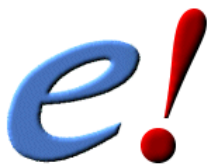


e!

Ensembl

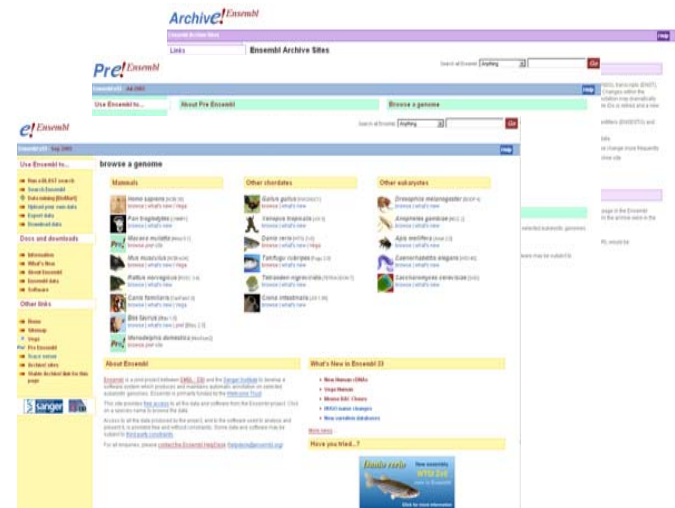
# Comparative genomics Ensembl

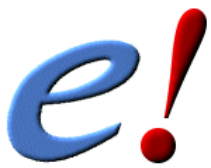


Ensembl

# Overview

- Rationale
- Species available
- Comparative genomics
- Genome-wide DNA alignments
- Outlook

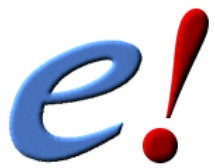




# Comparing different species

- From the Ensembl perspective joins species through
  - chromosome synteny links
- From a broader perspective
  - Where are syntenic regions located?
  - How many genes are conserved?
  - Is gene order conserved?

Ensembl



# Compara

The *Compara* database is one single multispecies database

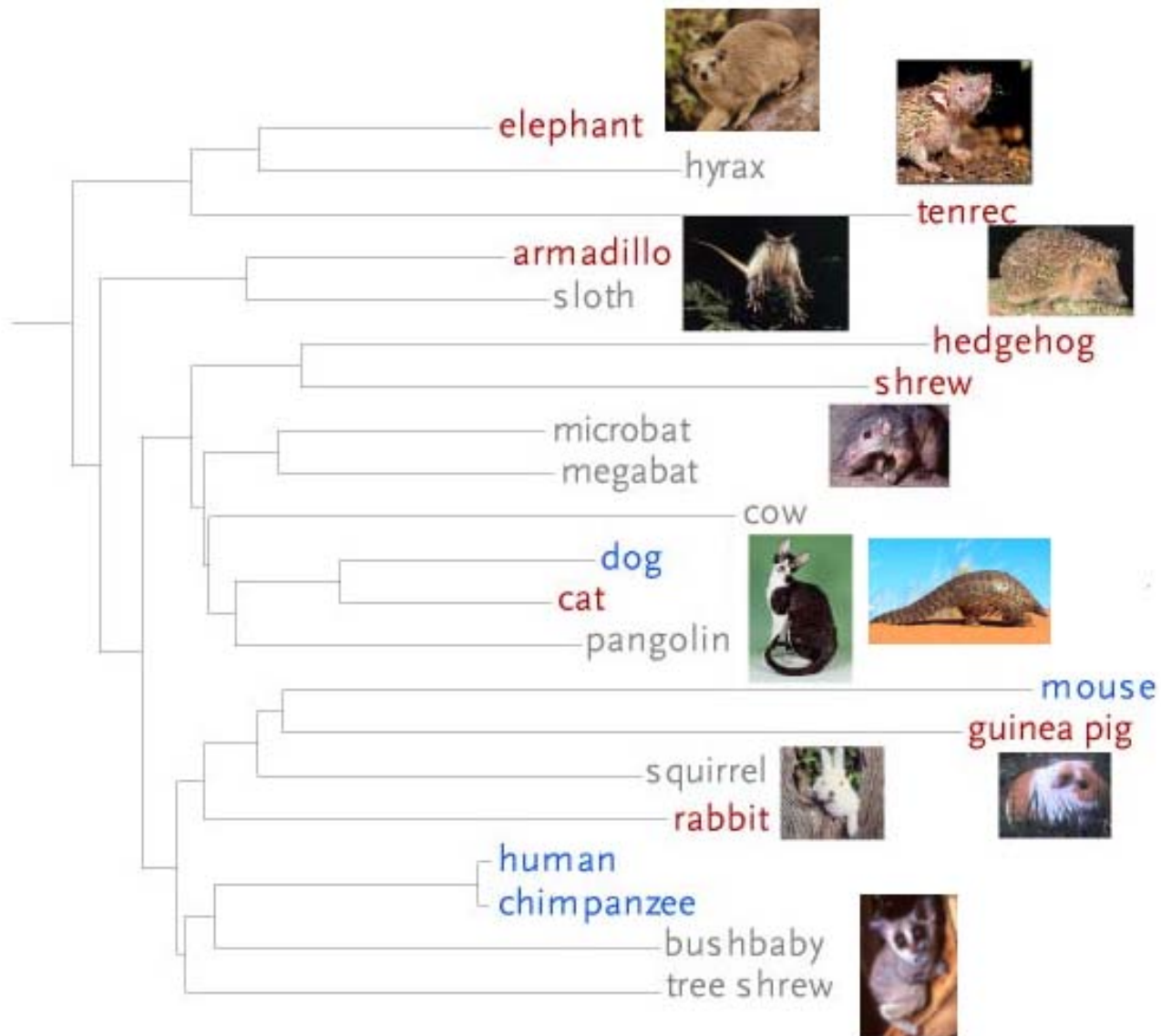
- Gene orthology/paralogy prediction
- Protein clustering
- Whole genome alignments
- Synteny regions

ENSEMBL  
ebi

e!

