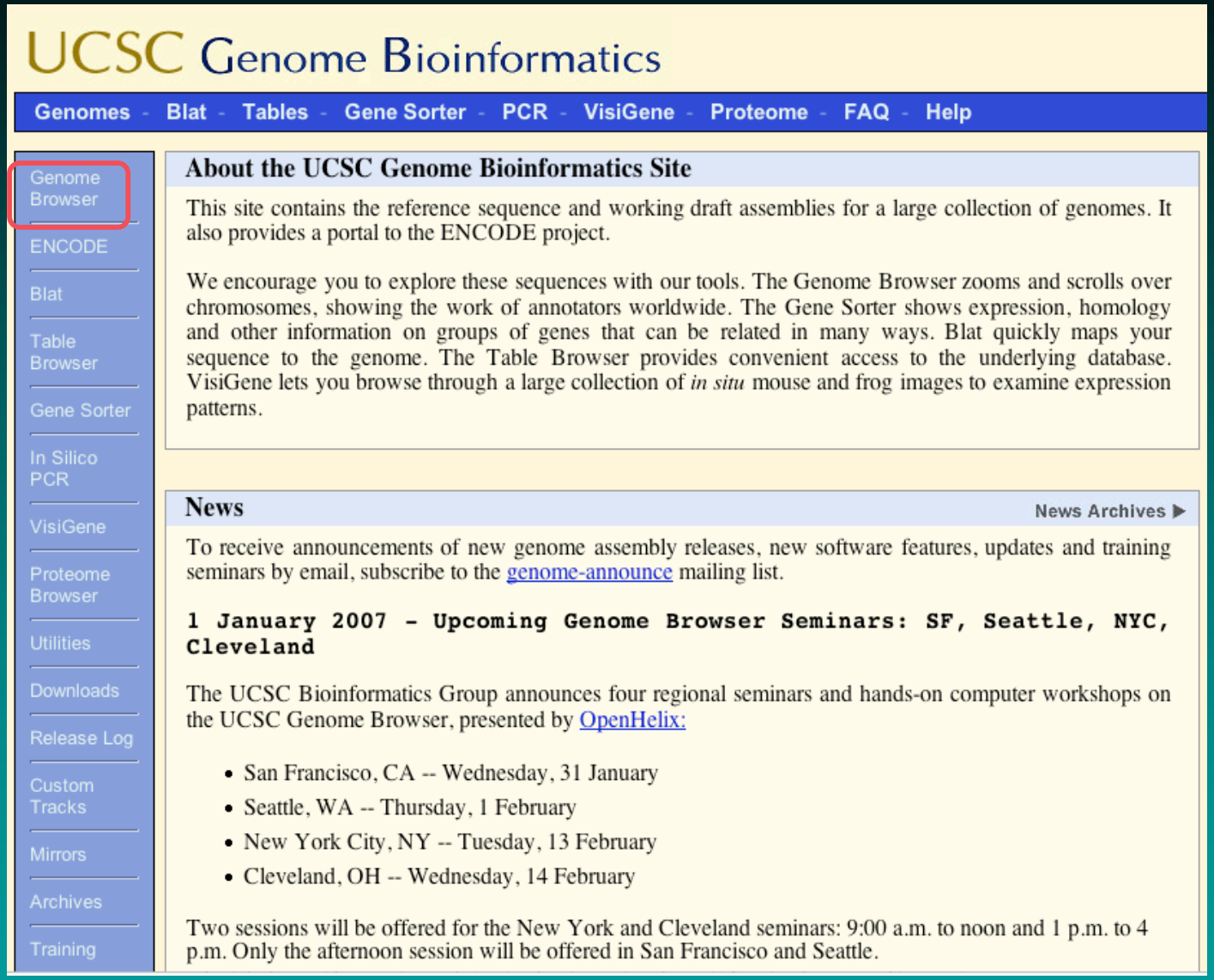
UCSC Genome Bioinformatics

Kate Rosenbloom

Center for Biomolecular Science and Engineering
University of California, Santa Cruz

GMOD User Interface Caucus

January 18, 2007



http://genome.ucsc.edu

The UCSC Genome Browser Presents Fully Annotated Genomes

Vertebrates

- human
- chimp
- rhesus macaque
- dog
- cow
- mouse
- rat
- opossum
- chicken
- tetraodon, fugu, zebrafish

Invertebrates

- sea squirt
- sea urchin
- fruitfly (12)
- honeybee
- mosquito
- worm (2)
- yeast

And coming soon…

- cat
- platypus
- medaka, stickleback

Hardware

- Public Site
  - -- 8 machines -- redundant
  - -- 64-bit
  - -- 8 Gb RAM
  - -- 1500 Gb storage
  - + 15 blat servers

- Under the hood

KiloKluster = 1000 CPUs

- -- Linux Red Hat 9, Apache, Parasol
- -- 10-Gigabit data transmission
- -- dual 866 MHz machines x 500
- -- 1 Gb RAM each

Smaller Clusters

- -- 100-node cluster: dual Xeon 2.6 GHz
- -- 400-node cluster

NFS

- -- 12 machines on RAID arrays
- -- 4 - 8 Gb RAM
- -- 20+ Tb storage

Data Contributors

- Human Genome Project
- Genbank/DDJ/EMBL contributors
- ENCODE Consortium
- Novartis GNF foundation
- Affymetrix, Perlegen, SNP Consortium
- SwissProt, Ensembl, EBI and NCBI
- Jackson Labs, RGD, Wormbase, Flybase
- Many contributors of gene prediction and other tracks.

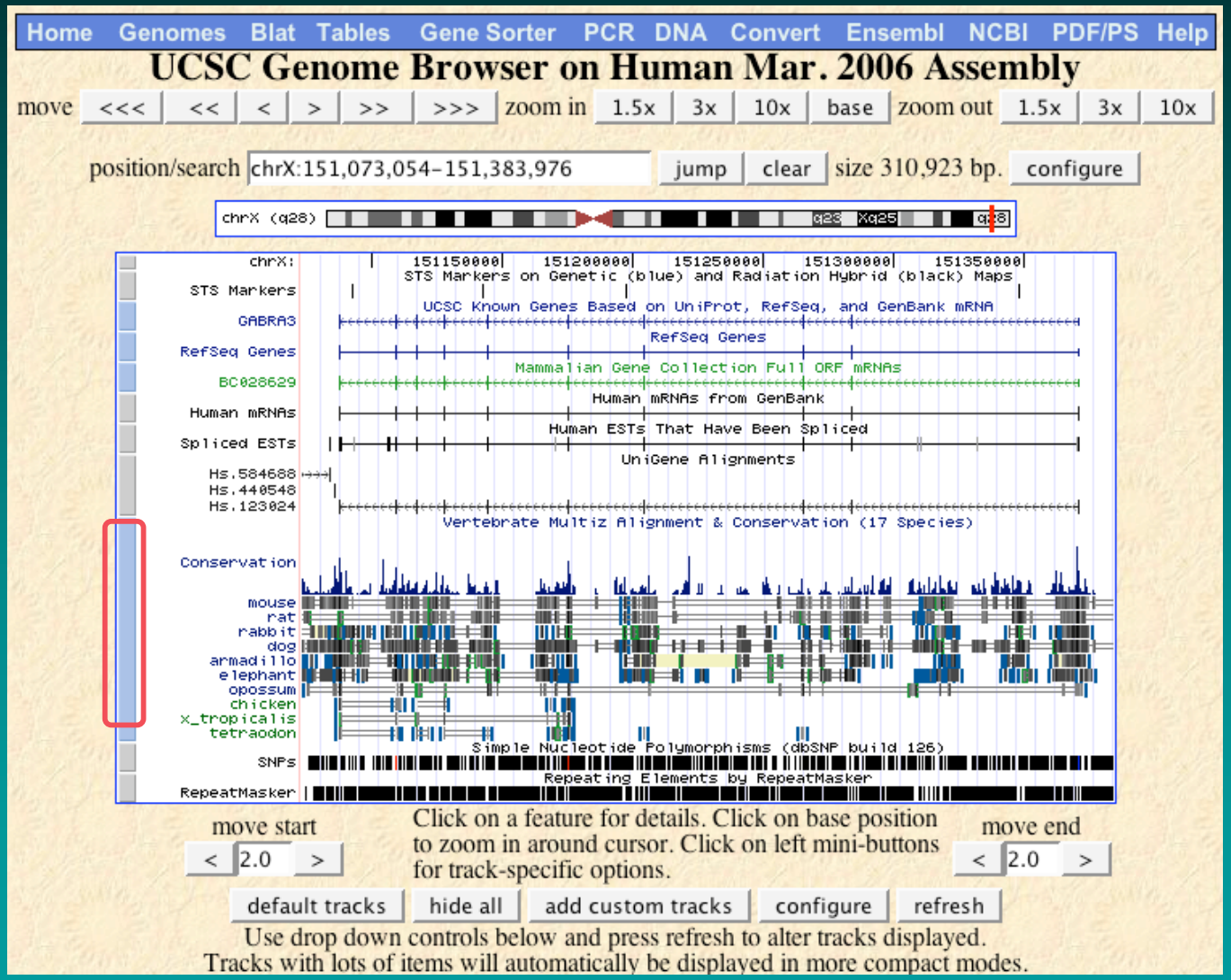
High volume data handling

- All Genbank mRNAs loaded and aligned to the genome nightly; all ESTs weekly (24-48 hours to process).
- At least 6000 - 7000 regular users (separate IP addresses daily).
- 2 - 3 million hits a week
- Consistently #1 or #2 user of bandwidth on the UCSC campus

UCSC Bioinformatics Tools

- Genome Browser
- Table Browser
- Gene Sorter
- VisiGene
- Custom Tracks
- BLAT
- Downloads server, DAS server, mySQL access

Genome Browser



Track configuration & description

Vertebrate Multiz Alignment & Conservation (17 Species)

Display mode: pack Submit

Pairwise alignments:

mammal

chimp rhesus mouse rat rabbit
dog cow armadillo elephant tenrec
opossum

vertebrate

chicken X. tropicalis zebrafish tetraodon fugu

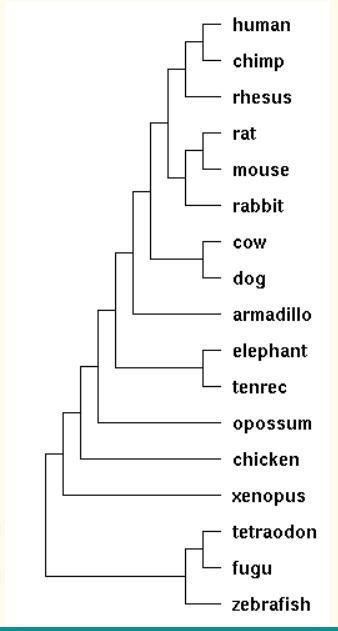
Multiple alignment base-level:

Display bases identical to reference as dots
Display chains between alignments

Codon Translation:

Default species to establish reading frame: hg18

No codon translation
Use default species reading frames for translation
Use reading frames for species if available, otherwise no translation
Use reading frames for species if available, otherwise use default species



Description

This track shows a measure of evolutionary conservation in 17 vertebrates, including mammalian, amphibian, bird, and fish species, based on a phylogenetic hidden Markov model, phastCons (Siepel *et al.*, 2005). Multiz alignments of the following assemblies were used to generate this track:

- human (Mar. 2006, hg18)
- chimp (Nov 2003, panTro1)
- macaque (Jan 2006, rheMac2)
- mouse (Feb 2006, mm8)

Codon translation uses the following gene tracks as the basis for translation, depending on the species chosen:

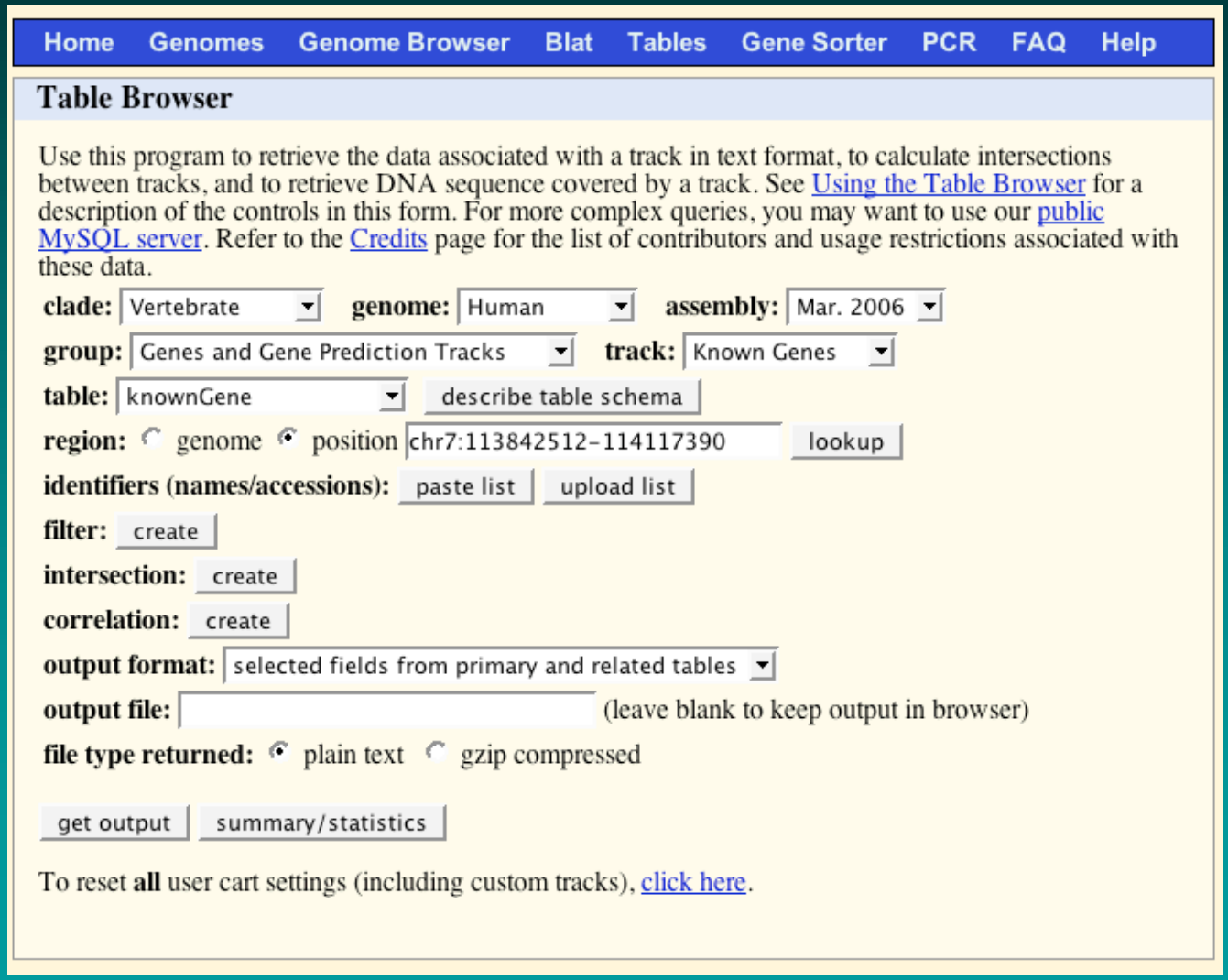
Gene Track	Species
Known Genes	human, mouse, rat
RefSeq Genes	chicken
MGC Genes	X. tropicalis
Ensembl Genes	Fugu, chimp
mRNAs	rhesus, rabbit, dog, cow, zebrafish
not translated	armadillo, elephant, tenrec, opossum, Tetraodon

Methods

Best-in-genome pairwise alignments were generated for each species using blastz, followed by chaining and netting. The pairwise alignments were then multiply aligned using multiz, following the ordering of the species tree diagrammed above. The resulting multiple alignments were then assigned conservation scores by phastCons, using a tree model with branch lengths derived from the ENCODE project Multi-Species Sequence Analysis group, September 2005 tree model. This tree was generated from TBA alignments over 23 vertebrate species and is based on 4D sites.

