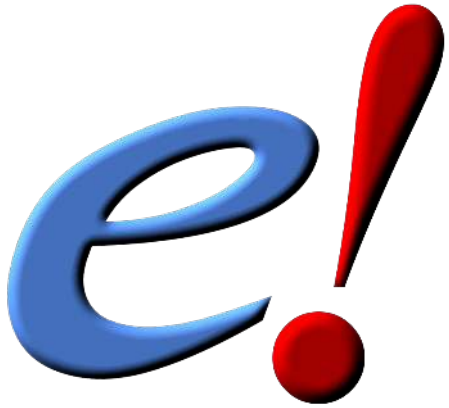
# Ensembl Genome Browser

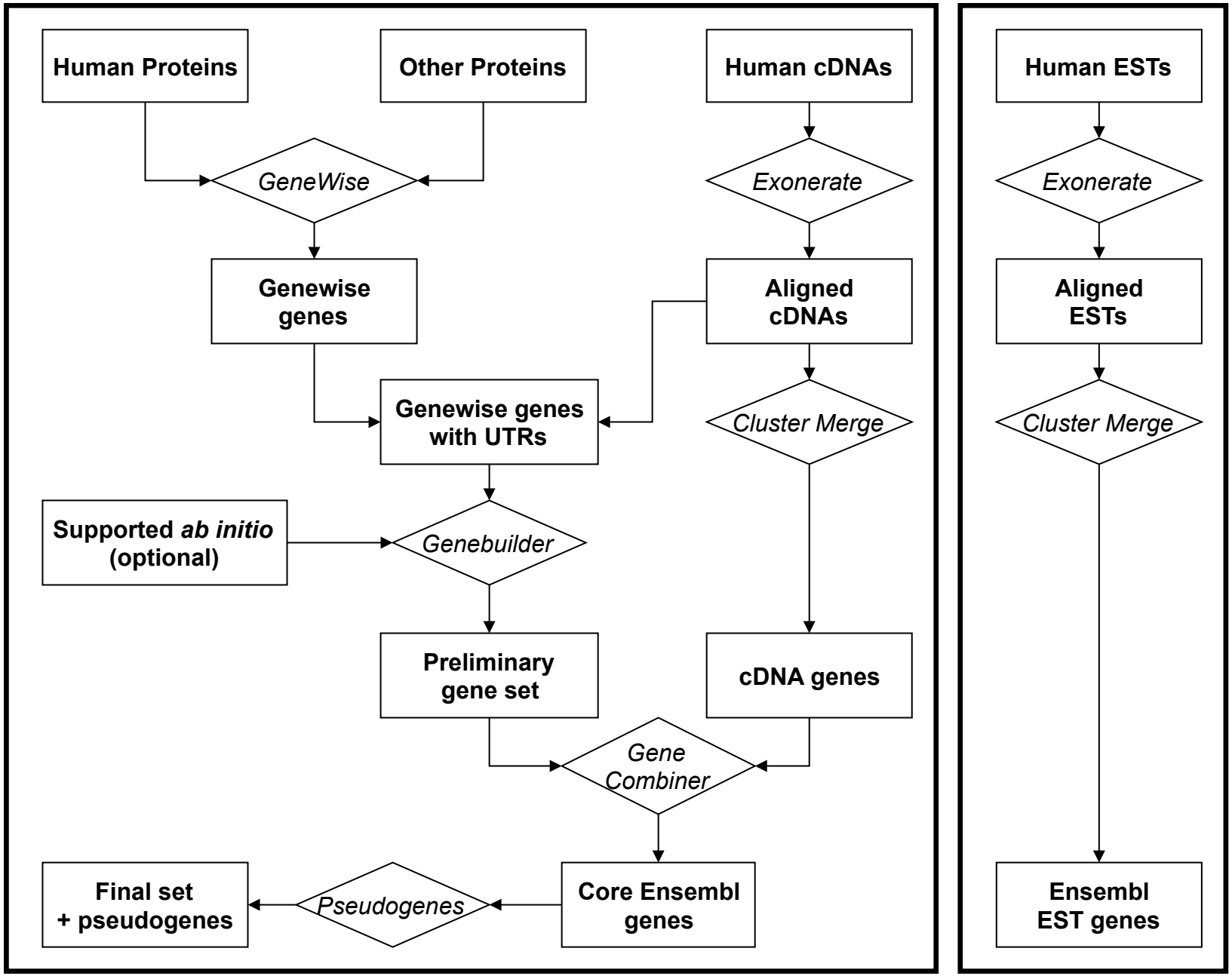
Genes and Transcripts

Genome Annotation Strategy

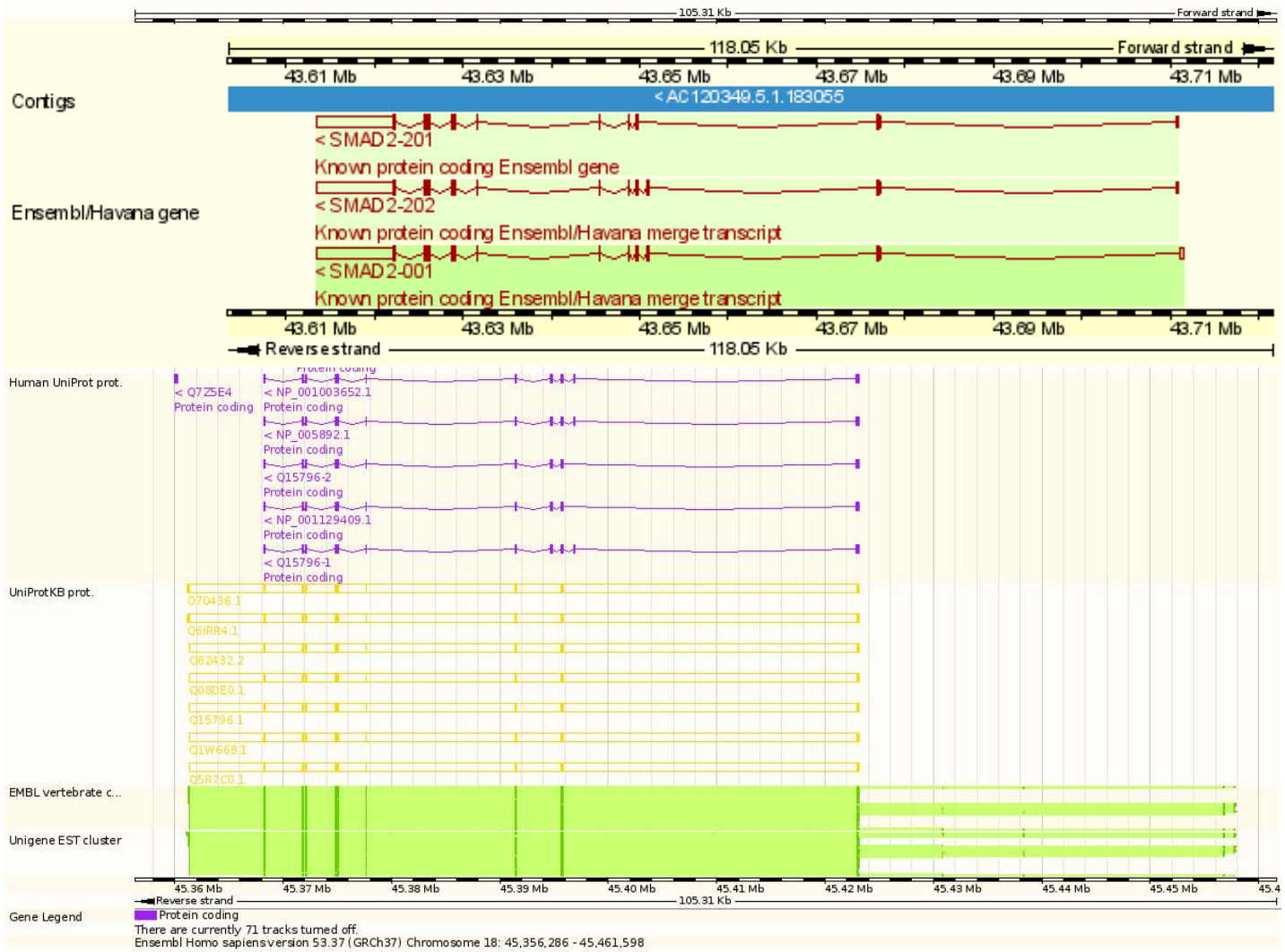



# Genes and Transcripts

- Automated annotation (Ensembl genes)
- Manual curation (Havana genes)
- Ensembl EST genes
- *Ab initio* predictions
- Non-coding RNA genes

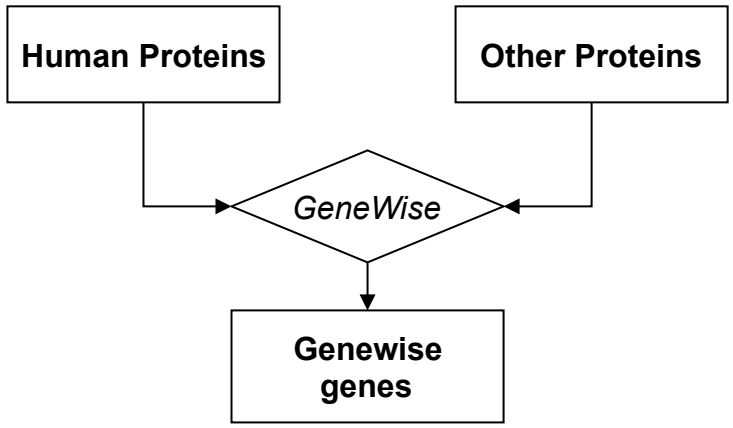

Human Proteins
Other Proteins
Human cDNAs
Human ESTs
GeneWise
Exonerate
Exonerate
Genewise genes
Aligned cDNAs
Aligned ESTs
Genewise genes with UTRs
Cluster Merge
Cluster Merge
Supported ab initio (optional)
Genebuilder
Preliminary gene set
cDNA genes
Gene Combiner
Final set + pseudogenes
Pseudogenes
Core Ensembl genes
Ensembl EST genes

# Non-redundant Set of Alternatively Spliced Transcripts



# The Ensembl Gene Build

- Align **species-specific proteins** and **cDNAs** to the genome to create initial transcript models (Targeted Build)
- Align similar proteins and cDNAs from **closely related species** to annotate additional transcripts (Similarity Build)
- Build additional transcripts using *ab initio* predictors and homology evidence if there is not enough support
- Collate redundant transcripts and cluster the final set of alternative transcripts into genes.


Human Proteins
Other Proteins
GeneWise
Genewise genes

# Gene Builds are Protein-based

- Simple mRNA to DNA alignments do **not** lead to translatable genes
