Large-scale comparative genomics/proteomics, examples from Ensembl

Abel Ureta-Vidal

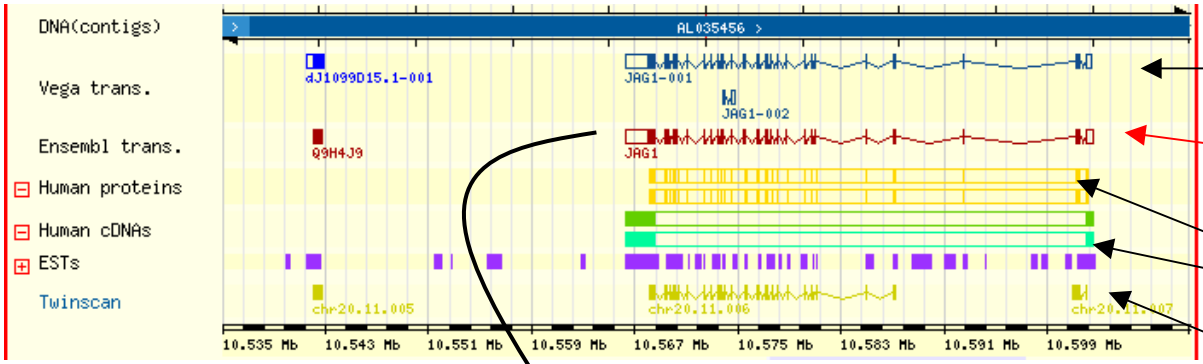
EMBL-EBI

Hinxton, Cambridge, UK

Overview

- Evaluating genes and transcripts
  - Ensembl gene set
  - Comparison with manual curation
- Comparative proteomics
  - Orthologues prediction
  - Protein clustering into families
- Comparative genomics
  - Genome-wide DNA alignments
  - Conserved synteny blocks
  - Multi-species view

Our aim

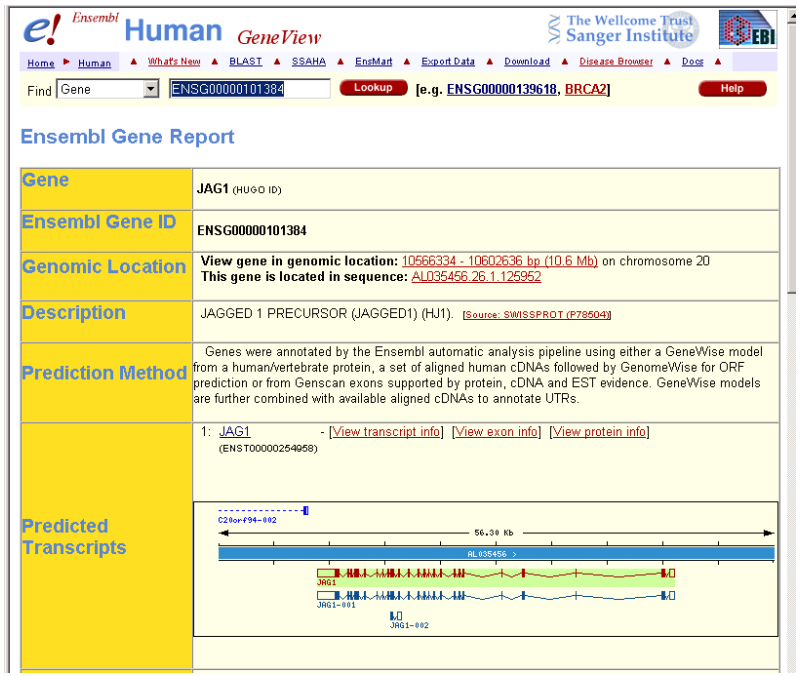


manual curation

Ensembl transcript predictions

evidence

other groups' models



Ensembl gene set

- Place all available species-specific proteins to make transcripts
- Place similar proteins to make transcripts
  - Use mRNA data to add UTRs
- Build genes using cDNA evidence

- Combine annotations to make genes with alternative transcripts

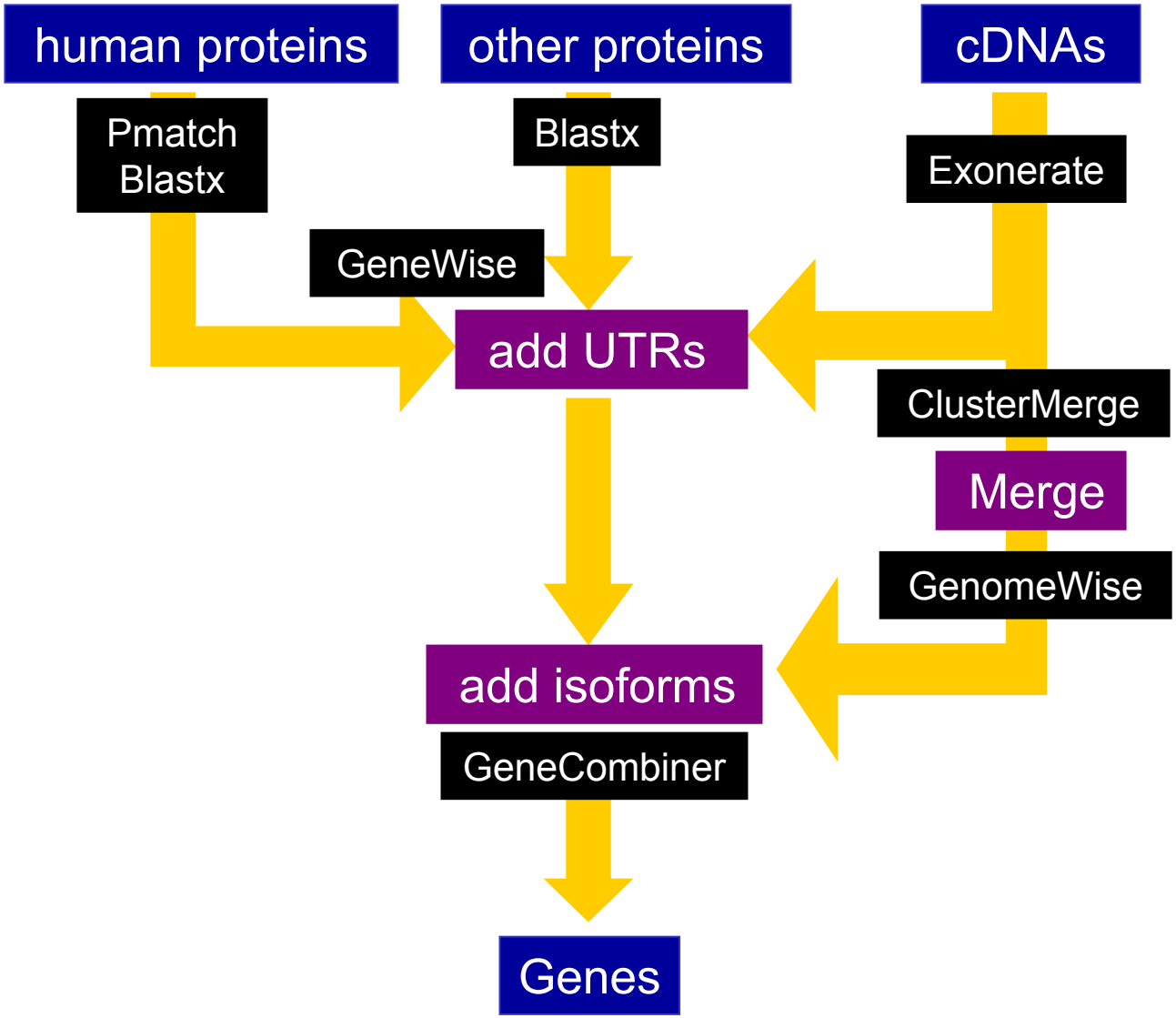
Gene build is massively protein based

- DNA-DNA alignments don't give us translatable genes

- Essential to align at the protein level allowing for frameshifts and splice sites

- Genewise (Ewan Birney)
  - Protein – genomic alignment
  - Has splice site model
  - Penalizes stop codons
  - Allows for frameshifts

Automatic Gene Annotation



Genes from known proteins

Human protein sequences
SwissProt/TrEMBL/RefSeq

↓

pmatch *v*. assembly

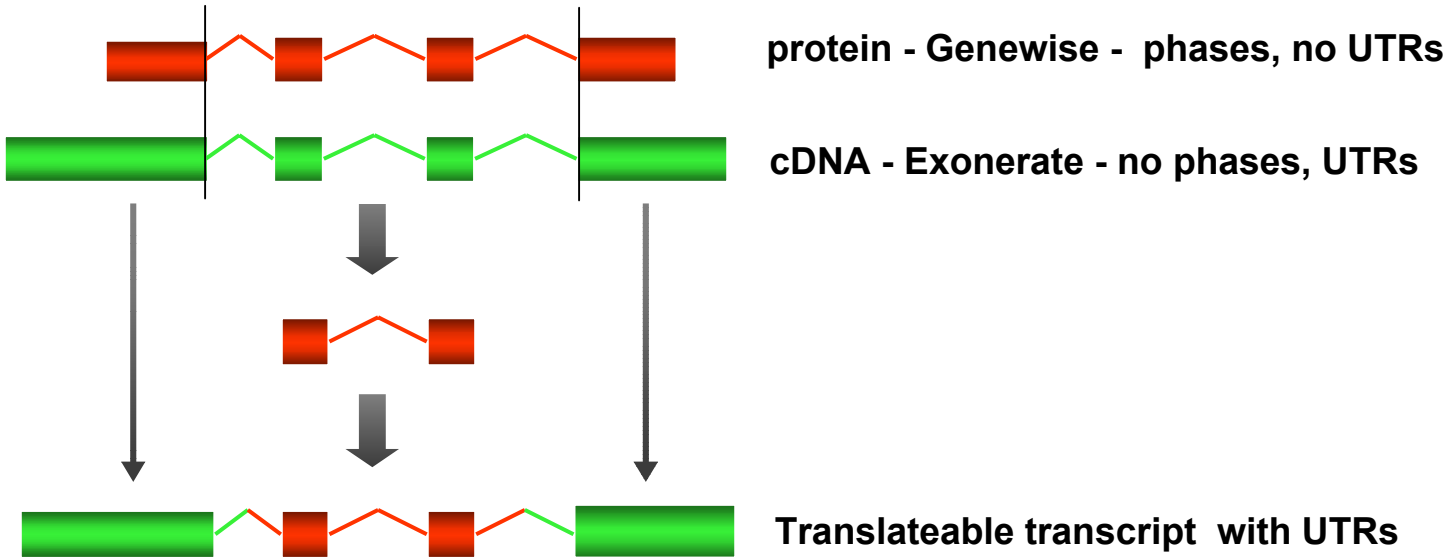
↓

blastx

↓

genewise

Add UTRs using mapped mRNAs



Full Human Build

- NCBI 33 build

- Ensembl genes: 24,261
- Ensembl transcripts: 32,997
- Ensembl exons: 226,669

- Input: 48,176 proteins; 86,918 cDNAs

- Transcripts made from:
  - Human proteins with (without) UTRs 68% (19%)
  - Non-human proteins with (without) UTRs 2% (9%)
  - cDNA alignment only 0.8%

Comparison to manual annotation

Genes

Sensitivity

~90% of manual genes are in e!