The phastCons program computes conservation scores based on a phylo-HMM, a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the way this process changes from one site to the next (Felsenstein and Churchill 1996, Yang 1995, Siepel and

Table Browser



Gene Sorter

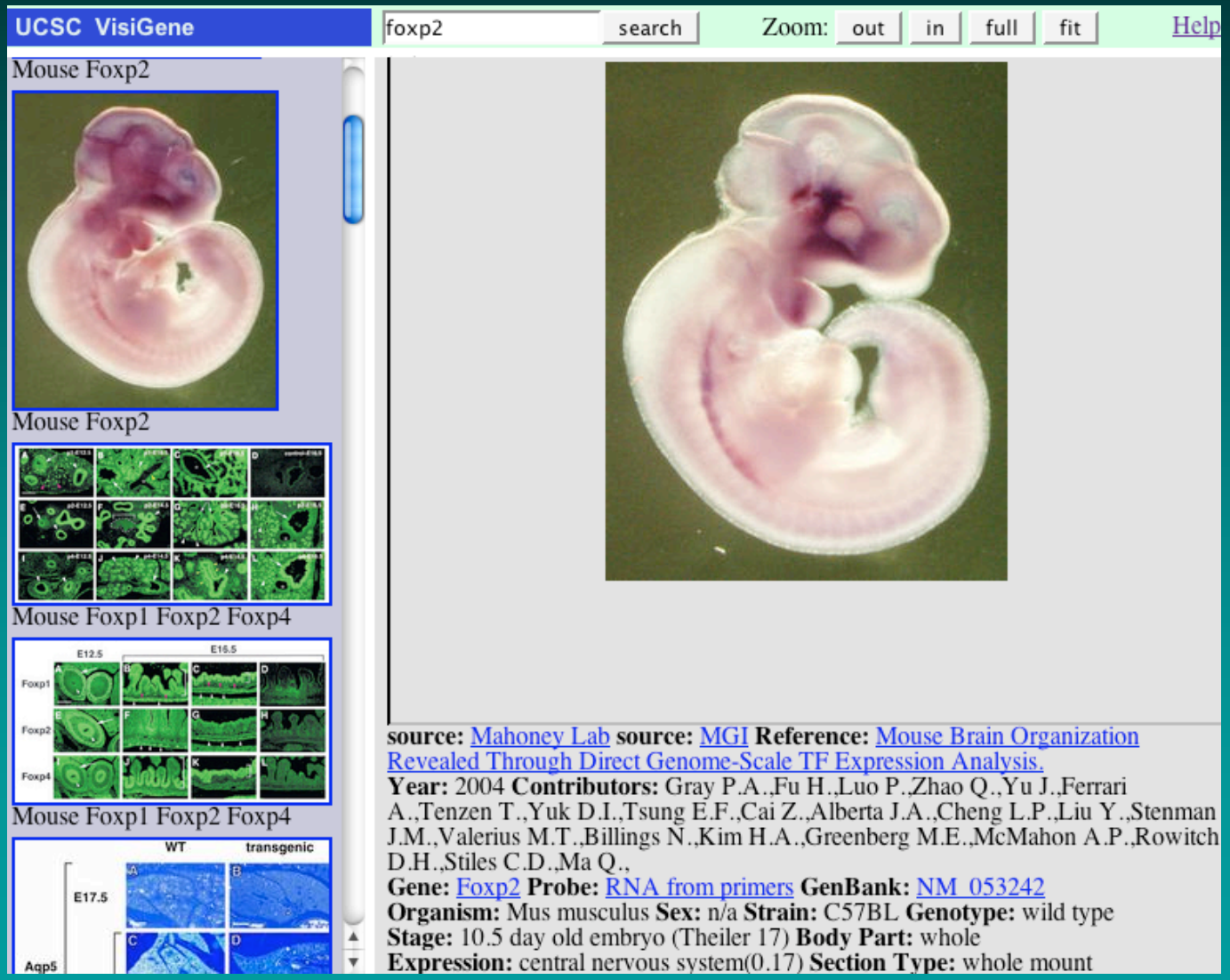
Home | UCSC Human Gene Sorter | Help

genome: Human | assembly: May 2004 | search: NM_007295 | Go!

sort by: Expression (GNF Atlas2) | configure | filter (now off) | display: 50 | output: sequence | text

#	Name	VisiGene	fetal brain	whole brain	amygdala	thymus	bone marrow	PB-CD4+ Tcells	skin	adipocyte	pancreatic islets	heart	lung	kidney	liver	ovary	testis	BLASTP E-Value	Rankprop Score	Genome Position	Description
1	BRCA1	1446																0	1	chr17 38,490,250	breast cancer 1, early onset isoform 1
2	ITGB3BP	97924																n/a	0.00234442	chr1 63,659,669	integrin beta 3 binding protein
3	RFC3	161999																n/a	0.210275	chr13 33,299,938	replication factor C 3 isoform 1
4	Y15758	n/a																n/a	n/a	chr13 51,938,044	LB1 protein.
5	RACGAP1	n/a																n/a	n/a	chr12 48,687,350	Rac GTPase activating protein 1
6	NUP107	n/a																n/a	n/a	chr12 67,394,868	nucleoporin 107kDa
7	MRPL39	99488																n/a	n/a	chr21 25,890,756	mitochondrial ribosomal protein L39 i
8	POLE2	100419																n/a	n/a	chr14 49,202,356	DNA polymerase epsilon subunit 2
9	RFC4	100881																n/a	0.21743	chr3 187,998,689	replication factor C 4
10	AF290612	102240																n/a	n/a	chr15 39,436,457	Nucleolar and spindle associated prote
11	DSCR2	83207																n/a	n/a	chr21 39,473,284	Down syndrome critical region protein
12	ZWINT	54581																n/a	n/a	chr10 57,789,122	ZW10 interactor isoform a
13	RBBP8	n/a																n/a	0.402232	chr18 18,814,521	retinoblastoma binding protein 8
14	CR749851	101368																n/a	n/a	chr1 47,459,826	TAL1 (SCL) interrupting locus (Fragr
15	GMNN	n/a																n/a	n/a	chr6 24,888,700	geminin
16	TRIP13	n/a																n/a	0.211022	chr5 958,581	thyroid hormone receptor interactor 13
17	KIF2C	n/a																n/a	n/a	chr1 44,888,575	kinesin family member 2C
18	KIF14	n/a																n/a	n/a	chr1 197,287,241	kinesin family member 14

Visigene (a "virtual microscope")



ENCODE Project at UCSC

Regions - Data Submission - Downloads - Tools - Terms - Help

- Regions (hg17)
- Regions (hg16)
- Data Status
- Downloads
- Submission
- Tools
- Release Log
- Contributors
- Terms of Use
- Genome Browser

About the ENCODE Project

This site contains information related to the ENCODE project at NHGRI. The UCSC Genome Bioinformatics Group manages the official repository of sequence-related data for the ENCODE consortium and supports the coordination of data submission, storage, retrieval, and visualization. A summary of the status of datasets submitted to UCSC by ENCODE contributors is available on the ENCODE data status page.

UCSC also has a special interest in comparative genomics, and we provide resources for the ENCODE multiple sequence alignment interest group. Ensembl also provides an ENCODE resource page.

We'd like to thank NHGRI and the contributors of annotations and analyses to this project. The team at UCSC that develops and maintains this ENCODE site is made up of Daryl Thomas, Kate Rosenbloom, Jim Kent, and the UCSC Genome Bioinformatics staff. Read more.

News

News Archives ▶

7 Oct. 2006 - Comparative Genomics Data Release

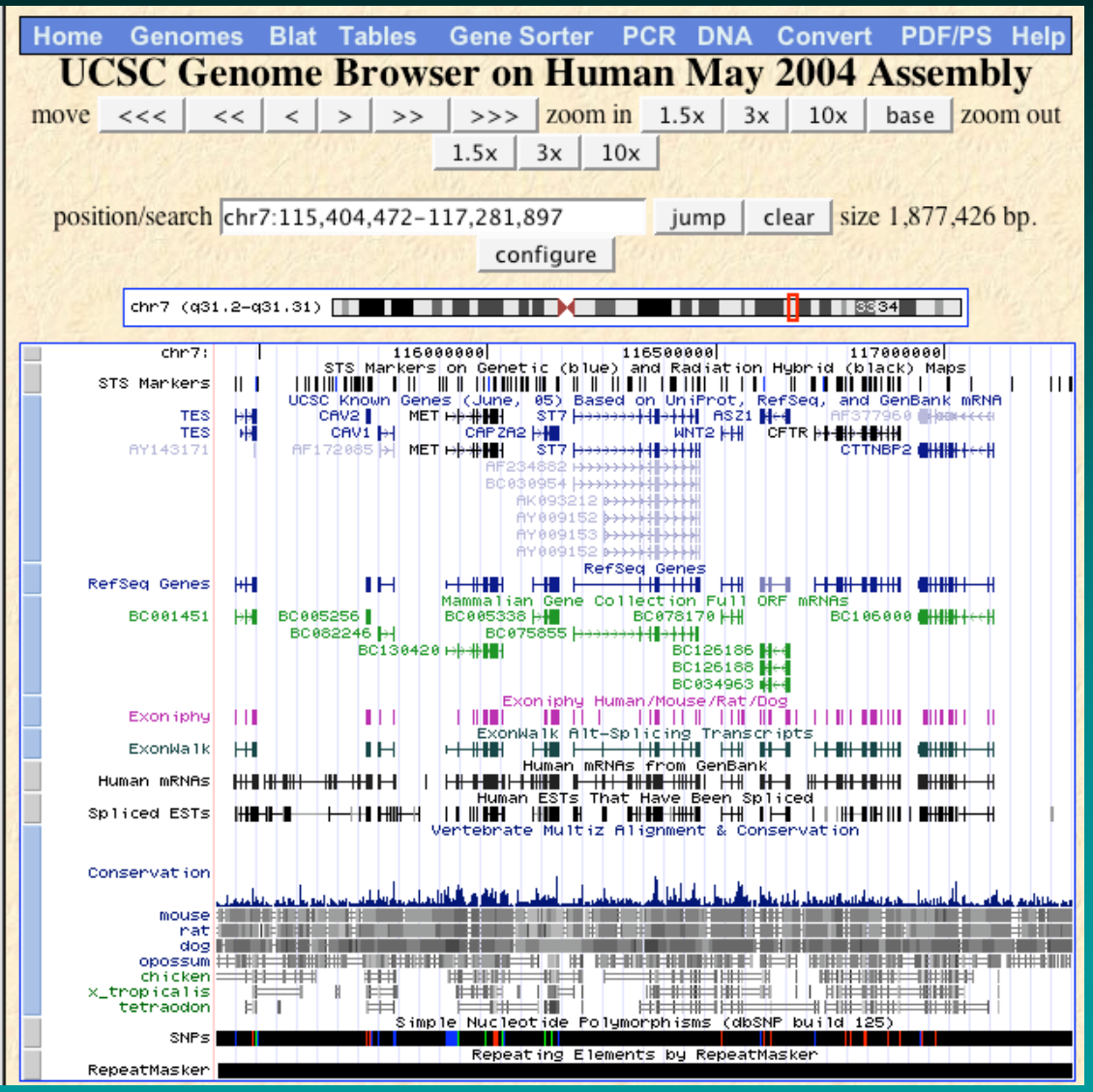
Twelve tracks of data produced by the ENCODE Multi-Species Sequence Analysis group have been released to the UCSC public server. These tracks contain multiple sequence alignments, conservation, and conserved (constrained) elements produced by four conservation methods (phastCons, binCons, GERP, SCONE) applied to three sequence alignments (TBA, MLAGAN, MAVID), and also an assessment of the agreement among the alignment methods. The alignments were based on genomic sequence in the ENCODE regions of 28 vertebrate species, as defined in the MSA September 2005 sequence freeze.

http://genome.ucsc.edu/ENCODE

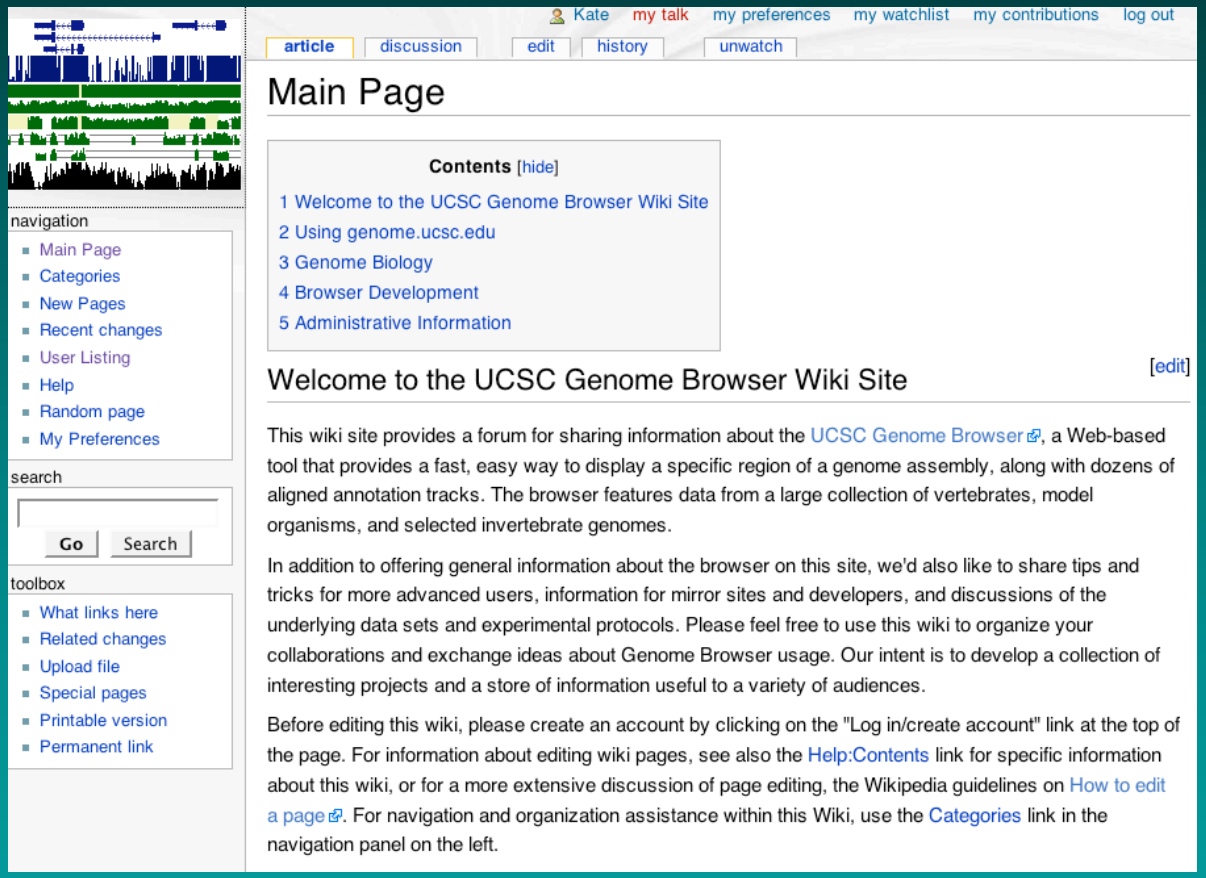
ENCODE Browser

hide Old (hg16)

