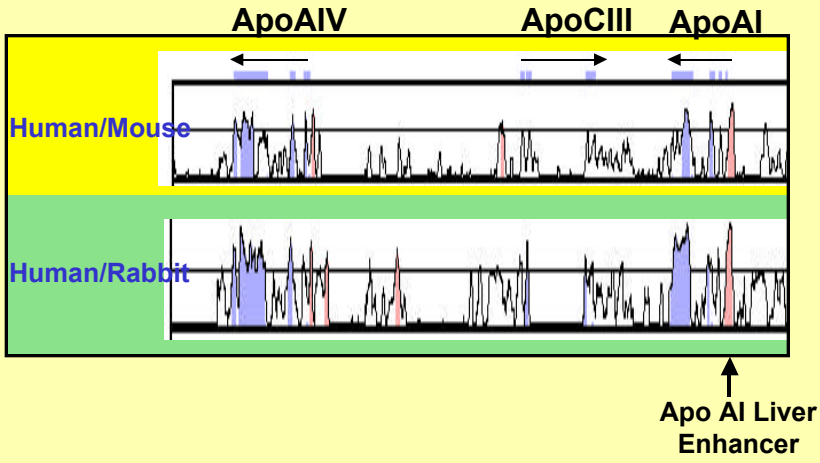
VISTA Plot of Human and Mouse ABCA1 Sequences

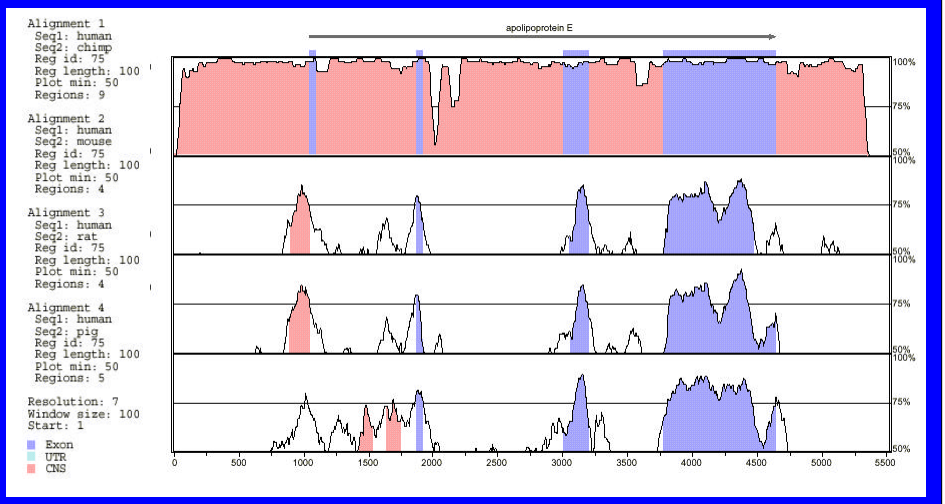


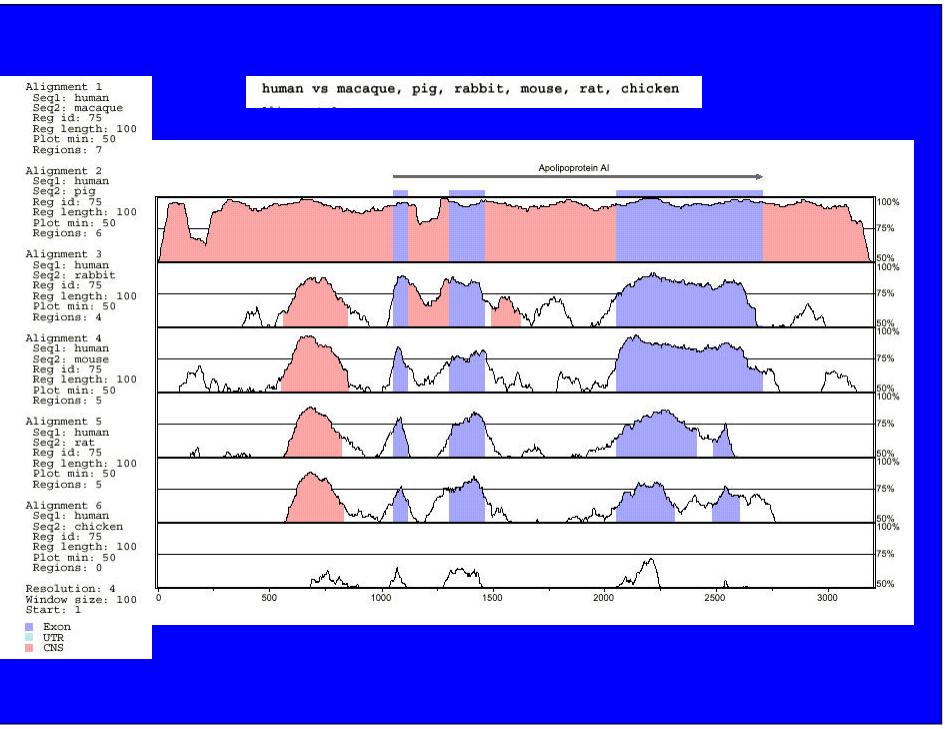
SNPs and Trascription Factor Binding Sites