- Protein to DNA alignments require **frameshift** and **splice site** models
- GeneWise (Ewan Birney)
  - Protein to Genome alignments
  - Splice site model
  - Penalises stop codons
  - Models frameshifts
  - E. Birney et al. Genome Research 14:988-995 (2004)

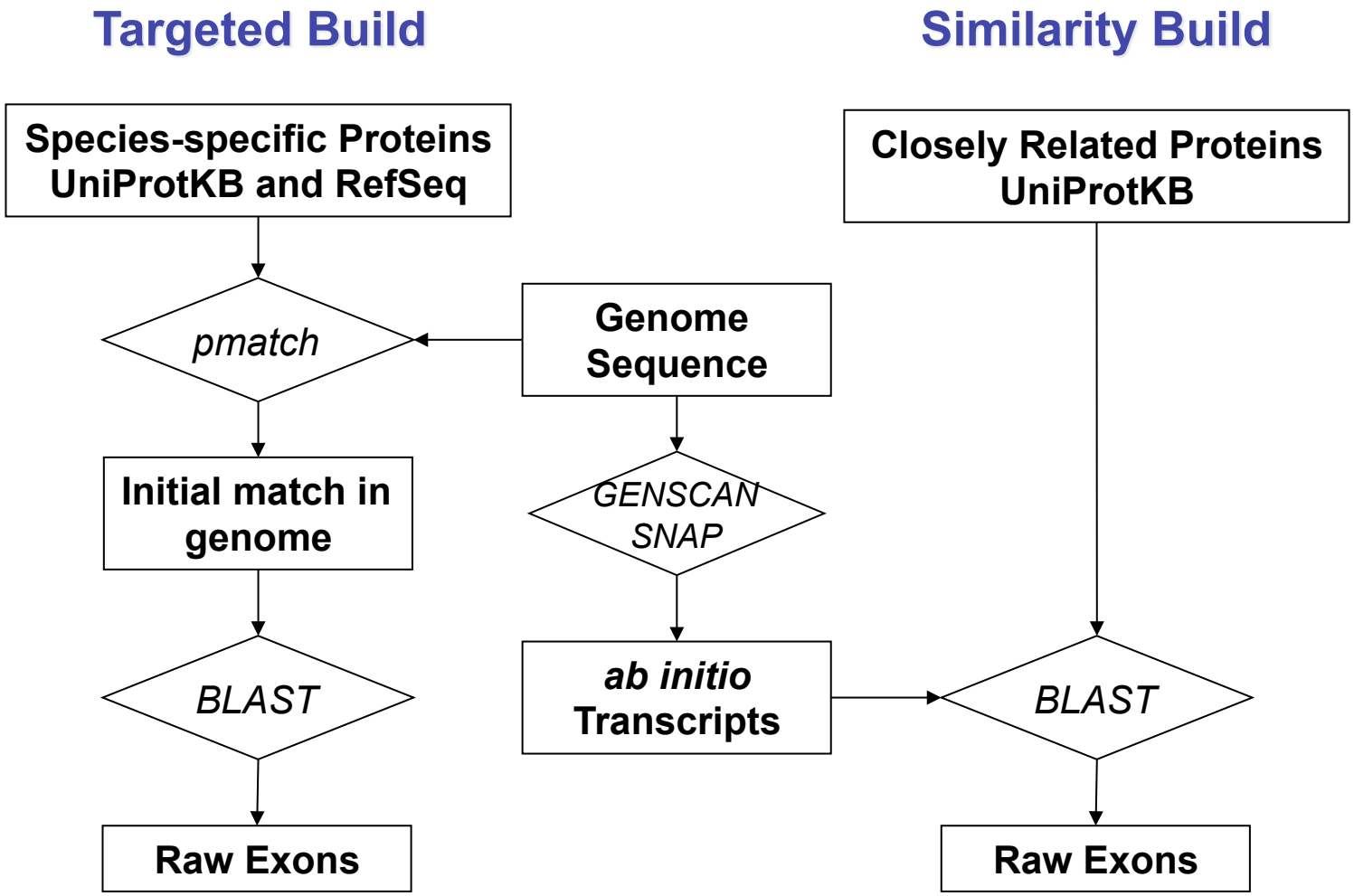
# BLAST - High-Scoring Pairs (HSPs)

```
>AC120349.5.1.183055 dna:contig chromosome:NCBI36:18:43590584:43739179:-1
        Length = 183,055

  Plus Strand HSPs:

 Score = 150 (51.5 bits), Expect = 6.1e-285, Sum P(11) = 6.1e-285
 Identities = 18/18 (100%), Positives = 18/18 (100%), Frame = +2

Query:     244 GSPAELSPTTLSPVNHSL 261
               GSPAELSPTTLSPVNHSL
Sbjct: 107483 GSPAELSPTTLSPVNHSL 107536

 Score = 133 (45.9 bits), Expect = 6.1e-285, Sum P(11) = 6.1e-285
 Identities = 16/16 (100%), Positives = 16/16 (100%), Frame = +3

Query:       1 MSSILPFTPPVVKRLL 16
               MSSILPFTPPVVKRLL
Sbjct:   62055 MSSILPFTPPVVKRLL 62102

 Query:   169 RVET 172
             RVET
 Sbjct: 89553 RVET 89564

  Score = 430 (142.7 bits), Expect = 6.1e-285, Sum P(11) = 6.1e-285
  Identities = 47/47 (100%), Positives = 47/47 (100%), Frame = +3

 Query:    333 GRGVRLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPG 379
               GRGVRLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPG
 Sbjct: 113010 GRGVRLYYIGGEVFAECLSDSAIFVQSPNCNQRYGWHPATVCKIPPG 113150
```

