

e!

Ensembl

# Evaluating genes and transcripts in Ensembl



# *Outline*

**Ensembl gene set**

**Pseudogenes**

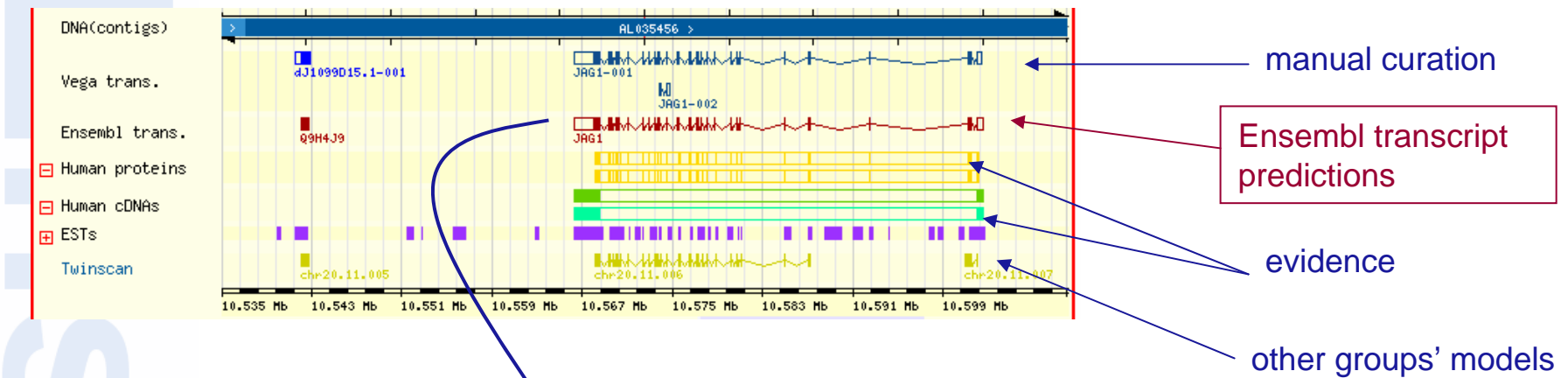
**Ensembl EST genes**

***Ab initio* predictions**

**Manual curation (Vega, CCDS)**

**Gene models from other groups**

# Overview



e! Ensembl Human GeneView

The Wellcome Trust Sanger Institute

Home Human What's New BLAST SSAHA EntMart Export Data Download Disease Browser Docs

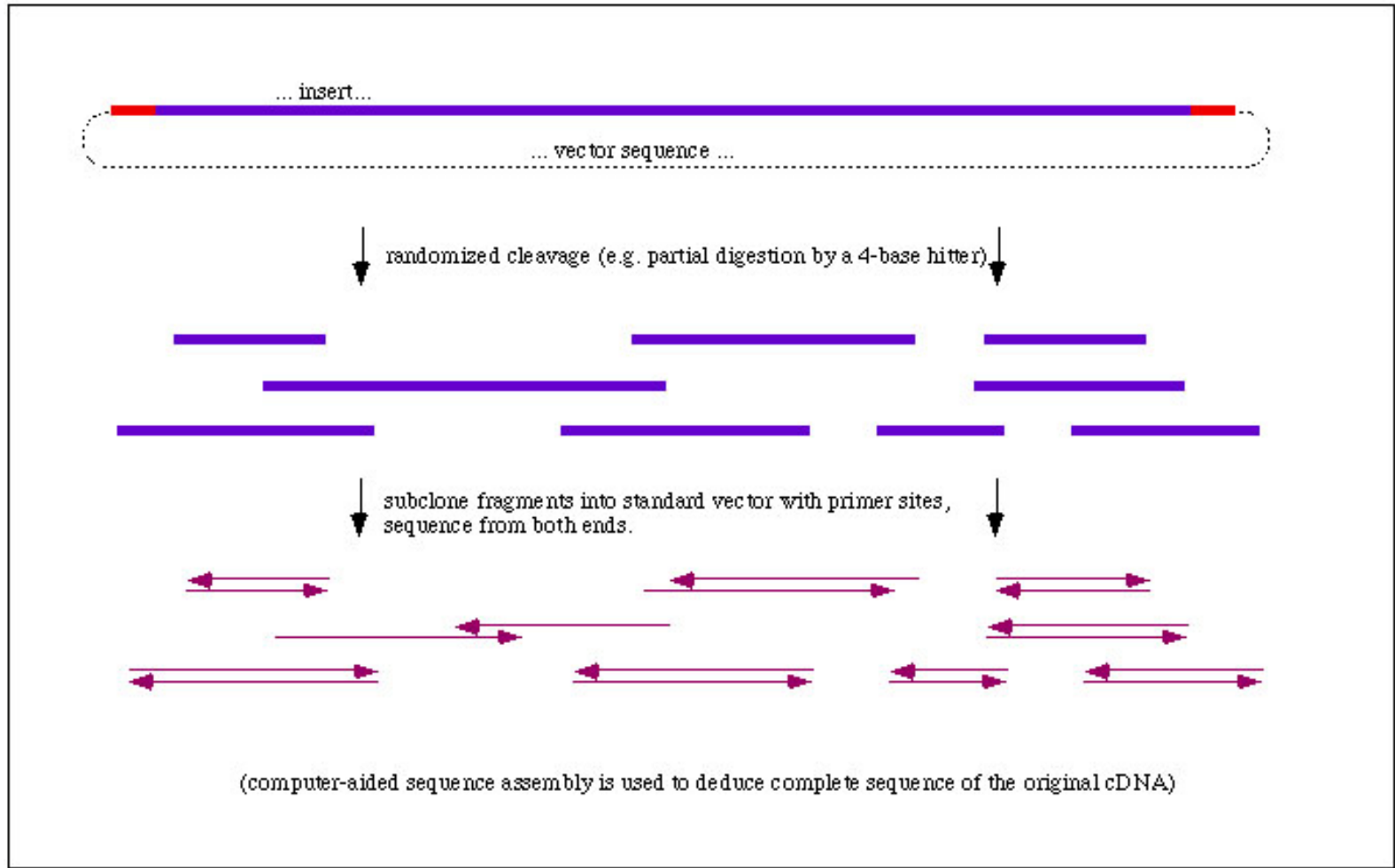
Find Gene   [e.g. ENSG00000139618, BRCA2]

### Ensembl Gene Report

<b>Gene</b>	JAG1 (HUGO ID)
<b>Ensembl Gene ID</b>	ENSG00000101384
<b>Genomic Location</b>	<b>View gene in genomic location:</b> <a href="#">10566334 - 10602636 bp (10.6 Mb)</a> on chromosome 20 <b>This gene is located in sequence:</b> <a href="#">AL035456.26.1.125952</a>
<b>Description</b>	JAGGED 1 PRECURSOR (JAGGED1) (HJ1). <a href="#">[Source: SWISSPROT (P78604)]</a>
<b>Prediction Method</b>	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomeWise for ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
<b>Predicted Transcripts</b>	1: JAG1 - <a href="#">[View transcript info]</a> <a href="#">[View exon info]</a> <a href="#">[View protein info]</a> (ENST00000254959)

sanger

# Gene Sequencing





# Traces

## e! Ensembl Trace Server

Find trace:  **Go**

e.g. [ml1B-a1798c05.etc\\_QTMZP1D038'](#)

### Trace Server

- Home page
- View trace statistics
- Search trace sequences
- Download traces (FTP)

### Links

- Ensembl
- NCBI trace archive

### Ensembl Trace Server

The Ensembl trace repository provides a permanent archive for single-pass DNA sequencing reads and associated traces and quality values. These data come from whole-genome shotgun projects, EST projects, and other large-scale sequencing projects. Data is regularly exchanged with the [NCBI](#) trace archive.

Current services include the ability to examine individual reads by name, to search the whole archive or subsets of the archive with another DNA sequence using [SSAHA](#) (a fast sequence search) and to download sets of read sequences in fasta format and associated quality values by [FTP](#).