Comparative Genomic Sequence Analysis of Human/Mouse/Rabbit ApoAI, CIII, AIV Cluster




Alignment 1
Seq1: human
Seq2: chimp
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 9
Alignment 2
Seq1: human
Seq2: mouse
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 4
Alignment 3
Seq1: human
Seq2: rat
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 4
Alignment 4
Seq1: human
Seq2: pig
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 5
Resolution: 7
Window size: 100
Start: 1
Exon
UTR
CNS
apolipoprotein E
100%
75%
50%
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500


human vs macaque, pig, rabbit, mouse, rat, chicken
Alignment 1
Seq1: human
Seq2: macaque
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 7
Alignment 2
Seq1: human
Seq2: pig
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 6
Alignment 3
Seq1: human
Seq2: rabbit
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 4
Alignment 4
Seq1: human
Seq2: mouse
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 5
Alignment 5
Seq1: human
Seq2: rat
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 5
Alignment 6
Seq1: human
Seq2: chicken
Reg id: 75
Reg length: 100
Plot min: 50
Regions: 0
Resolution: 4
Window size: 100
Start: 1
Exon
UTR
CNS
Apolipoprotein AI
100%
75%
50%
0
500
1000
1500
2000
2500
3000

Finding conserved regions with percentage and length cutoffs

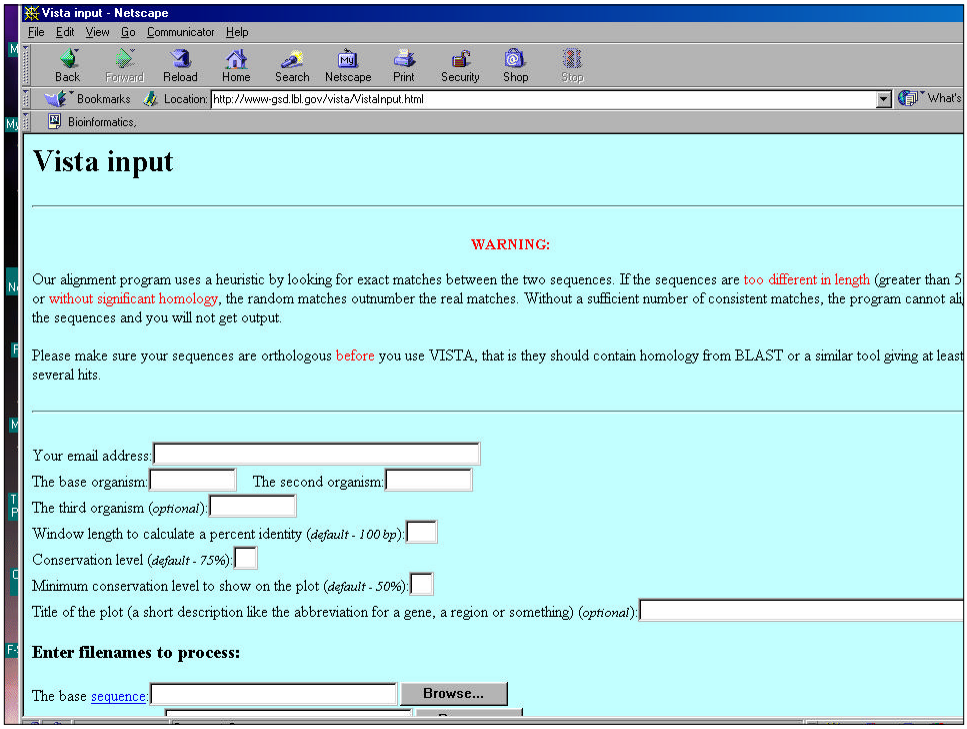
Conserved segments with percent identity X and length Y are defined to be regions in which every contiguous subsegment of length Y was at least X% identical to its paired sequence. These segments are merged to define the conserved regions.

Output:

11054 - 11156 = 103bp at 77.670%	NONCODING
13241 - 13453 = 213bp at 87.793%	EXON
14698 - 14822 = 125bp at 84.800%	EXON

Main features of VISTA

- Clear , configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Effective handling of gaps in the alignment
- Available source code