Region	Description	Chr	Size (~Mb)
ENm001	CFTR	7	1.9
ENm002	Interleukin	5	1.0
ENm003	Apo Cluster	11	0.5
ENm004	Chr22 Pick	22	1.7
ENm005	Chr21 Pick	21	1.7
ENm006	ChrX Pick	X	1.2
ENm007	Chr19 Pick	19	1.0
ENm008	Alpha Globin	16	0.5
ENm009	Beta Globin	11	1.0
ENm010	HOXA Cluster	7	0.5
ENm011	1GF2/H19	11	0.6
ENm012	FOXP2	7	1.0
ENm013	Manual	7	1.1
ENm014	Manual	7	1.2
ENr111	Random	13	0.5
ENr112	Random	2	0.5
ENr113	Random	4	0.5
ENr114	Random	10	0.5
ENr121	Random	2	0.5
ENr122	Random	18	0.5
ENr123	Random	12	0.5
ENr131	Random	2	0.5

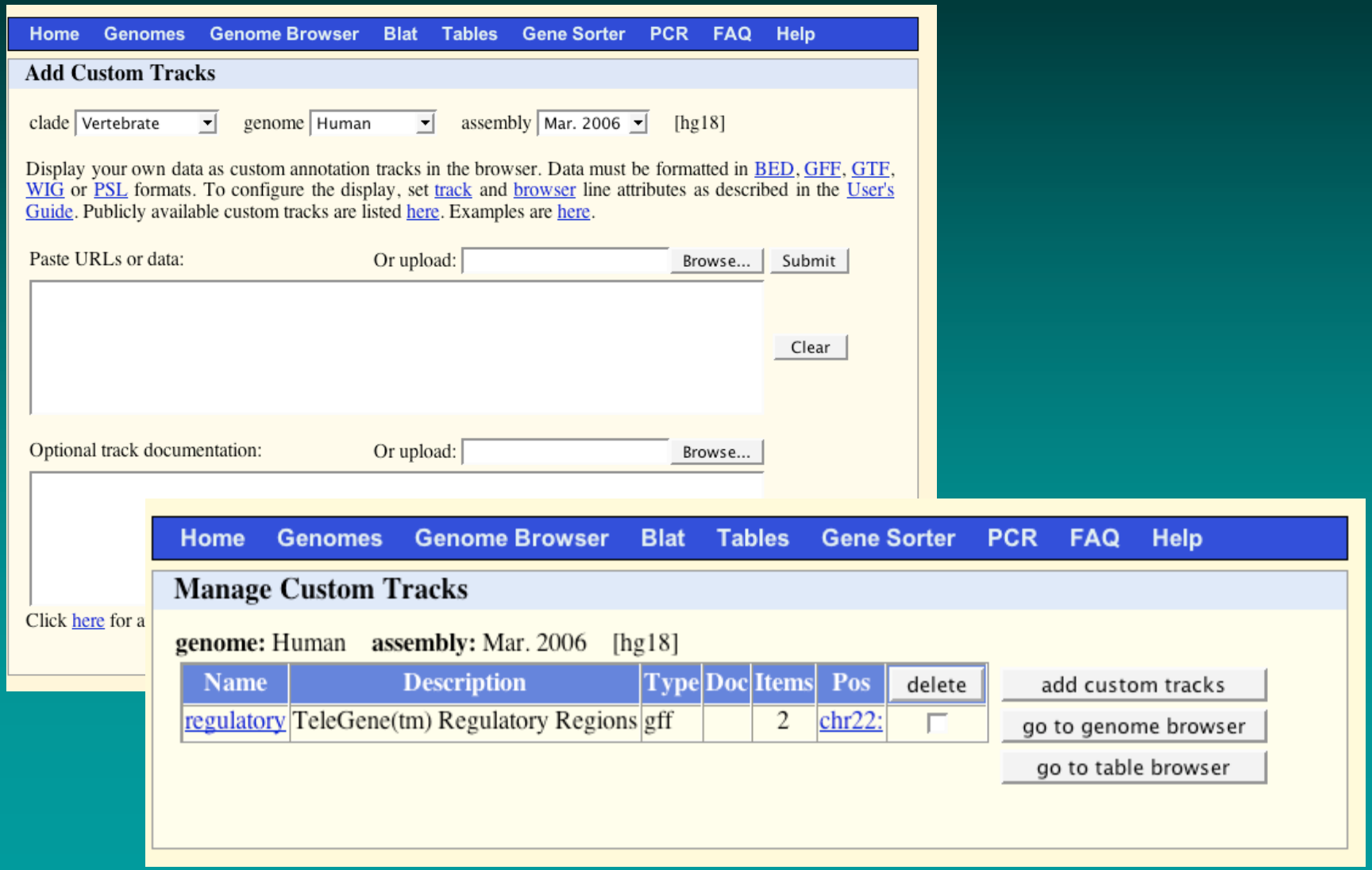


New features: Genomewiki

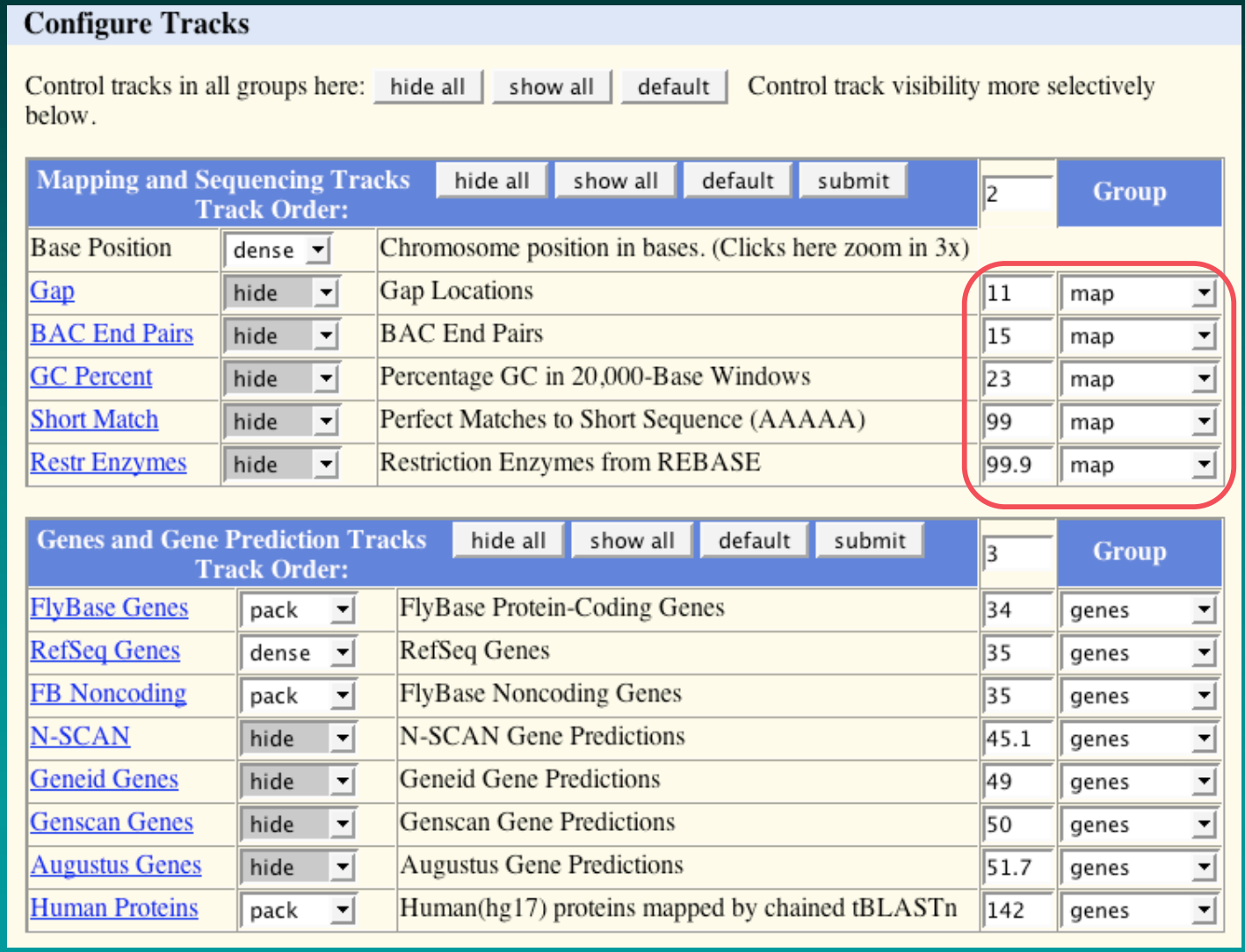


http://genomewiki.cse.ucsc.edu

New features: Custom track manager



New feature: Track reordering



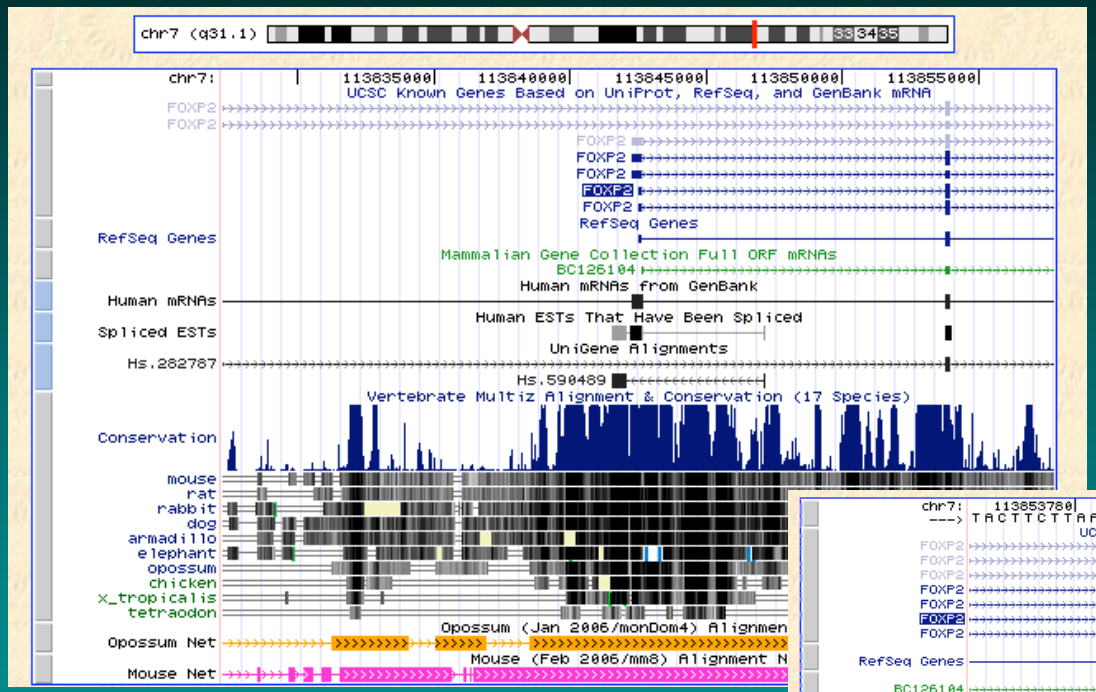
Configure Tracks

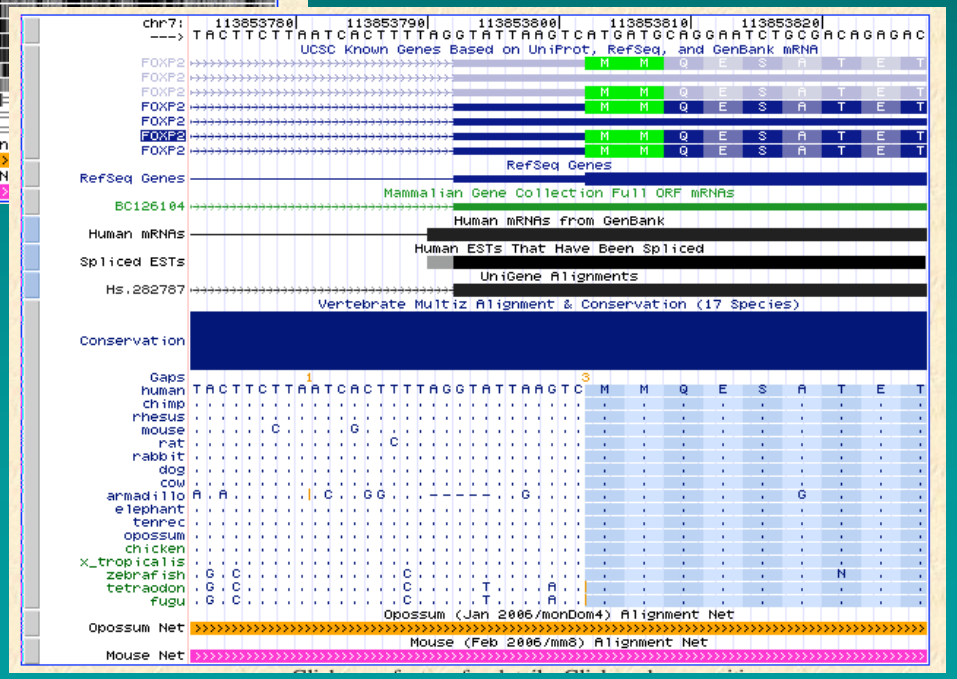
Control tracks in all groups here: hide all | show all | default Control track visibility more selectively below.