Requests for large data sets of complete trace information to be sent by tape should be made to [trace-request@ensembl.org](mailto:trace-request@ensembl.org)

### Trace Totals

The trace repository currently contains **1,077,149,386** traces from 745 species. [\[trace statistics\]](#)

## Electropherogram

View Options

Help

Goto base

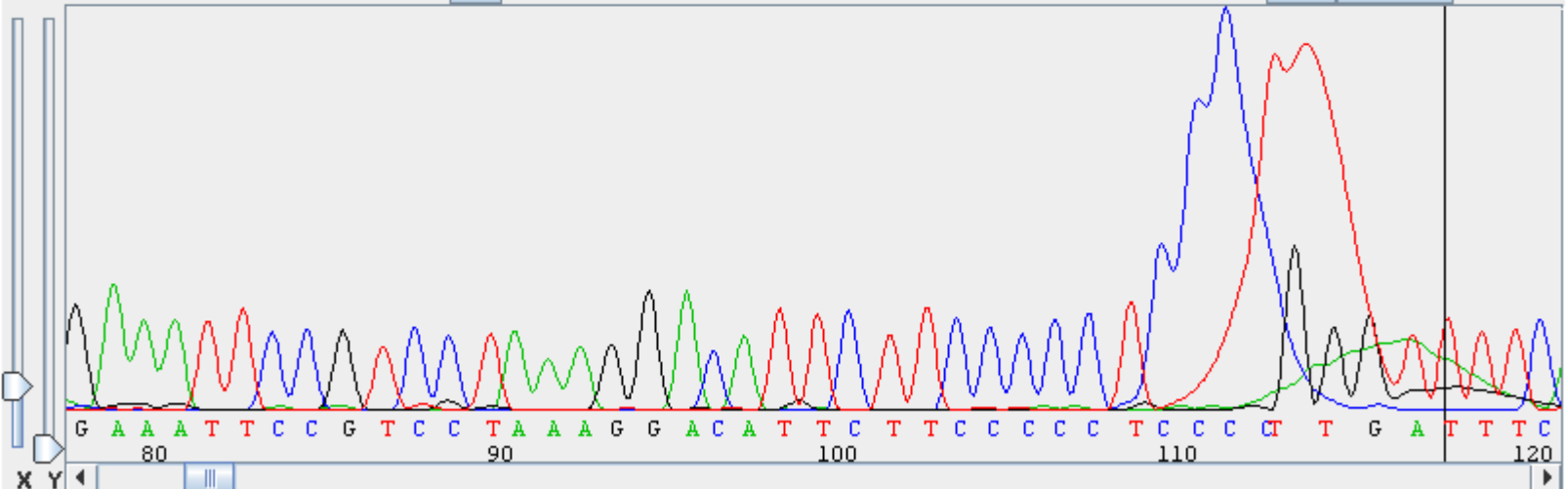
**OK**

Search

**Next**

**Previous**

**51025108**



X Y

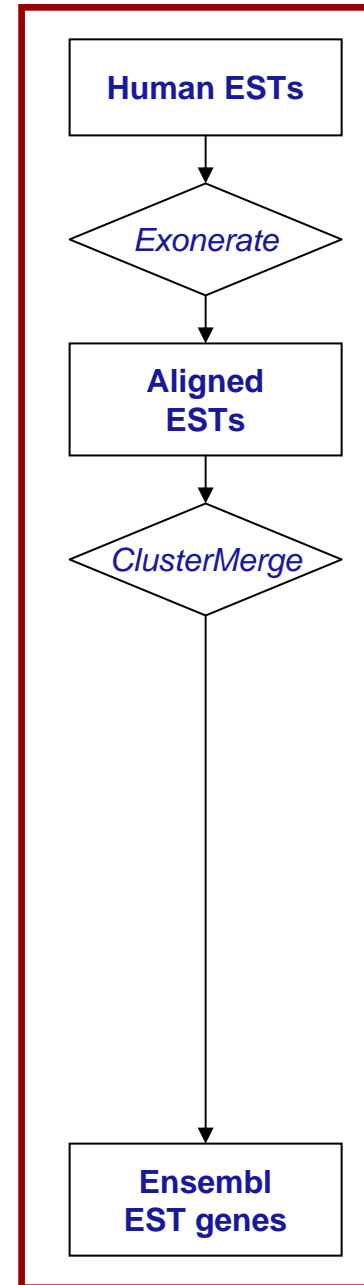
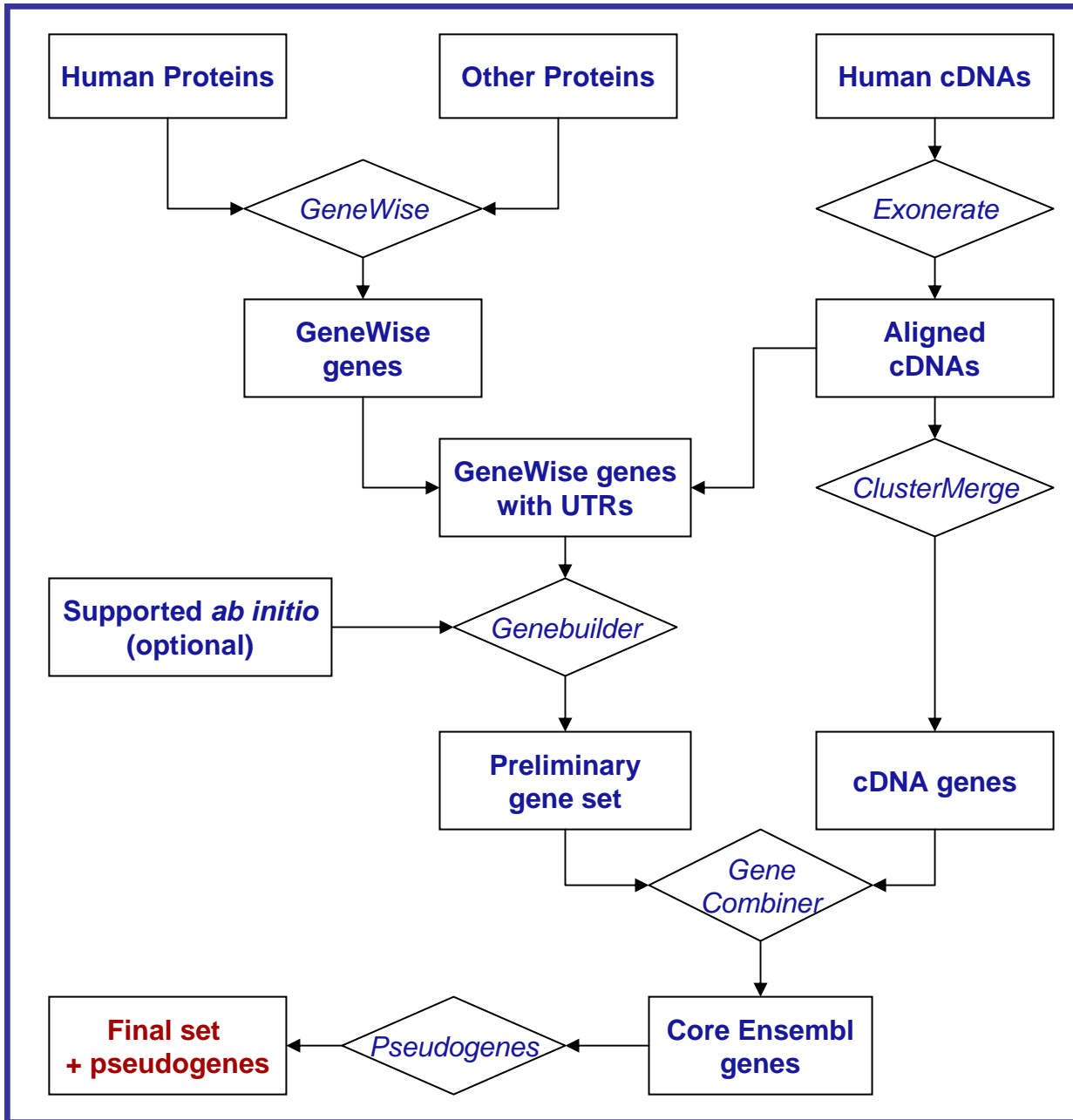
Ensembl

# ***Annotation process***

- **Automated analysis**
  - **Repeat masking**
    - RepeatMasker (Smit), tandem, inverted
  - **Gene prediction**
    - Genscan (Burge), FGENESH (Solovyev)...
  - **Database searches**
    - initial protein and DNA matches using BLAST
    - refined protein matches using *GeneWise*
    - refined EST matches using *EST2GENOME*, spangle
    - *Pfam* annotation using *GeneWise*.