Specificity

~75% of e! genes are in the manual sets

Exon bps

Sensitivity

~70% of manual bps are in e! exons

(90% of coding bps)

Specificity

~80% of e! bps are in manual exons

Alternative transcripts per gene

manual 3 e! 1.3

Figures are for the gene build on NCBI 33 (human) and manual annotation for chromosomes 6, 14 & 20

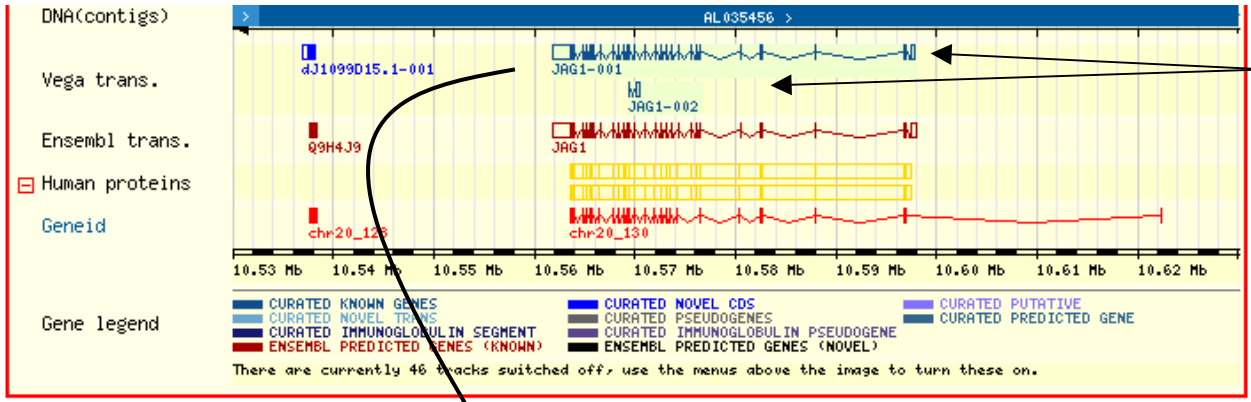
Manual curation
(human, mouse, zebrafish)

- Manual annotation of finished clones/chromosomes
- Vega database (vega.sanger.ac.uk) at Sanger
  - Uses ensembl schema database and web display

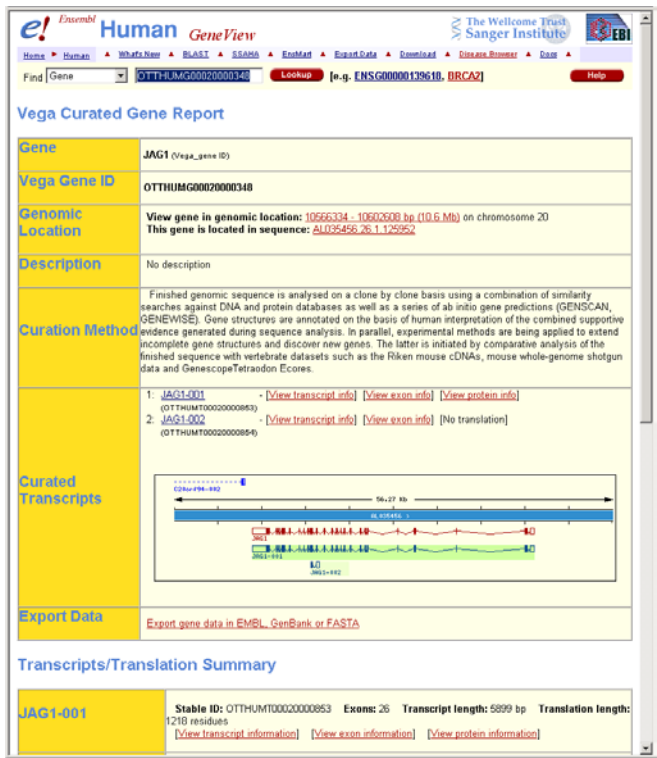
- Currently has human 6, 13, 20, 22 from Sanger and 14 from Genoscope, 7 from University of Washington
- Other groups will also contribute to Vega