Mapping and Sequencing Tracks Track Order:	hide all	show all	default	submit	2	Group
Base Position	dense	Chromosome position in bases. (Clicks here zoom in 3x)				
Gap	hide	Gap Locations	11	map		
BAC End Pairs	hide	BAC End Pairs	15	map		
GC Percent	hide	Percentage GC in 20,000-Base Windows	23	map		
Short Match	hide	Perfect Matches to Short Sequence (AAAAA)	99	map		
Restr Enzymes	hide	Restriction Enzymes from REBASE	99.9	map		

Genes and Gene Prediction Tracks Track Order:	hide all	show all	default	submit	3	Group
FlyBase Genes	pack	FlyBase Protein-Coding Genes	34	genes		
RefSeq Genes	dense	RefSeq Genes	35	genes		
FB Noncoding	pack	FlyBase Noncoding Genes	35	genes		
N-SCAN	hide	N-SCAN Gene Predictions	45.1	genes		
Geneid Genes	hide	Geneid Gene Predictions	49	genes		
Genscan Genes	hide	Genscan Gene Predictions	50	genes		
Augustus Genes	hide	Augustus Gene Predictions	51.7	genes		
Human Proteins	pack	Human(hg17) proteins mapped by chained tBLASTn	142	genes		

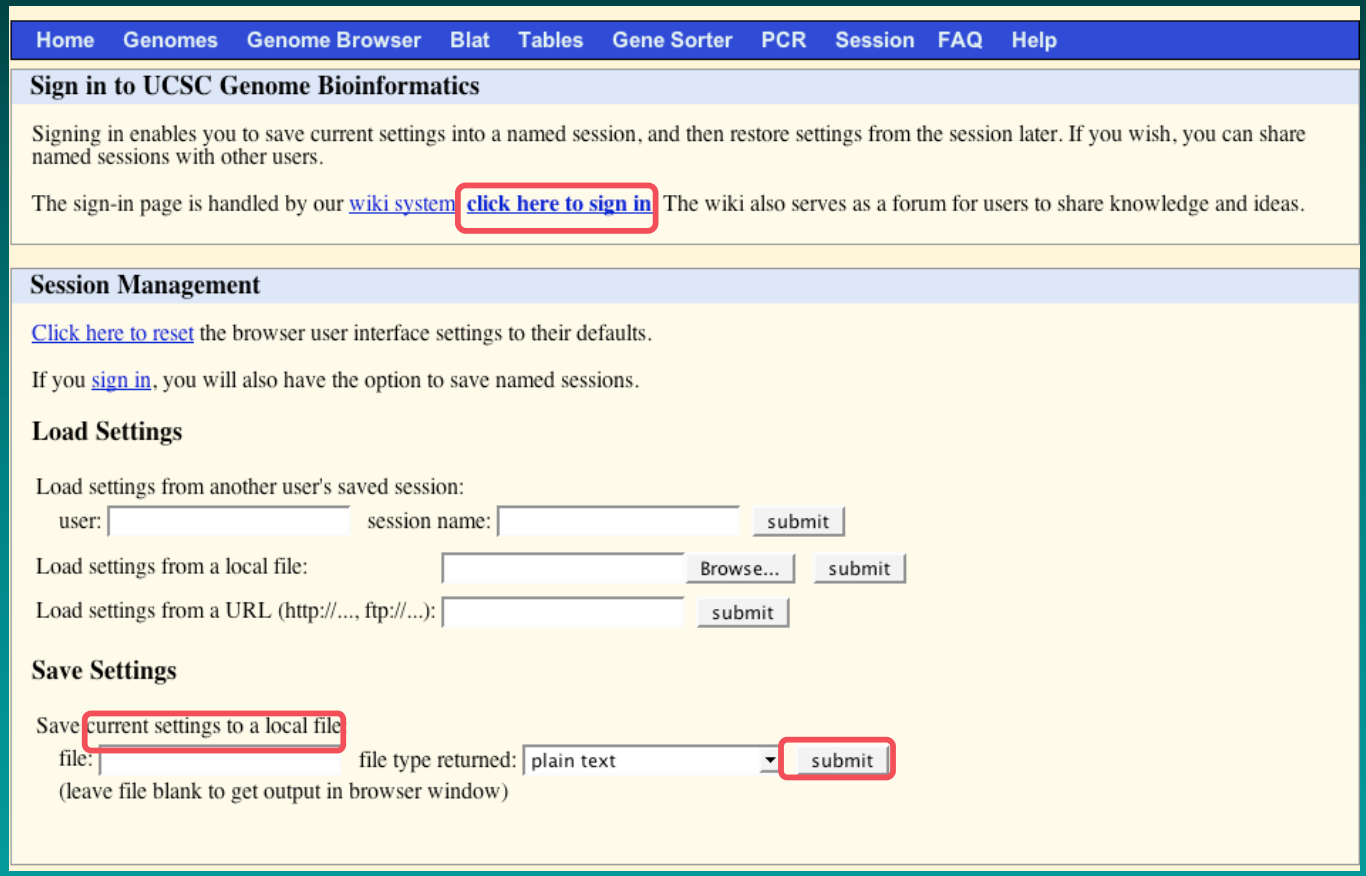
New features: Comparative genomics



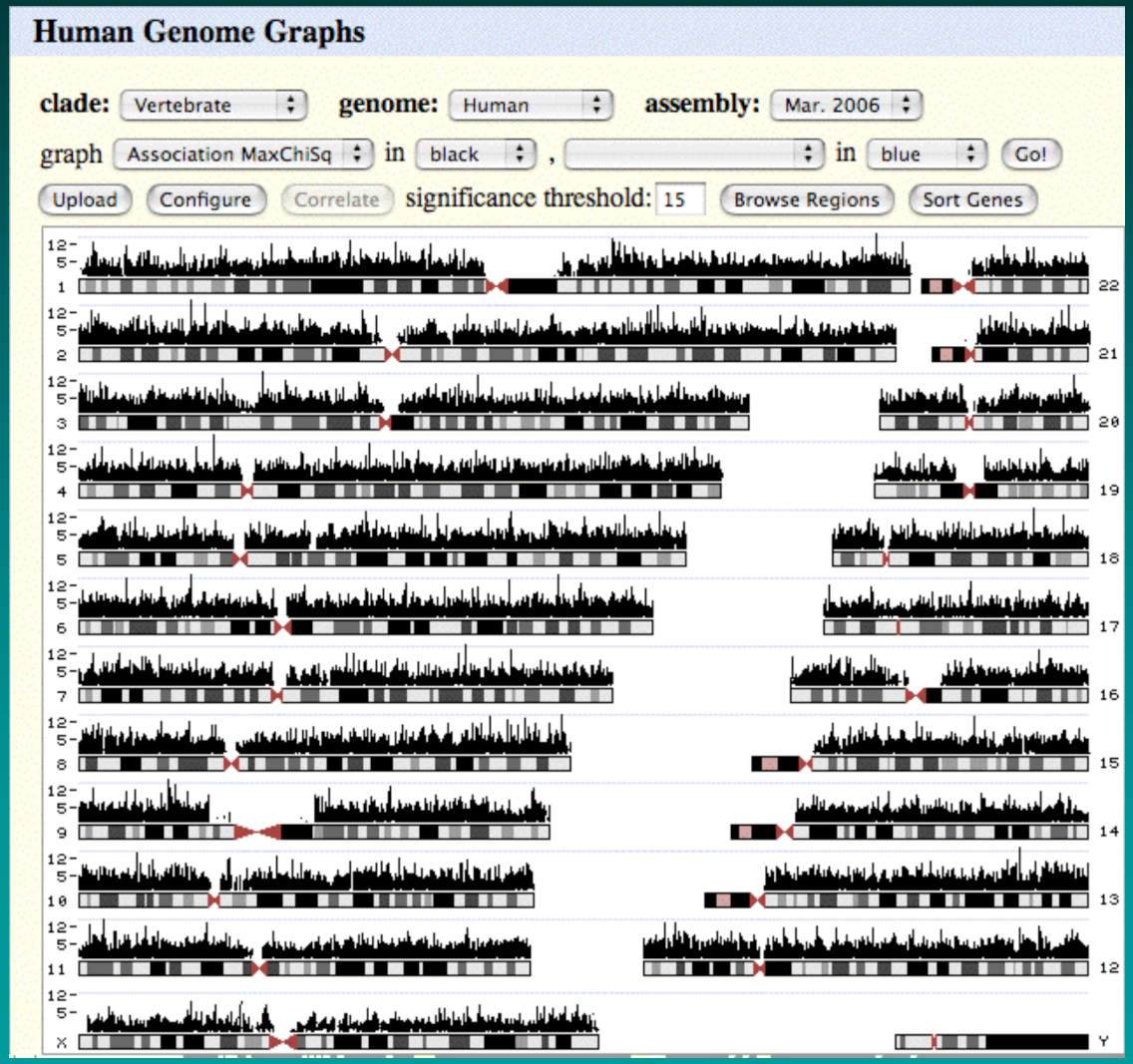


- Gap annotation
- Genomic breaks
- Codon translation at base level

New features (under review): Saving user sessions

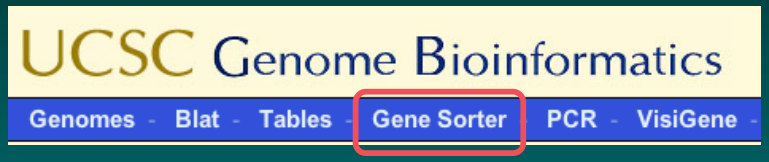


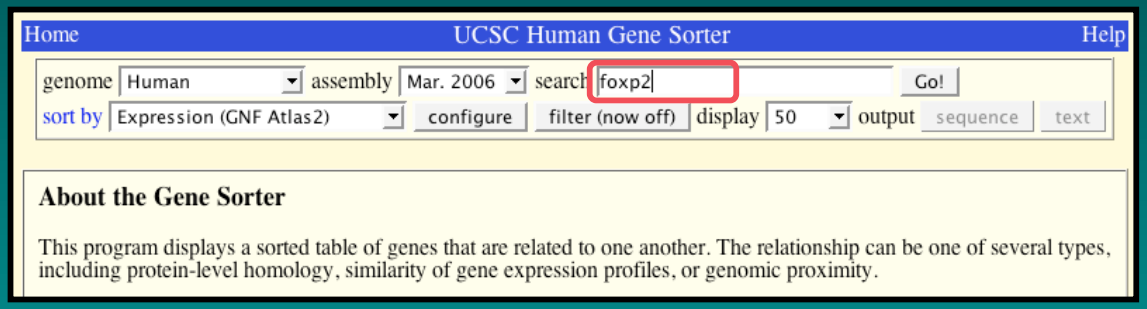
New features (in development): Whole genome graphing

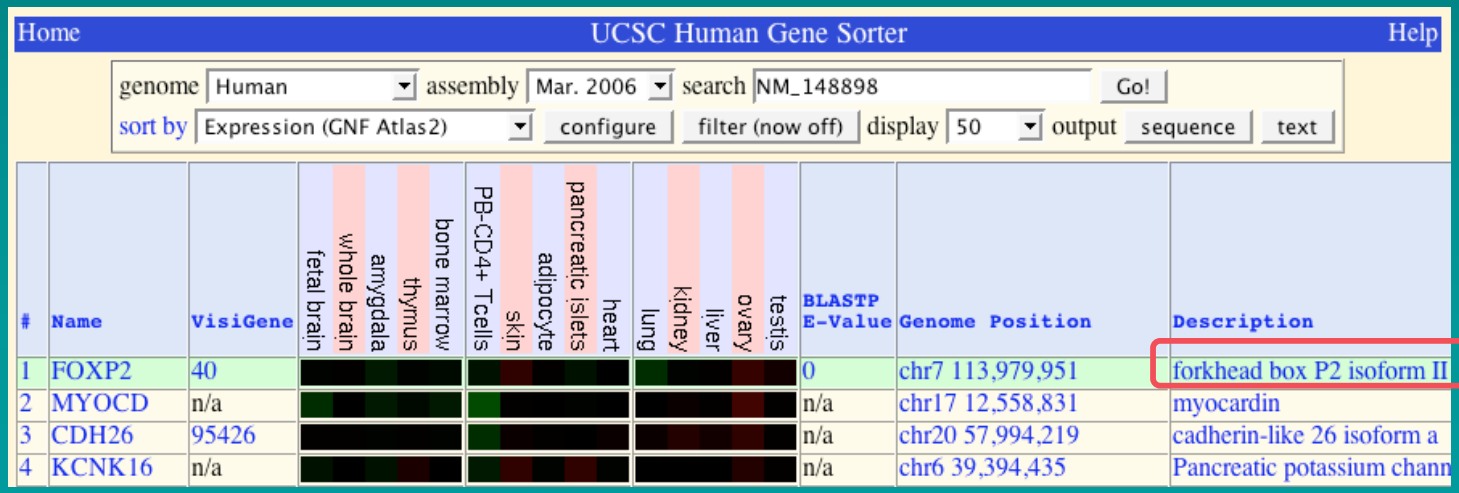


SNP association study, prepublication data

GMOD Scenario #1:
Search for gene by name…







GMOD Scenario #1:
… and view information page

Human Gene FOXP2 Description and Page Index

Description: forkhead box P2 isoform II
Alternate Gene Symbols: CAGH44, TNRC10
Representative Refseq: NM_148898 **Protein:** O15409 (aka FOXP2_HUMAN or FXP2_HUMAN)
RefSeq Summary: This gene encodes an evolutionarily conserved transcription factor expressed in fetal and adult brain. This transcription factor is a member of the forkhead/winged-helix (FOX) family of transcription factors, and contains a FOX DNA-binding domain and a large polyglutamine tract. Members of the FOX family of transcription factors are regulators of embryogenesis. The product of this gene is thought to be required for proper development of speech and language regions of the brain during embryogenesis. Although a point mutation in this gene has been associated with the KE pedigree segregating developmental verbal dyspraxia, no association between mutations in this gene and another speech disorder, autism, has been found. Four alternative transcripts encoding three different isoforms have been identified.
Position: chr7:113842512-114117390
Strand: +
Genomic Size: 274879
Exon Count: 18 **CDS Exon Count:** 17

Page Index	Quick Links	UniProt Comments	Sequence	Microarray	RNA Stru
Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Methods	

Quick Links to Tools and Databases

Gene Sorter	Genome Browser	Proteome Browser	Table Schema	VisiGene	Allen Brain Atlas
CGAP	Ensembl	Entrez Gene	ExonPrimer	GeneCards	GeneLynx
HGNC	HPRD	Jackson Labs	OMIM	PubMed	Stanford SOURCE
UniProt	Gepis Tissue				

Comments and Description Text from UniProt (Swiss-Prot/TrEMBL)