Ensembl





# *The Ensembl Gene Build*

**Align** species-specific proteins

**Align** similar proteins from **closely related species**

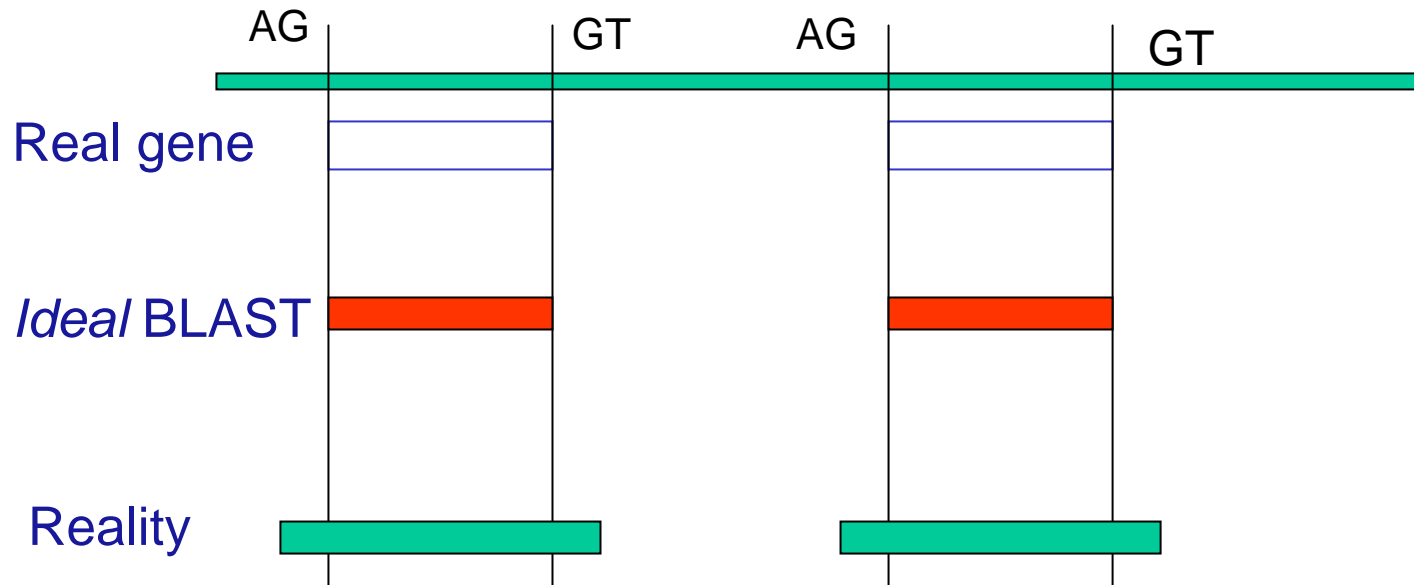
**Use** mRNA information to add UTRs

**Build** transcripts using mRNA evidence

**Build** additional transcripts using *ab initio* predictors and homology evidence

**Combine** annotations to make genes with alternative transcripts

# *The trouble with BLAST*



**BLAST is good for finding possible exon positions  
In large genomic sequences.**

# ***BLAST 'replacements'***

## **Exonerate\* (Guy Slater)**

**Fast gapped DNA-DNA matcher  
10,000 x faster than BLAST**

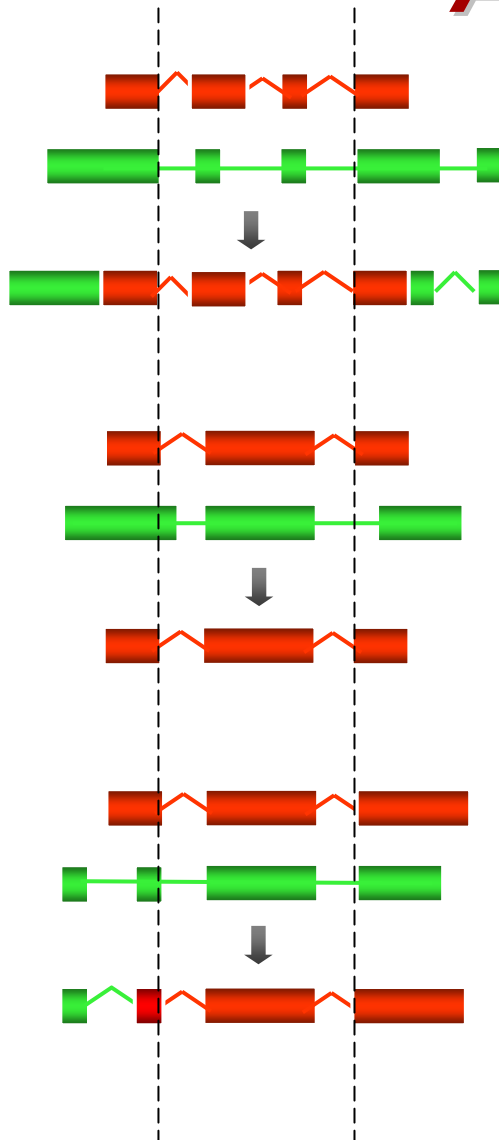
## **Pmatch (Richard Durbin)**

**Fast exact protein-dna matcher  
>10,000 x faster than BLAST**

\*BMC Bioinformatics 6: 31 (2005)



# Adding UTRs



protein - GeneWise (phases, no UTRs)  
cDNA - exonerate (UTRs, no phases)

Combined prediction

protein - GeneWise (phases, no UTRs)  
cDNA - exonerate (UTRs, no phases)

GeneWise prediction

protein - GeneWise (phases, no UTRs)  
cDNA - exonerate (UTRs, no phases)