- Displayed in Ensembl when available

Vega genes

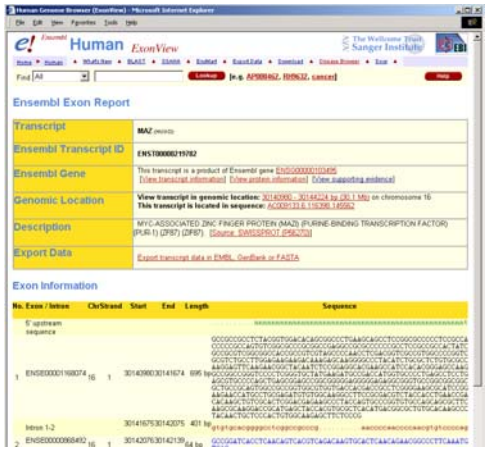


Vega manual curation



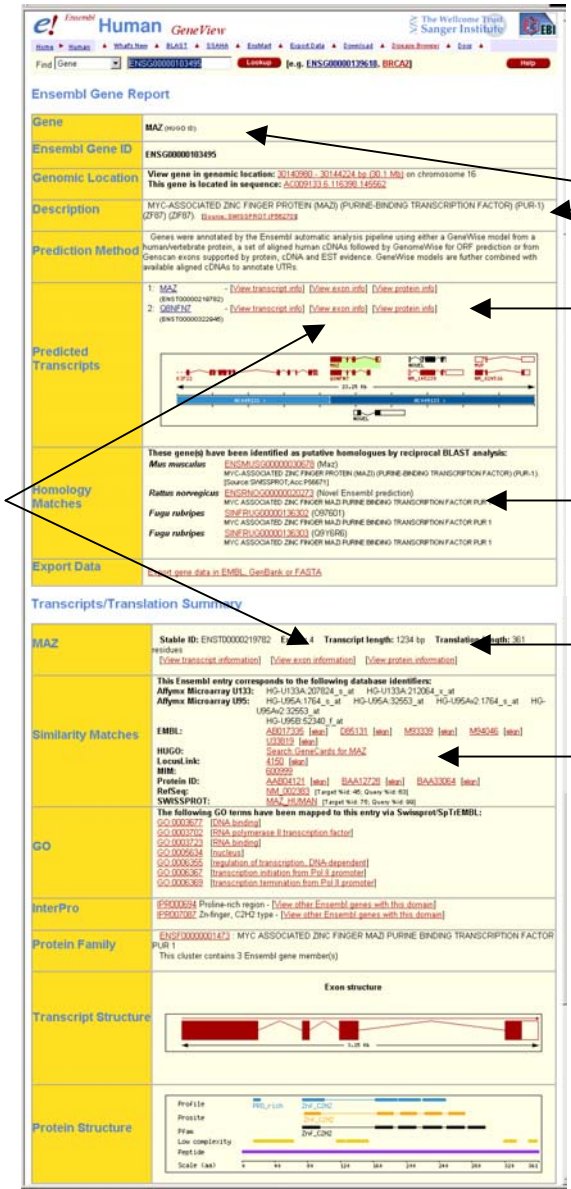
GeneView

ExonView



Evidence used to build the transcript

links to Exonview



Gene name & description

Alternative transcripts

Links to putative orthologues

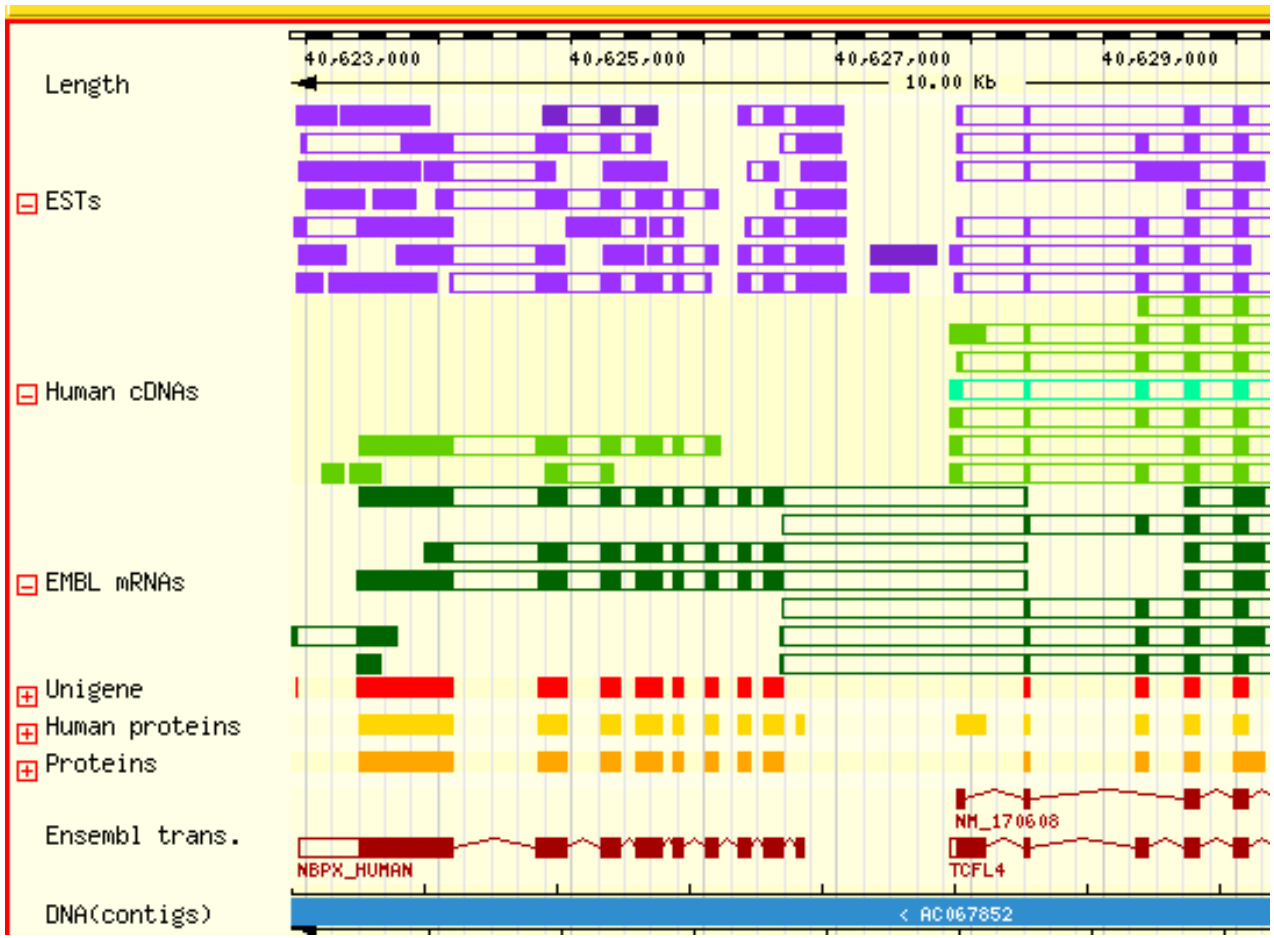
Transcript name

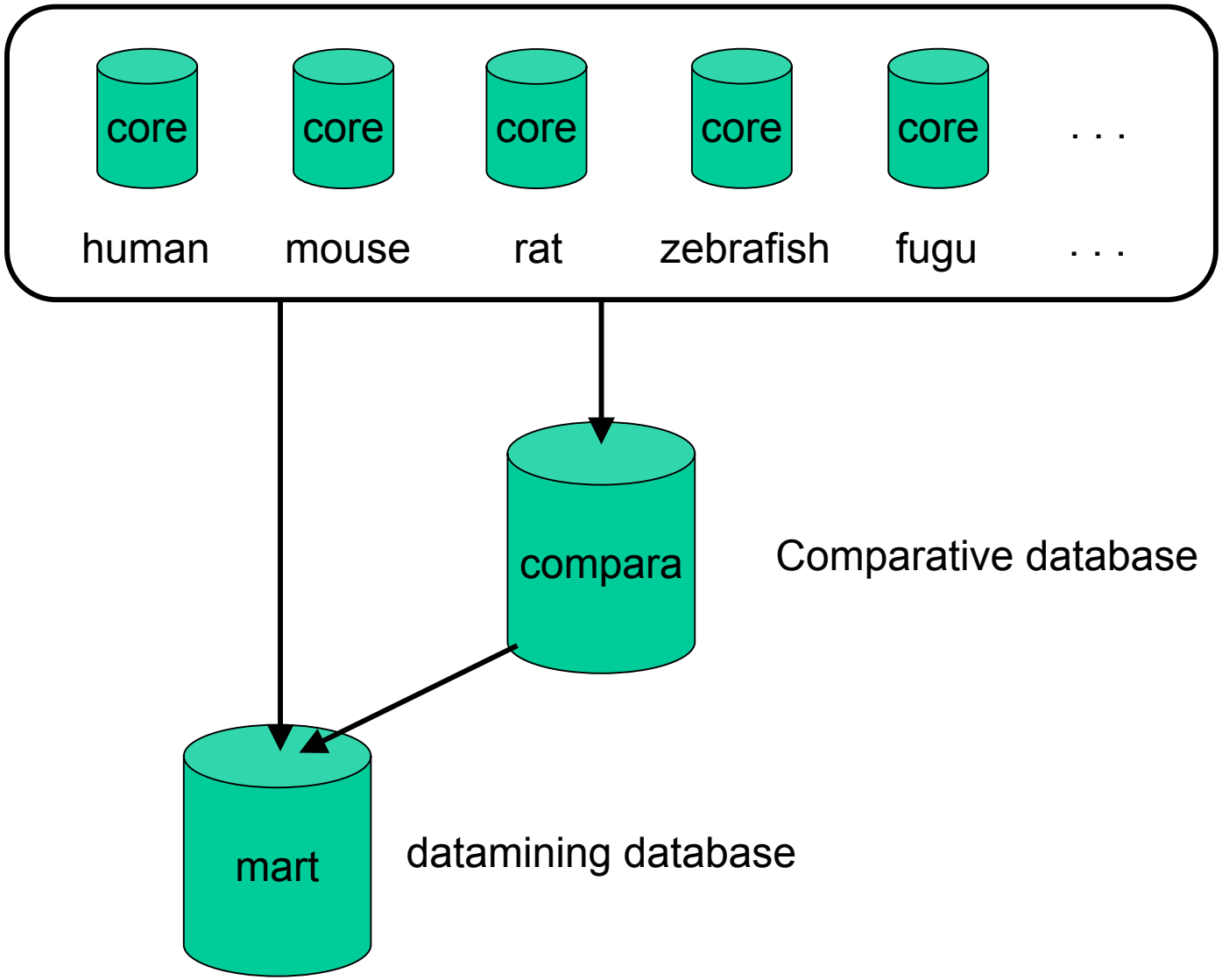
Mapping to external databases

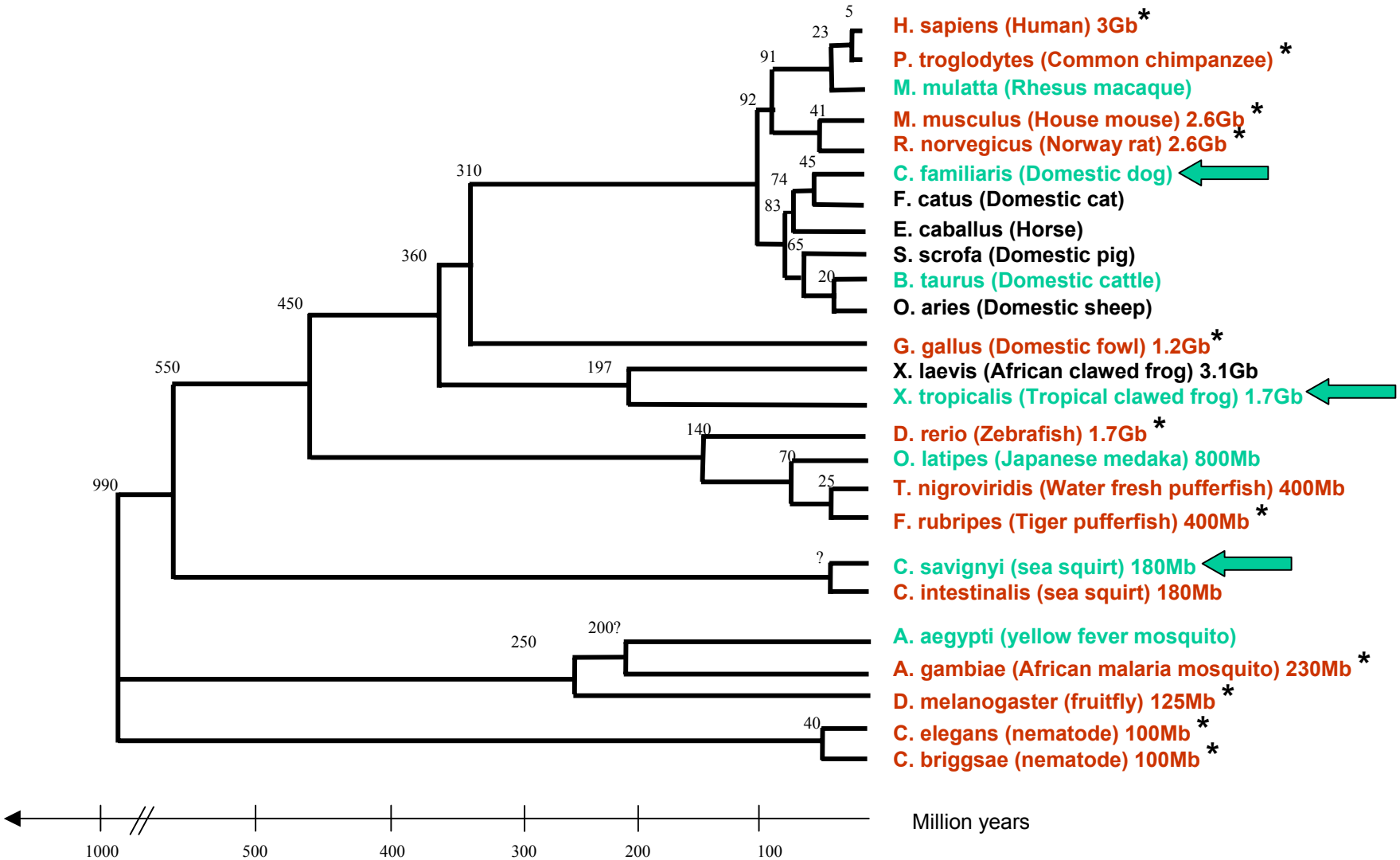
Evidence tracks in ContigView

Expanded tracks

Compressed tracks




core
core
core
core
core
. . .
human
mouse
rat
zebrafish
fugu
. . .
compara
Comparative database
mart
datamining database



Red : whole genome assembly available (also Honey bee)

Green : whole genome assembly due in the next 2 years

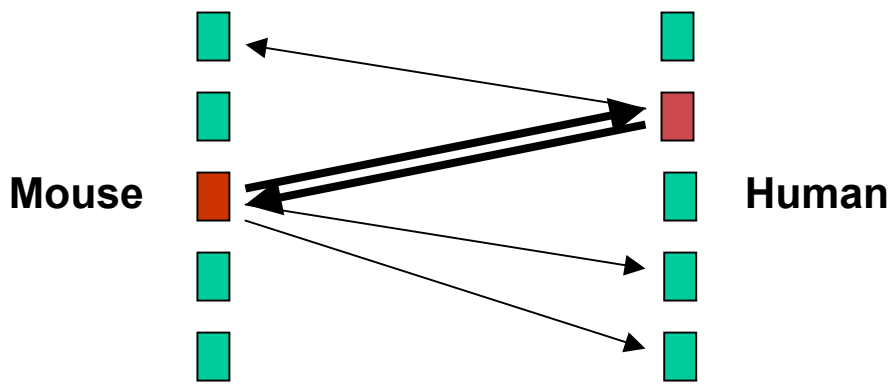
*** 11 species currently in Ensembl**

Overview

- Evaluating genes and transcripts
  - Ensembl gene set
  - Comparison with manual curation
- Comparative proteomics
  - Orthologues prediction
  - Protein clustering into families
- Comparative genomics
  - Genome-wide DNA alignments
  - Conserved synteny blocks
  - Multi-species view