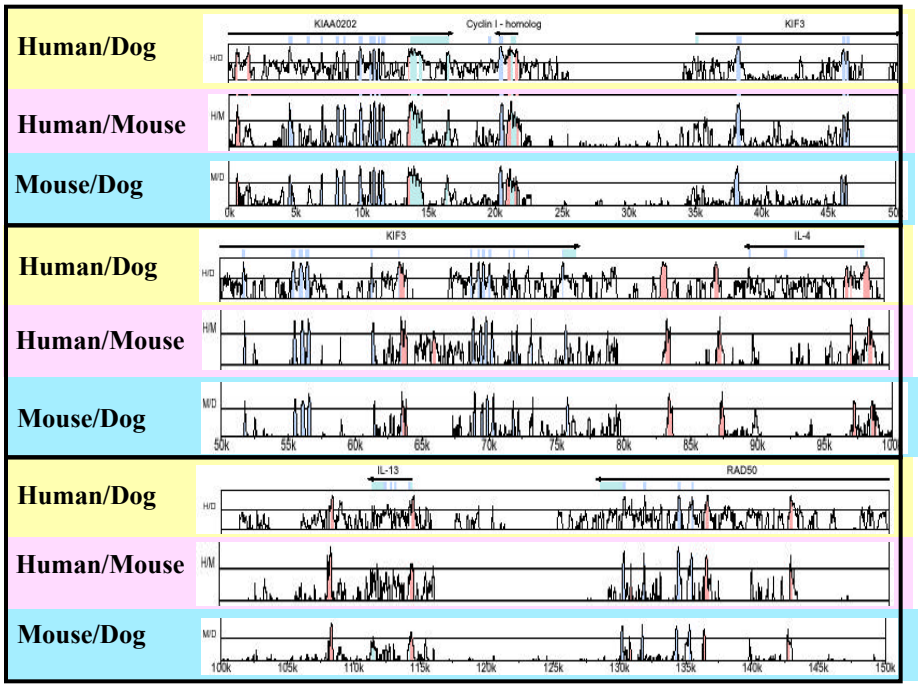


VISTA server output files

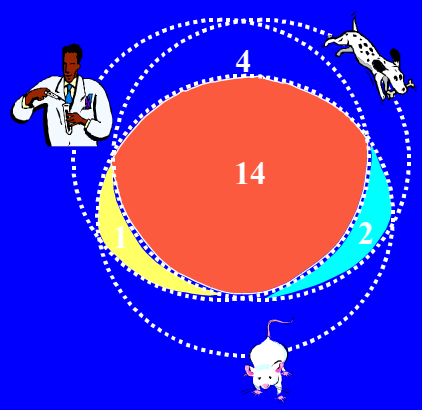
- All pair wise global alignments of the sequences
- VISTA plot in pdf format
- The list of conserved regions at predefined length and conservation cutoffs

Additional modules of VISTA

- Analyzing cutoffs to define actively conserved non-coding sequences
- Finding potential transcription factor binding sites by two-species comparison
- Comparing two closely related species



Active conservation of noncoding sequences – present in more than two mammals



% Cutoff

sum of three pair wise Intersection/Union values is maximal

Over 120 basepairs:

H/D > 92%

H/M > 80%

D/M > 77%

_Vista (complementary Vista)

Visualizes and analyzes alignments of genomic fragments of **closely related** species like human/primates

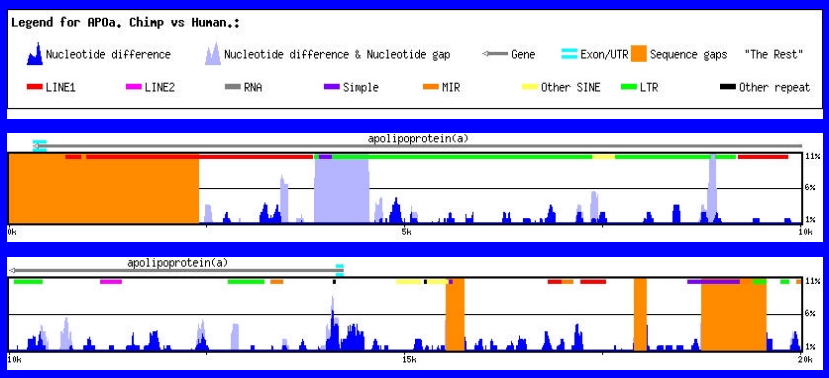
cVista shows patterns of DNA, which differ, while Vista is aimed at representing DNA patterns with a high conservation level

- represents location and magnitude of main **sequence differences.**
- compares 3 DNA sequences simultaneously.
- separates sequence differences into two parts:

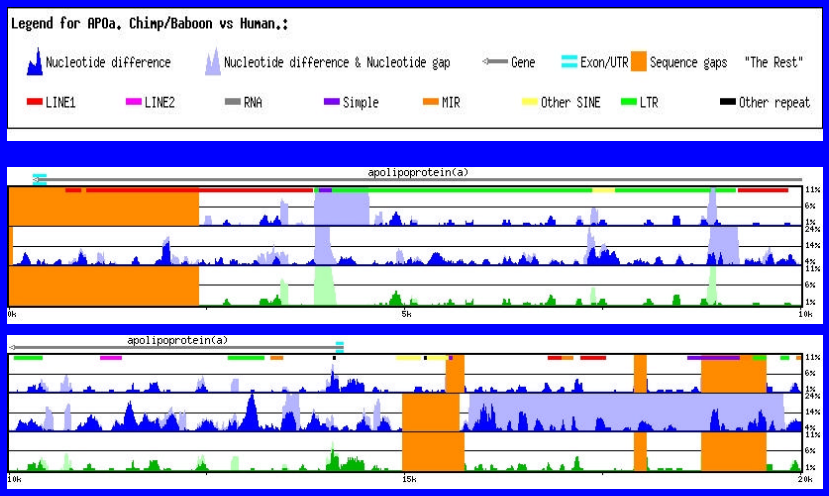
1) when nucleotides are present in both sequences, but differ