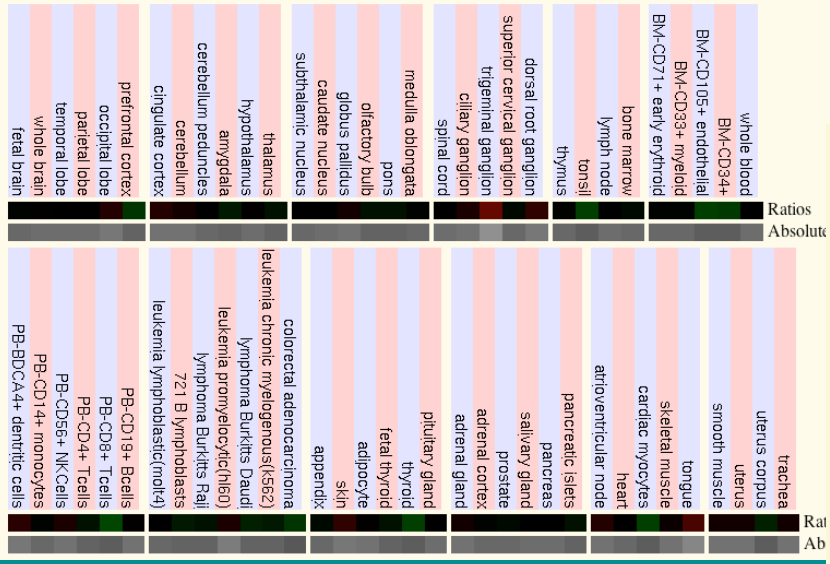
ID: FOXP2_HUMAN
DESCRIPTION: Forkhead box protein P2 (CAG repeat protein 44) (Trinucleotide repeat- containing gene 10 protein).
FUNCTION: Transcriptional repressor that may play a role in the specification and differentiation of lung epithelium. May also play a role in developing neural, gastrointestinal and cardiovascular tissues. Involved in neural mechanisms mediating the development of speech and language.
SUBUNIT: Forms homodimers and heterodimers with FOXP1 and FOXP4. Interacts with CTBP1 (By similarity).
SUBCELLULAR LOCATION: Nuclear (Probable).
TISSUE SPECIFICITY: Isoform 1 and isoform 6 are expressed in adult and fetal brain, caudate nucleus and lung.
DEVELOPMENTAL STAGE: Expressed in the brain at 15 and 22 weeks of gestation, with a pattern of strong cortical, basal ganglia, thalamic and cerebellar expression. Highly expressed in the head and tail of nucleus caudatus and putamen. Restricted expression within the globus pallidus, with high levels in the pars interna, which provides the principal source of output from the basal ganglia to the nucleus centrum medianum thalami (CM) and the major motor relay nuclei of the thalamus. In the thalamus, present in the CM and nucleus medialis dorsalis thalami. Lower levels are observed in the nuclei anterior thalami, dorsal and ventral, and the nucleus parafascicularis thalami. Expressed in the ventrobasal complex comprising the nucleus ventralis posterior lateralis/medialis. The ventral tier of the thalamus exhibits strong expression, including nuclei ventralis anterior, lateralis and posterior lateralis pars oralis. Also expressed in the nucleus subthalamicus bilaterally and in the nucleus ruber.
DISEASE: Defects in FOXP2 are the cause of speech-language disorder 1 (SPCH1) [MIM:602081]; also known as autosomal dominant speech and language disorder with orofacial dyspraxia. Affected individuals have a severe impairment in the selection and sequencing of fine orofacial movements, which are necessary for articulation. They also show deficits in several facets of language processing (such as the ability to break up words into their constituent phonemes) and grammatical skills.
DISEASE: A chromosomal aberration disrupting FOXP2 is a cause of severe speech and language impairment. Translocation t(5;7)(q22;q31.2).
SIMILARITY: Contains 1 C2H2-type zinc finger.
SIMILARITY: Contains 1 fork-head DNA-binding domain.
DATABASE: NAME=Protein Spotlight; NOTE=Issue 51 of October 2004; WWW="http://www.expasy.org/spotlight/back_issues/sptlt051.shtml".

GMOD Scenario #1:
… and view information page (2)

Microarray Expression Data

Expression ratio colors: red high/green low Submit

GNF Expression Atlas 2 Data from U133A and GNF1H Chips



mRNA Secondary Structure of 3' and 5' UTRs

Region	Fold Energy	Bases	Energy/Base	Display As		
5' UTR	-25.70	150	-0.171	Picture	PostScript	Text
3' UTR	-33.29	173	-0.192	Picture	PostScript	Text

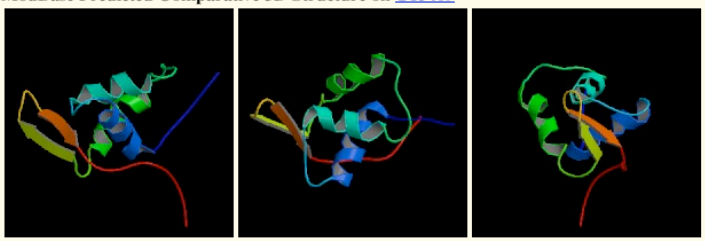
The RNAfold program from the Vienna RNA Package is used to perform the secondary structure predictions and folding calculations. The estimated folding energy is in kcal/mol. The more negative the energy, the more secondary structure the RNA is likely to have.

Protein Domain and Structure Information

InterPro Domains: Graphical view of domain structure
IPR001766 - Fork head transcription factor
IPR011991 - Winged helix repressor DNA-binding
IPR007087 - Zinc finger, C2H2-type

Pfam Domains:
PF00250 - Fork head domain

ModBase Predicted Comparative 3D Structure on O15409



Front Top Side

GMOD Scenario #1:
… and view information page (3)

Homologous Genes in Other Species (BLASTP Best Hit)

Mouse	Rat	Zebrafish	D. melanogaster	C. elegans	S. cerevisiae
Genome Browser	Genome Browser	Genome Browser	Genome Browser	Genome Browser	Genome Browser
Gene Details	Gene Details		Gene Details	Gene Details	Gene Details
Gene Sorter	Gene Sorter		Gene Sorter	Gene Sorter	Gene Sorter
Jackson Lab	RGD	Ensembl	FlyBase	WormBase	SGD
Protein Sequence	Protein Sequence		Protein Sequence	Protein Sequence	Protein Sequence
Alignment	Alignment		Alignment	Alignment	Alignment

Gene Ontology (GO) Annotations with Structured Vocabulary

Molecular Function:
GO:0003676 nucleic acid binding
GO:0003677 DNA binding
GO:0003700 transcription factor activity
GO:0008270 zinc ion binding
GO:0043565 sequence-specific DNA binding
GO:0046872 metal ion binding

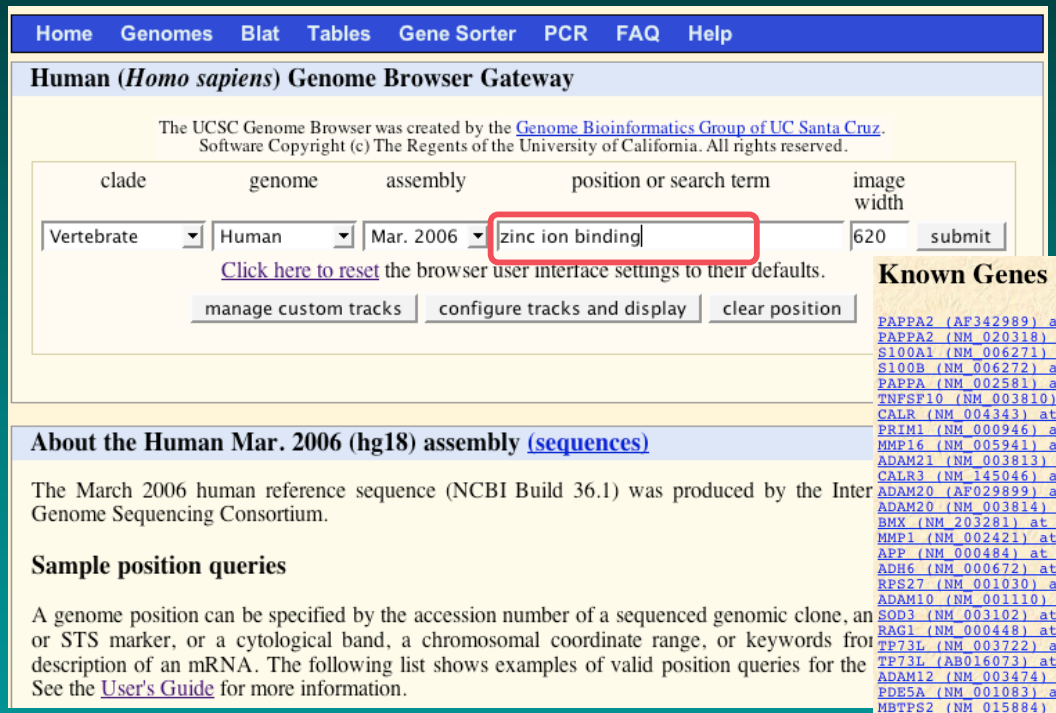
Biological Process:
GO:0006350 transcription
GO:0006355 regulation of transcription, DNA-dependent

Cellular Component:
GO:0005634 nucleus

Descriptions from all associated GenBank mRNAs

AK131266 - Homo sapiens cDNA FLJ16201 fis, clone CTONG2008721, highly similar to Homo sapiens CAGH44 mRNA.
AF454830 - Homo sapiens forkhead transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
CR749236 - Homo sapiens mRNA; cDNA DKFZp686H1726 (from clone DKFZp686H1726).
AF467252 - Homo sapiens clone FCA1 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
BC018016 - Homo sapiens forkhead box P2, mRNA (cDNA clone IMAGE:4285527), complete cds.
AY144615 - Homo sapiens brain forkhead/winged helix transcription factor FOXP2 isoform mRNA, complete cds; alternatively spliced.
AF493430 - Homo sapiens FOXP2 short isoform (FOXP2) mRNA, complete cds.
AF337817 - Homo sapiens putative forkhead/winged-helix transcription factor (FOXP2) mRNA, complete cds.
AF467257 - Homo sapiens clone HF2B2 forkhead/winged helix transcription factor (FOXP2) mRNA, complete cds; alternatively spliced.
AF467253 - Homo sapiens clone P22 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
U80741 - Homo sapiens CAGH44 mRNA, partial cds.
AF467258 - Homo sapiens clone AMYG 2a.2 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
AF467259 - Homo sapiens clone AMYG 4a.4 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
BC126104 - Homo sapiens forkhead box P2, mRNA (cDNA clone MGC:161382 IMAGE:8991820), complete cds.
AF467255 - Homo sapiens clone 3RACE700 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
AF467256 - Homo sapiens clone BA4 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
AF467254 - Homo sapiens clone STR7 forkhead/winged helix transcription factor (FOXP2) mRNA, partial cds; alternatively spliced.
AF086040 - Homo sapiens full length insert cDNA clone YX52E07.
DQ778626 - Homo sapiens forkhead box P2 variant 3 mRNA, complete cds, alternatively spliced.

GMOD Scenario #2 (sort of): Search by keyword



Known Genes

PAPPA2 (AF342989) at chr1:174698930-175078591 - pappalysin 2
PAPPA2 (NM_020318) at chr1:174698930-175078591 - pappalysin 2 isoform 1
S100A1 (NM_006271) at chr1:151867497-151871137 - S100 calcium binding protein A1
S100B (NM_006272) at chr21:46842959-46849424 - S100 calcium-binding protein, beta
PAPPA (NM_002581) at chr9:117955892-118204420 - pregnancy-associated plasma protein A
TNFSF10 (NM_003810) at chr3:173706159-173723963 - tumor necrosis factor (ligand) superfamily,
CALR (NM_004343) at chr19:12910423-12916303 - calreticulin precursor
PRIM1 (NM_000946) at chr12:55411637-55432413 - DNA primase small subunit, 49kDa
MMP16 (NM_005941) at chr8:89118578-89408825 - matrix metalloproteinase 16 isoform 1
ADAM21 (NM_003813) at chr14:69993970-69996374 - ADAM metallopeptidase domain 21 preproprotein
CALR3 (NM_145046) at chr19:16450888-16468003 - calreticulin 3
ADAM20 (AF029899) at chr14:70058832-70061255 - ADAM metallopeptidase domain 20
ADAM20 (NM_003814) at chr14:70058832-70071485 - ADAM metallopeptidase domain 20 preproprotein
BMX (NM_203281) at chrX:15428873-15484572 - BMX non-receptor tyrosine kinase
MMP1 (NM_002421) at chr11:102165861-102174104 - matrix metalloproteinase 1 preproprotein
APP (NM_000484) at chr21:26174733-26465003 - amyloid beta A4 protein precursor, isoform a
ADH6 (NM_000672) at chr4:100344903-100359424 - class V alcohol dehydrogenase 6
RPS27 (NM_001030) at chr1:152229853-152231247 - ribosomal protein S27
ADAM10 (NM_001110) at chr15:56675802-56829469 - ADAM metallopeptidase domain 10
SOD3 (NM_003102) at chr4:24405153-24411561 - superoxide dismutase 3, extracellular
RAG1 (NM_000448) at chr11:36546139-36557871 - recombination activating gene 1
TP73L (NM_003722) at chr3:190831910-191097756 - tumor protein p73-like
TP73L (AB016073) at chr3:190831971-191095181 - tumor protein p73-like
ADAM12 (NM_003474) at chr10:127693415-128067055 - ADAM metallopeptidase domain 12 isoform 1
PDE5A (NM_001083) at chr4:120634999-120769429 - phosphodiesterase 5A isoform 1
MBTPS2 (NM_015884) at chrX:21767675-21810794 - membrane-bound transcription factor protease,
TP53 (NM_000546) at chr17:7512465-7531642 - tumor protein p53
S100A3 (NM_002960) at chr1:151786433-151788358 - S100 calcium binding protein A3
STAMBP (NM_213622) at chr2:73909655-73943518 - STAM binding protein
RIMS1 (NM_014989) at chr6:72653448-73167137 - regulating synaptic membrane exocytosis 1
TP73 (NM_005427) at chr1:3558989-3639716 - tumor protein p73
EWSR1 (NM_005243) at chr22:27994283-28026505 - Ewing sarcoma breakpoint region 1 isoform EWS

Non-Human RefSeq Genes

AT1G22440 at chr6:66603562-66604647 - (NM_102093) oxidoreductase/ zinc ion binding
AT4G22110 at chr6:66603859-66604647 - (NM_118332) oxidoreductase/ zinc ion binding
AT4G22110 at chr6:66603859-66604647 - (NM_202860) oxidoreductase/ zinc ion binding
AT5G24760 at chr1:79759536-79760470 - (NM_203102) oxidoreductase/ zinc ion binding
AT5G24760 at chr1:79759536-79760470 - (NM_122385) oxidoreductase/ zinc ion binding
AT5G42250 at chr6:66603952-66604491 - (NM_123590) oxidoreductase/ zinc ion binding
AT2G32600 at chr19:2194439-2196550 - (NM_128819) nucleic acid binding / zinc ion binding
APC11 at chr17:77445680-77451230 - (NM_111461) APC11; protein binding / zinc ion binding
AT1G32780 at chr6:66603790-66604488 - (NM_103012) oxidoreductase/ zinc ion binding
AT1G32780 at chr4:100216469-100568747 - (NM_103012) oxidoreductase/ zinc ion binding

GMOD Scenario #3:
Customized report on aspects of gene

Home — Gene Sorter Filter

On this page you can restrict which genes appear in the main table based on the values in any column. Click the *submit* button to return to the main Gene Sorter page with the current filter settings applied.

submit | clear filter | save filter | load filter

Quickly obtain a list of gene names that pass the filter: list names

Filter Controls for Displayed Columns:

Name - Gene Name/Select Gene

Name search (including * and ? wildcards): foxp2

Include if any words in search term match.