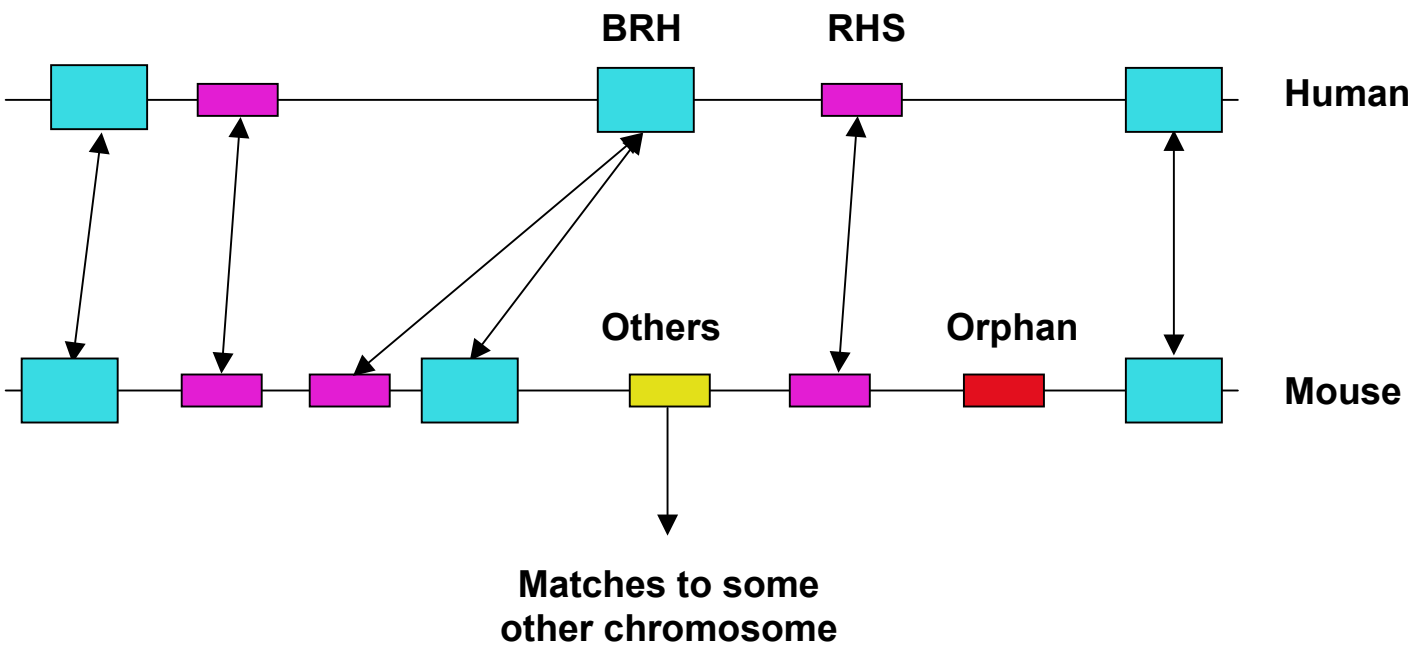
Orthologues prediction

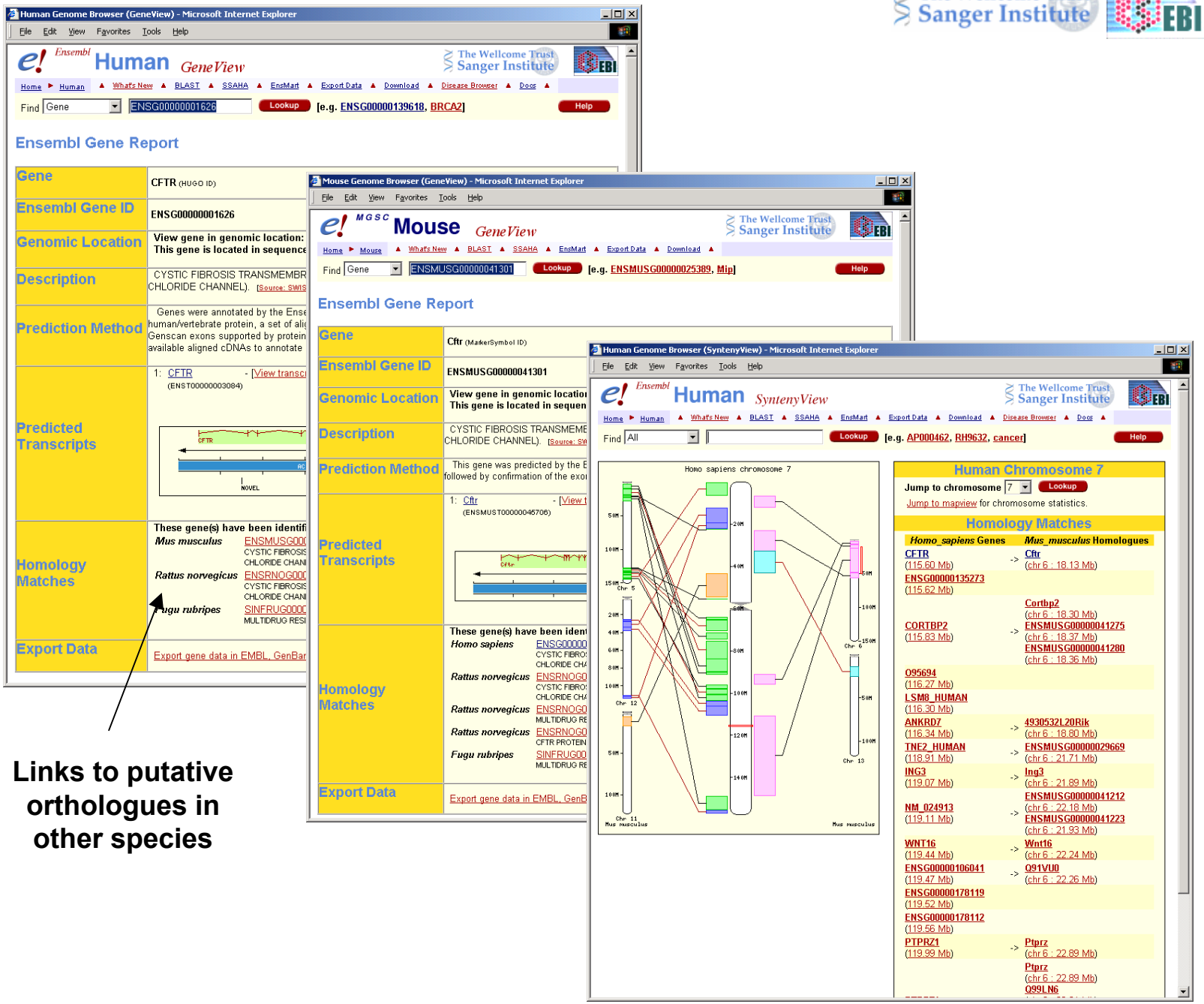
- Find orthologous genes by comparing the protein sets of two species (only the longest peptide considered).
- blastp+sw all *versus* all (on a paired species basis)
- Best Reciprocal Hit as putative orthologues (named "BRH")



RHS, Orphans and Others

Based on BRH genomic coordinates in both species compared and gene order conservation, we identify additional orthologues or RHS for Reciprocal Hit supported by Synteny.





Links to putative orthologues in other species

For each orthologous gene pair

- We store
  - %identity, %positivity, %coverage, alignment, type (BRH, RHS), dN, dS
- Using the compara perl API (soon from the web site)
  - Protein or cDNA alignment
  - 4D, 2D sites can also be easily retrieved
- On going developments
  - Build clusters of orthologues
  - Multiple alignments and phylogenies
  - Consider all isoforms for each gene
  - Include information on orphans and non-BRH/non-RHS pairs as well as provide the full blastp results

Protein clustering into families

- Cluster proteins from different organisms that may share the same function
- Obtain some for ‘novel’ genes/proteins
- Locate family members over the whole genome
- Identify possible orthologues and paralogues in other species

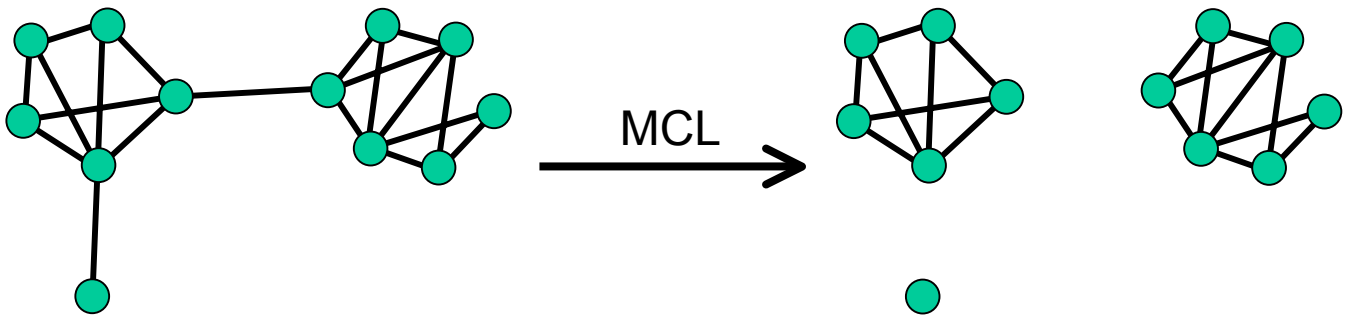
Dataset used and comparisons

- Half a million proteins clustered:
  - All Ensembl proteins from all species in Ensembl
    - 233,000 predicted proteins
  - All metazoan (animal) proteins in SWISSPROT/SPTrEMBL
    - 40,000 SWISSPROT
    - 230,000 SPTREMBL