GeneWise prediction

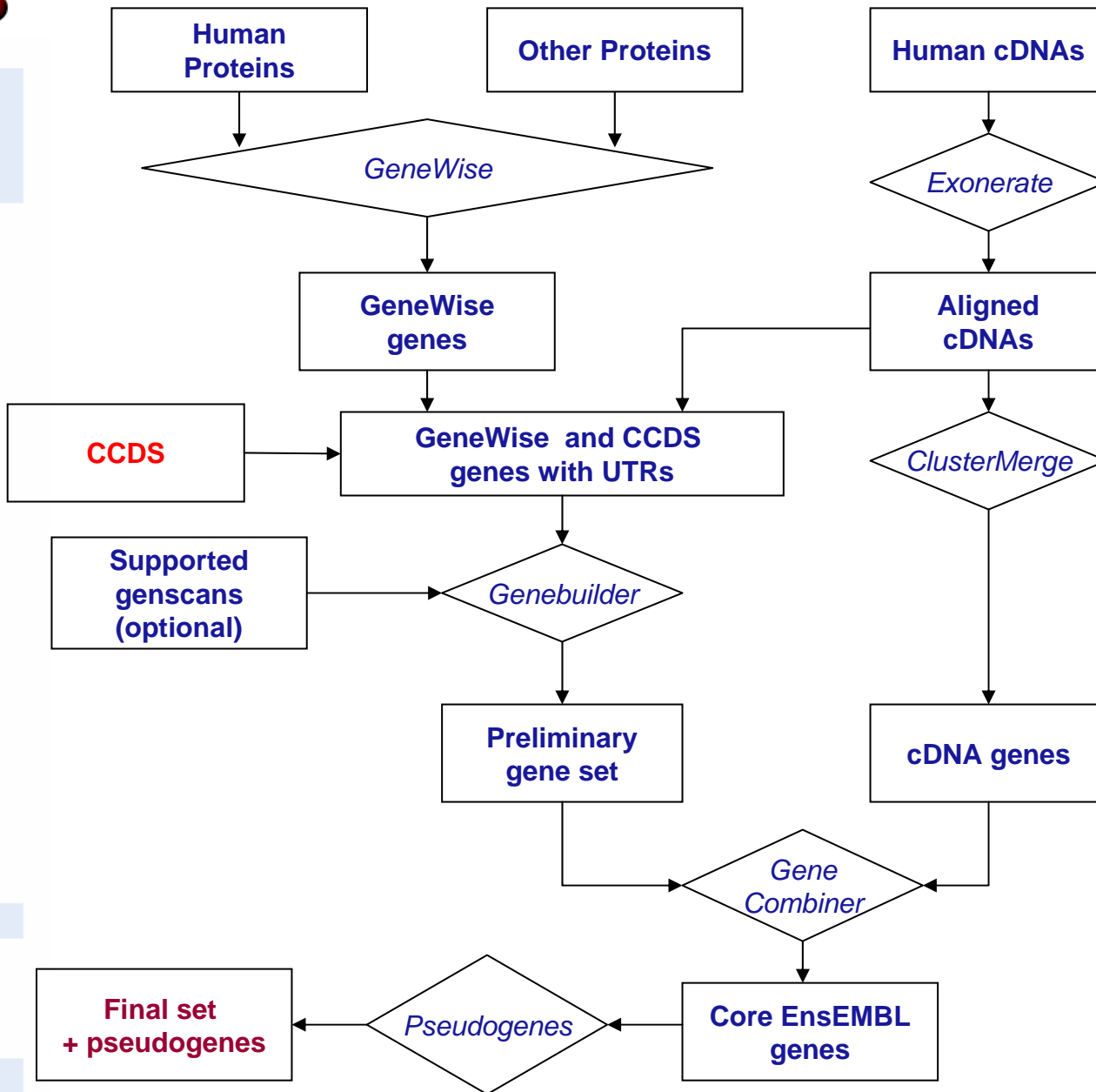
## ***Gene Builder***

**Combines results after GeneWise and eventually *ab initio* predictions.**

- **Clusters transcripts into genes by genomic exon overlap.**
- **Groups transcripts, which share exons**
- **Rejects non-translating transcripts**
- **Removes duplicate exons**
- **Attaches supporting evidence**
- **Writes genes to database**



# Genebuild Summary



ensembl

# Evidence Tracks in Contig View

Ensembl

Expanded tracks

Compressed tracks



# *ncRNAs*

Functional RNAs

Families share conserved secondary structure

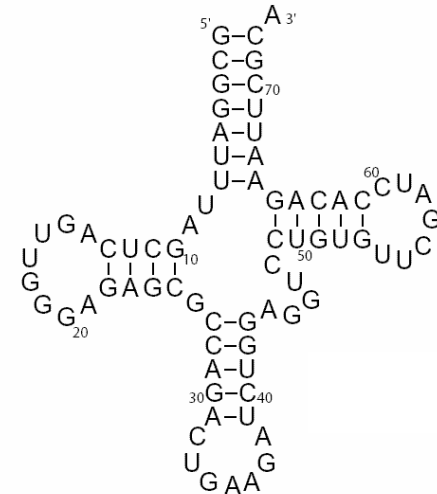
Low sequence identity

Ribosome

Spliceosome

tRNAs

miRNA



# RFAN

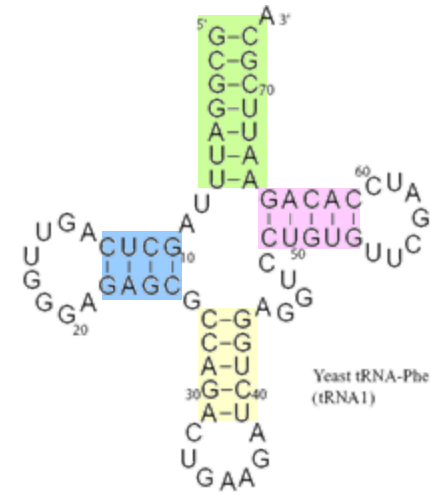
Hand made alignments

Use Infernal to make Covariance Models

Scan models over subset of EMBL to build family alignments

```
tRNA1      GCGGAUUUAGCUCAGUUGGG . AGAGCGCCAGACUGAAGAUCUGGAGGUCC
tRNA2      UCCGAUAUAGUGUAAC . GGCUAUCACAUCACGCUUUCACCGUGGAGA . CC
tRNA3      UCCGUGAUAGUUUAAU . GGUCAGAAUGGGCGCUUGUCGCGUGCCAGA . UC
tRNA4      GCUCGUAUGGCGCAGU . GGU . AGCGCAGCAGAUUGCAAUUCUGUUGGUCC
tRNA5      GGGCACAUGGCGCAGUUGGU . AGCGCGCUUCCCUUGCAAGGAAGAGGUCA
#=GC SS_cons <<<<<<...<<<<.....>>>>.<<<<<<.....>>>>.....<
```

```
tRNA1      UGUGUUCGAUCCACAGAAUUCGCA
tRNA2      GGGGUUCGACUCCCCGUAUCGGAG
tRNA3      GGGGUUCAAUUCCCCGUCGCGGAG
tRNA4      UUAGUUCGAUCCUGAGUGCGAGCU
tRNA5      UCGGUUCGAUCCGGUUGCGUCCA
#=GC SS_cons <<<<.....>>>>>>>>>>>>.
```





e!