Ensembl

# Comparing different species

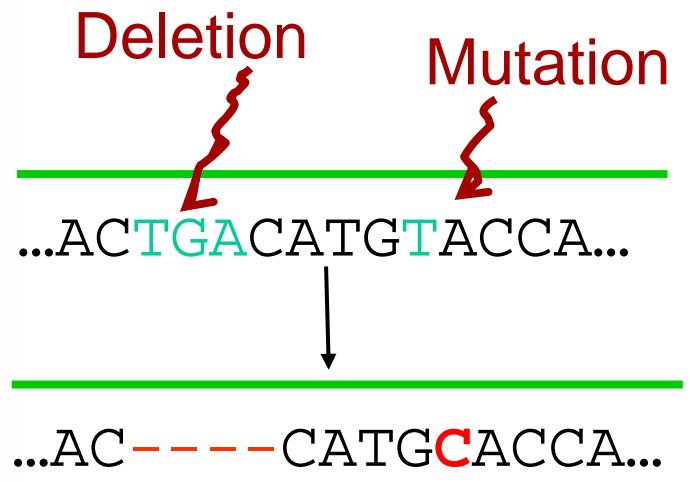


e!

Ensembl

# Aligning complete genomes

# Evolution at the DNA level



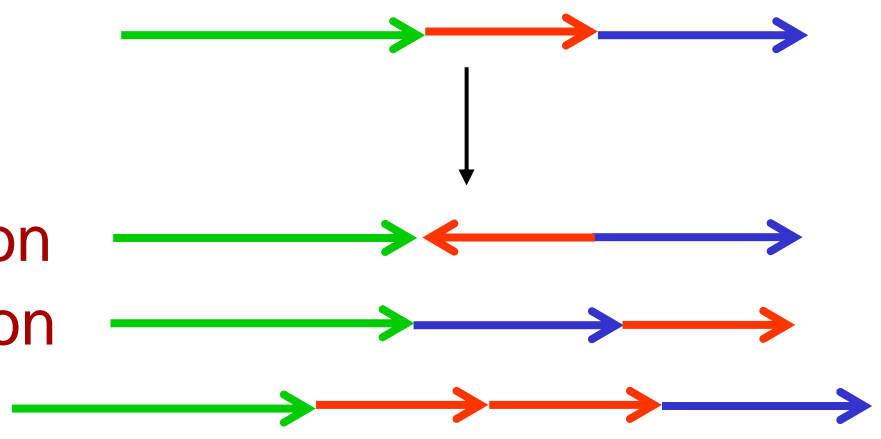
Sequence edits

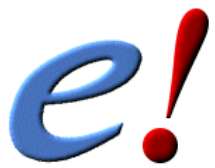
Rearrangements

Inversion

Translocation

Duplication

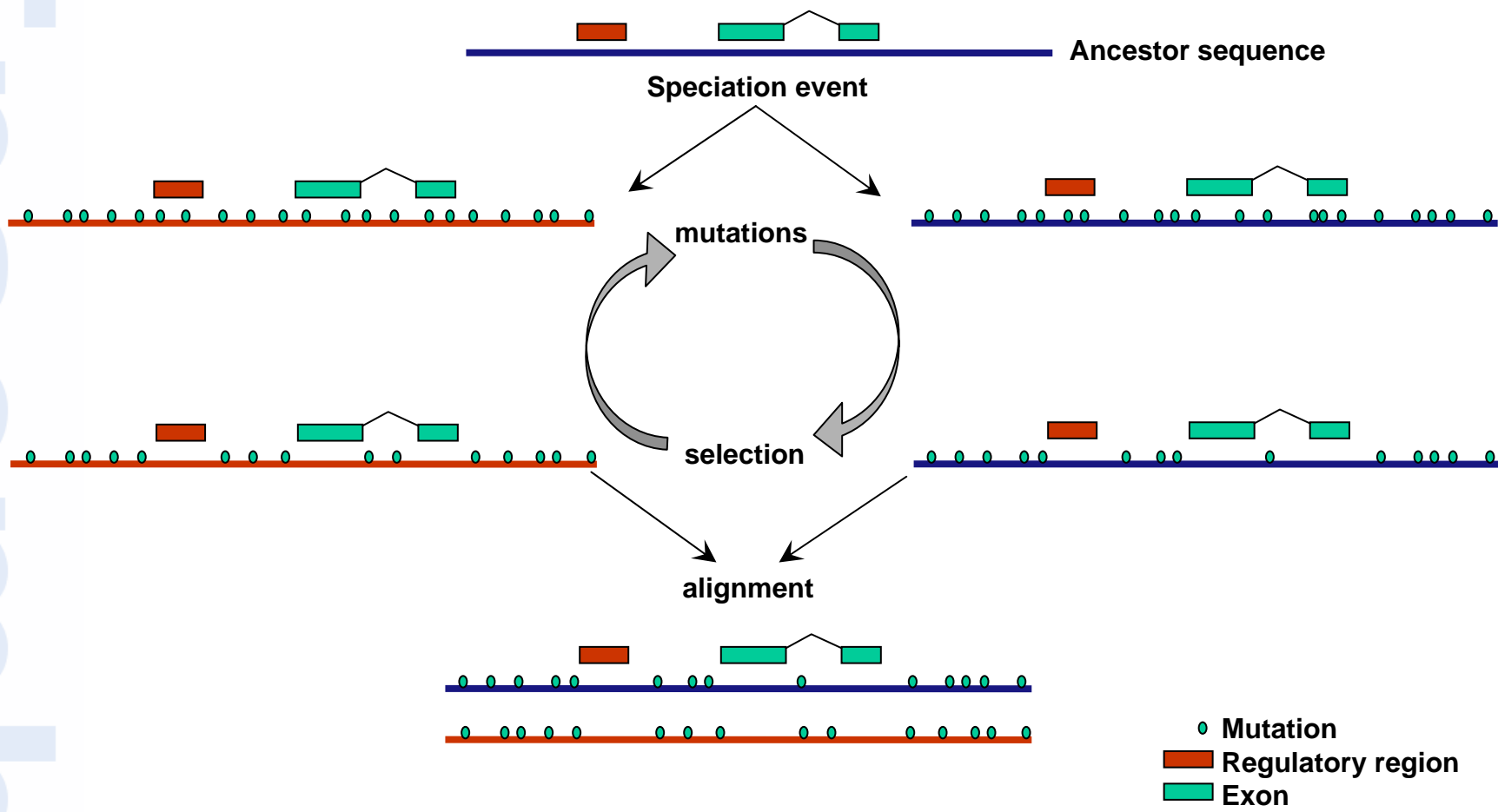




# Aligning genomes, why?

- Understand what evolution has done on the species compared, after their speciation
- Define syntenic regions, those long regions of DNA sequences where order and orientation is 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/monkey), may help understanding speciation

# Basic ideas

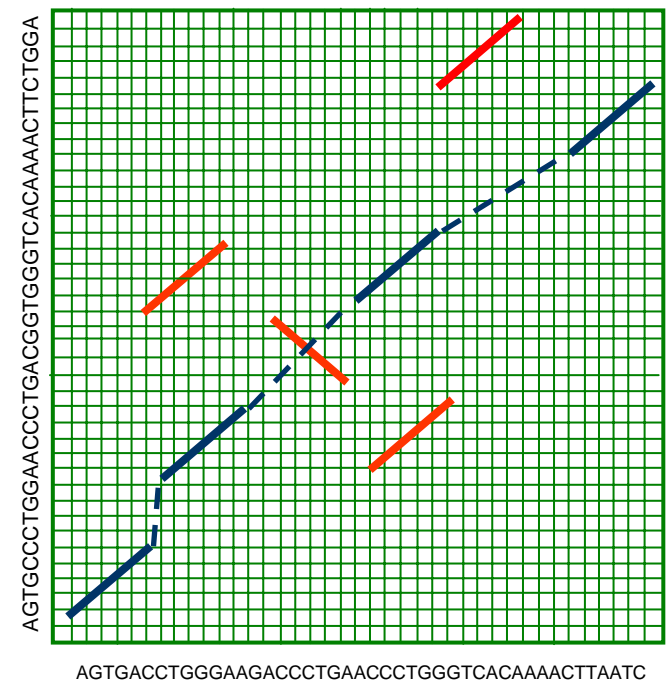
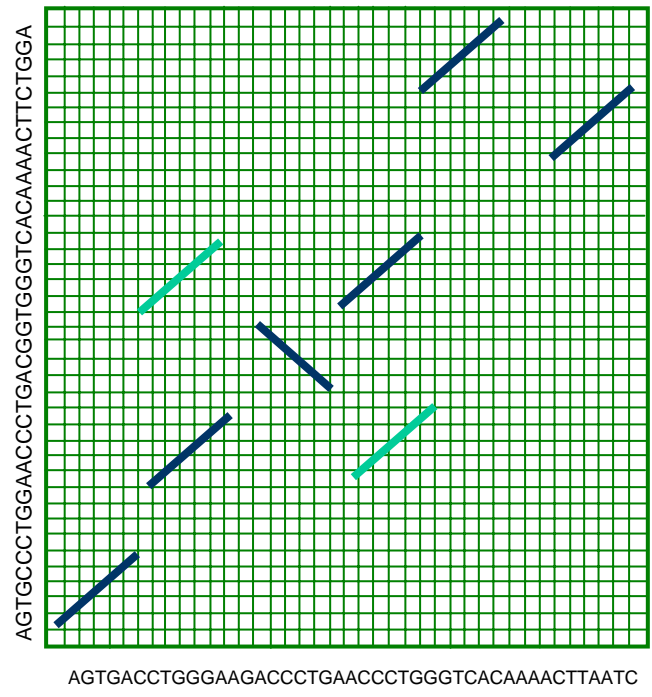


EMBL  
Sanger  
Institute

# 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 v Global Alignment



	Local	Global
Advantages	<p>Compares large genomic regions (requires syntenic maps)</p> <p>Can detect, rearrangements like <b>translocations, inversions and duplications (!)</b></p>	<p>Detects <b>insertions and deletions</b></p>
Disadvantages	<p>Fails to identify <b>insertions or deletions</b></p>	<p>Fails to detect rearrangements (<b>inversions</b>)</p>



e!

# 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

EMBL  
Sanger  
institute

e!

Ensembl

## **all *versus* all approach using BLASTZ (collaboration with UCSC)**

- Can handle large sequences
- Used 2-weighted spaced seeding strategy
- Dynamic masking
- Makes distinction between repeat and non-repeat sequences (soft masking)
- Try aligning inside repeats
- One iterative step with lower threshold to expand alignments

e!

ENSEMBL

# Blastz strategy

- 10Mb Human fragments (3000)
- 30Mb Mouse fragments (100)
- Lineage-specific repeats removed
  
- 48 hours on 1024 CPUs
  
- Generates 9Gb of output
  
- When filtered for Best hit on Human, reduced to 2.5Gb
- 10Mb Human fragments (3000)
- 30Mb Mouse fragments (100)

e!