- Blastp all *versus* all, then clustering with MCL

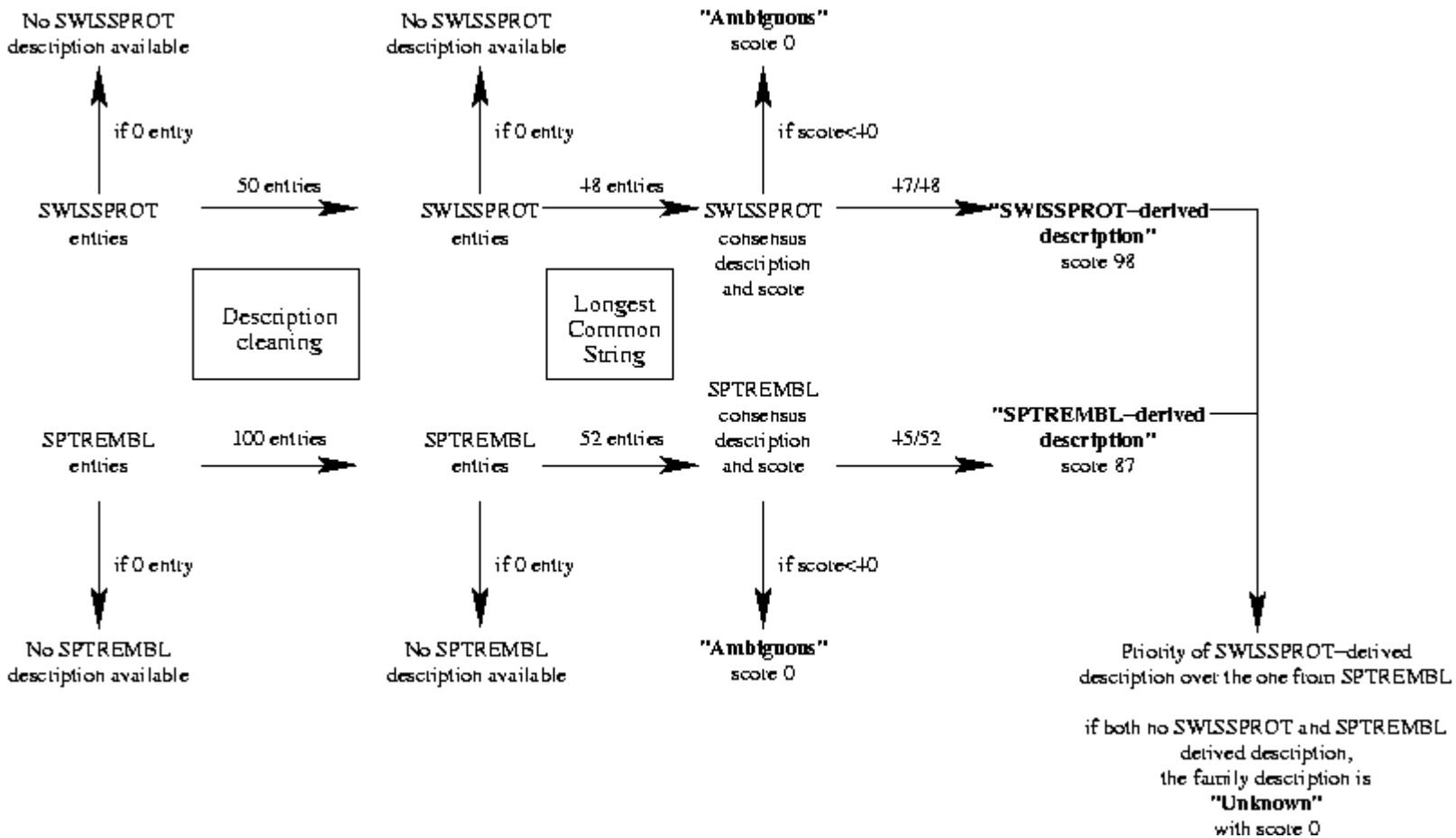
Clustering with MCL

- MCL for Markov CLustering algorithm, based on flow simulation in graphs (http://micans.org/mcl/)
- Keeps into the same graph/cluster only very well inter-connected nodes/protein



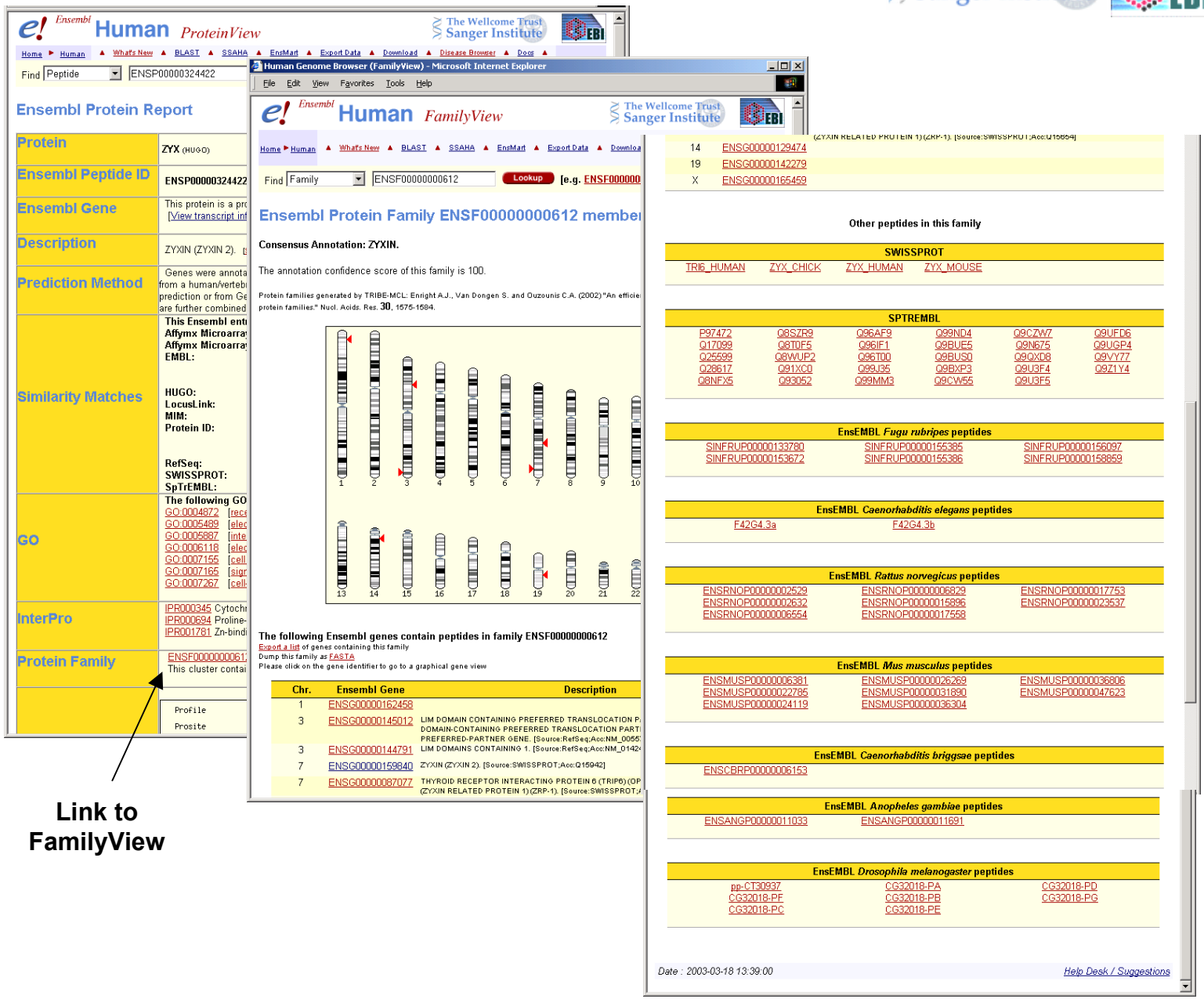
- Allows rapid and accurate detection of protein families on large-scale.
- Automatic description and clustalw multiple alignment applied on each cluster

Automatic description and scoring



For each cluster

- We store
  - Description and score
  - Multiple alignment
- Future extensions
  - Improving descriptions
  - Multiple alignment assessment
    - t-coffee
    - Protein domain information consistency
  - Build phylogeny on each cluster
    - Using the multiple alignment
    - Using dS values (mainly inside mammals)
    - Identify intra/inter-species orthologue/paralogues



Link to FamilyView

Addition of protein domain information

- Introduction of protein domain
  - Help for internal data QC by checking consistancy between orthologues, protein clusters and domains information.
  - Provide this kind of cross-check data to the user