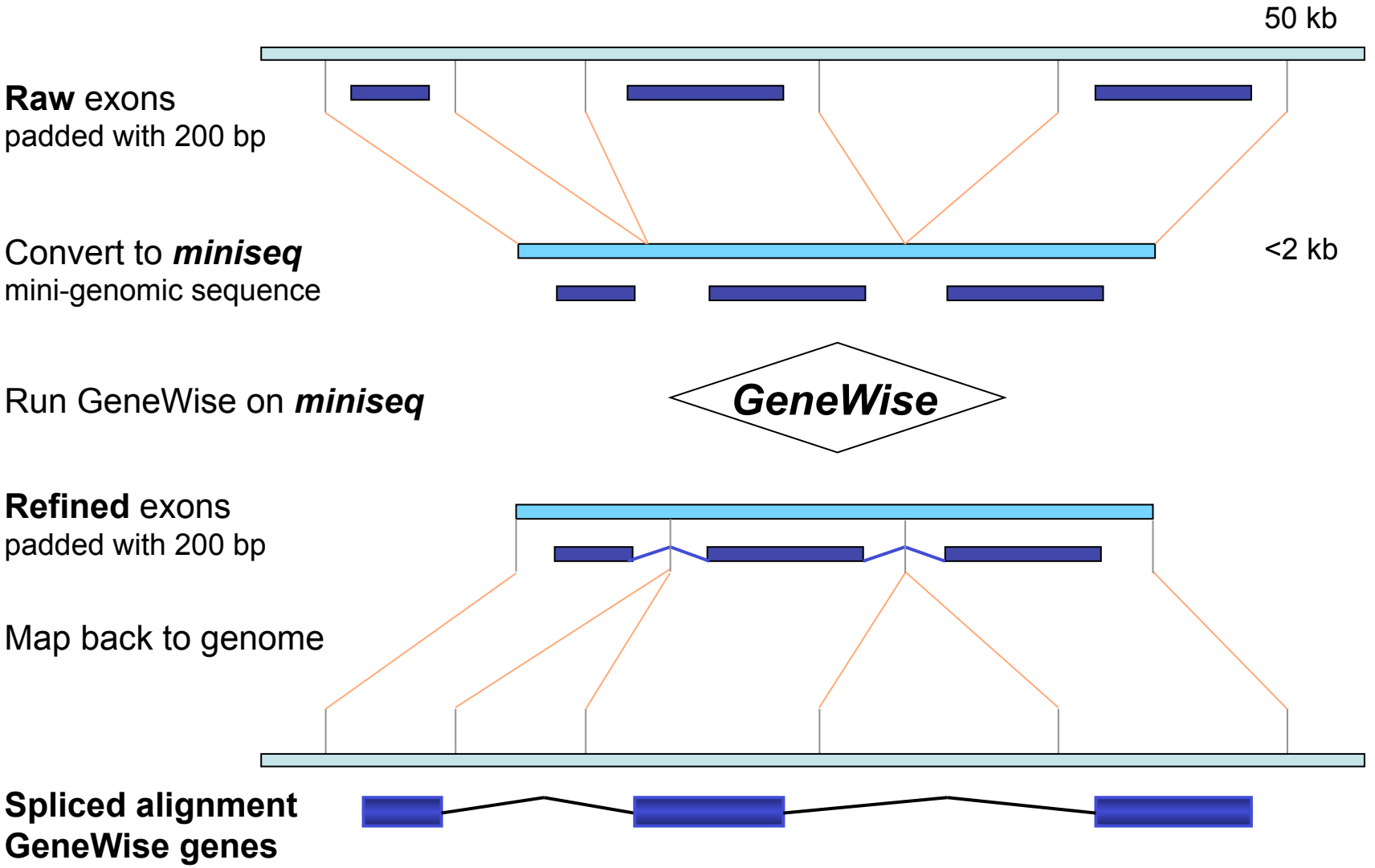
# Exonerate - Splice Site Models

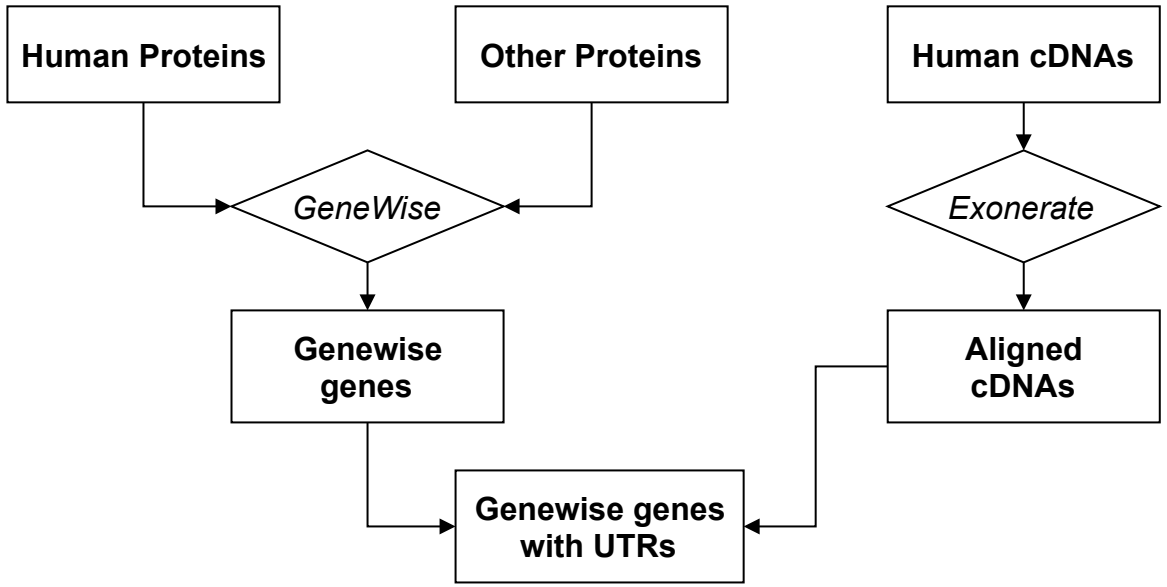
```
Query: SMAD2_HUMAN Q15796 Mothers against decapentaplegic homolog 2 (SMAD 2).
Target: AC120349 AC120349.5 Homo sapiens chromosome 18, clone RP11-380M21, complete sequence.

     1 : MetSerSerIleLeuProPheThrProProValValLysArgLeuLeuGlyTrpLysL :     20
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
         MetSerSerIleLeuProPheThrProProValValLysArgLeuLeuGlyTrpLysL
 62055 : ATGTCGTCCATCTTGCCATTCACGCCGCCAGTTGTGAAGAGACTGCTGGGATGGAAGA :  62112

    21 : ysSerAlaGlyGlySerGlyGlyAlaGlyGlyGlyGluGlnAsnGlyGlnGluGluLy :     39

    79 : ro{Se}  >>>> Target Intron 1 >>>>  {r}ThrCysSerGluIle
         ||{||}++          25956 bp          ++{|}||||||||||||||||
         ro{Se}                                {r}ThrCysSerGluIle
 62287 : CA{AG}gt.........................ag{C}ACTTGCTCTGAAATT

    59 : AspGluLeuGluLysAlaIleThrThrGlnAsnCysAsnThrLysCysValThrIleP :     78
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
         AspGluLeuGluLysAlaIleThrThrGlnAsnCysAsnThrLysCysValThrIleP
 62227 : GATGAGCTTGAGAAAGCCATCACCACTCAAAACTGTAATACTAAATGTGTTACCATAC :  62286

    79 : ro{Se}  >>>> Target Intron 1 >>>>  {r}ThrCysSerGluIleTrpGl :     86
         ||{||}++          25956 bp          ++{|}||||||||||||||||||
         ro{Se}                                {r}ThrCysSerGluIleTrpGl
 62287 : CA{AG}gt.........................ag{C}ACTTGCTCTGAAATTTGGGG :  88266

    87 : yLeuSerThrProAsnThrIleAspGlnTrpAspThrThrGlyLeuTyrSerPheSer :    105
         ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
         yLeuSerThrProAsnThrIleAspGlnTrpAspThrThrGlyLeuTyrSerPheSer
 88267 : ACTGAGTACACCAAATACGATAGATCAGTGGGATACAACAGGCCTTTACAGCTTCTCT :  88323

   106 : GluGlnThr{Ar}  >>>> Target Intron 2 >>>>  {g}SerLeuAspGlyA :    114
         ||||||||||{||}++          1038 bp          ++{|}|||||||||||||||
         GluGlnThr{Ar}                               {g}SerLeuAspGlyA
 88324 : GAACAAACC{AG}gt.........................ag{G}TCTCTTGATGGTC :  89388
```

# Initial Protein Placement

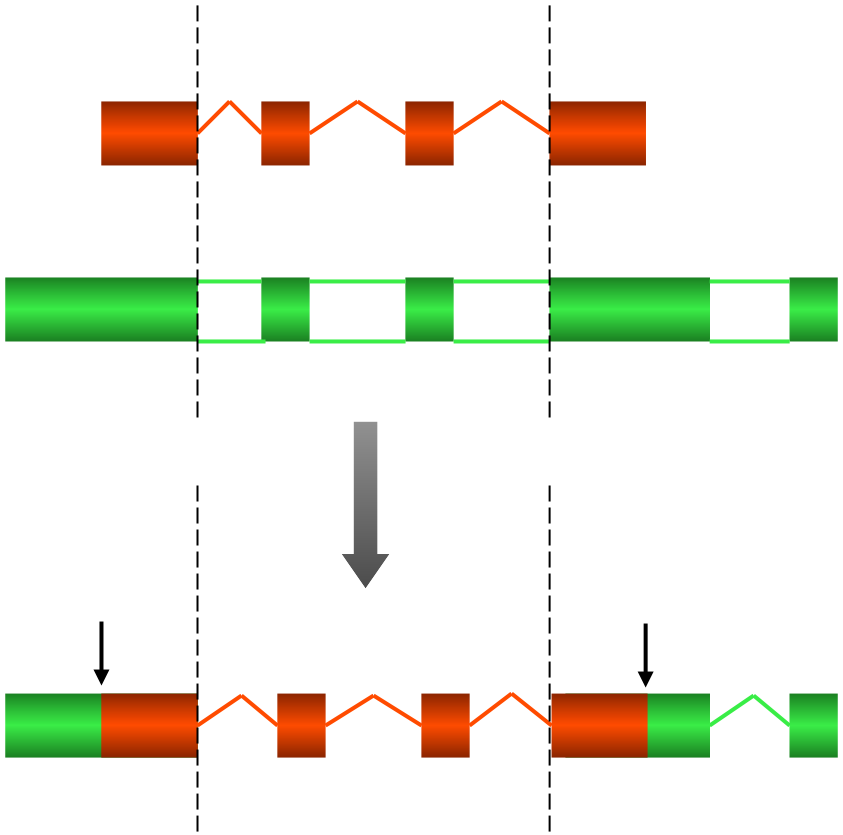


# Sequence Space Reduction




Human Proteins
Other Proteins
Human cDNAs
GeneWise
Exonerate
Genewise genes
Aligned cDNAs
Genewise genes with UTRs