ENSEMBL

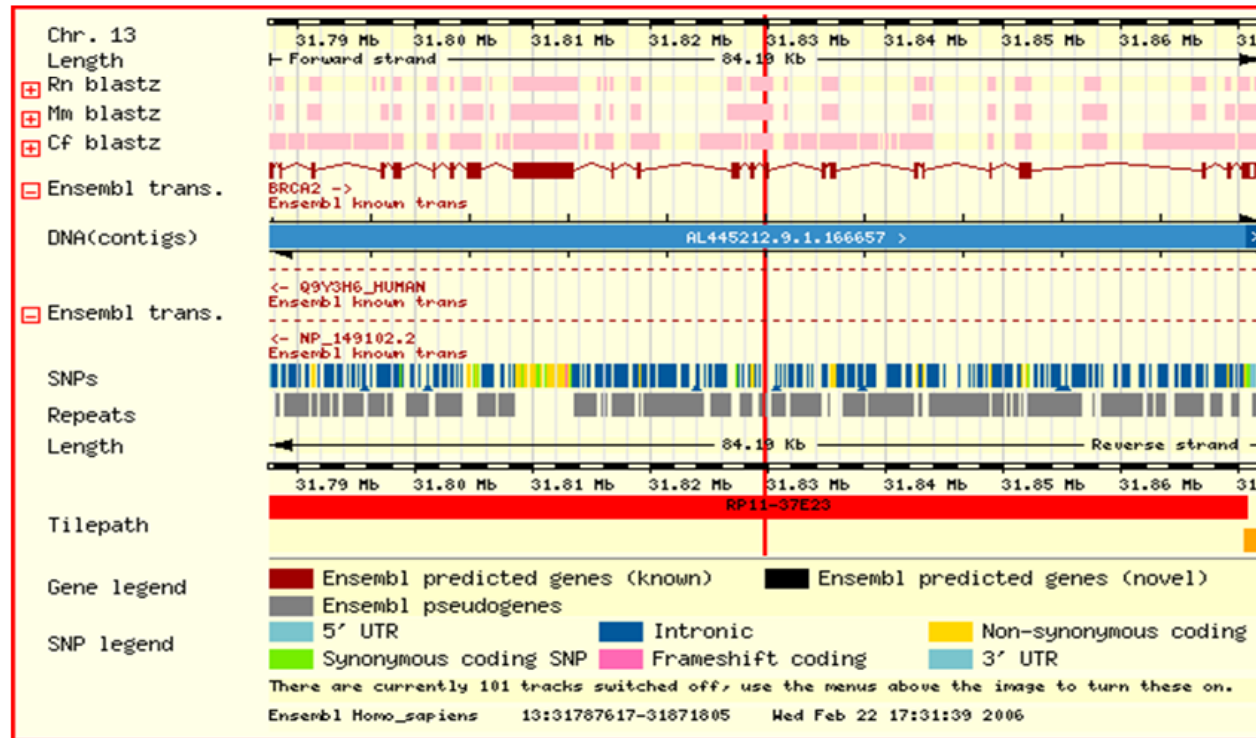
# Blastz human genome coverage

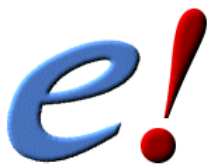
- 40% of the human genome is covered by an alignment of mouse sequences