Limit to items (no wildcards) in list: paste list | upload list

Home — Configure Gene Sorter — Help

submit | Columns: hide all | show all | default | Settings: save | load

Expression ratio colors: red high/green low | Show all splicing variants: | custom columns

Name	On	Position	Description	Configuration
#	☐	▼	Item Number in Displayed List/Select Gene	n/a
Name	☑	▲▼	Gene Name/Select Gene	n/a
UniProt	☐	▲▼	UniProt (SwissProt/TrEMBL) Protein Display ID	n/a
UniProt Acc	☐	▲▼	UniProt (SwissProt/TrEMBL) Protein Accession	n/a
PDB	☐	▲▼	Protein Data Bank	n/a
Gene Ontology	☑	▲▼	Gene Ontology (GO) Terms Associated with Gene	n/a
M. Vidal P2P	☐	▲▼	Human Protein-Protein Interaction Network from Marc Vidal	n/a
E. Wanker P2P	☐	▲▼	Human Protein-Protein Interaction Network from Erich Wanker	n/a
HPRD P2P	☐	▲▼	Human Protein-Protein Interaction Network from the Human Reference Protein Database	n/a
Description	☑	▲	Short Description Line/Link to Details Page	n/a

- Exon count
- GO terms
- Description

Home — UCSC Human Gene Sorter — Help

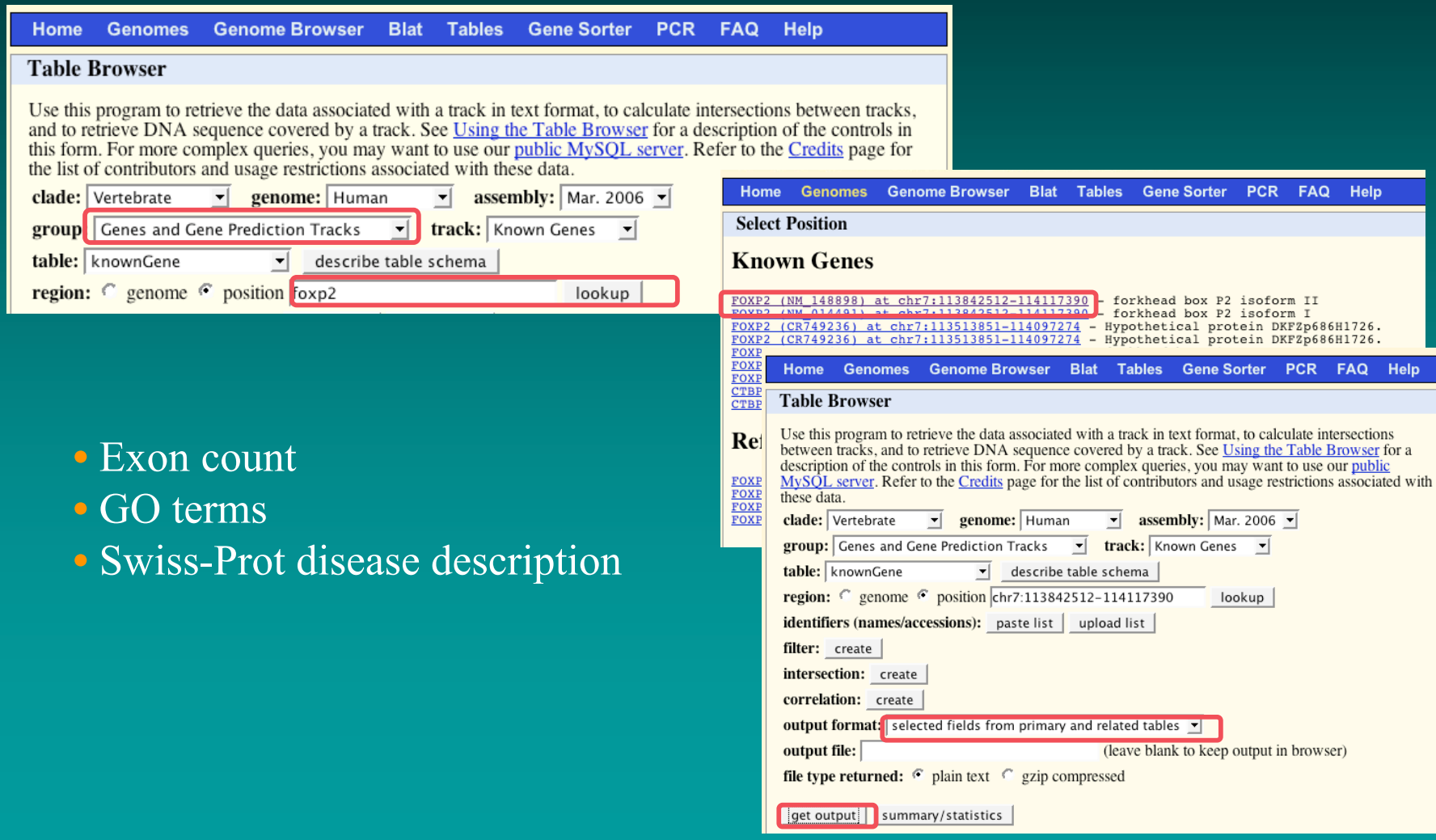
genome Human | assembly Mar. 2006 | search NM_148898 | Go!

sort by Expression (GNF Atlas2) | configure | filter (now on) | display 50 | output sequence | text

#	Name	Exon Count	Gene Ontology
1	FOXP2	18	'nucleic acid binding' 'DNA binding' 'transcription factor activity' 'nucleus' 'transcription' 'regulation of transcription, DNA-dependent'

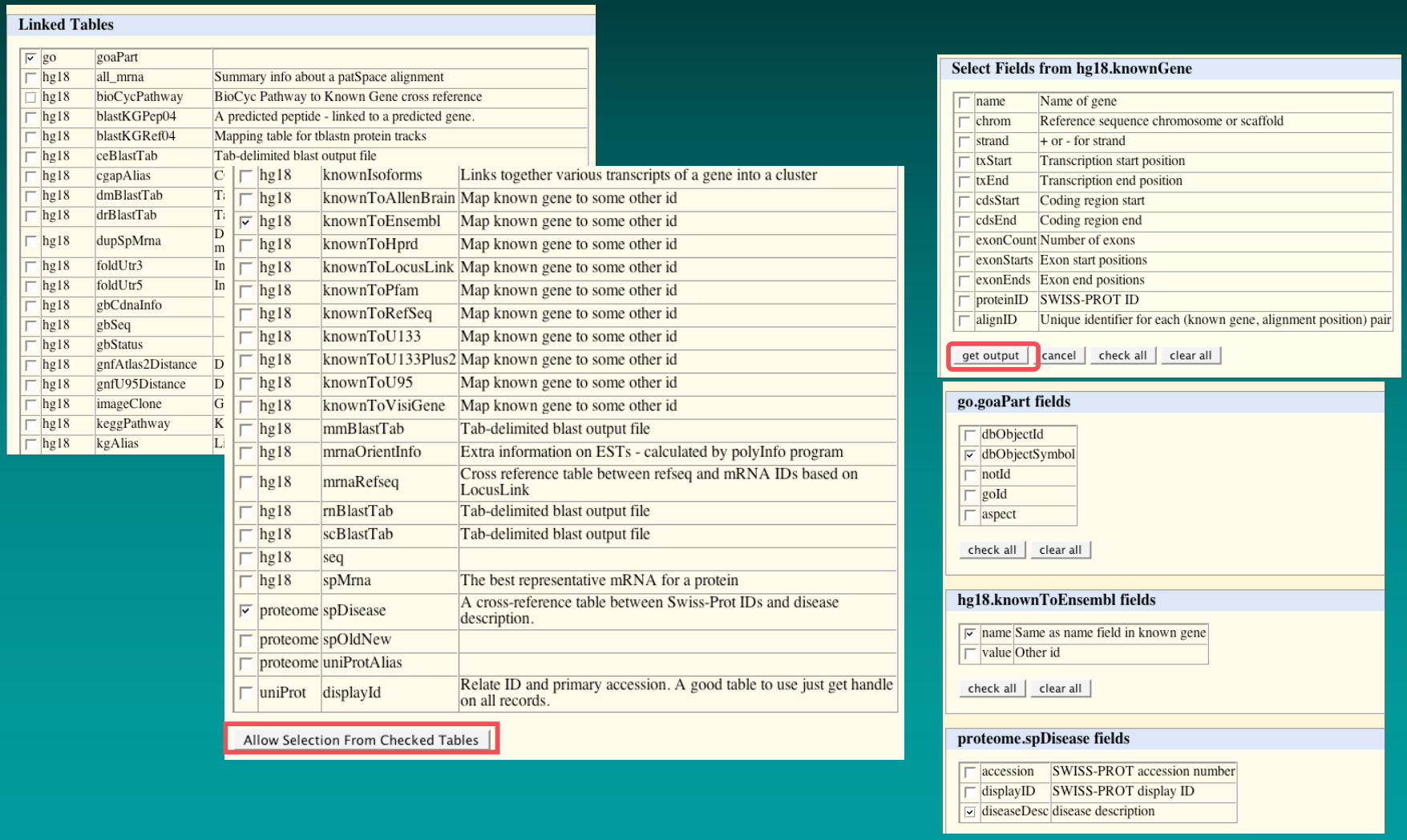
```
#num    name    exonCount       go      description
1       FOXP2   18      'nucleic acid binding','DNA binding','transcription factor activity','nucleus','transcription','regulation of transcription, DNA-dependent',
```

GMOD Scenario #3 Alternate: Customized report on aspects of gene



- Exon count
- GO terms
- Swiss-Prot disease description

GMOD Scenario #3:
Customized report on gene, cont.



Linked Tables

☑	go	goaPart	
☐	hg18	all_mrna	Summary info about a patSpace alignment
☐	hg18	bioCycPathway	BioCyc Pathway to Known Gene cross reference
☐	hg18	blastKGPep04	A predicted peptide - linked to a predicted gene.
☐	hg18	blastKGRef04	Mapping table for tblastn protein tracks
☐	hg18	ceBlastTab	Tab-delimited blast output file
☐	hg18	cgapAlias	C
☐	hg18	dmBlastTab	T
☐	hg18	drBlastTab	T
☐	hg18	dupSpMrna	D m
☐	hg18	foldUtr3	In
☐	hg18	foldUtr5	In
☐	hg18	gbCdnaInfo	
☐	hg18	gbSeq	
☐	hg18	gbStatus	
☐	hg18	gnfAtlas2Distance	D
☐	hg18	gnfU95Distance	D
☐	hg18	imageClone	G
☐	hg18	keggPathway	K
☐	hg18	kgAlias	L
☐	hg18	knownIsoforms	Links together various transcripts of a gene into a cluster
☐	hg18	knownToAllenBrain	Map known gene to some other id
☑	hg18	knownToEnsembl	Map known gene to some other id
☐	hg18	knownToHprd	Map known gene to some other id
☐	hg18	knownToLocusLink	Map known gene to some other id
☐	hg18	knownToPfam	Map known gene to some other id
☐	hg18	knownToRefSeq	Map known gene to some other id
☐	hg18	knownToU133	Map known gene to some other id
☐	hg18	knownToU133Plus2	Map known gene to some other id
☐	hg18	knownToU95	Map known gene to some other id
☐	hg18	knownToVisiGene	Map known gene to some other id
☐	hg18	mmBlastTab	Tab-delimited blast output file
☐	hg18	mrnaOrientInfo	Extra information on ESTs - calculated by polyInfo program
☐	hg18	mrnaRefseq	Cross reference table between refseq and mRNA IDs based on LocusLink
☐	hg18	rnBlastTab	Tab-delimited blast output file
☐	hg18	scBlastTab	Tab-delimited blast output file
☐	hg18	seq	
☐	hg18	spMrna	The best representative mRNA for a protein
☑	proteome	spDisease	A cross-reference table between Swiss-Prot IDs and disease description.
☐	proteome	spOldNew	
☐	proteome	uniProtAlias	
☐	uniProt	displayId	Relate ID and primary accession. A good table to use just get handle on all records.

Allow Selection From Checked Tables

Select Fields from hg18.knownGene

☐	name	Name of gene
☐	chrom	Reference sequence chromosome or scaffold
☐	strand	+ or - for strand
☐	txStart	Transcription start position
☐	txEnd	Transcription end position
☐	cdsStart	Coding region start
☐	cdsEnd	Coding region end
☐	exonCount	Number of exons
☐	exonStarts	Exon start positions
☐	exonEnds	Exon end positions
☐	proteinID	SWISS-PROT ID
☐	alignID	Unique identifier for each (known gene, alignment position) pair

get output | cancel | check all | clear all

go.goaPart fields

☐	dbObjectId
☑	dbObjectSymbol
☐	notId
☐	goId
☐	aspect

check all | clear all

hg18.knownToEnsembl fields

☑	name	Same as name field in known gene
☐	value	Other id

check all | clear all

proteome.spDisease fields

☐	accession	SWISS-PROT accession number
☐	displayID	SWISS-PROT display ID
☑	diseaseDesc	disease description