2) the difference due both to nucleotide difference and gaps

cVISTA for two species



Three closely related species

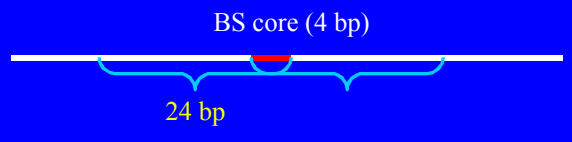


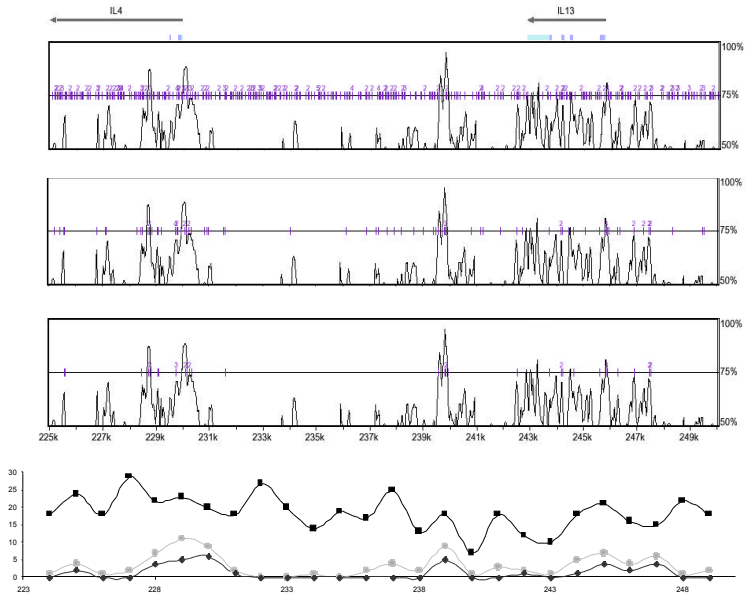
Prediction of transcription factor binding sites

- Simultaneous searches of the major transcription factor binding site database (Transfac) and the use of global sequence alignment to sieve through the data.
- Combination of database searches with comparative sequence analysis reduces the number of predicted transcription factor binding sites by several orders of magnitude.

~1 Meg region, 5q31

	Coding	Noncoding
Human interval Transfac predictions for GATA sites	839	20654
Aligned with the same predicted site in the mouse seq.	450	2618
Alligned sites conserved at 80% / 24 bp dynamic window	303	731
Random DNA sequence of the same length	29280	





Sequence motif recognition

+

multiple sequence alignment of syntenic regions,

↓

a high throughput strategy for filtering and prioritizing putative DNA binding sites

↓

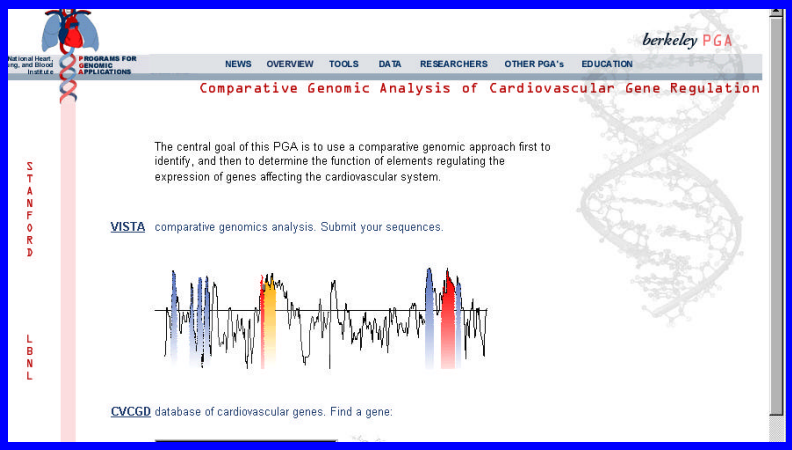
genomically informed starting place for globally investigating the detailed regulation

Reproduction of VISTA

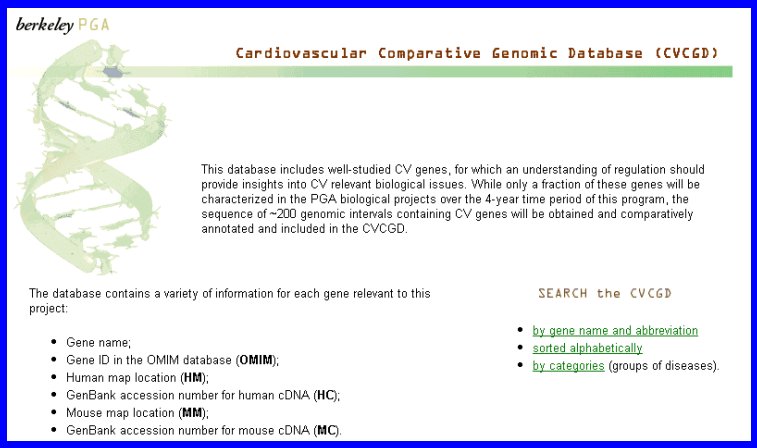
- VISTA
- multiVISTA
- cVISTA
- regulatoryVISTA
- activeVISTA or conservationVISTA
- audioVISTA

Information in Sequence

http://pga.lbl.gov



http://pga.lbl.gov/cvcgd.html



Example of CVCGD entry

Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1) - Netscape

File Edit View Go Communicator Help

CVCGD search results | Reload | Home | Search | Netscape | Print | Security | Shop | Stop