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%

# DNA/DNA matches web display

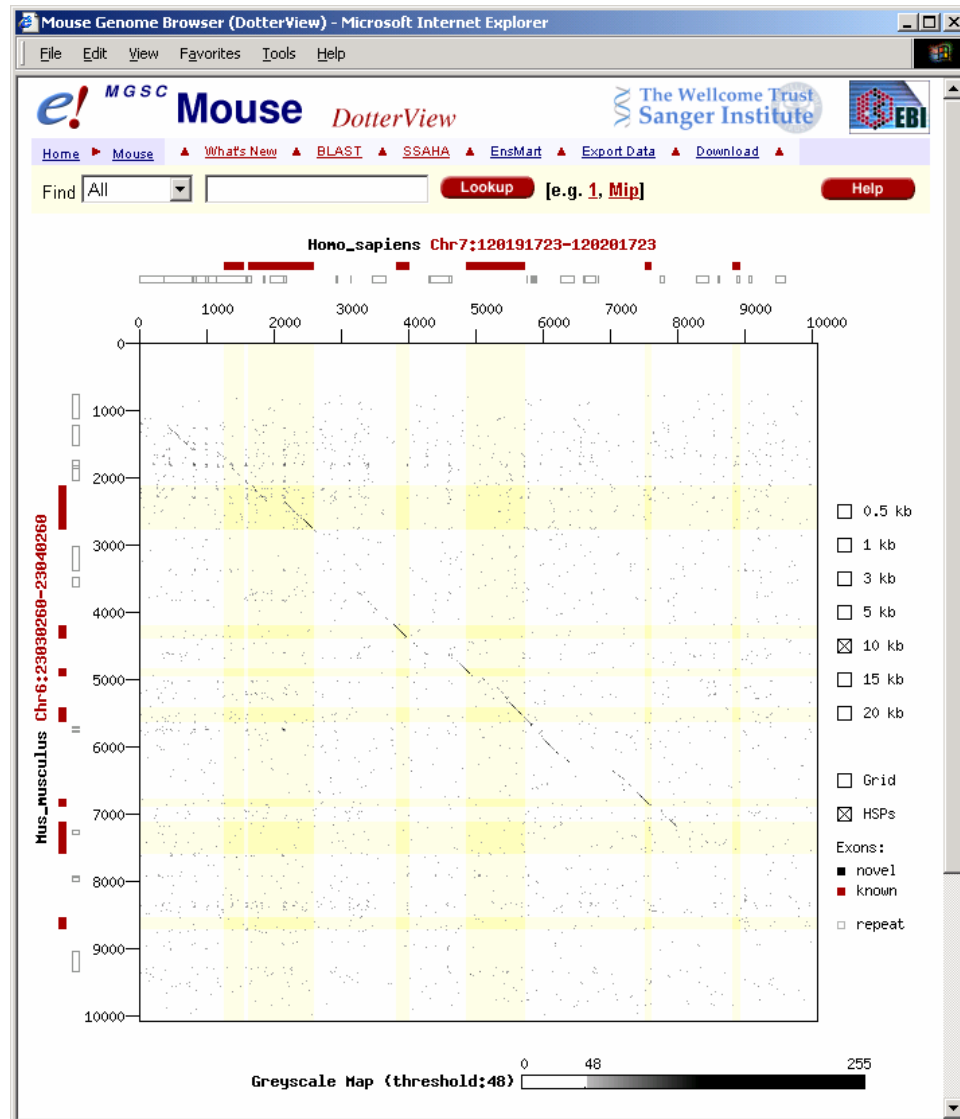
## ContigView human BRCA2



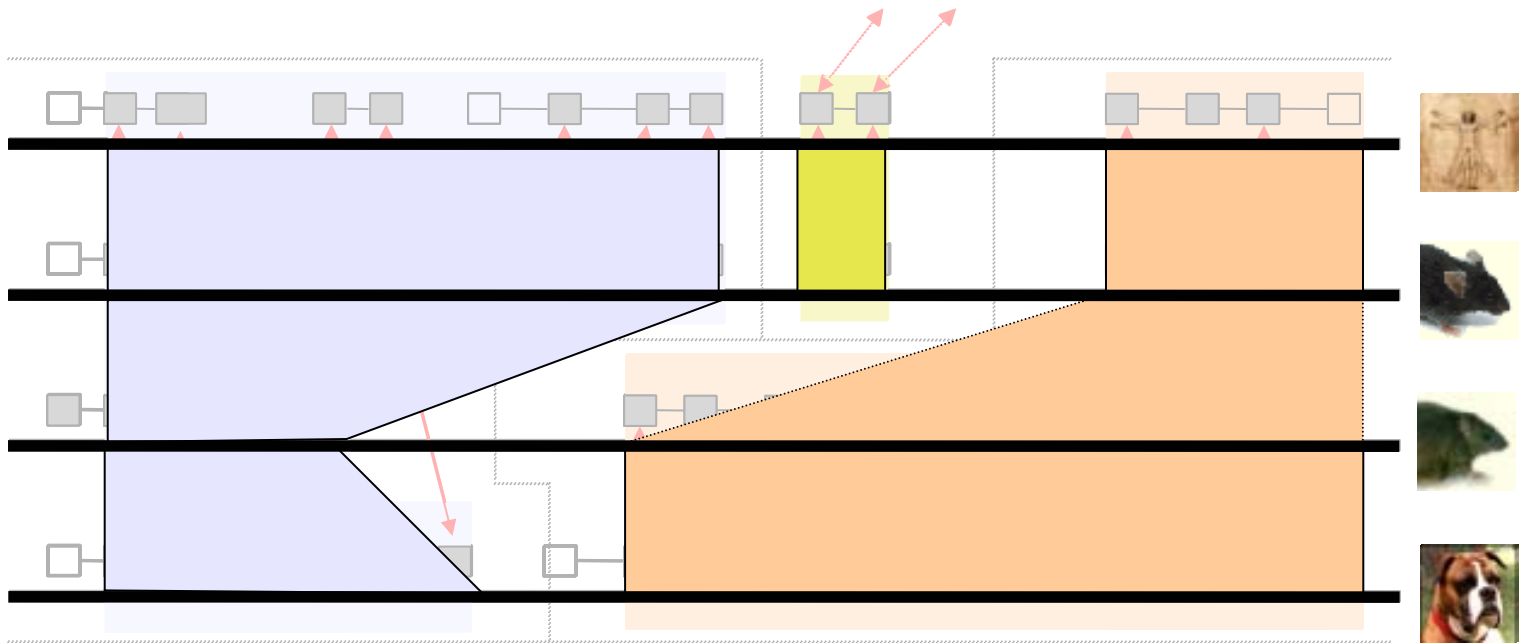


# DotterView

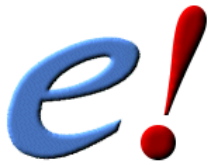
ensembl



# Strategy



- Use all coding exons
- Get sets of best reciprocal hits
- Create orthology maps
- **Build multiple global alignments**



EMBL

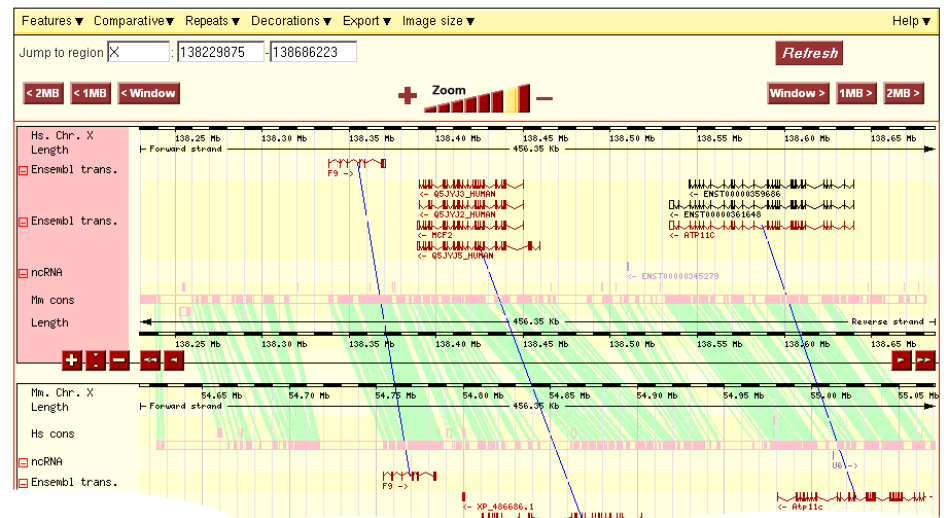
# MultiContigView

Chromosome X  
138,229,875 - 138,686,223

- View of Chromosome X
- Graphical view
- Graphical overview
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region
- View alignment
- View S...
- View r...
- View r...
- View r...
- Use Ensembl
- Run a...
- Search...
- Data n...
- Upload...
- Downl...
- Export...