GMOD Scenario #3:
Report on aspects of gene, cont.(2)

hg18.kgXref fields

☐ kgID	Known Gene ID
☐ mRNA	mRNA ID
☐ spID	SWISS-PROT protein Accession number
☐ spDisplayID	SWISS-PROT display ID
☑ geneSymbol	Gene Symbol
☐ refseq	RefSeq ID
☐ protAcc	NCBI protein Accession number
☐ description	Description

check all | clear all

proteome.spDisease fields

☐ accession	SWISS-PROT accession number
☐ displayID	SWISS-PROT display ID
☑ diseaseDesc	disease description

check all | clear all

Select Fields from hg18.knownGene

☑ name	Name of gene
☐ chrom	Reference sequence chromosome or scaffold
☐ strand	+ or - for strand
☐ txStart	Transcription start position
☐ txEnd	Transcription end position
☐ cdsStart	Coding region start
☐ cdsEnd	Coding region end
☑ exonCount	Number of exons
☐ exonStarts	Exon start positions
☐ exonEnds	Exon end positions
☐ proteinID	SWISS-PROT ID
☐ alignID	Unique identifier for each (known gene, alignment position) pair

get output | cancel | check all | clear all

go.term fields

☐ id
☑ name
☐ term_type
☐ acc
☐ is_obsolete
☐ is_root

check all | clear all

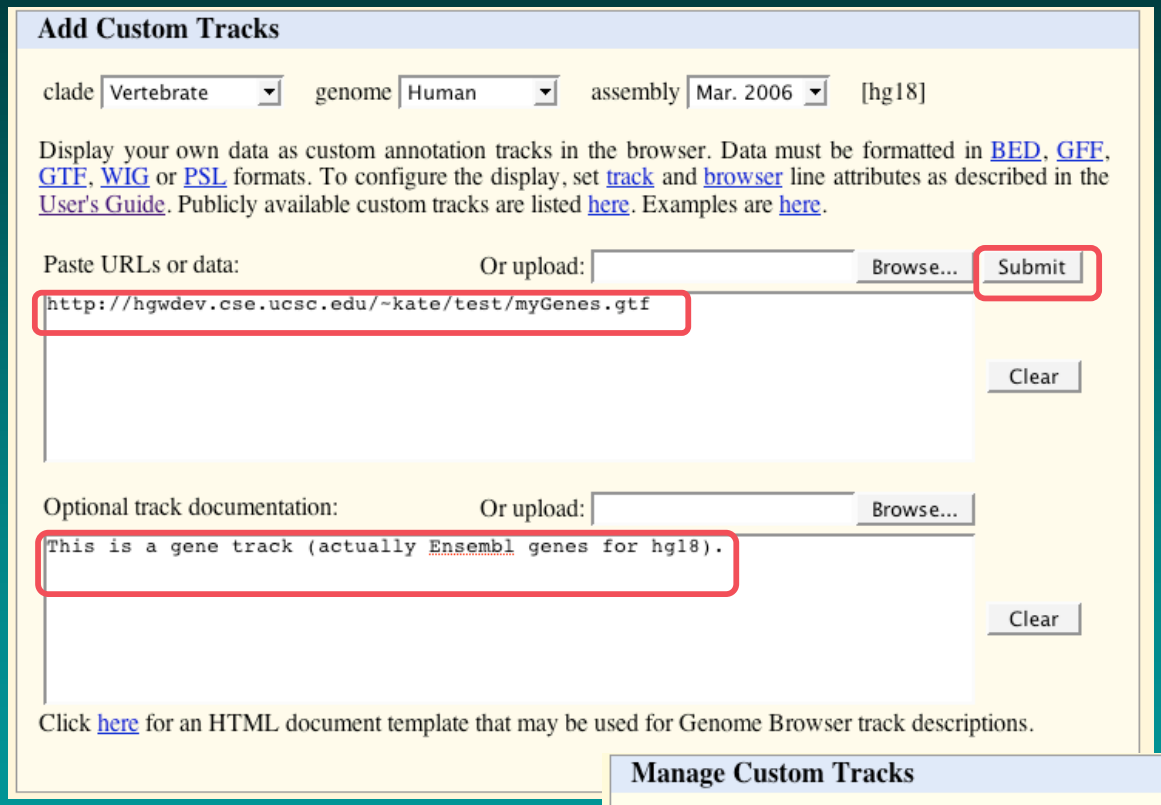
- Exon count
- GO terms
- Swiss-Prot disease description

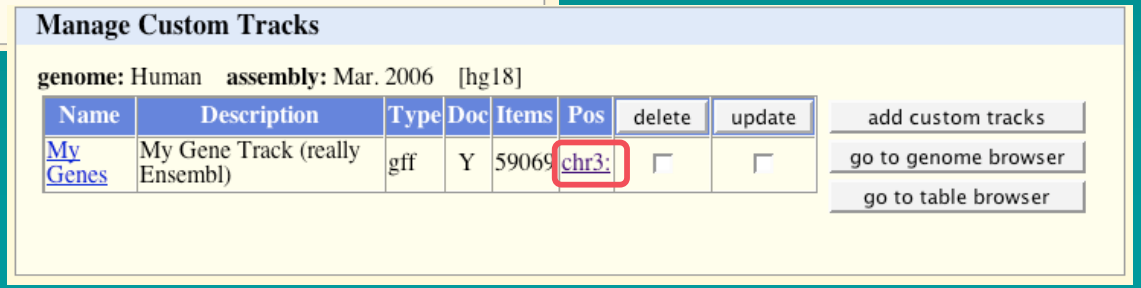
#hg18.knownGene.name	hg18.knownGene.exonCount	go.term.name	hg18.kgXref.geneSymbol	proteome.spDisease.diseaseDesc
CR749236	23	n/a	FOXP2	n/a
CR749236	23	nucleic acid binding,nucleus,zinc ion binding,	FOXP2	n/a
BC018016	3	n/a	FOXP2	n/a
AY144615	18	nucleic acid binding,DNA binding,transcription factor activity,nucleus,transcription,regulation of transcription, DNA-dependent,zinc ion binding,s		
AY144615	18	n/a	FOXP2	n/a
NM_148898	18	nucleic acid binding,DNA binding,transcription factor activity,nucleus,transcription,regulation of transcription, DNA-dependent,zinc ion binding,s		
NM_014491	17	nucleic acid binding,DNA binding,transcription factor activity,nucleus,transcription,regulation of transcription, DNA-dependent,zinc ion binding,s		

GMOD Scenarios 4 & 5:
Bulk queries and external data integration;
Compare user gene set to UCSC Known Genes

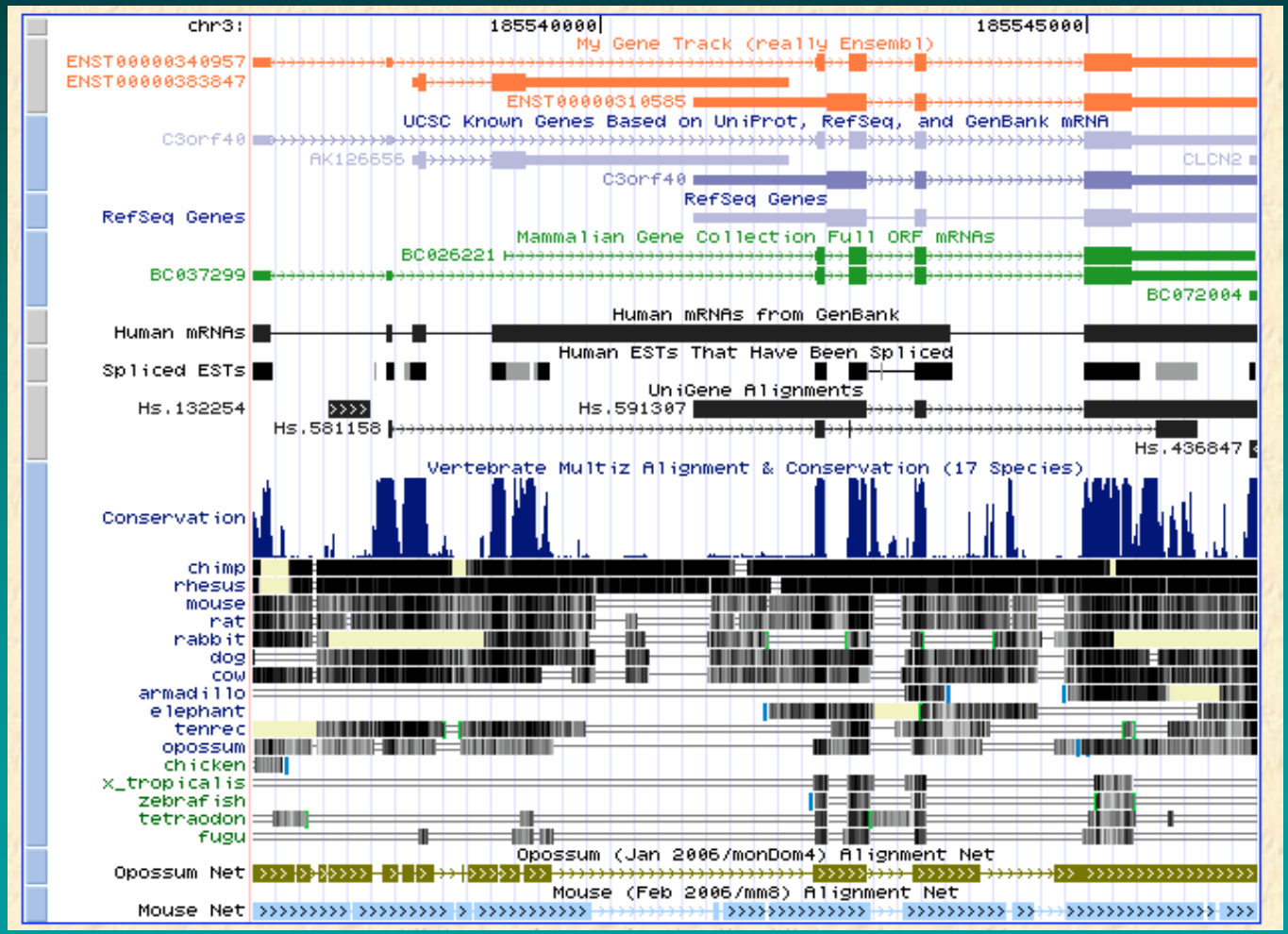
- How many user genes are not in Known Genes ?
- How well conserved across different species are the genes unique to the user gene set ?

GMOD Scenarios 4 & 5: Loading external data

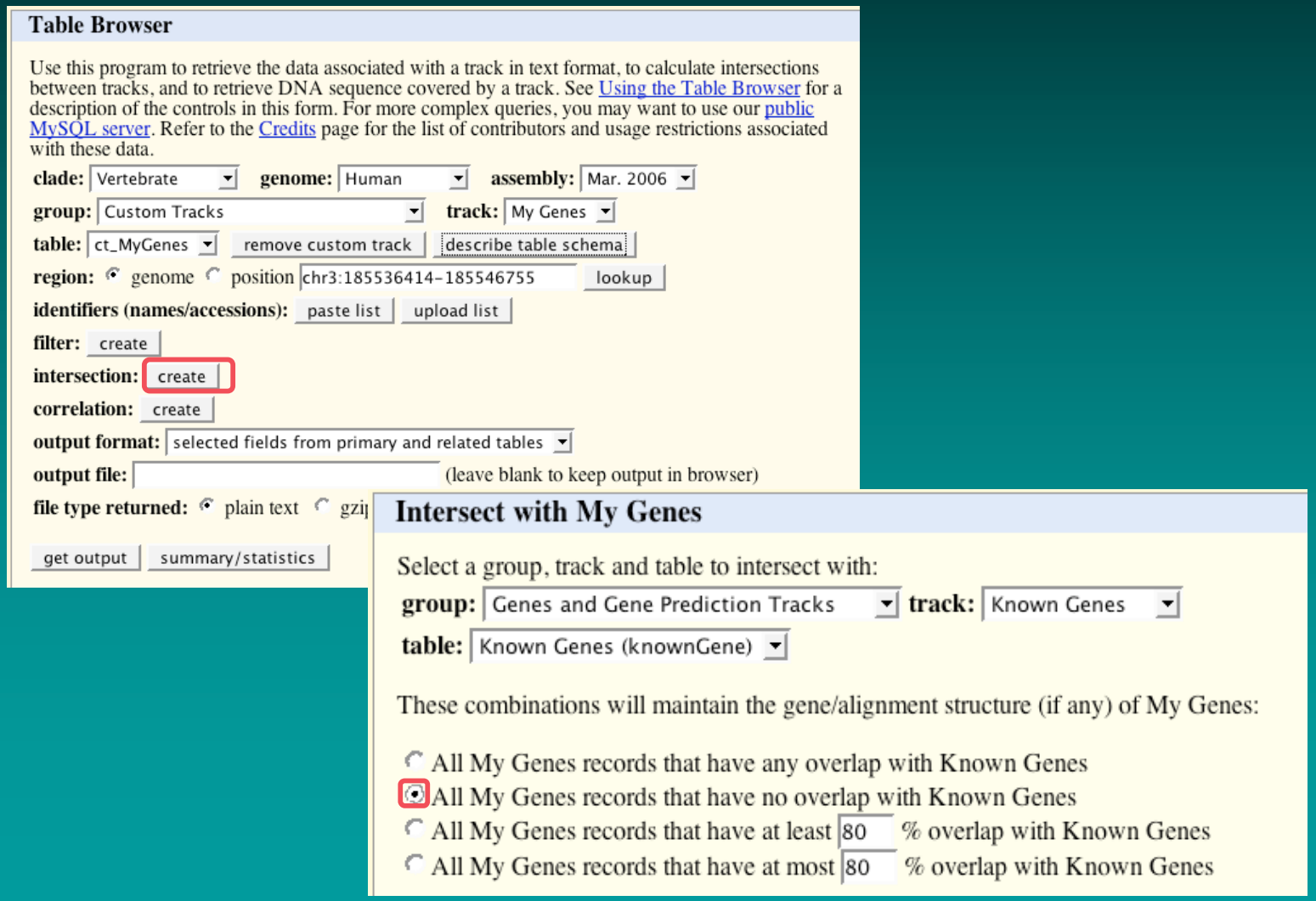




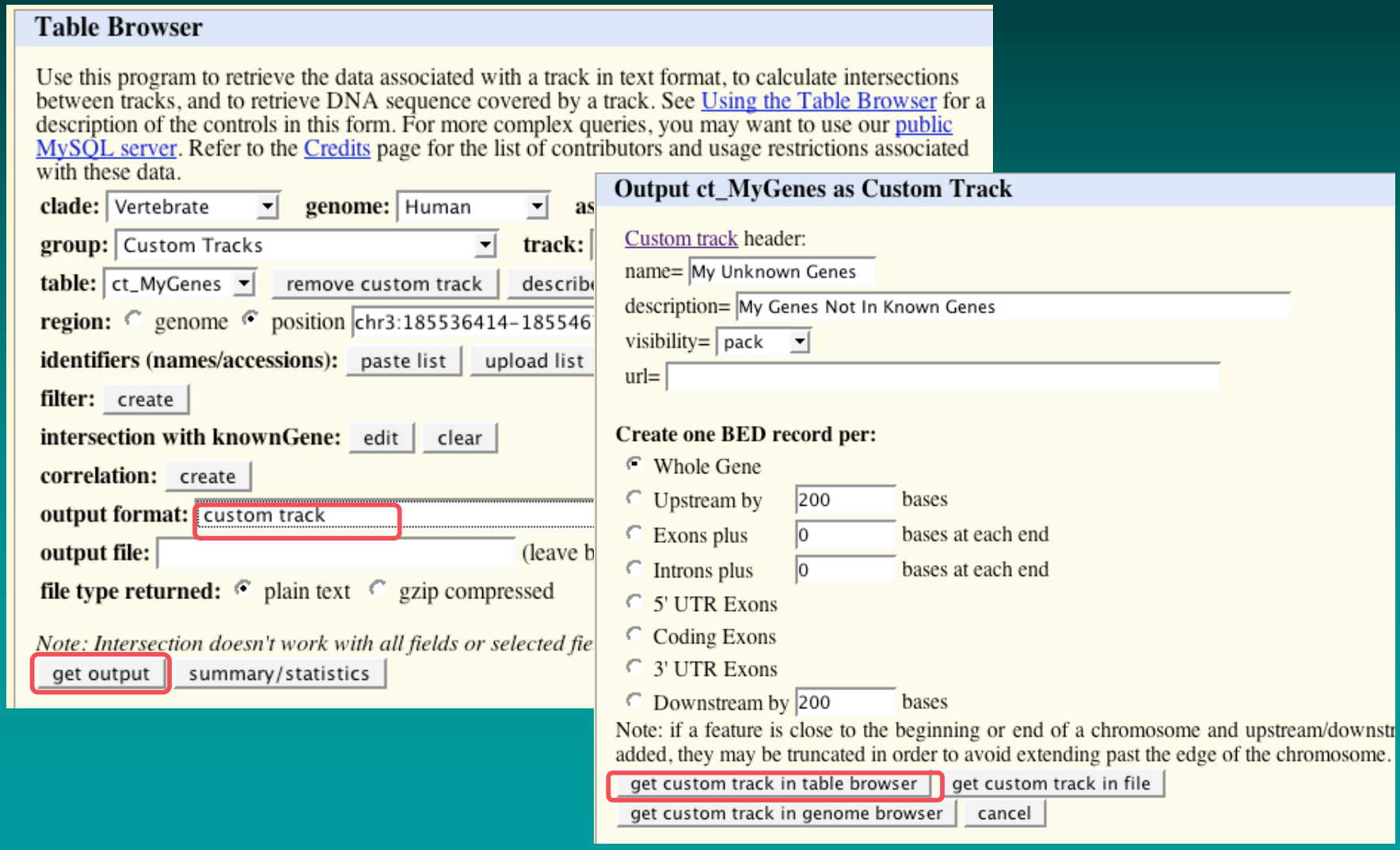
GMOD Scenarios 4 & 5:
Loading external data, cont.