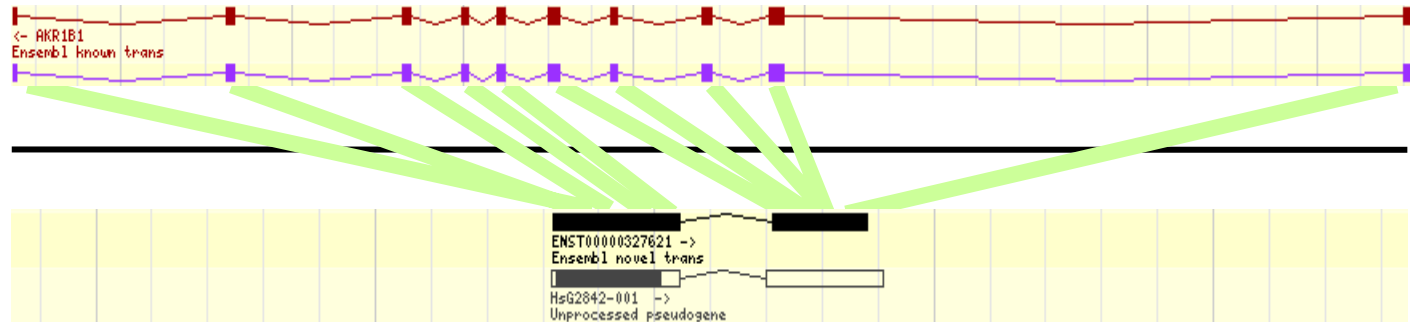
Ensembl

***Pseudogenes and ncRNA and  
ncRNA Pseudogenes***

# Pseudogenes

## Eliminate retro-transposed (processed) pseudogenes

Chromosome 7:133,584,367-133,601,097



Chromosome 18:8,535,635-8,536,757

```

Query: 3  SRLLLNNGAKMPILGLGTWKSPPGQVTEAVKVAIDVGYRHIDCAHVYQNEVEGVVAIQEK 62
          S ++LNNNG K  +LGLGTWKSPPGQV  EAVKVAI+  YRHIDC+HV+QN++      QE+
Sbjct: 2  SHIMLNNGTKTDMGLGLGTWKSPPGQVAEAVKVAINTVYRHIDCSHVHQNKD-----QEQ 55

Query: 63  LREQVVKREELFIVSKLWCTYHEKGLVKGACQKTLSDLKLDYLDLYLIHWPTGFKPGKEF 122
          L+EQVV+RE LFI+SK W   H K LV+G+C+K LS L+LDYLDL+LIHWPTG  PGKEF
Sbjct: 56  LKEQVVRREWLFIIISKPWGICHKRCLVRGSCRKVLSGLELDYLDLHLIHWPTGCHPGKEF 115

Query: 123  FPLDESGNVVPSDTNILDWAAMEELVDEGLVKAIGISNFNHLQVEMILNKPGLKYKPAV 182
          LDESG +                +GLVKA GISNF HLQ E  LNK GLK
Sbjct: 116  SFLDESGLI-----QGLVKAAGISNF-HLQAERTLNKSGLKLSATG 155

Query: 183  NQIECHPYLTQEKLIQYCQSKGIVVTAYSPLGSPDRPWAKPEDPSLLEDPRIKAIAAKHN 242
          LTQE LIQY QSK   VTAYSPLGSPDRP AKPEDPSLLEDPRIK IAAKHN
Sbjct: 156  RS-----LTQENLIQYYQSKA-AVTAYSPLGSPDRPRAKPEDPSLLEDPRIKVIAAKHN 208

Query: 243  KTTAQVLRIFPMQRNLVVIPKSVTPERIAENFKVDFELSSQDMTLLSYNRN 295
          + T+QVL+   QRNLVV P SVT +RIAENFKVDFELSSQDMT+LLS NRN
Sbjct: 209  E-TSQVLMWLLTQRNLVVTPTSVTLDRIAENFKVDFELSSQDMTSLLSNRN 260

```

✓ Mostly dead-on-arrival

✓ Intronless, poly-A tail, short direct repeats

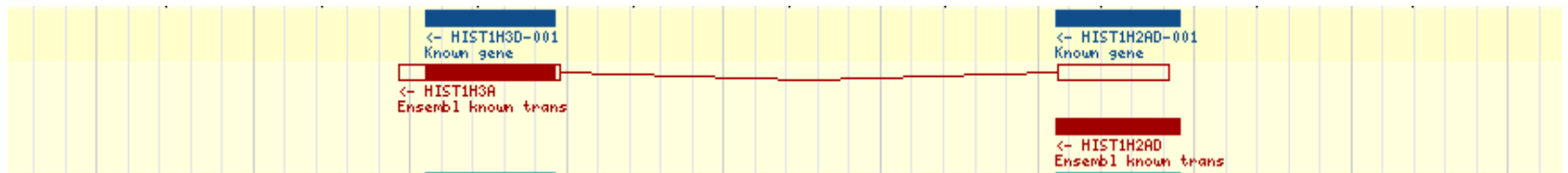
# Spliced Elsewhere

BLASTs single exon genes against a database of the multi-exon genes.

Span of real gene > 3x span of retro gene

Finds an additional ~ 600 pseudogenes in human

False positives can occur where gene predictions join together neighbouring genes in a cluster



Questions remain over the wisdom of calling all of these genes pseudogenes as some may be functional.

*Single exon transcripts with frameshifts*

*Single exon transcripts with a spliced gene model elsewhere*

*Transcripts which introns contain more than 80% repeat sequences*

*A gene is labeled as a pseudogene if all transcripts in that gene are labeled as pseudo-transcripts*

# *ncRNAs*

Functional RNAs

Families share conserved secondary structure

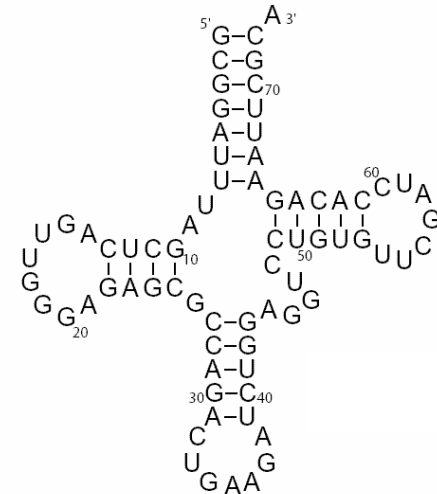
Low sequence identity

Ribosome

Spliceosome

tRNAs

miRNA



# RFAN

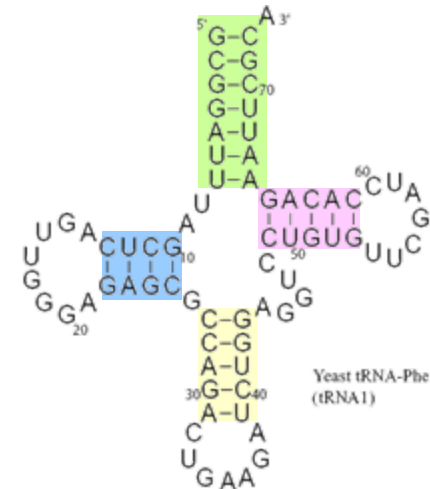
Hand made alignments

Use Infernal to make Covariance Models

Scan models over subset of EMBL to build family alignments

```
tRNA1      GCGGAUUUAGCUCAGUUGGG . AGAGCGCCAGACUGAAGAUCUGGAGGUCC
tRNA2      UCCGAUAUAGUGUAAC . GGCUAUCACAUCACGCUUUCACCGUGGAGA . CC
tRNA3      UCCGUGAUAGUUUAAU . GGUCAGAAUGGGCGCUUGUCGCGUGCCAGA . UC
tRNA4      GCUCGUAUGGCGCAGU . GGU . AGCGCAGCAGAUUGCAAUUCUGUUGGUCC
tRNA5      GGGCACAUGGCGCAGUUGGU . AGCGCGCUUCCCUUGCAAGGAAGAGGUCA
#=GC SS_cons <<<<<<...<<<<.....>>>>.<<<<<<.....>>>>.....<
```

```
tRNA1      UGUGUUCGAUCCACAGAAUUCGCA
tRNA2      GGGGUUCGACUCCCCGUAUCGGAG
tRNA3      GGGGUUCAAUUCCCCGUCGCGGAG
tRNA4      UUAGUUCGAUCCUGAGUGCGAGCU
tRNA5      UCGGUUCGAUCCGGUUGCGUCCA
#=GC SS_cons <<<<.....>>>>>>>>>>.
```









# ***Human Build Statistics***

**NCBI 35 assembly, released June 2004**

<b>Ensembl genes:</b>	<b>24,194</b>
<b>protein-coding genes</b>	<b>22,242</b>
<b>pseudogenes</b>	<b>1,978</b>
<b>ncRNA genes</b>	<b>6,501</b>
<b>Ensembl coding transcripts:</b>	<b>33,838</b>
<b>non-coding transcripts</b>	<b>1,978</b>
<b>Ensembl exons:</b>	<b>245,215</b>
<b>Human input sequences:</b>	
<b>54,272 proteins, redundant set</b>	
<b>120,421 cDNAs</b>	

e!

Ensembl

# *Evaluating Genes and Transcripts*

Ensembl gene set

Pseudogenes

**Ensembl EST genes**

***Ab initio* predictions**

**Manual curation (Vega, CCDS)**

**Gene models from other groups**



# *EST Analysis*

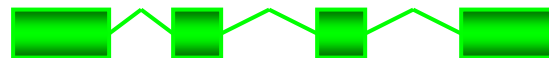
## Map ESTs with Exonerate

(determine coverage, % identity and location in genome)

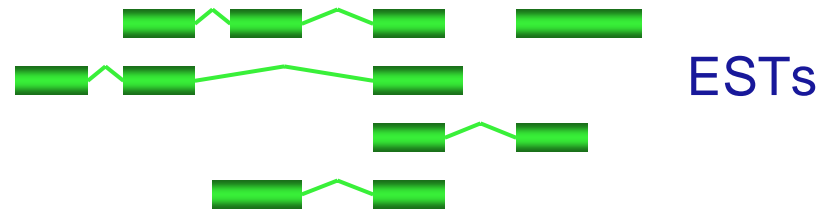


## Filter on % identity and depth

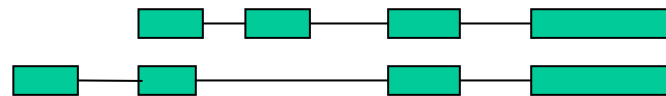
(5.5 million ESTs from dbEST – we map about  
1/3)