Bookmarks | Location: http://pga.lbl.gov/cgi-bin/get_gene?id=234

Bioinformatics,

Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

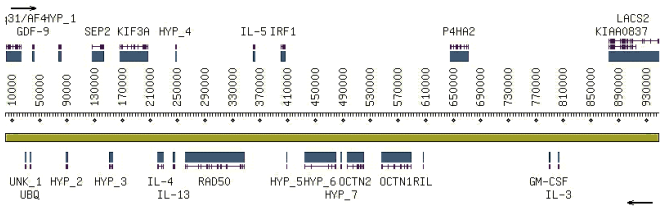
- Category: Atherosclerosis
- Gene ID in the OMIM database: 604190
- Human map location: 5q31
- GenBank accession number for human cDNA: NM_003059
- Mouse map location: 11
- GenBank accession number for mouse cDNA: NM_019687
- Annotation of the human sequence
- Human_mouse alignment: Whole sequence | 1-100000 | 100001-200000 | 200001-300000 | 300001-400000 | 400001-500000 | 500001-600000 | 600001-700000 | 700001-800000 | 800001-900000 | 900001-967696 *(see important note below)* | Printable version (PDF)
- List of conserved regions

Note: *If your browser hangs or crashes on the alignment page you can try this link instead.*

Short annotation of the region

Annotation of the VA5q31 region *

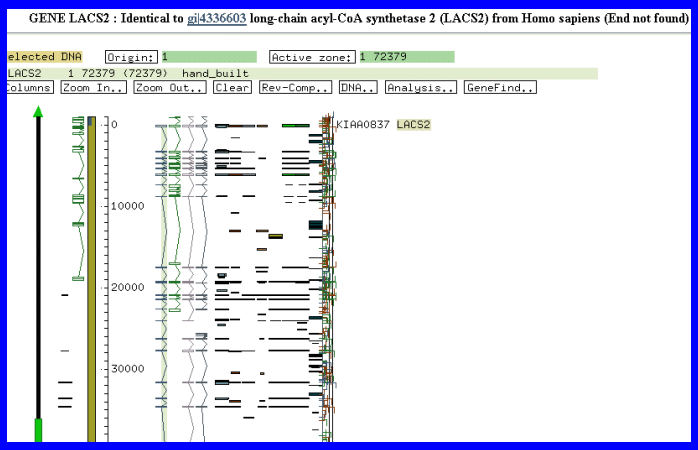
948869 bp



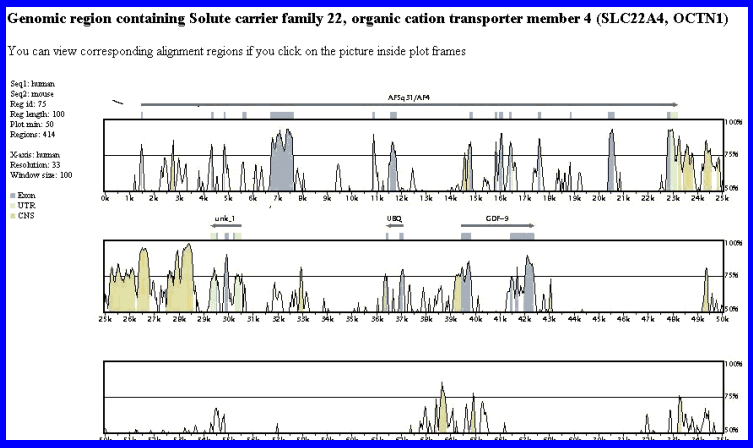
*Assembly contains a deletion of 18822 bp after the first exon of the RIL gene

Gene Name	Identity/Similarity
AF5q31/AF4	Identical to gi|6601437 Homo sapiens AF5q31 protein (AF5q31) (Start not found)
GDF-9	Identical to gi|488526 growth differentiation factor 9 from Homo sapiens
HYP_1	hypothetical
SEP2	Identical to gi|1503987 KIAA0202/Septin 2 gene from Homo sapiens

Detailed annotation in AceDB format



VISTA plot of the region



Alignment

Genomic region containing Solute carrier family 22, organic cation transporter membe

```
seq1 = human
seq2 = mouse

                 6990      7000      7010      7020      7030      7040
seq1     CAGAGTGACAGCACAACACAGAGAAGAACTGTAGGCAAAAAACAACCCAAAAAGGCTGAG
         ||||| ||||| || || ||||| |||||||| |||||||||||||||||||  |||||
seq2     CAGAGCGACAGTACCACTCAGAGGAGAACTGTCGGCAAAAAACAACCCAAAAAACCTGAG
            8130      8140      8150      8160      8170      8180

                 7050      7060      7070      7080      7090      7100
seq1     AAGGCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAGTGAAACCCCTGTAGAC
         ||| ||||||||||||||||||||||||||||||||||||||||||| |||||||| |||
seq2     AAGTCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAGTGAGACCCCTGTGGAC
            8190      8200      8210      8220      8230      8240

                 7110      7120      7130      7140      7150      7160
seq1     TTGGCTAGCAGCATGCCCTCCAGCAGACACAAAGCAGCCACCAAAGGCTCAAGGAAACCC
          |||||  |||||||||||||||||| |||||||||||||||| ||||| |||||||||
seq2     ATGGCTGCCAGCATGCCCTCCAGCAGGCACAAAGCAGCCACCAAGGGCTCGAGGAAACCC
            8250      8260      8270      8280      8290      8300
```