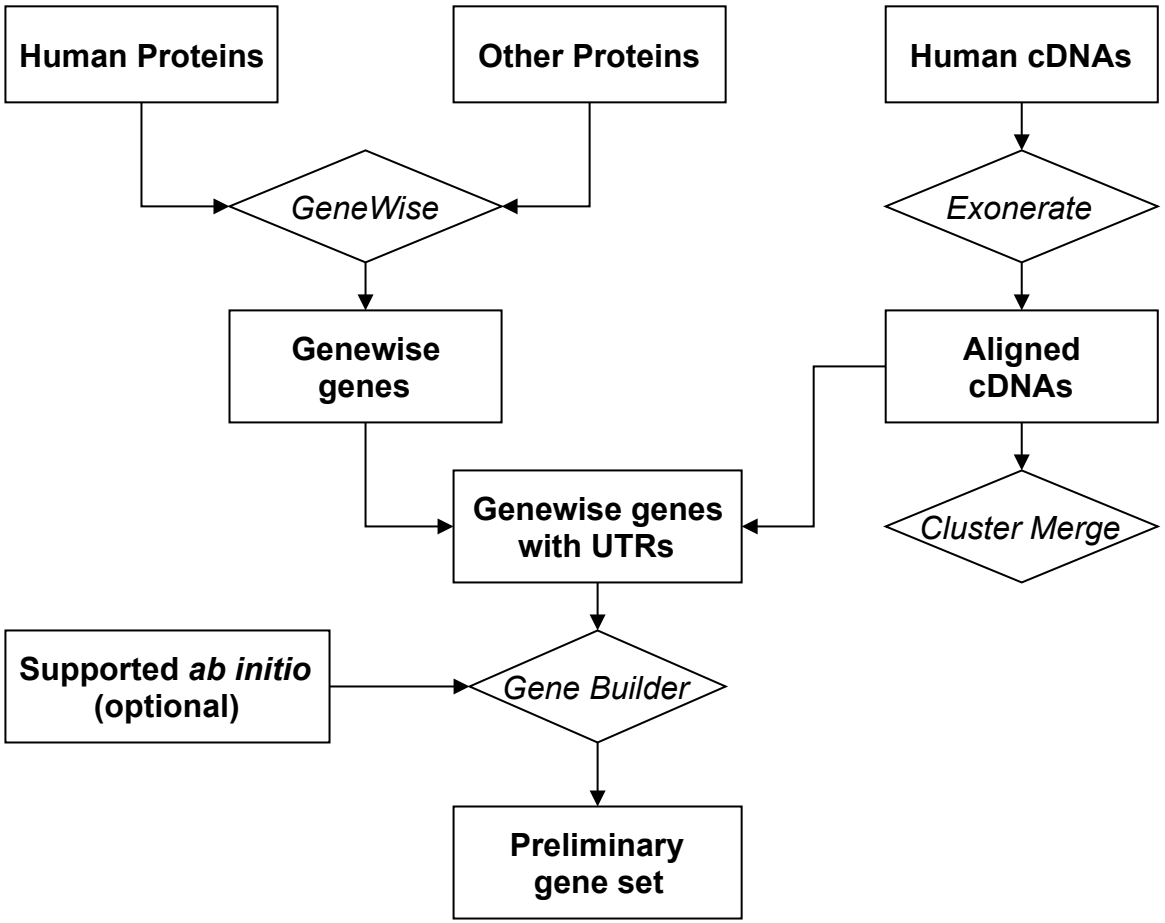
# Attaching UTRs



Proteins aligned with *GeneWise* have **phases** (splice sites), but **no UTRs**

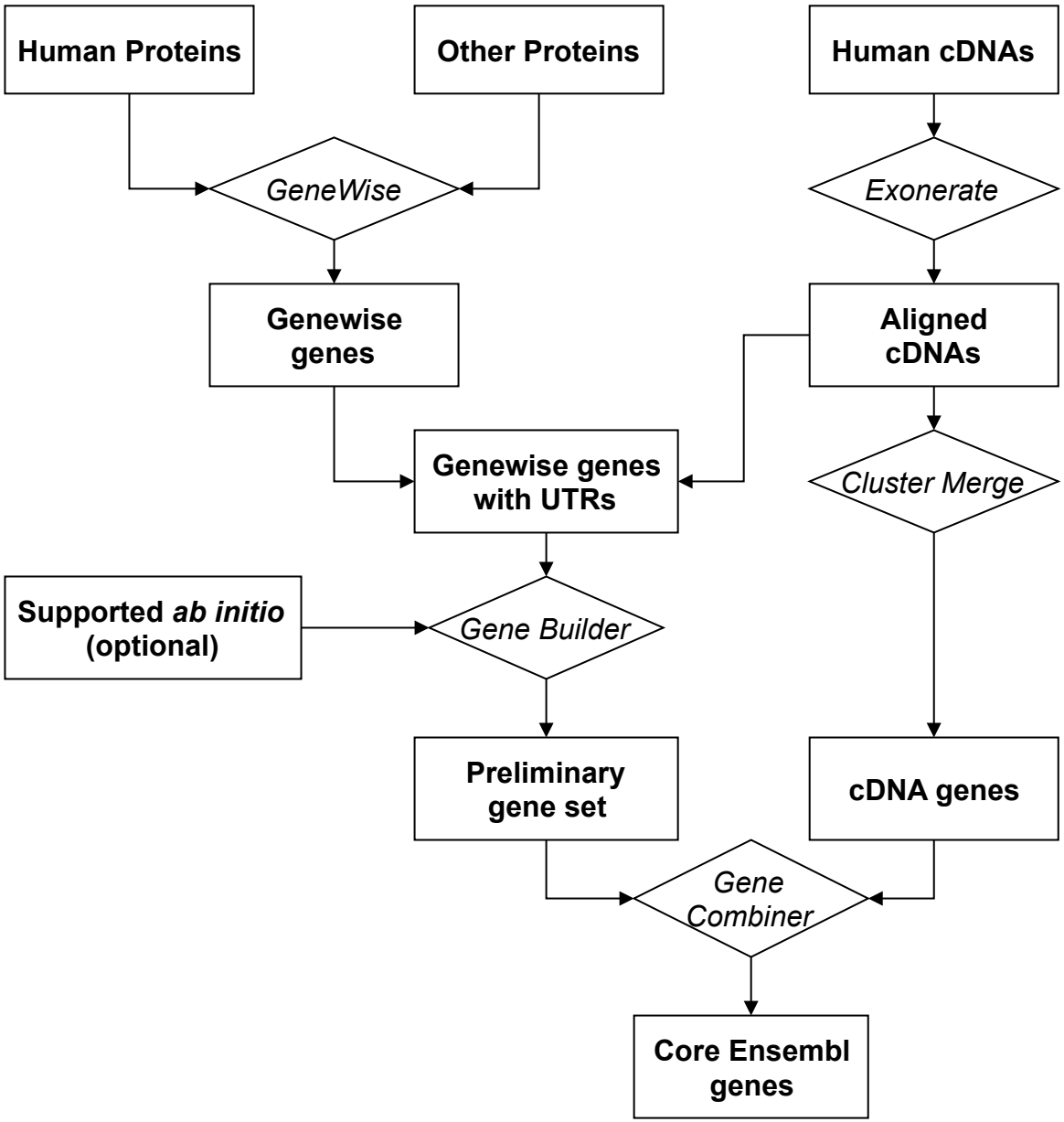