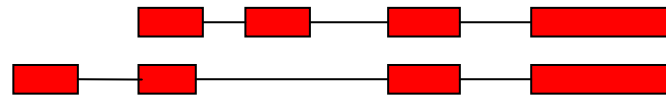
# Alternative Splicing Forms



Merge ESTs according to consecutive exon overlap and set splice ends

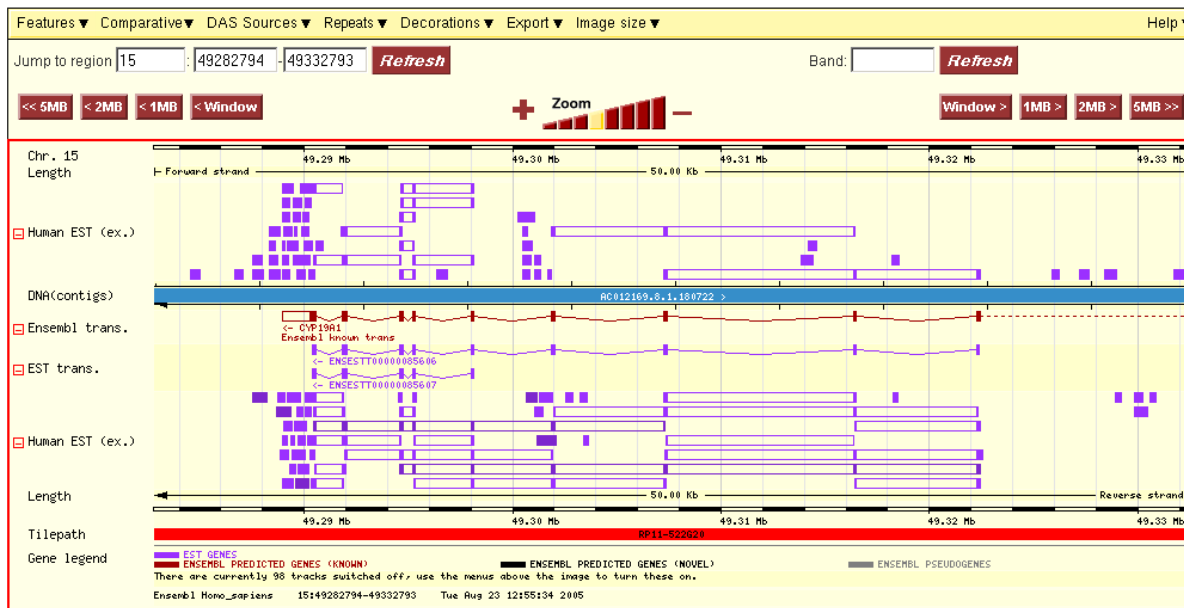


Assign translation



Alternative transcripts with translation and UTRs

# EST Genes and ESTs



Ensembl transcript  
←  
EST transcripts  
←  
Human ESTs

Latest Human Build  
NCBI 35 assembly  
released July 2004

EST Genes: 28,636  
EST Transcripts: 56,351

e!

Ensembl

# *Evaluating Genes and Transcripts*

Ensembl gene set

Pseudogenes

Ensembl EST genes

***Ab initio* predictions**

**Manual curation (Vega, CCDS)**

**Gene models from other groups**





Ensembl

# Ab initio Predictions

Features ▾ Comparative ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Image size ▾ Help ▾

Jump to region    **Refresh** Band:  **Refresh**

<< 5MB < 2MB < 1MB < Window **+** Zoom **-** Window > 1MB > 2MB > 5MB >>

Chr. 18  
Length 200.00 Kb

Forward strand

Genscan  
GENSCAN0000023541  
Ab-initio Genscan trans

DNA(contigs)  
AC120349.5.1.183055

Ensembl trans.  
0725E4\_HUMAN  
Ensembl known trans  
SMAD2  
Ensembl known trans  
SMAD2  
Ensembl known trans  
SMAD2  
ENST00000356825  
Ensembl novel trans

Genscan  
Length 200.00 Kb  
Reverse strand

Tilepath  
RP11-767C9  
GTD-2549E12

Gene legend  
ENSEMBL PREDICTED GENES (KNOWN)  
There are currently 98 tracks switched off; use the menus above

Ensembl Homo\_sapiens 18:43566415-43766414 Tue Aug 23 14:00:00 2005

Genscan transcript

Ensembl Transcript Report	
Transcript ID	GENSCAN0000023541
Transcript information	Exons: 17 Transcript length: 2,492 bps Translation length: 830 residues
Genomic Location	This transcript can be found on Contig AC120349.5.1.183055 at location: <a href="#">2,313-124,099</a> This start of this transcript is located in <a href="#">Contig AC120349.5.1.183055</a> .
Description	No description
Prediction Method	This transcript was predicted by the Ensembl pipeline analysis system, using Chris Burge's Genscan program Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 76-94. The splice site models used are described in more detail in Burge, C. B. (1999) Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam, 127-163.
Transcript structure	
Transcript neighbourhood	
Transcript sequence	<pre> GTAACTGGGAAACCAGATCCATTCTCAGAAGCTTGCTGAATGAGCCGACAGCCGGAATGA AAACACTCTCTGGATGTCATGGCTTTCTCAGAAGTGGTAGTGCAGATGAATCTCTGCT CCGGCTCTCTCATCTGGACTGTGACGGCTCTCTCACTGGCTGGGGCCCTAAGCTTTCTT TAATCACACAGTCCCTTGAAGCTGCTGAGGTTATGGCTTTCTTCAAGCTCTGATCCAA GCCTAAGGATAGCCACTGCTCCCTGGACAGAAATGTCGACAGCAAGTACTGTCGCCAGG ACAATGGATTTGGACACTGGCTCTCTGAGACCACTGGACACCCCTGGAGAGGACA GCCTGGAGACAAACTTGAAGAAATATCCAAAAATCCATACCACTTTTGGTCACTG ACTGGCTGAAGGGCTTAAAGGATCTCCAAACCAACCCCTGATTAGACTCACTTTGCTC AGCCTGAGGCAAGAGGGCGCTCTCCGGCTGCGGGGCGCTCCCGGTGGATGACACCG TGTCCGCTGCCGGCTCACACGGACGGTGGACGAGGGCGGACGGCCCTACAGGCCCCAC </pre>

e!

Ensembl

Ensembl gene set

Ensembl EST genes

*Ab initio* predictions

**Manual curation (Vega, CCDS)**

**Gene models from other groups**

## *Manual Curation*

**Manual annotation of finished clones**

**Vega Genome Browser**

<http://vega.sanger.ac.uk/>

**Currently only chromosomes**

**1, 6, 9, 10, 13, 20, 22, X and Y (Sanger Institute)**

**7 (Washington University)**

**14 (Genoscope)**

**18 (Broad Institute)**

**16, 19 (DOE Joint Genome Institute)**

**Other groups will also contribute to Vega**

**Displayed in Ensembl when available**

# *Manual Curation*

## Manually-curated gene sets in Ensembl

**WormBase** (data import)

*Caenorhabditis elegans*

**FlyBase** (data import)

*Drosophila melanogaster*

**Génoscope** (data import)

*Tetraodon nigroviridis*

**IMCB, Singapore** (data import)

*Takifugu rubripes*

**SGD** (data import)

*Saccharomyces cerevisiae*

**Vega includes some manually-curated  
finished clones from**

*Danio rerio, Mus musculus and Canis familiaris*



ensembl

# Manually-curated Vega Genes

Features ▾ Comparative ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Image size ▾ Help ▾

Jump to region  :  -  **Refresh** Band:  **Refresh**

<< 5MB < 2MB < 1MB < Window + Zoom - Window > 1MB > 2MB > 5MB >>

Chr. 7  
Length  
Forward strand  
Ensembl trans. CFTR => Ensembl known trans.  
Vega trans. CFTR-001 => Curated predicted gene  
DNA(contigs) AC000111.1.1.149554 >  
Vega trans. AC000111.1.2-003 Curated novel Trans. AC000111.1-1745 Curated pseudogenes  
Ensembl trans. ENST00000356121 Ensembl novel trans.  
Reverse strand  
Tilepath  
Encode regions  
Gene legend