## Detailed View



e!

Ensembl

# Multiple alignments

- Currently 2 sets:
  - MLAGAN-mammals:

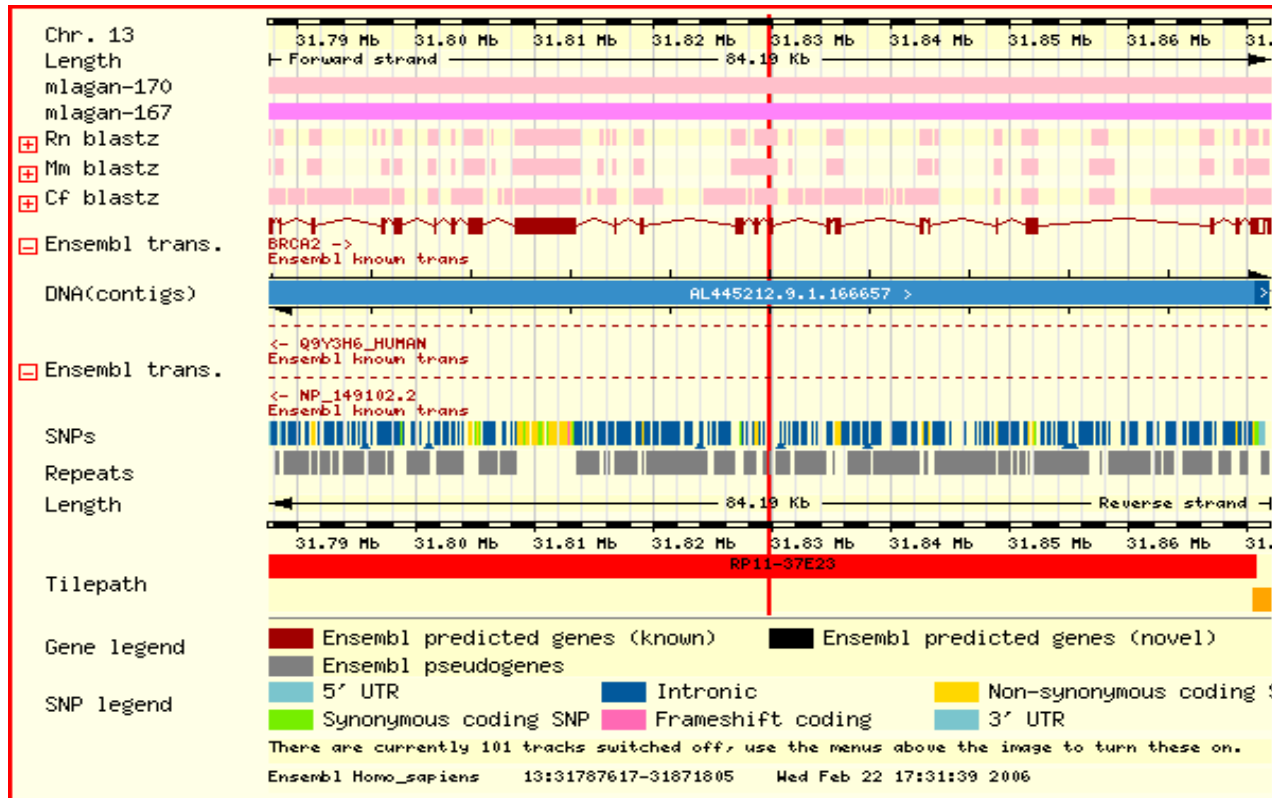


- MLAGAN-vertebrates:



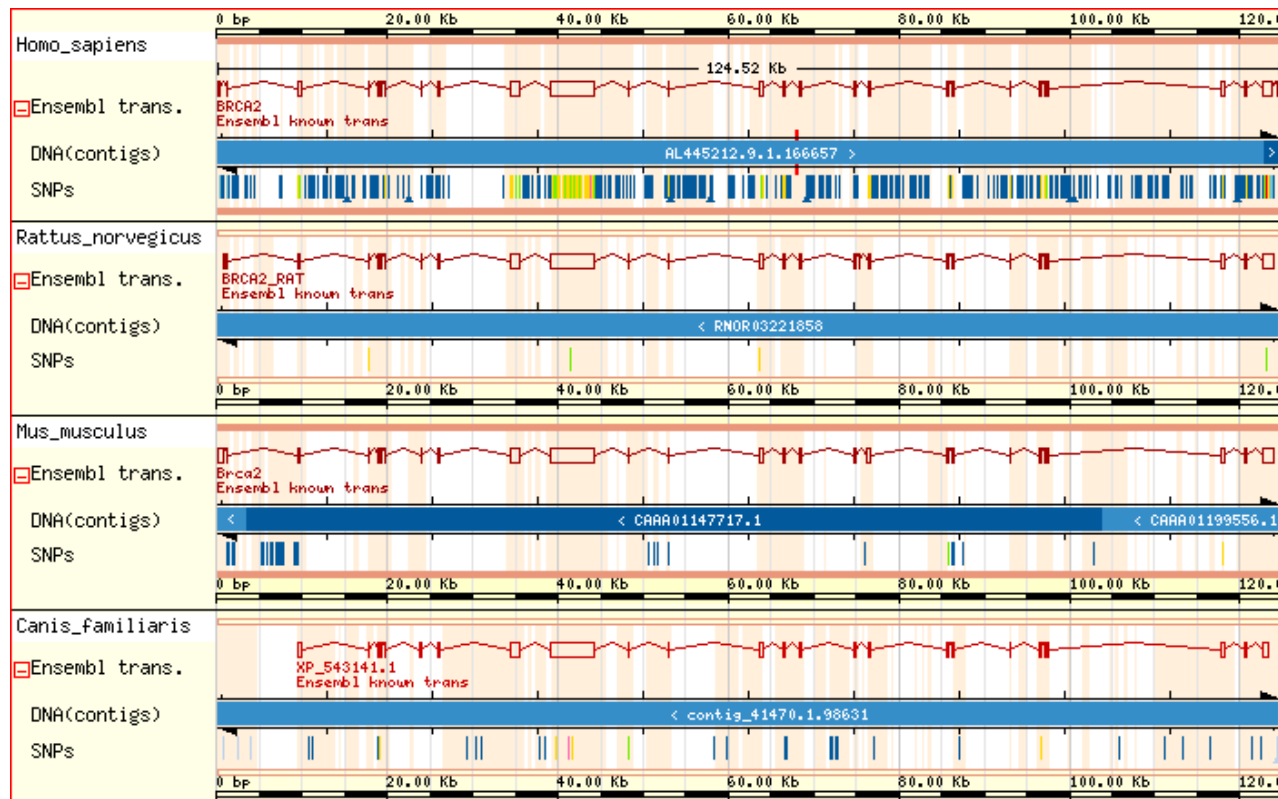
# Multiple alignments

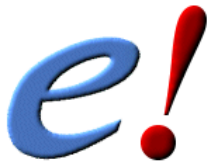
## ContigView human BRCA2



# Multiple alignments

## AlignSliceView human BRCA2 v other mammals





# Multiple alignments

## GeneSeqalign View human BRCA2 v other mammals

EMBL  
Sanger  
Bioinformatics

```

H= GGCTAGTCTCGAACTCCTGGGCTCAAGCAGTCTTCTGCTCAGCCTCCAAAAGTCTGAGATTACAGCCATGAGCCACTGTGCCAAA
Rn -----CTGTACTCTTTATTTTAAA
Mn -----CGAACTCACAGAGATCCGCTGCTTCTGCTCCAAAGTCTGGGAGCGCCACCACCTGTACTCTTTATTTTAAA
C= -----ATGTAACTCTTGCTTCAGA

H= CACTACCTTTTAACTTAGTGA AAAATATTAGTGAATGTGTTGTTGTTACTTTAATTTTGTCACTTTGTGTTTTATGTTTATAGCTTAA
Rn CACTCTT-----AACTTGGGAAAAGCACTGAGTAGATG-----CGTACTCTAACTTTGTCACTCTGTATTCTTATGCTTAGCTATAA
Mn CACTAT-----AACTTGGGAAAAGCACTGAGTGGATGCAMTTAATCGTACTCTAACTTTATCATCTGTGTTTTATGCTATAGCTATAA
C= TATTTT-----TACTTAGTGA AAAATACTTAGTGAATGTGATTGATGGCAGCTTAATTT-GTCACTTTGTCTTTCATGCTTAGCTTAA

H= TTGCATTCTTCTGTGAAAAGAGCTGTTTACAGAAATGATTTCTGAAGAACCAACTTTGTCCTTAACTAGCTCTTTTGGGCAAACTTCTGAGG
Rn CCAGATTCTTCAATCAAAAAGAGCAATTTACCAAAATGATCCTGAAGAGCCACTTTTGTCTTGAACCAACTCTTTTGTGACTGC-----
Mn CCAGATTCTTCTGACAAAAGAGCTGTTTACCAAAATGATCCTGAAGAGCCA-----TCTTGAACCAACTCTTTTGGGACTGC-----
C= TTGCATTCTTCTATCAAAAAGAGCTGTTTACAGAAATGATTTCTGAAGAACCAACTTTGTCCTTAACTAGCTCTTTTGGGCAAACTTCTGAGA

H= AAATGTTCTAGAATGAACATGTTCTAATAATACAGTAMTCTCTCAGGATCTTGATTAAAGAAGCAAAATGTAAAGGAAAAAAGT
Rn -----TGCCAGTAAAGAAATAGTTATATTCAATGCATTGATATCTCAGGATCTAATGACAAGAAGCAATACTCAGTGAAGAAAAAGCA
Mn -----TACCAGTAAAGAAATAGTTATATTCAATGCATTGATATCTCAGGATCTAATGACAAGAAGCAATAGTCAATGAAGAAAAAAGCA
C= AAAGTTCCAGTAAAGGAGCAATGTTAATAAATAAATATCTCAGGATCTGATTAAAGAAGCAAAATATGAAGAAAAAAGT

H= CAGTATTATTACCCAGAAGCTGATTTCTGTCTGCTGCAGGAAGGACAGTGTGAAAATGATCAAAAAGCAAAAAGTTTCAGAT
Rn CAGCCATATACGCTCCAGAAGCTGATTTCTGTCTGCTGCAGGAAGATCATGTGAAAATGATCAAAAAGTCAAAAAGTTTCCGAC
Mn CAGCCATATACGCTCCAGAAGCTGATTTCTGTATGCTTGCAGAAAGAACATGTGAAAATGATCAAAAAGTCAAAAAGTTTCCAMT
C= CAGTATTATTACCCAGAAGCTGATTTCTGTCTGCTGCAGGAAGAACATTTGGGAAGATGATGCAAAAAGCAAAAAGTTTTCAGAT

H= ATAAAAGGAGAGCTTGGCTGCAGCATGTCACCCAGTACAACATCAAAAGTGAATACAGTGAATCACTGACTTTCAATCCAGAAAAGT
Rn CGAAAAAGAAAAGTCTTAGTCTCAGCATGTCGCTCCTTCAGGAAGGGCAGCAGTGCAGCTCAGCAGCATTTAGTCTTCACTCAGGAAGAC
Mn GGAAGAAGAAAAGTCTTAGTCTCAGCATGTCGCTCCTT-----CAGCAGTGCAGCTCAGCAGCATTTAGTCTTCAATCAGGAAGAAC
C= ATAAAAGGAAAAGTCTTGGCTGCAGCATGTCACCCAGTACAACATCAAAAGTGAATACAGTGAATCACTGACTTTCAATCCAGAAAAGT

H= CTTETATATGATCATGAAATGCAGCAGTCTTATTTAACTCCTACTTCCAAAGATGTTCTGTCAAACTAGTCACTGATTTAGAGGGC
Rn CTTCTTGGTACCCACCAAGCTAAGCAAGTACTCTTAAATAACTCCAGCCCGAAGACACTCTGTCAAAAGCAGTGTGTTTTCTAGAGGGG