cDNAs aligned with *Exonerate* have **UTRs**, but **no phases**

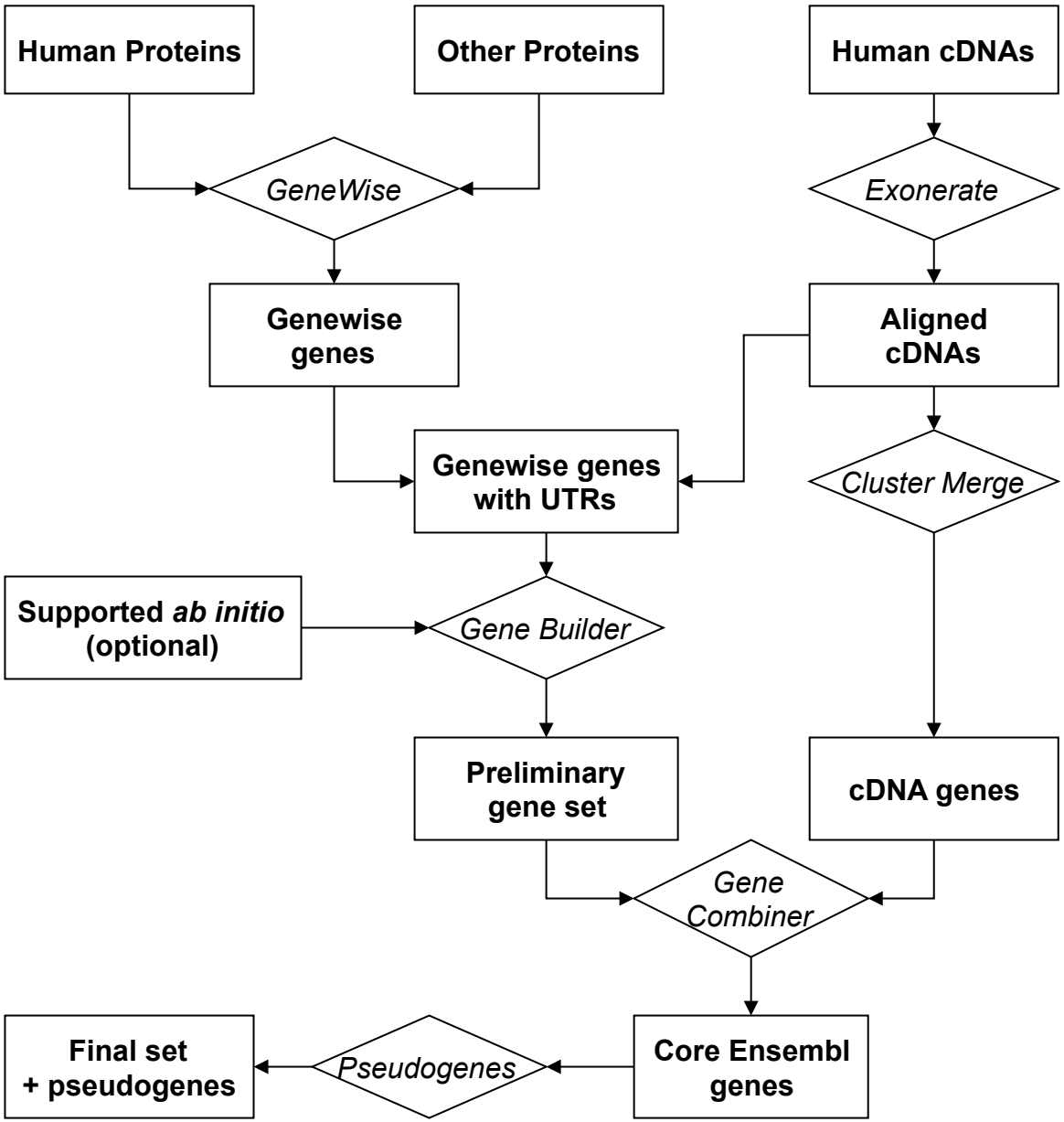
**Translatable genes in phase with UTRs**