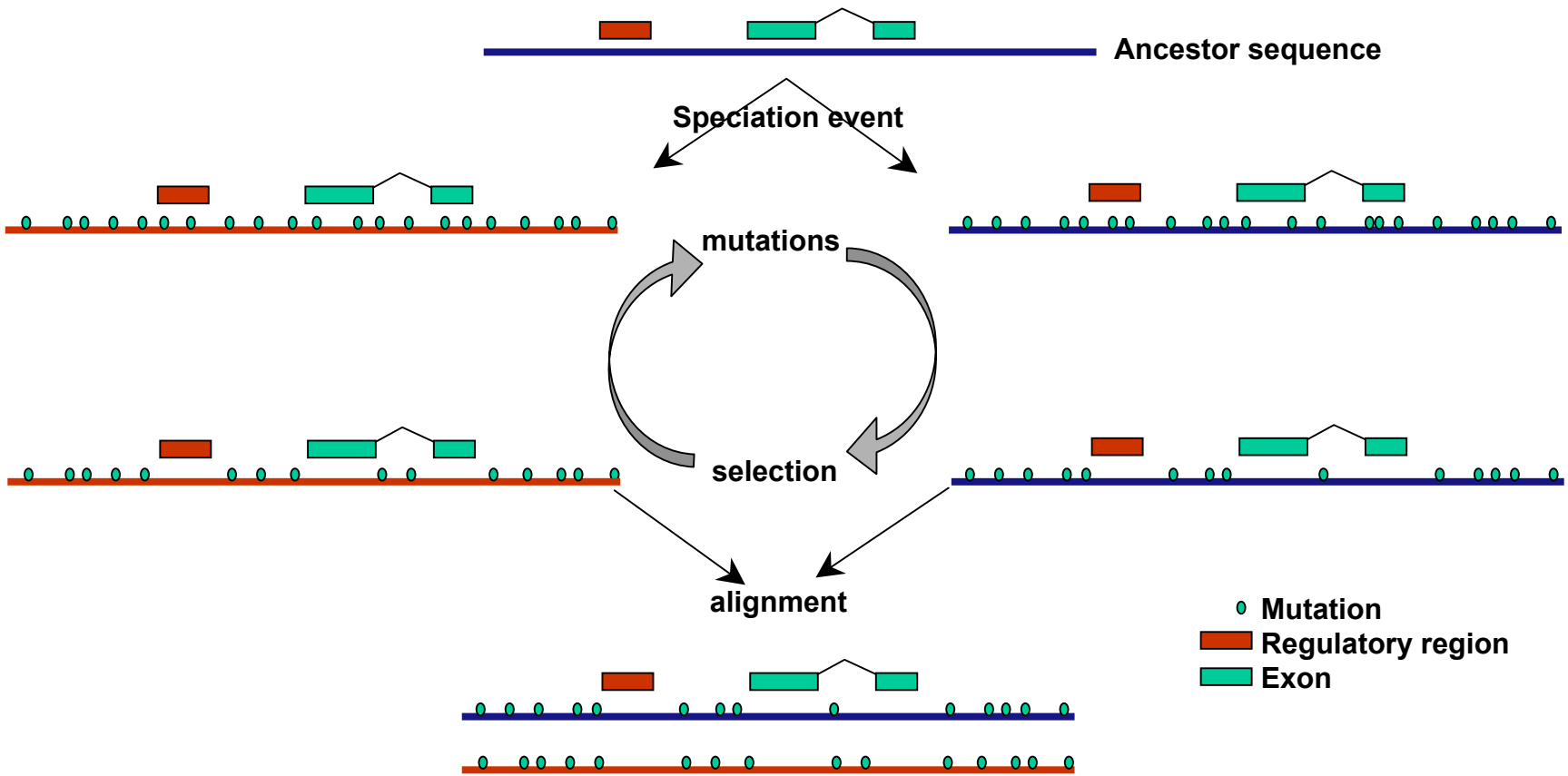
Overview

- Evaluating genes and transcripts
  - Ensembl gene set
  - Comparison with manual curation
- Comparative proteomics
  - Orthologues prediction
  - Protein clustering into families
- Comparative genomics
  - Genome-wide DNA alignments
  - Conserved synteny blocks
  - Multi-species view

Aligning genomes, why?

- How genomes of the species considered have been rearranged since their divergence by speciation.
- Define syntenic regions, long regions of DNA sequences were order and orientation of functional elements are highly conserved
- Finding conserved non coding regions
  - Good guides to find and test putative regulatory regions
- What is missing in one species, present only in another?
- Differences between closely related species (human/chimpanzee, human/macaque), may help understanding the speciation mechanisms

Basic concepts (1)



Basic concepts (2)

Functional sequences (coding exons, regulatory regions) are generally highly conserved

Conserved sequences can be functionnaly important

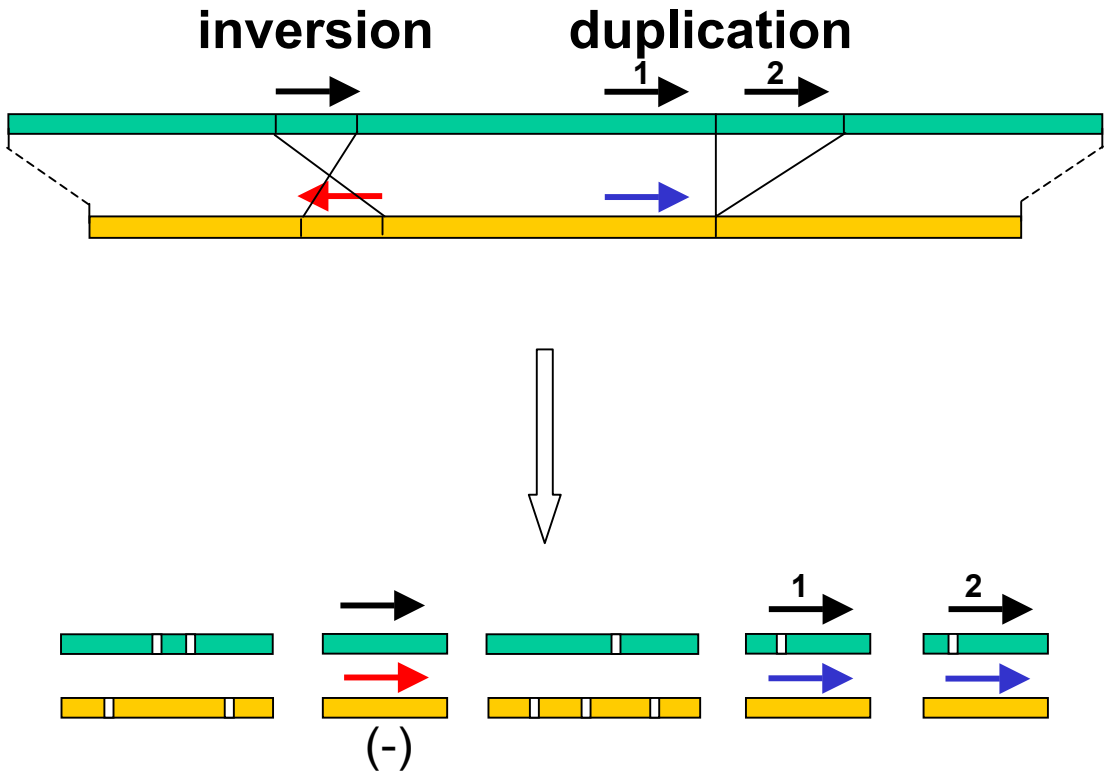
Conservation ⟷ Function

Comparing DNA sequences from different species can help to find biological functions

Using a local aligner

- Local alignment
  - Find all highly similar regions over 2 sequences
    - Find the orthologous as well as all the paralogous sequences
  - Separated by segments without alignment
  - Can handle rearranged sequences
  - Need post- filtering to limit too much overlapping alignments

Local alignment



Aligning large genomic sequences

- Independent from protein/gene predictions

- Issues
  - Heavy process
  - Computes run only by few dedicated groups
  - Scalability (more and more species available)
  - Time constraint

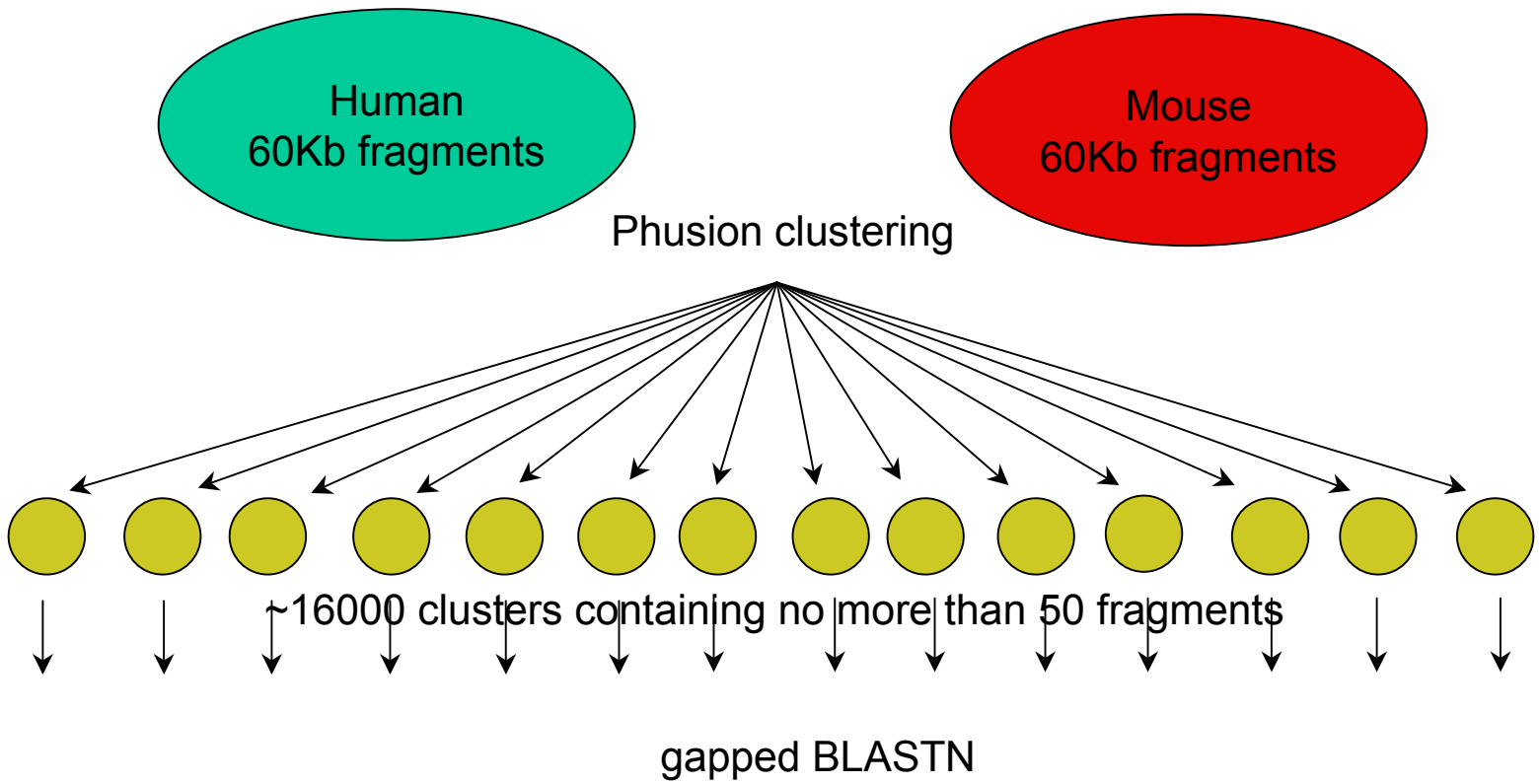
- As the « true » alignment is not known, then difficult to measure the alignment accuracy and apply the right method

Trying to avoid the all *versus* all comparison

- Phusion shotgun assembler-gapped BLASTN combinaison

  (Jim Mullikin and Zemin Ning, Sanger Institute)

Phusion - gapped BLASTN



Phusion - gapped BLASTN

- Fast but speed comes at a cost

- Only 22% of human genome coverage

- Good enough for generating orthologous links between the 2 species aligned, so that can be used either
  - in the web site for moving from one species to another
  - calculate synteny regions

- Not good enough for serious genome-wide post-analysis because not comprehensive enough

All vs all approach, key features of BLASTZ (collaboration with UCSC)

- Can handle large sequences

- Used 2-weighted spaced seeding strategy
1110100110010101111 (12of19)

- Makes distinction between repeat and non-repeat sequences (soft masking)
- Dynamic masking
- Try aligning inside repeats