Conserved regions

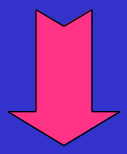
Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

Criteria: 75% identity over 100 bp

********** Conserved Regions - human (mouse) **********

1469	(580)	to	1515	(626)	=	47bp	at	85.1%	exon
2668	(2043)	to	2817	(2191)	=	153bp	at	80.4%	noncoding
4316	(4531)	to	4370	(4585)	=	55bp	at	100.0%	exon
4816	(6136)	to	4853	(6173)	=	38bp	at	97.4%	exon
6717	(7860)	to	7634	(8777)	=	918bp	at	87.8%	exon
10839	(10749)	to	10927	(10837)	=	89bp	at	91.0%	exon
11553	(12627)	to	11793	(12873)	=	247bp	at	81.8%	exon
14508	(15706)	to	14622	(15823)	=	119bp	at	76.5%	noncoding
14671	(15886)	to	14783	(16003)	=	118bp	at	74.6%	noncoding
14784	(16004)	to	14878	(16098)	=	95bp	at	89.5%	exon
15797	(17526)	to	15860	(17589)	=	64bp	at	93.8%	exon
15975	(17703)	to	16111	(17839)	=	137bp	at	90.5%	exon
16365	(18045)	to	16436	(18116)	=	72bp	at	91.7%	exon
16437	(18117)	to	16535	(18217)	=	101bp	at	75.2%	noncoding
17554	(18914)	to	17647	(19007)	=	94bp	at	87.2%	exon

Demonstrated the validity of a comparative genomic approach to study gene regulation in a pilot 1 mb study



for the entire human genome

Whole genome approach
http:/pipeline.lbl.gov

(A PIPELINE FOR REAL-TIME COMPARATIVE ANALYSIS OF THE HUMAN AND MOUSE GENOMES)

Description of the project:

- Introduction
- Examples
- Contact us

Pipeline results:

- Localization of Mouse Contigs on Human Genome
- Database of Conserved Noncoding Sequences (CNS)
- Annotation of Conserved Noncoding Sequences & **VISTA** visualization data (MySQL Tables)
- Faux paired mouse reads (coming soon)
- Mouse Contaminations in the Human Genome (Long stretches of mouse sequence exactly matching fragments in Human Genome)
- Overlapping Mouse Contigs
- 8 testing regions

Comments?

Localization of Mouse Contigs

Mouse Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y

Mouse contig	Length	Approximate Reverse Complement Mixture	VISTA		Conserved regions			AVID's Top Hit Localization		Effective Length of Top 3 Collapsed BLAT Hits		
NT_002597	204618 bp	0.03%	main	rc	main	rc	–	HChr4	59033900 – 59357429 :: 321529 bp	87836 bp	14513 bp	1329 bp
NT_026598	180686 bp	27.07%	main	rc	main	rc	+	HChr7	67215844 – 67551382 :: 335538 bp	57690 bp	31602 bp	15615 bp
NT_014920	469165 bp	7.51%	main	rc	main	rc	+	HChr7	67531287 – 68780650 :: 1249363 bp	194667 bp	14626 bp	7984 bp
NT_026533	296820 bp	0.30%	main	rc	main	rc	+	HChr7	102809830 – 103061093 :: 251263 bp	178267 bp	12670 bp	6872 bp
NT_026101	180673 bp	95.95%	main	–	main	–	–	HChr13	30633825 – 30940397 :: 306572 bp	37802 bp	36271 bp	765 bp
NT_026540	617309 bp	61.80%	main	rc	main	rc	–	HChr13	30633884 – 31080619 :: 446735 bp	89145 bp	55093 bp	1074 bp
NT_026539	216988 bp	0.94%	main	–	main	–	–	HChr13	30740275 – 31002789 :: 262514 bp	82505 bp	882 bp	775 bp
NT_025616	182534 bp	73.15%	main	rc	main	rc	+	HChr18	7365055 – 7500181 :: 135126 bp	730 bp	534 bp	443 bp