Human Proteins
Other Proteins
Human cDNAs
GeneWise
Exonerate
Genewise genes
Aligned cDNAs
Genewise genes with UTRs
Cluster Merge
Supported ab initio (optional)
Gene Builder
Preliminary gene set

# Gene Builder

- Combines results after GeneWise and eventually *ab initio* predictions
- Collates exon-sharing transcripts
- Removes redundant exons
- Re-attaches supporting evidence
- Rejects non-translating transcripts
- Clusters transcripts into genes by
  - Genomic exon overlap
  - Coding sequence overlap


Human Proteins
Other Proteins
Human cDNAs
GeneWise
Exonerate
Genewise genes
Aligned cDNAs
Genewise genes with UTRs
Cluster Merge
Supported ab initio (optional)
Gene Builder
Preliminary gene set
cDNA genes
Gene Combiner
Core Ensembl genes

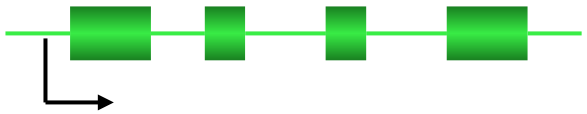

Human Proteins
Other Proteins
Human cDNAs
GeneWise
Exonerate
Genewise genes
Aligned cDNAs
Genewise genes with UTRs
Cluster Merge
Supported ab initio (optional)
Gene Builder
Preliminary gene set
cDNA genes
Gene Combiner
Final set + pseudogenes
Pseudogenes
Core Ensembl genes

# Pseudogene Annotation

- Ensembl does not actively characterise and annotate pseudogenes
- However, proteins and cDNAs align to pseudogene regions in the genome
- Accidental annotation of pseudogenes would lead to spurious gene copies leading to false paralogs in gene trees
- Therefore, once annotated, pseudogenes need to be flagged as such

# Pseudogene Types

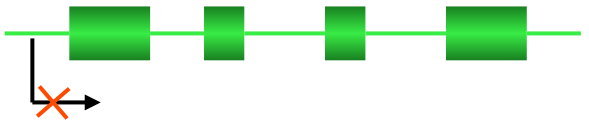
## Processed



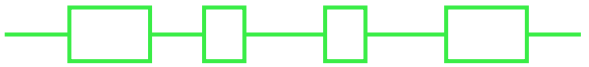
Transcription
and
RNA Processing


Reverse Transcription
and
Retro-transposition


## Unprocessed



Duplication
or
Rearrangement



Stochastic Mutation

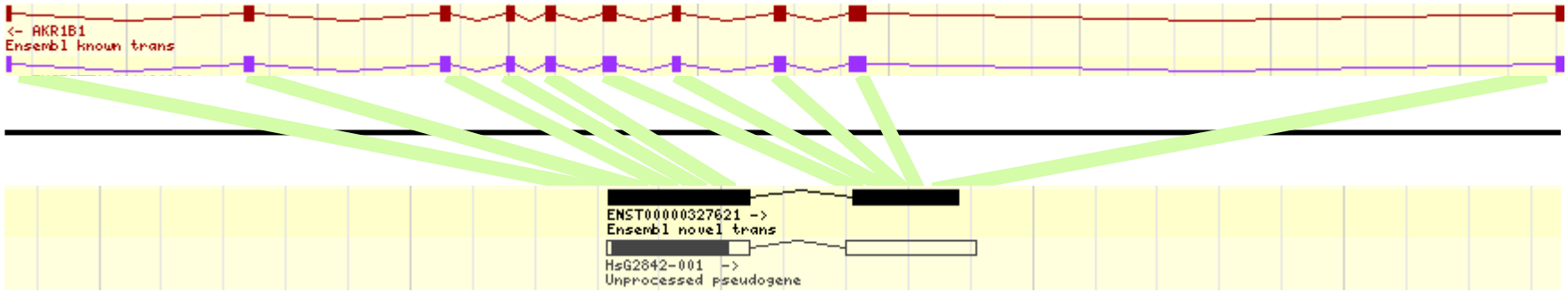
# Pseudogene Analysis

- Processed Pseudogenes
  - Single exon transcripts with a spliced gene model elsewhere in the genome
- Unprocessed Pseudogenes
  - Transcripts with several frameshifts
  - 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
- Polymorphic or segregating pseudogenes
  - Coding in some populations or individuals
  - Mis-sense mutation or nonsense-mediated decay in others
  - 27 cases in the human reference assembly GRCh37
  - More to be expected from the 1000 Genomes project

# Processed Pseudogenes

- Eliminate retro-transposed (processed) pseudogenes

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




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

```
Query: 3   SRLLLNNGAKMPILGLGTWKSPPGQVTEAVKVAIDVGYRHIDCAHVYQNENEVGVAIQEK 62
           S ++LNNG K  +LGLGTWKSPPGQV EAVKVAI+  YRHIDC+HV+QN++      QE+
Sbjct: 2   SHIMLNNGTKTDMLGLGTWKSPPGQVAEAVKVAINTVYRHIDCSHVHQNKD------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  LKEQVVRREWLFIISKPWGICHRKCLVRGSCRKVLSGLELDYLDLHLIHWPTGCHPGKEF 115

Query: 123 FPLDESGNVVPSDTNILDTWAAMEELVDEGLVKAIGISNFNHLQVEMILNKPGLKYKPAV 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 KTTAQVLIRFPMQRNLVVIPKSVTPERIAENFKVFDFELSSQDMTTLLSYNRN 295
           + T+QVL+    QRNLVV P SVT +RIAENFKVFDFELSSQDMT+LLS NRN
Sbjct: 209 E-TSQVLMWLLTQRNLVVTPTSVTLDRIAENFKVFDFELSSQDMTSLLSCNRN 260
```