- One iterative step with lower threshold to expand alignments

How Blastz was used

- 10Mb Human fragments (3000)
- 30Mb Mouse fragments (100)
- Lineage-specific repeats removed

- 48 hours on 1024 CPUs

- Generates 9Gb of ouput

- When filtered for Best hit on Human genome, it is reduced to 2.5Gb

Blastz human genome coverage

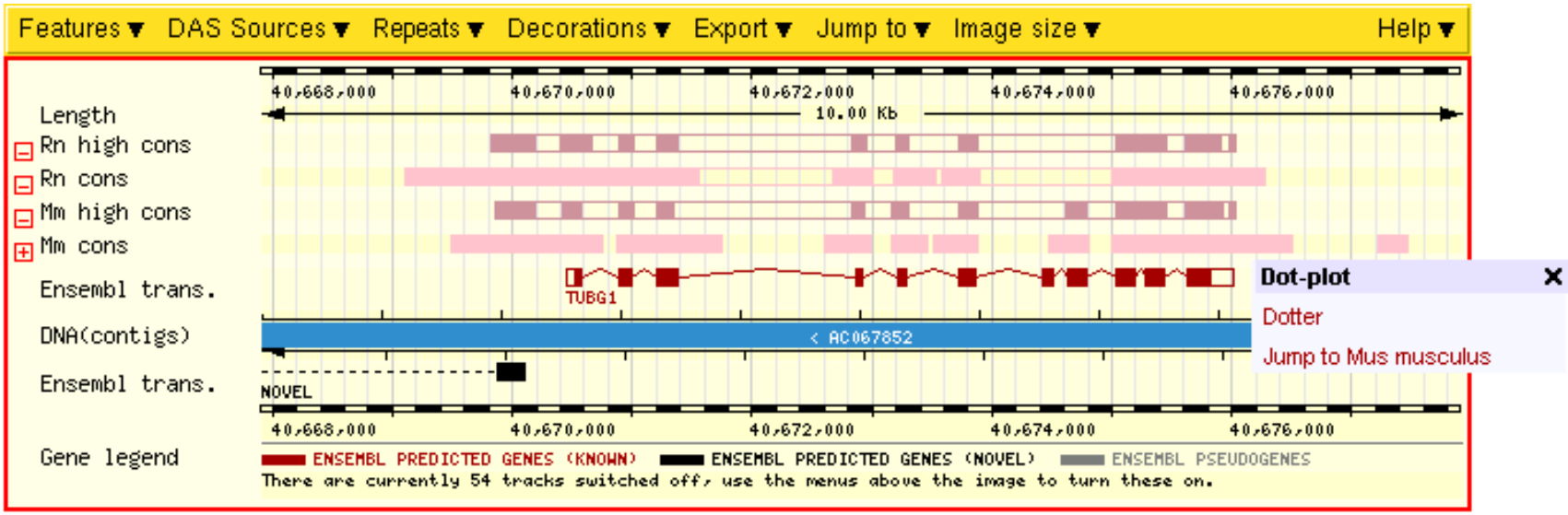
- 40% of the human genome is covered by an alignment of mouse sequences
- By rescoring the alignment over a “tight” matrix that is very stringent and look for high conservation (>70% identity), the coverage goes down to 6%

Genome alignment summary

- "cons" track
  - blastz from UCSC : human/mouse, human/rat, mouse/rat, human/chimpanzee, human/chicken
  - phusion-blastn : elegans/briggsae
- "high cons" track

  Obtained by rescoring the raw alignments over a "tight" matrix
- "trans BLAT " track
  - translated BLAT : human/fugu, human/zebrafish, human/chicken, drosophila/anopheles, drosophila/honey bee, anopheles/honey bee, elegans/briggsae

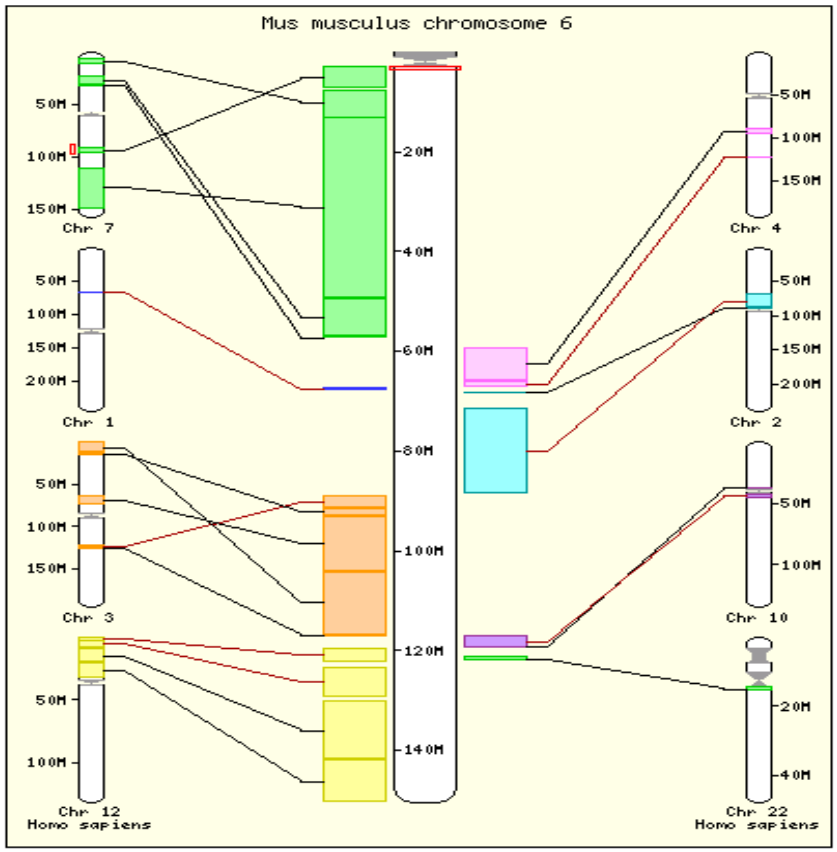
DNA/DNA matches web display



Defining large syntenic regions

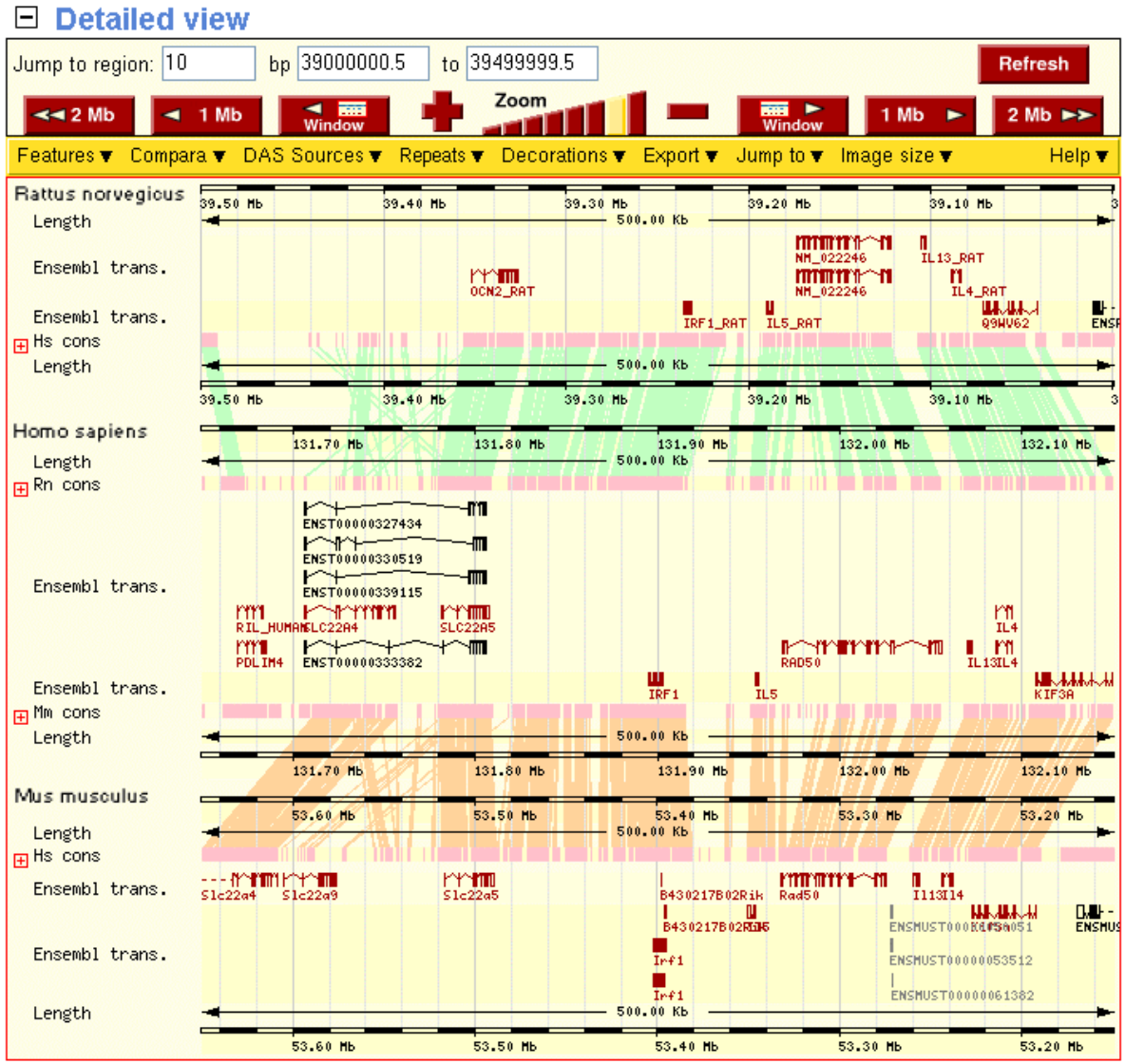
- genome alignments are refined into large syntenic regions.
- Alignments are clustered together when the relative distance between them is less than 100kb and order and orientation are consistent.
- Any clusters less than 100kb are discarded.