Global status: done=173, in progress=22

```
Criteria: 70% identity over 100 bp

*************** Conserved Regions - human (mouse) ***************

    1371  (110614)  to     1417  (110660)  =      47bp  at  85.1%  exon
    2251  (111775)  to     2429  (111943)  =     179bp  at  68.7%  noncoding
    2557  (112066)  to     2719  (112225)  =     166bp  at  78.3%  noncoding
    2808  (112500)  to     2955  (112645)  =     152bp  at  69.1%  noncoding
    3098  (112777)  to     3262  (112923)  =     165bp  at  68.5%  noncoding
    3684  (112997)  to     3869  (113173)  =     190bp  at  68.9%  noncoding
    4055  (114232)  to     4216  (114381)  =     162bp  at  67.9%  noncoding
    4217  (114382)  to     4272  (114437)  =      56bp  at 100.0%  exon
    4717  (115675)  to     4755  (115713)  =      39bp  at  97.4%  exon
    4756  (115714)  to     4897  (115868)  =     156bp  at  66.7%  noncoding
    5466  (116080)  to     5629  (116243)  =     164bp  at  89.6%  exon
    5938  (116887)  to     6044  (116991)  =     110bp  at  70.0%  noncoding
    6618  (117432)  to     7536  (118350)  =     919bp  at  87.8%  exon
    9280  (119788)  to     9440  (119953)  =     170bp  at  71.2%  noncoding
   10740  (120321)  to    10829  (120410)  =      90bp  at  91.1%  exon
   11341  (122070)  to    11453  (122198)  =     131bp  at  65.6%  noncoding
   11454  (122199)  to    11695  (122446)  =     248bp  at  81.9%  exon
   14407  (124121)  to    14540  (124262)  =     143bp  at  71.3%  noncoding
   14549  (124283)  to    14684  (124423)  =     141bp  at  70.9%  noncoding
   14685  (124424)  to    14780  (124519)  =      96bp  at  89.6%  exon
   15698  (125723)  to    15762  (125787)  =      65bp  at  93.8%  exon
   15698  (125723)  to    16013  (126037)  =     322bp  at  79.5%  exon
   15876  (125900)  to    16013  (126037)  =     138bp  at  90.6%  exon
   16266  (126242)  to    16338  (126314)  =      73bp  at  91.8%  exon
   16339  (126315)  to    16573  (126537)  =     237bp  at  68.8%  noncoding
   17455  (127111)  to    17549  (127205)  =      95bp  at  87.4%  exon
   18253  (127705)  to    18351  (127803)  =     100bp  at  71.0%  noncoding
   18739  (128759)  to    18783  (128803)  =      45bp  at  86.7%  exon
   20298  (129679)  to    20519  (129900)  =     222bp  at  92.3%  exon
   20520  (129901)  to    20664  (130044)  =     145bp  at  69.7%  noncoding
   22672  (133231)  to    22800  (133359)  =     129bp  at  93.8%  exon
   22672  (133231)  to    23206  (133751)  =     539bp  at  86.3%  exon
   23207  (133752)  to    23779  (134276)  =     574bp  at  72.5%  noncoding
   24031  (134619)  to    24598  (135172)  =     577bp  at  73.7%  noncoding
   24602  (135173)  to    24838  (135393)  =     237bp  at  68.8%  noncoding
   25046  (135567)  to    26171  (136647)  =    1144bp  at  75.7%  noncoding
   26182  (136657)  to    26759  (137244)  =     589bp  at  84.6%  noncoding
```

CNSes in the Human Genome (Oct. 7, 2000)

Human Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Mouse contig	Starting pos	No of CNSes	CNSes
NT_025617	44978245	19	1–679 680–1096 7476–7763 8099–8209 8210–8408 9990–10220 10692–10942 21674–21818 31817–31980 39077–40445 52343–53220 13735–13901 13902–13995 17218–17899 21044–21377 24819–25986 37113–38045 38066–38902 53148–53648
NT_026597	52985836	45	11603–11851 11907–12066 12426–12655 12691–12805 12912–13014 206601–206788 207045–207168 207958–208070 208862–209067 209344–209495 212811–212910 212965–213292 224829–225165 225435–225947 228334–228518 228560–228661 230305–230442 242466–242643 243977–244074 244087–244196 245176–245418 246511–246854 246969–247106 247237–247855 248618–248881 251687–252037 254344–254529 255197–255340 255517–255729 255765–255904 257892–257994 260305–260544 261364–261483 261620–261753 263295–263485 264185–264316 279093–279229 280865–280963 284226–284337 288574–288764 347–471 2220–2397 3080–3199 36302–36411 36739–36838
NT_025576	131374500	44	117–356 953–1069 6512–7047 9046–9173 10442–10560 11238–11708 14454–14580 18584–18846 22198–22332 22652–22776 23391–23957 24065–24350 24383–24530 25241–25356 25837–26327 39448–39576 41185–41641 41649–41774 42285–42974 43074–43207 43443–43540 43971–44091 44518–44618 44857–45320 46559–46765 47163–47382 47387–47806 64020–64127 65988–66289 70052–70175 70412–71042 71410–71762 71993–72135 72442–72541 72568–72737 72755–73199 73780–73883 75783–75930 98742–98909 99644–99796 99880–100358 102362–102460 104795–104982 105138–105326
NT_026035	131987513	131	150–258 2124–2903 3018–3383 3538–3703 3716–4029 7214–7306 7715–8315 8416–8608 9733–10041 10256–10418 10623–11052 11610–11743 13530–13638 13685–13794 13825–13926 14622–14786 15061–15197 15269–15396 16523–16730 16769–16981 17151–17591 17625–18178 20002–20122 20586–21068 21079–21202

Mouse Contaminations in the Human Genome (Oct. 7, 2000)