Ensembl Gene Report for OTTHUMG00000023076

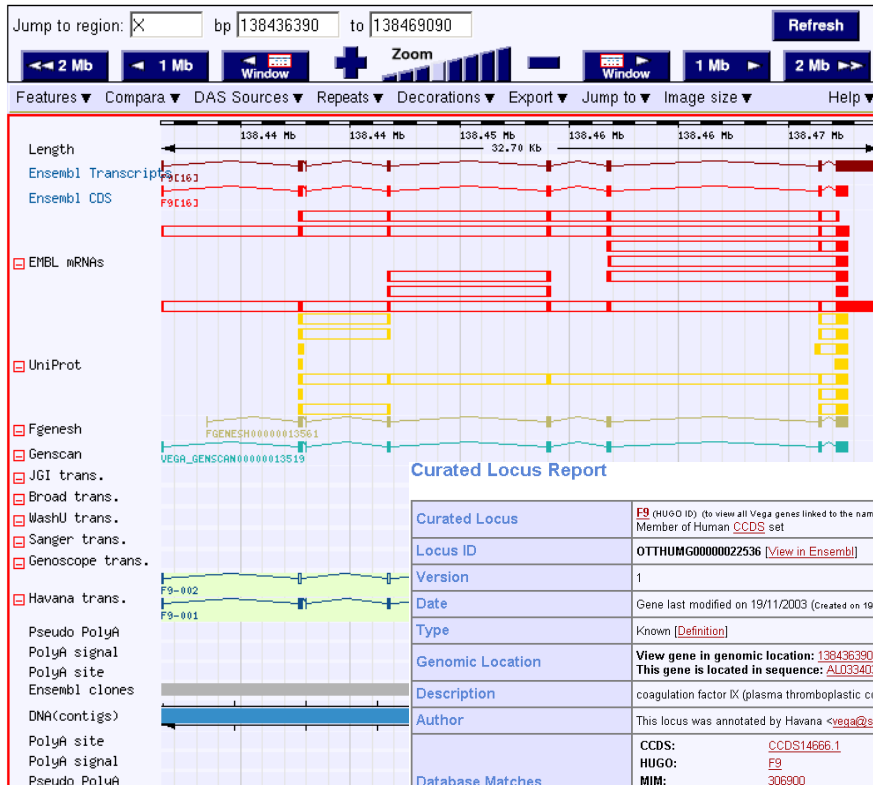
Vega manual curation

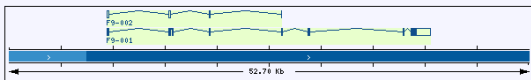
Gene	<a href="#">CFTR</a> (HUGO ID) (to view all Ensembl genes linked to the name <a href="#">click here</a> )
Vega Gene ID	OTTHUMG00000023076 ( <a href="#">View Gene OTTHUMG00000023076 in Vega</a> )
Genomic Location	This gene can be found on Chromosome 7 at location: <a href="#">116,713,968-116,902,666</a> This start of this gene is located in <a href="#">Contig AC000111.1.1.149554</a> .
Description	cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7)
Curation Method	No Curation Method defined in database
Transcripts	<p>OTTHUMT00000059397 OTTHUMP00000024694 CFTR-001 <a href="#">[Transcript info]</a> <a href="#">[Exon info]</a> <a href="#">[Peptide info]</a></p>
Disease Matches	<p>This Ensembl entry corresponds to the following OMIM disease identifiers:</p> <p><b>Congenital bilateral absence of vas deferens, 277180 (3)</b>  <a href="#">[Omim ID: 602421] - View disease information</a></p> <p><b>Cystic fibrosis, 219700 (3)</b>  <a href="#">[Omim ID: 602421] - View disease information</a></p> <p><b>Sweat chloride elevation without CF (3)</b>  <a href="#">[Omim ID: 602421] - View disease information</a></p>



# Manually-curated Vega Genes

## Detailed view



Curated Locus	
Curated Locus	F9 (HUGO ID) (to view all Vega genes linked to the name <a href="#">click here</a> ) Member of Human <a href="#">CCDS</a> set
Locus ID	OTTHUMG00000022536 <a href="#">[View in Ensembl]</a>
Version	1
Date	Gene last modified on 19/11/2003 (Created on 19/11/2003)
Type	Known <a href="#">[Definition]</a>
Genomic Location	<b>View gene in genomic location:</b> <a href="#">138436390 - 138469090 bp (138.4 Mb)</a> on chromosome X <b>This gene is located in sequence:</b> <a href="#">AL033403.1.1.158557</a>
Description	coagulation factor IX (plasma thromboplastic component, Christmas disease, hemophilia B)
Author	This locus was annotated by Havana < <a href="mailto:vega@sanger.ac.uk">vega@sanger.ac.uk</a> >
Database Matches	CCDS: <a href="#">CCDS14666.1</a> HUGO: <a href="#">F9</a> MIM: <a href="#">306900</a> RefSeq dna: <a href="#">NM_000133</a> Uniprot/SWISSPROT: <a href="#">P00740</a>
Sequence Markup	<a href="#">View genomic sequence for this gene with exons highlighted</a>
Export Data	<a href="#">Export gene data in EMBL, GenBank or FASTA</a>
Curated Transcripts	<p>1: <a href="#">F9-001</a> (OTTHUMT00000058557) <a href="#">[Transcript information]</a> <a href="#">[Exon information &amp; supporting evidence]</a> <a href="#">[Protein information]</a></p> <p>2: <a href="#">F9-002</a> (OTTHUMT00000058558) <a href="#">[Transcript information]</a> <a href="#">[Exon information &amp; supporting evidence]</a> <a href="#">[No translation]</a></p> 
<input type="checkbox"/> <a href="#">GAD</a> (Genetic Association Database) <input type="checkbox"/> <a href="#">HUGO text</a> (PubMed text-mining via HUGO symbol) <input type="checkbox"/> <a href="#">EMBL text</a> (EMBL text-mining via EMBL symbol)	

ensembl

## ***CCDS (consensus CDS)***

**Collaboration between RefSeq (NCBI ), UCSC, Ensembl and Havana to produce a set of stable, reliable, complete (ATG->stop) CDS structures for human**

**NCBI and Ensembl guarantee to retain these structures in their gene sets**

**Long term aim is to get to a single gene set for human.**

## ***Ensembl/Vega merged set***

**On chromosomes with Vega annotation**

**Map Vega annotation to current assembly**

**Prune partial CDSs from Vega Known and Novel\_CDS**

**Add unique Ensembl complete CDSs to Novel\_CDS and Known genes**

**Add unique Ensembl genes**

**Add Ensembl genes which overlap Vega Novel\_Transcript ONLY when the Ensembl has a complete CDS**

**Remove Vega putative genes**

**On chromosomes without Vega annotation**

**Take Ensembl set (including pseudogenes)**

# Comparison of CDSs to NCBI

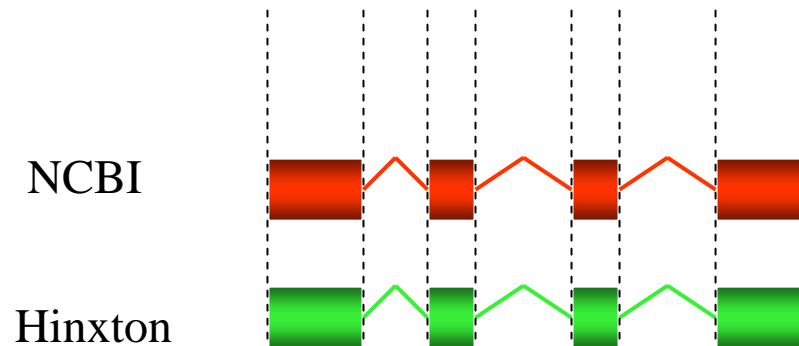
Exact matching CDS on the genome with:

Complete CDS (ATG->stop)

No frameshifts

No phase problems

No internal stop codons



## ***CCDS release (March 2005)***

**Conservative first set, so the following have been removed:**

**All CDSs which match XMs**

**CDSs with large cDNA v genomic discrepancies**

**CDSs with non consensus splice sites**

**Set contains:**

**14795 different CDSs**

**16085 transcripts (in Ensembl)**

**13031 genes**

**The genebuild pipeline has been modified to retain these 'blessed' CDSs (stored in a database for incorporation in the build)**

e!