Synteny web display

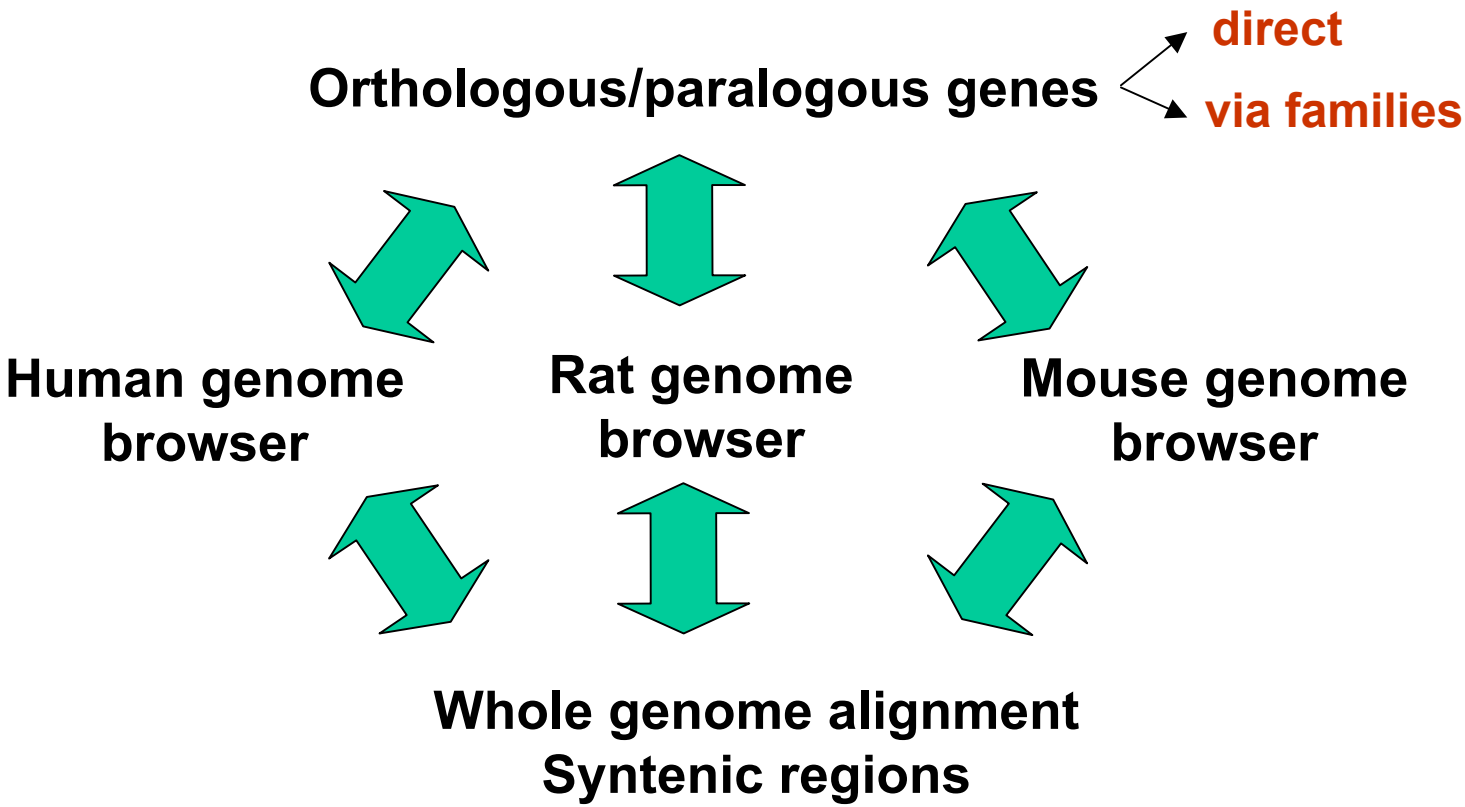


- 347 syntenic regions
- Coverage
  - 87.5% human
  - 92.4% mouse
- Size range
  - human
  - [104.4Kb - 57.3Mb]
  - mouse
  - [100.2Kb - 51.4Mb]

Multi-species display



Integrated multigenome browser



Code and data fully accessible

Website
www.ensembl.org

MySQL server
mysql -h ensembldb.ensembl.org -u anonymous

CVS repository
See documentation section and tutorial
at www.ensembl.org

Mailing list and user support
ensembl-dev@ebi.ac.uk
HelpDesk

Ewan Birney (EBI), Tim Hubbard (Sanger)

Pipeline/Genebuild
Val Curwen
Steve Searle
Vivek Iyer
Laura Clarke
Simon Potter
Dan Andrews

Zebrafish
Kerstin Jekosch
Mario Caccamo

Anopheles
Martin Hammond

Data Mining
Arek Kasprzyk
Damien Keefe
Damien Smedley
Darin London
Craig Melsopp

PhD students
Laurence Ettwiller
Ben Paten
Michael Hoffman

Core API and schema
Arne Stabenau
Graham Cameron
Glenn Proctor
Ian Longden

Exonerate
Guy Slater

SNPs
Yuan Chen
Hekki Lehvaslaiho

Helpdesk
Xose Fernandez-Suarez
Michael Schuster

Vega
James Gilbert
Stephen Keenan

Comparative
Abel Ureta-Vidal
Cara Woodwark
Jessica Severin

Systems
Tim Cutts
Guy Coates

Web team
Jim Stalker
James Smith
Brian Gibbins
Will Spooner

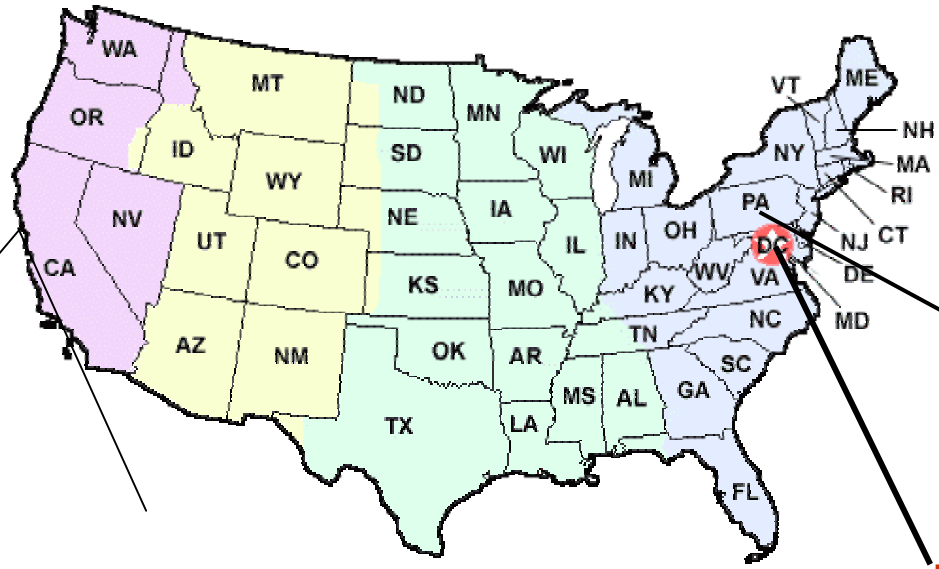
DAS
Tony Cox
Andreas Kahari

UCSC
Jim Kent

Past members
Michele Clamp, James Cuff, Emmanuel Mongin

Who, where, what?

Ensembl
Cambridge, UK



Penn State
(Blastz/Multiz)

NCBI
MapViewer

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

UC Berkeley (AVID/MAVID)

LBNL (VISTA viewer)

LLNL (dcode.org)

Stanford University (LAGAN/MLAGAN)

UCSC

URLs

Genome browser and visualization tools
http://www.ensembl.org
http://genome.ucsc.edu
http://www.dcode.org/
http://gsd.lbl.gov/vista/index.shtml
http://hanuman.math.berkeley.edu/cgi-bin/kbrowser2

BLASTZ/MultiPipMaker/Multiz/TBA
http://www.bx.psu.edu/

LAGAN/MLAGAN
http://lagan.stanford.edu/lagan_web/index.shtml

AVID/MAVID
http://gsd.lbl.gov/vista/mvista/download.shtml
http://baboon.math.berkeley.edu/mavid/

Other URLs can be found in a review now a slightly out of date as the field is evolving so fast.
Ureta-Vidal A et al. "Comparative genomics: genome-wide analysis in metazoan eukaryotes" 2003 Nature Reviews Genetics 4, 251-262.

References (1)

BLASTZ
Schwartz S at al "Human-mouse alignments with BLASTZ" GR 2003, 13,103-107.
used in
Waterston, RH et al. "Initial sequencing and comparative analysis of the mouse genome" Nature 2002, 420, 520-562.
Kent, WJ et al. "Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes" PNAS 2003, 100, 11484-11489.

MultiPipMaker
Schwartz S et al. "MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences" NAR 2003, 31, 3518-3524.
used in
Thomas JW et al. "Comparative analyses of multi-species sequences from targeted genomic regions" Nature 2003, 424, 788-793.
Margulies EH et al. "Identification and characterization of multi-species conserved sequences" GR 2003, 13, 2507-2518.

Multiz/TBA
Blanchette M et al. "Aligning multiple genomic sequences with the Threaded Blockset Aligner" GR 2004, 14, 708-715.
used in
Rat Genome Sequencing Project Consortium "Genome sequence of the Brown Norway rat yields insights" into mammalian evolution. Nature 2004, 428, 493-521.

References (2)

LAGAN/MLAGAN

Brudno M et al. "LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA" GR 2003, 13, 721-731.

Brudno M. et al. "Glocal alignment: finding rearrangment during alignment" Bioinformatics 2003, 19, S1, i54-i62.

Brudno M. et al. "Automated whole-genome multiple alignment of rat, mouse, and human" GR 2004, 14, 685-692.

AVID/MAVID

Bray N et al. "AVID: a global alignment program" GR 2003, 13, 97-102.

Couronne O et al. "Strategies and tools for whole-genome alignments" GR 2003, 13, 73-80.

Bray N et al. "MAVID multiple alignment server" NAR 2003, 31, 3525-3526.

Bray N et al. "MAVID: constrained ancestral alignment of multiple sequences" GR 2004, 14, 693-699.

Chakrabarti K et al. "Visualization of multiple genome annotations and alignments with the K-browser" GR 2004, 14, 716-720.

OTHERS

Delcher AL et al. "Alignment of whole genomes" NAR 1999, 27, 2369-2376.

Delcher AL et al. "Fast algorithms for large-scale genome alignment and comparison" NAR 2002, 30, 2478-2483.

Kent WJ "BLAT - the BLAST-like alignment tool" GR 2002, 12, 656-664.

Ning Z et al. "SSAHA: a fast search method for large DNA databases" GR 2001, 11, 1725-1729.

Ma, B et al. "PatternHunter: faster and more sensitive homology search" Bioinformatics 2002, 18,440-445.

Kent WJ et al. "Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment" GR 2000, 10, 1115-25.