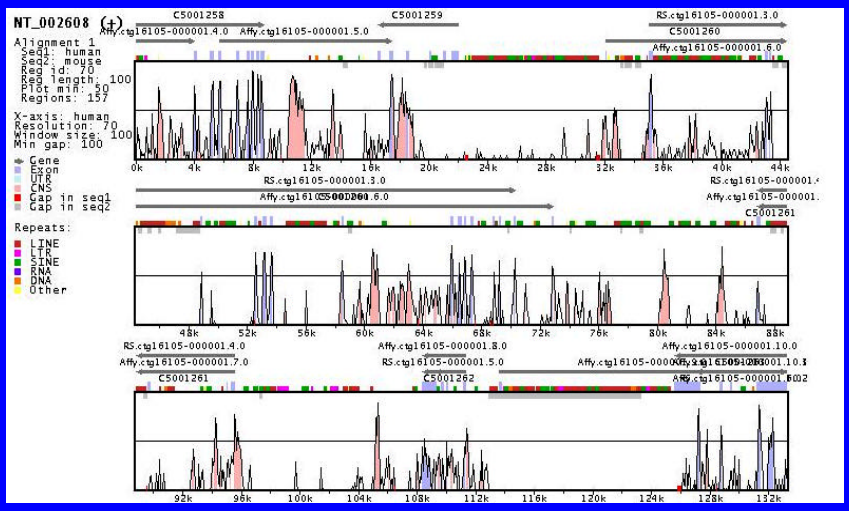
Human Chr: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Mouse contig	Mouse Chr	VISTA	Human Starting pos	No of Regions	Regions
NT_025623	Chr11	main rc	57850628	2	38841–44431 128484–140122
NT_026038	Chr11	main rc	58316024	8	12996–17449 19704–22884 41156–46746 46647–50632 99048–110244 130799–142537 143269–158764 190620–206117

* – **red** color corresponds to an opposite strand (rc)

Overlaping Locations of Mouse Contigs

Mouse contig	Mouse Chr	Human Position	Mouse contig	Mouse Chr	Human Position
NT_002625	Chr1	**Chr1** : 180211168 – 180466235	NT_002594	Chr1	**Chr1** : 180244673 – 180466202
NT_025539	Chr1	**Chr1** : 233498303 – 233975111	NT_025524	Chr1	**Chr1** : 233838288 – 233975111
NT_004180	Chr7	**Chr11** : 765063 – 2144767	NT_004181	Chr7	**Chr11** : 1951768 – 2160440
NT_019187	Chr2	**Chr11** : 32256933 – 32372132	NT_014930	Chr2	**Chr11** : 32256933 – 32368972
NT_026101	Chr5	**Chr13** : 30633825 – 30940397	NT_026540	Chr5	**Chr13** : 30633884 – 31080619
NT_026540	Chr5	**Chr13** : 30633884 – 31080619	NT_026539	Chr5	**Chr13** : 30740275 – 31002789
NT_026040	Chr12	**Chr14** : 105946386 – 106697482	NT_026138	Chr12	**Chr14** : 105979694 – 106697594
NT_026182	Chr11	**Chr17** : 18773846 – 19040117	NT_025583	Chr11	**Chr17** : 18959912 – 19329481
NT_026181	Chr11	**Chr17** : 44350536 – 44861111	NT_026137	Chr11	**Chr17** : 44372372 – 44588259
NT_026038	Chr11	**Chr17** : 58360959 – 58567075	NT_025623	Chr11	**Chr17** : 58363274 – 58503395
NT_002604	Chr7	**Chr19** : 47242229 – 47624454	NT_002601	Chr7	**Chr19** : 47598203 – 47627237
NT_026531	Chr1	**Chr2** : 233929296 – 234017066	NT_026094	Chr1	**Chr2** : 234016777 – 234137611
NT_026094	Chr1	**Chr2** : 234016777 – 234137611	NT_025615	Chr1	**Chr2** : 234076545 – 234364896
NT_003207	Chr6	**Chr22** : 14512703 – 14735210	NT_004233	Chr6	**Chr22** : 14681645 – 14927944
NT_004233	Chr6	**Chr22** : 14681645 – 14927944	NT_004170	Chr6	**Chr22** : 14896389 – 15302624
NT_003164	Chr8	**Chr22** : 30656456 – 30874219	NT_002627	Chr8	**Chr22** : 30656460 – 30921164
NT_026678	Chr8	**Chr22** : 32263472 – 32442880	NT_002628	Chr8	**Chr22** : 32314862 – 32402177
NT_026654	Chr14	**Chr3** : 69974000 – 70201345	NT_026652	Chr14	**Chr3** : 70200851 – 70383724
NT_026659	Chr18	**Chr5** : 158008330 – 158350642	NT_014989	Chr18	**Chr5** : 158128511 – 158526520
NT_014989	Chr18	**Chr5** : 158128511 – 158526520	NT_026660	Chr18	**Chr5** : 158429094 – 158942330
NT_026660	Chr18	**Chr5** : 158429094 – 158942330	NT_004235	Chr18	**Chr5** : 158521376 – 158746789



Acknowledgments

Biology

Poulabi Banerjee
Dario Boffelli
Kelly Frazer
Gaby Loots
Len Pennacchio
Eddy Rubin

Bioinformatics

Michael Brudno
Alexander Fabrikant
Jim Lord
Moshe Malkin
Chris Mayor
Ivan Ovcharenko
Lior Pachter
Alexander Poliakov
Jody Schwartz