Mn CTTCTTGGTACCCACCAAGCAAGTACTCTTAAATAACTCCAGCTCAAAAGTACTCTGTCAAAAGCAGCAGTGTGTTTTCTAGAGGAG
C= TTTTCAATGACTGTGATATACAG-----TCTGTTAACTCCTAGCTCTAGGATTTCTCCATCAGCTAGTGTGATGCTAGAGGAA

H= AAGGATCATACAAAATGTGAGCAGCTCAAGGTAACAAATATGATCTGATGTTGAATTAACGAAAATATTCATGGGAAAGAAAT
Rn AAAA---TCTGTAAATGCCAGAGAACTGCAATGTAAAGTGTAAAGATATATTGAATTAAGCAAAAATCACTCCTCGGGGGGTAAT
Mn AAAA---TCTGTAAATGCCAGAGAACTGCAATGTGAGCTGTAAAGTATATTGAATTAAGCAAAAATCACTCCTCGGGGGGTAAT
C= AAGGATCATATAAAATTCAGAGAAACTAAAATGTAAAGATCATGAAACTGTTTTGAATTAACGAAAATATTCATGGGAAAGAAAT

H= CAGGCTGTGTGCTTTAATGAAAATATAAAACGTTGAGCTGTTGCCACTGAAAATACATGAGATAGCACTCACCTCTAGAAAAG
Rn ---GAATCTGTGCTTAACTGAAAATTTCAAAACCTGAGCTCTGCCACTCTGAAATATATACAAAGAGTCCCTCTCAGTGAAG
Mn ---GAATATGTATCTTAACTGAAAATTTCAAAACTCTGGGCTTCTGCCACTGGTGAATATATACAAAGAGTCCCTCTCAGTGAAG
C= CAGGATCATATGTTTTAATGCAATTTCAAAAATGCTAACTGTTGTCAACTGAAAACATATACAGTAGCATCATCTTCAAGTAAAG

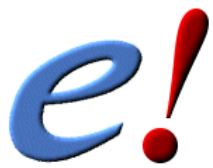
H= GTACAAATCAACCAAAACAAAATCTAAGACTAATCCAAAAAATCAAGAAAGAACTACTCAATTTCAAAAATAACTGTCAATCCAGAC
Rn TCACAGTTCAATCAAAATCAAAAATAGCACTGCTACAAAAGGACCAAAAAGACTCACTTTTATTTCAAGAAATCAACTATGAT
Mn TCACAGTTCAATCAAAATCAAAAAT---AATCATACAGAAAGACCAAAAAGGCTCACTTTTATTTCAAGAAATCAACTCAATGAA
C= GTTCAGTTCAACCAAAATCAAAATCTCACCACAAATCCAAAAGACCAAAAAGAACTACTTTAATTTCAAAAATAACTGTTAATCCAAAC

H= TCTGAAGAACTTTTCTCAGCAGTGAAGAAATATTTTTCTTCCAAAGTACTGATGAAGAAATATTTGCTTTAGGAAATATAGGAA
Rn TCTGAAGAACTTTTCCAGCAAGGAGAAATATTTTTGCTTTTCAAGTAACTAATGAAGCAATAAACCAATATAGGAACTACTGTGAA
Mn TCTGAAGAACTTTTCCAGCAAGTGGAAATATTTTTGCTTTTCAAGTAACTAATGAAGCAATAAAGCTGATAGGAACTCAAGTGAAG
C= TCTGAAGAACTTTTCCAGCAAGTGAAGAAATATTTTTCTTAAAGAACTAATGAAGAAATATACTCTGTTTTAGGAAATACTAGGAA

H= TTTCAATGAACAGACTTCACTCTGTAAACCAACTTTTCAAGAACTCACTCACTGTTTTATATGAGGACAGCTGATATAACAGGCA
Rn TTCGAAAGAAAGACTCTCCAGCACACAAAAGGCAATAGTCTCAAGAACTCTCCATGACAGTGAAGCACTAGTATGATCAGCAAGCA
Mn TTCGAAAGAAAGACTCTCCAGCACACAAAAGGCAATAGTCTCAAGAACTCTCCATGACAGTGAAGCAATAGTATGATGCGCATGCA
C= TTAATGATTTCAAACTCTGTTGTGAAGAACTCTGCTCAAGAACTCACTCACTGATAGTATGACAGACTCTGGATGACAAACAAAC

H= ACCCAAGTCTCAATTA AAAAAGATTTGTTTTATGTTCTTGCAGAGGAGAACAAAATACCTCAAAAGCAACTATAAAATCACTAGT
Rn GCCCAAGTCTGATTAACAGGAGACTTAACTCAATGATTAACAAAGGAGCAGAAAATACTATAGGCAAGCTCTAGAAAGCACTGCGAG
Mn GCCCAAGTCTGATTAACAGGAGACTTAACTCAATGATTAACAAAGGAGCAGAAAATACTATAGGCAAGCTCTAGAAAGCACTGCGAG
C= GCCAAAGTCTGATTAACAGGAGACTTAACTCAATGATTAACAAAGGAGCAGAAAATACTATAGGCAAGCTCTAGAAAGCACTGCGAG

```



Ensembl

# Acknowledgements

- **Abel Ureta-Vidal**
- **Javier Herrero**
- **Kathryn Beal**
- **Albert Vilella**
  
- **Ensembl team**