# Classification of Transcripts

- Naming takes place **after** the gene build has been completed
- Based on a mapping procedure on the sequence level
- **Ensembl Transcripts or Proteins** are mapped back to **UniProtKB/Swiss-Prot**, **NCBI RefSeq** and **UniProtKB/TrEMBL** entries
- Require **high sequence similarity**, but allow some ambiguity to accommodate genetic variation and sequencing errors
- Allow **incomplete coverage** due to assembly gaps and partially sequenced cDNA clones
- Assign external references with links to other databases

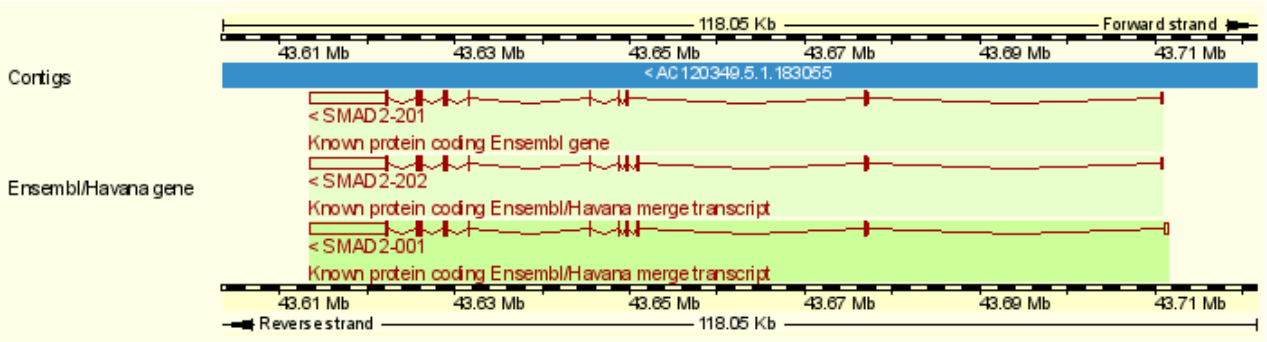
# Known and Novel Transcripts

- **Known** transcripts map to **species-specific** protein records. (targeted build)
- **Novel** transcripts do **not** map to **species-specific** protein records. (similarity build)

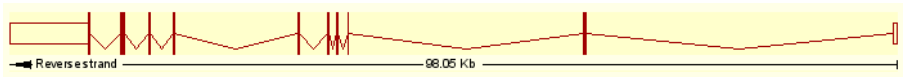
- But still, **all** Ensembl gene, transcript and protein predictions are based on **experimental evidence**.

# Supporting Evidence Display

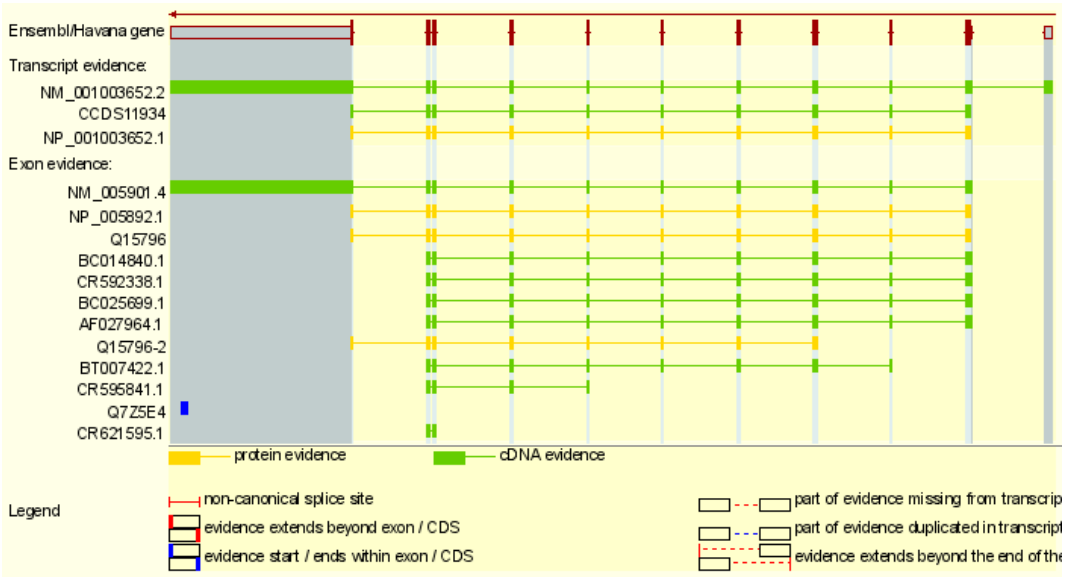
Gene-Summary




Transcript-Summary




Transcript-Supporting Evidence



# Supporting Evidence Tracks

# Gene Symbols and Names

- Naming is inferred from sequence mapping results
- Official gene symbols are assigned if available:
  - Human Genome Nomenclature Committee (HGNC)
  - Mouse Genome Informatics (MGI)
  - The Zebrafish Information Network (ZFIN)
  - Rat Genome Database (RGD)
  - Other species-specific nomenclature committees
- Otherwise named after the BAC clone
- Genes named after ‘best-named’ transcript
- Gene description is inferred from mapped database entries; the source is always given

# Orthologue Projections

- For species without nomenclature committees gene names and descriptions are projected via orthologue predictions
- Further Hints:
  - Orthologues can provide useful confirmation
  - If no description, check for any family description

# Adjusting the Gene Build

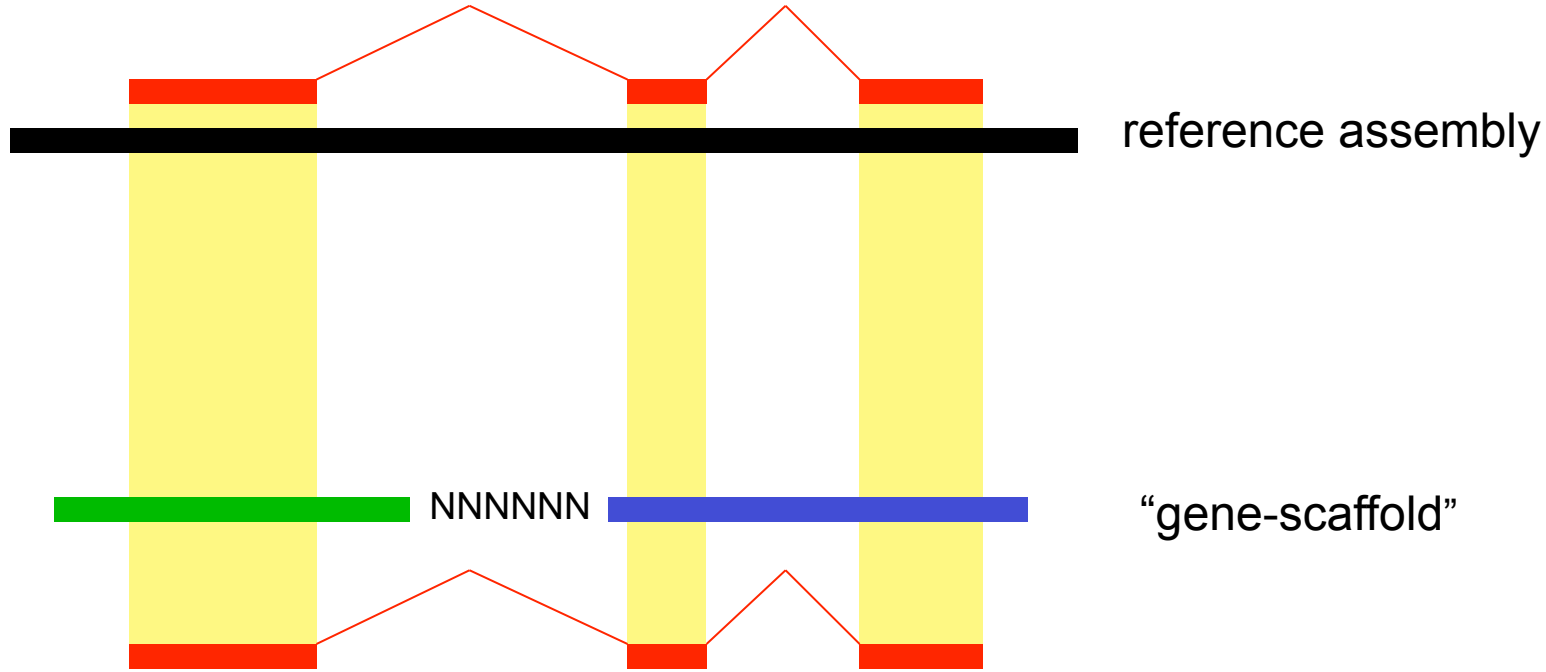
- Data availability e.g. (full-length) cDNA, protein, EST sets
- Targeted build most useful in human and mouse
- Similarity build more important in other species
- Structural Issues
  - *Danio rerio* (zebrafish)
    - Many similar genes near each other
    - Genome from different haplotypes
  - *Takifugu rubripes* (green-spotted puffer fish)
    - Very dense genome
    - Short introns
  - *Anopheles gambiae* (mosquito)
    - Many single-exon genes
    - Genes within genes
- Configuration files and experience provide flexibility