GMOD Scenarios 4 & 5:
Intersection on whole dataset



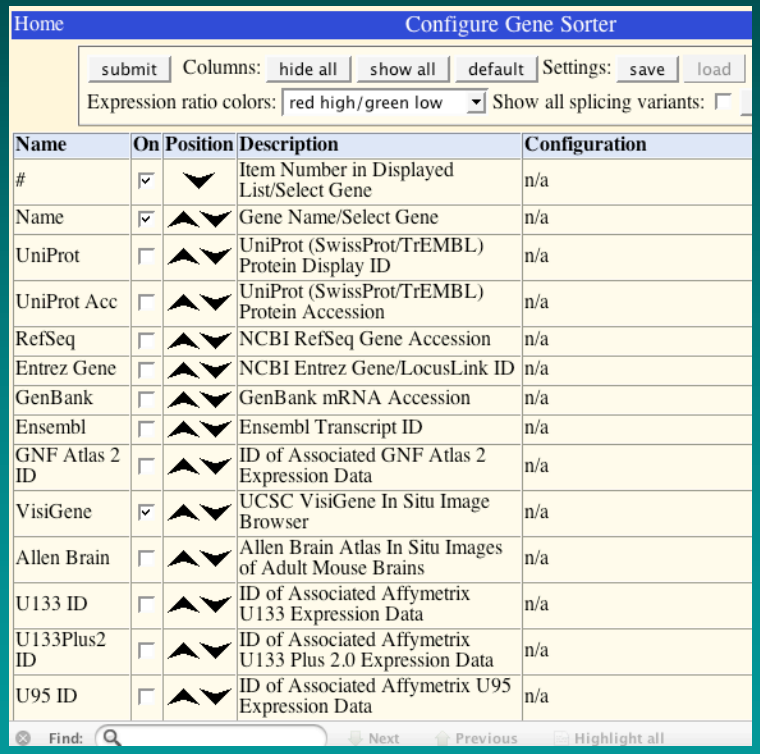
GMOD Scenarios 4 & 5:
Intersection on whole dataset, cont.

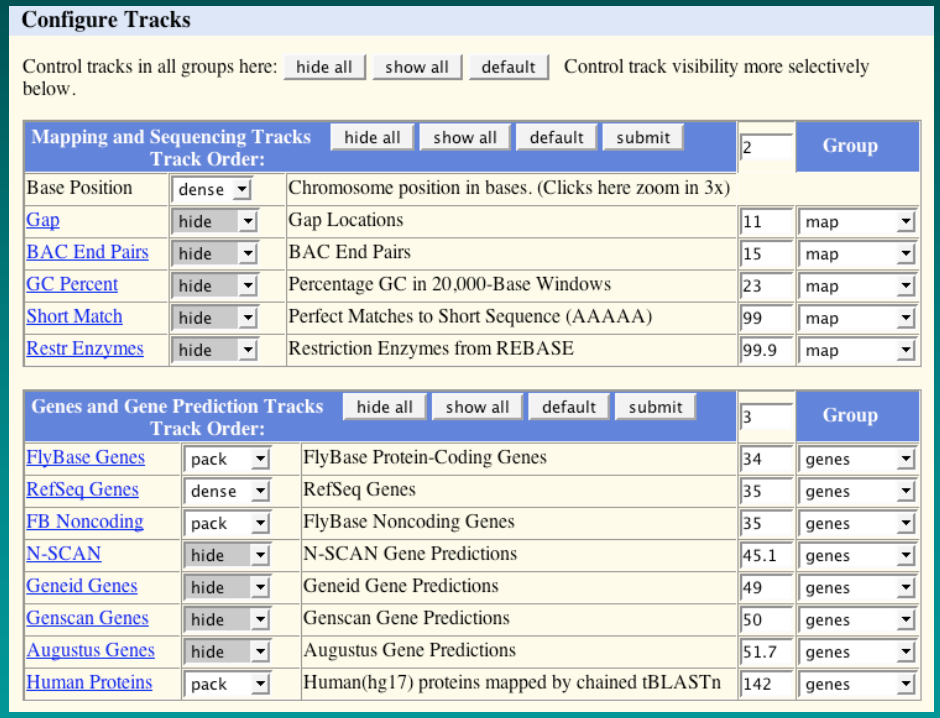


Kent's UI Guidelines

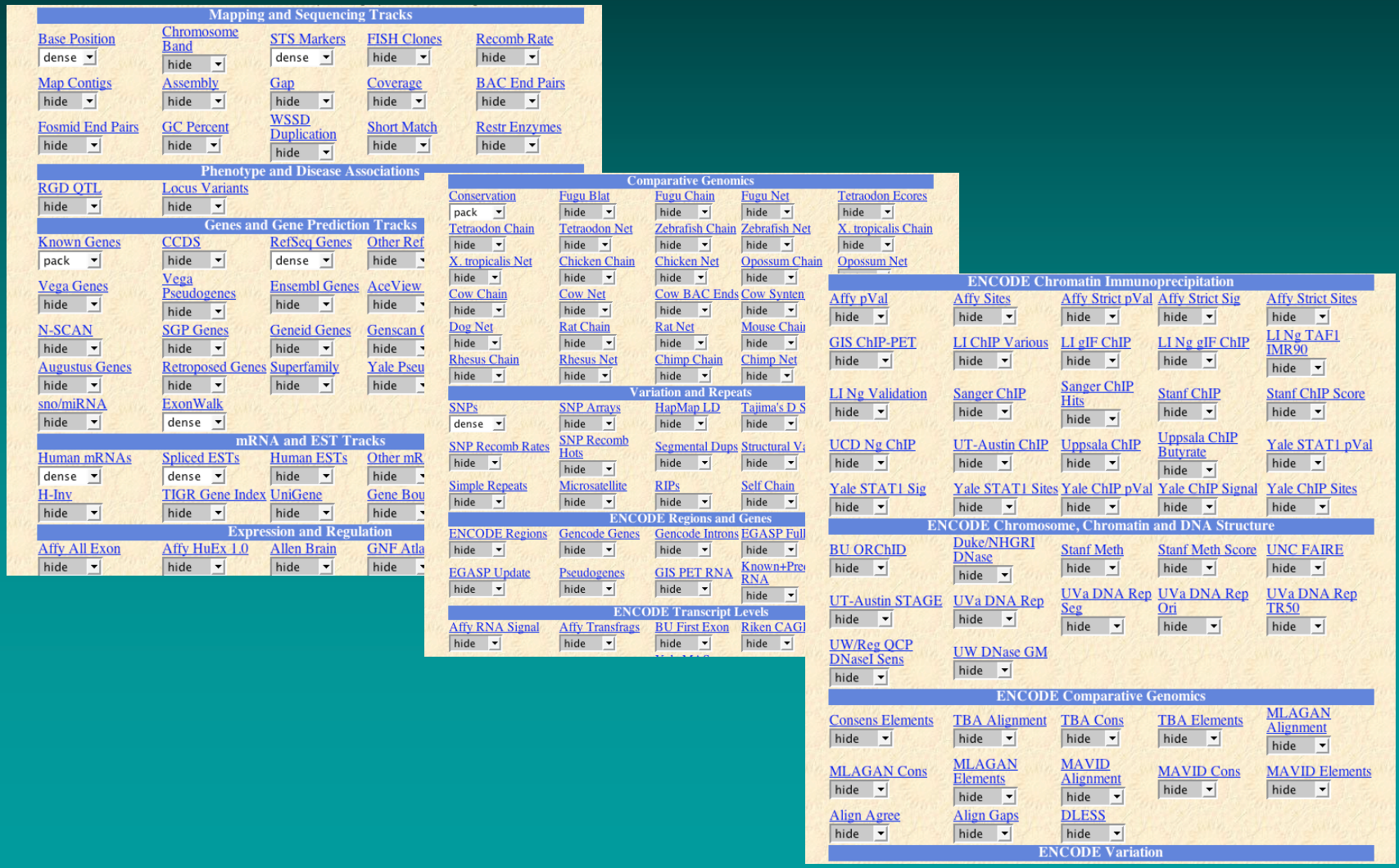
- Keep it reliable
- Keep it fast
- Label everything in plain English
- Put the most commonly used controls on the top of the page
- Keep it as simple as possible (but no simpler)
- Try to make options work together in an orthogonal way
- Remember your users are *intelligent* professionals. Don't dumb things down; complexity comes with the territory
- Don't change the site unnecessarily once people have gotten used to it.

User interface challenges:
User-configurable ordering





User interface challenges:
Track grouping to avoid overload



User interface challenges: Composite tracks to group similar data

Affymetrix ChIP/Chip (retinoic acid-treated HL-60 cells) Sites

Display mode: hide Submit

Select subtracks:

All add clear

Timepoint

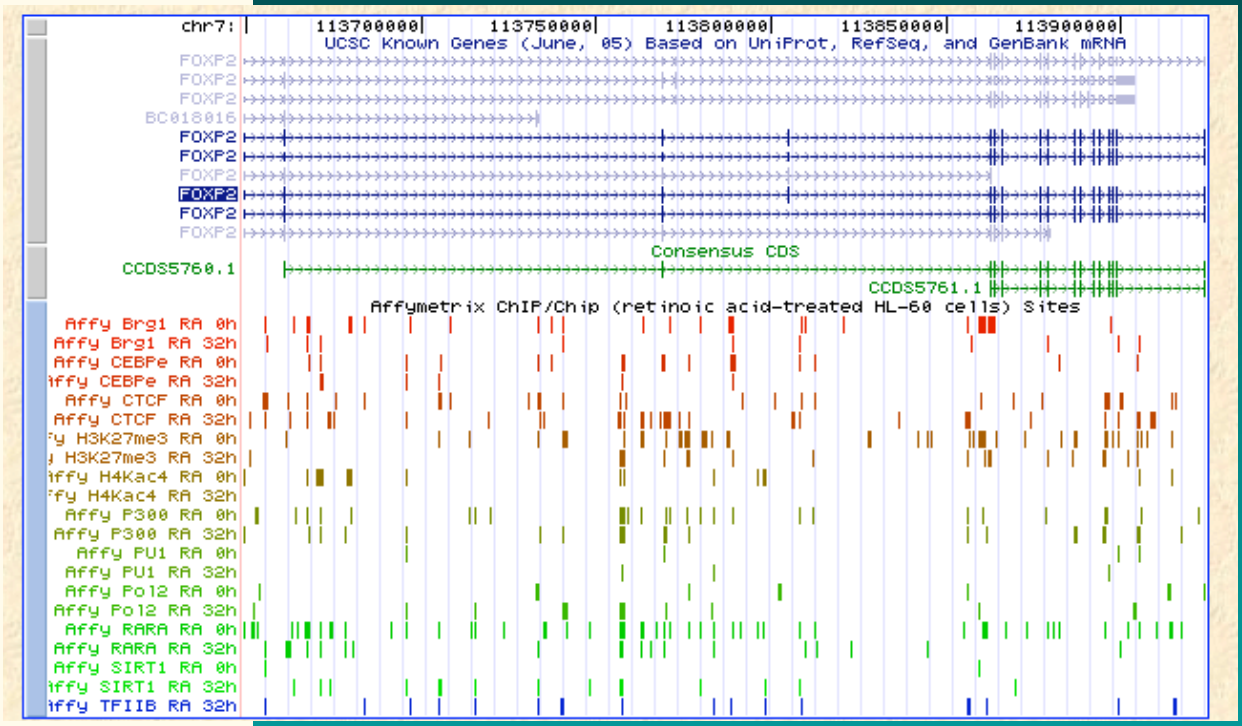
0hrs add clear
2hrs add clear
8hrs add clear
32hrs add clear

Factor

Brg1 add clear
CEBPe add clear
CTCF add clear
H3K27me3 add clear
H4Kac4 add clear
P300 add clear
PU1 add clear
RARA add clear
Pol2 add clear
SIRT1 add clear
TFIIB add clear

Show checkboxes for: Only selected subtracks All subtracks

Affymetrix ChIP/Chip (Brg1 retinoic acid-treated HL-60, 0hrs) Sites
Affymetrix ChIP/Chip (Brg1 retinoic acid-treated HL-60, 2hrs) Sites
Affymetrix ChIP/Chip (Brg1 retinoic acid-treated HL-60, 8hrs) Sites



User Support and Training

- FAQs: http://genome.cse.ucsc.edu/FAQ/
- questions? genome@soe.ucsc.edu

  archived answers:

  http://genome.ucsc.edu/contacts.html
- OpenHelix: http://www.openhelix.com/
  - Classes, seminars
  - Free online tutorial
  - Quick reference cards

Thanks!

- UCSC Genome Browser Team:
  - David Haussler – PI
  - Jim Kent – Browser Concept, BLAT, Team Leader
  - Donna Karolchik – Engineering Mgr, Docs & Training
  - Mark Diekhans, Fan Hsu, Angie Hinrichs, Kate Rosenbloom, Hiram Clawson, Rachel Harte, Heather Trumbower, Galt Barber, Andy Pohl - Engineering
  - Robert Kuhn (mgr), Ann Zweig, Kayla Smith, Brooke Rhead, Archana Thakkapallayil – QA/Support
  - Jorge Garcia, Chester Manuel, Victoria Lin, Erich Weller, Paul Tatarsky – KiloKluster, Sys-admin
- Funding:
  - National Human Genome Research Institute
  - Howard Hughes Medical Institute
  - National Cancer Institute