Ensembl

# *Evaluating Genes and Transcripts*

Ensembl gene set

Pseudogenes

Ensembl EST genes

*Ab initio* predictions

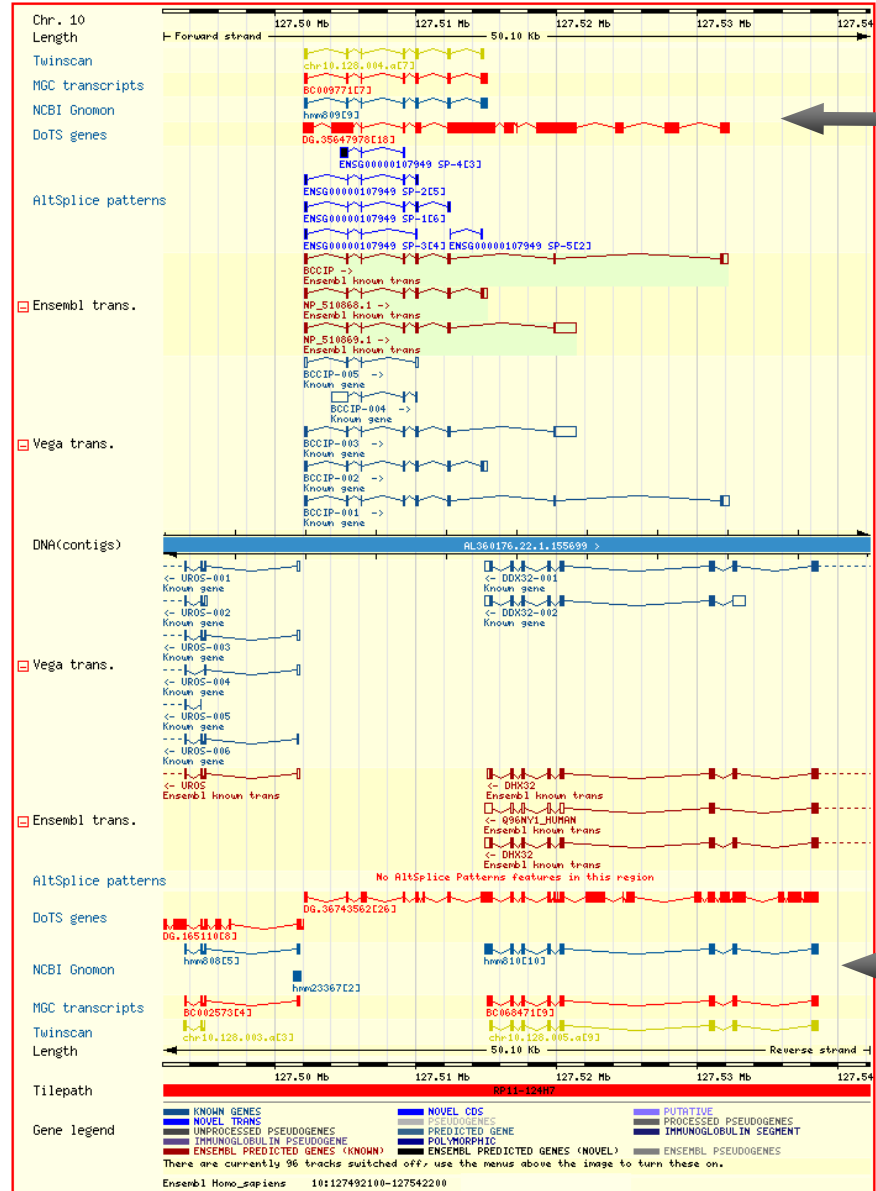
Manual curation (Vega)

**Gene models from other groups**



ensembl

# Other Gene Models



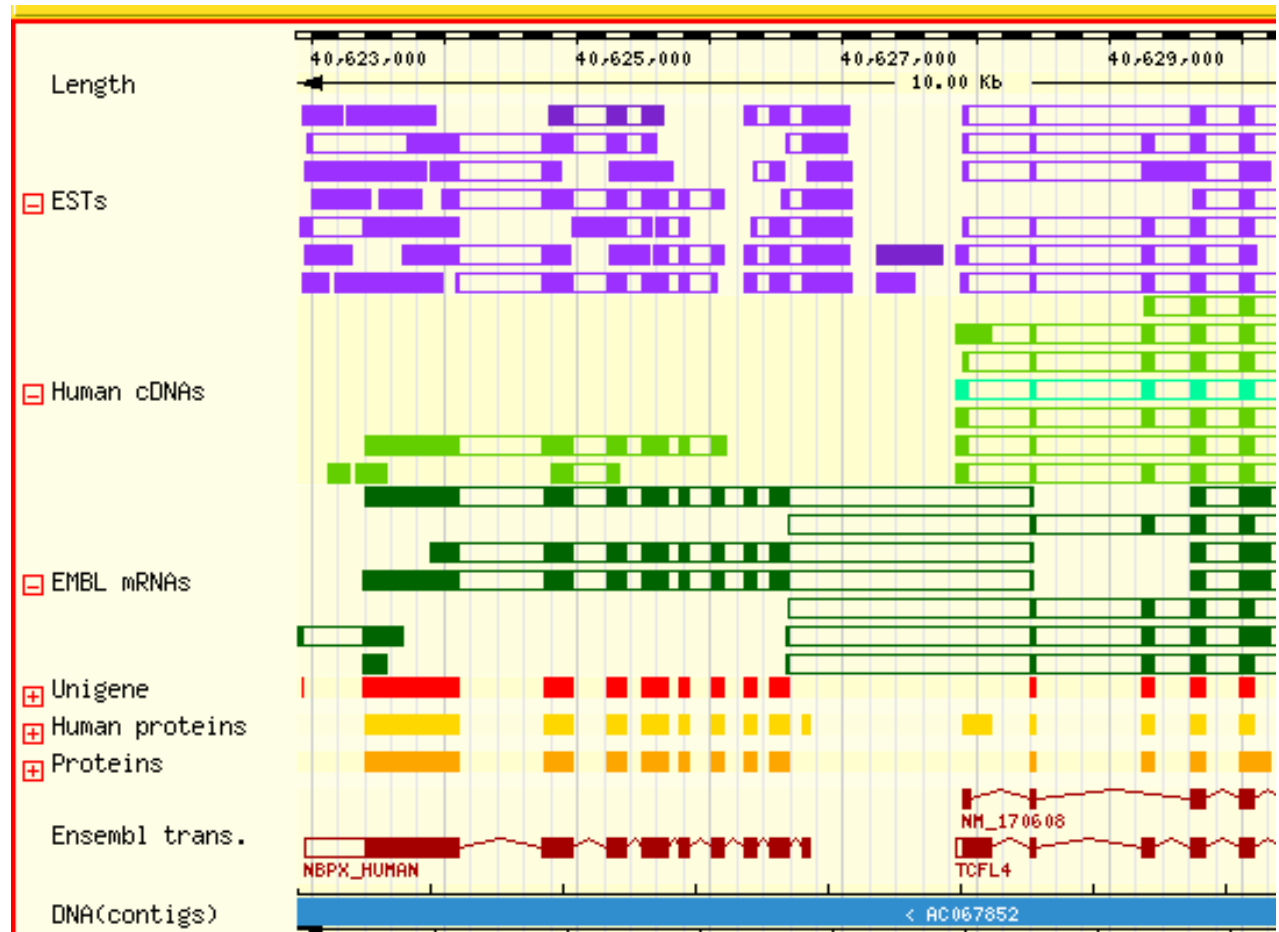


# Evidence Tracks in Contig View

Ensembl

Expanded  
tracks

Compressed  
tracks



e!

Ensembl

Q&A