# Low-Coverage Genomes

- High-coverage sequencing is time-consuming and expensive
  - Limited to popular research organisms
  - BAC clones (>10x): human, mouse, zebrafish
  - Whole Genome Shotgun (6x): chimp, rat, chicken,...
- Other genomes would provide insights into genome evolution
- Low (~2x) coverage genome sequencing
  - Faster, cheaper, but only useful when annotated
- Assembled into lots of “scaffolds”
- “Classic” Ensembl gene-build would result in many partial and fragmented genes

# Low-Coverage Gene-Build

- Whole genome alignment to an annotated high-quality reference genome (i.e. human)
- Guided re-ordering of scaffolds
- Annotation of longer, more complete gene structures



# Genes and Transcripts

- Automated annotation (Ensembl genes)
- Manual curation (Havana genes)
- Ensembl EST genes
- *Ab initio* predictions
- Non-coding RNA genes

# Annotation Methods

## Automatic Annotation

- Ensembl pipeline
- Quick whole genome analysis ~ weeks
- Consistent annotation
- Use unfinished sequence or shotgun assembly
- No PolyA sites or signals, pseudogene
- Predicts ~ 70% loci

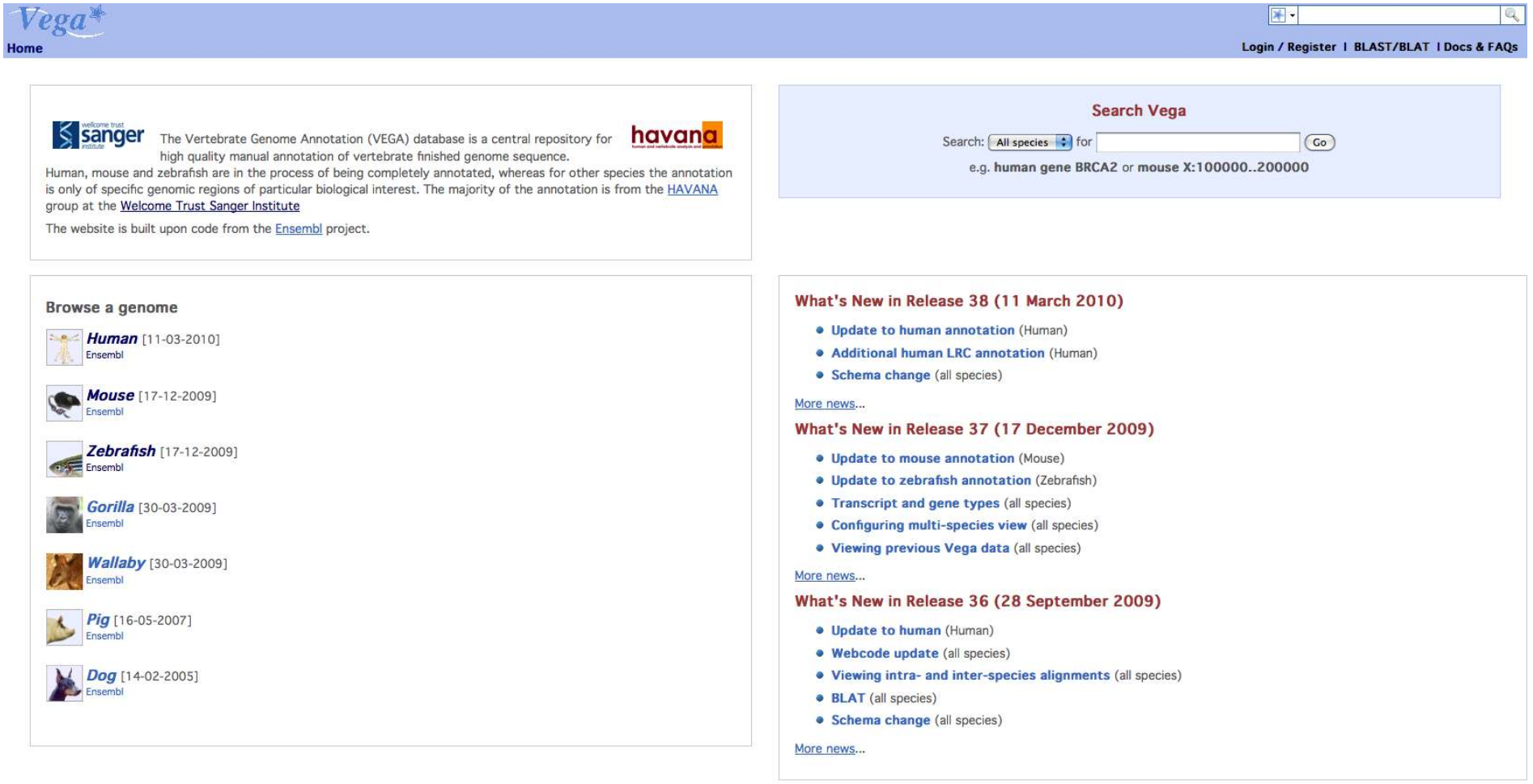
## Manual Annotation

- Havana group (WTSI)
- Extremely slow
  ~ 3 months for Chr 6
- Need finished sequences
- Flexible, can deal with inconsistencies
- Most rules have exception
- Consult publications as well as databases

# Manually Curated Human Chromosomes

- Manual annotation of finished clones
- Currently limited to chromosomes
  - 1, 6, 9, 10, 13, 20, 21, 22, X and Y
    (Havana - Wellcome Trust Sanger Institute)
  - 3, 12 (Baylor College)
  - 7 (Washington University)
  - 8, 15, 17, 18 (Broad Institute)
  - 14 (Génoscope)
  - 16 and 19 (DOE Joint Genome Institute)
- Vega Genome Browser http://vega.sanger.ac.uk/
- Other groups would contribute to Vega,
  but annotation never materialised

# Vega Genome Browser

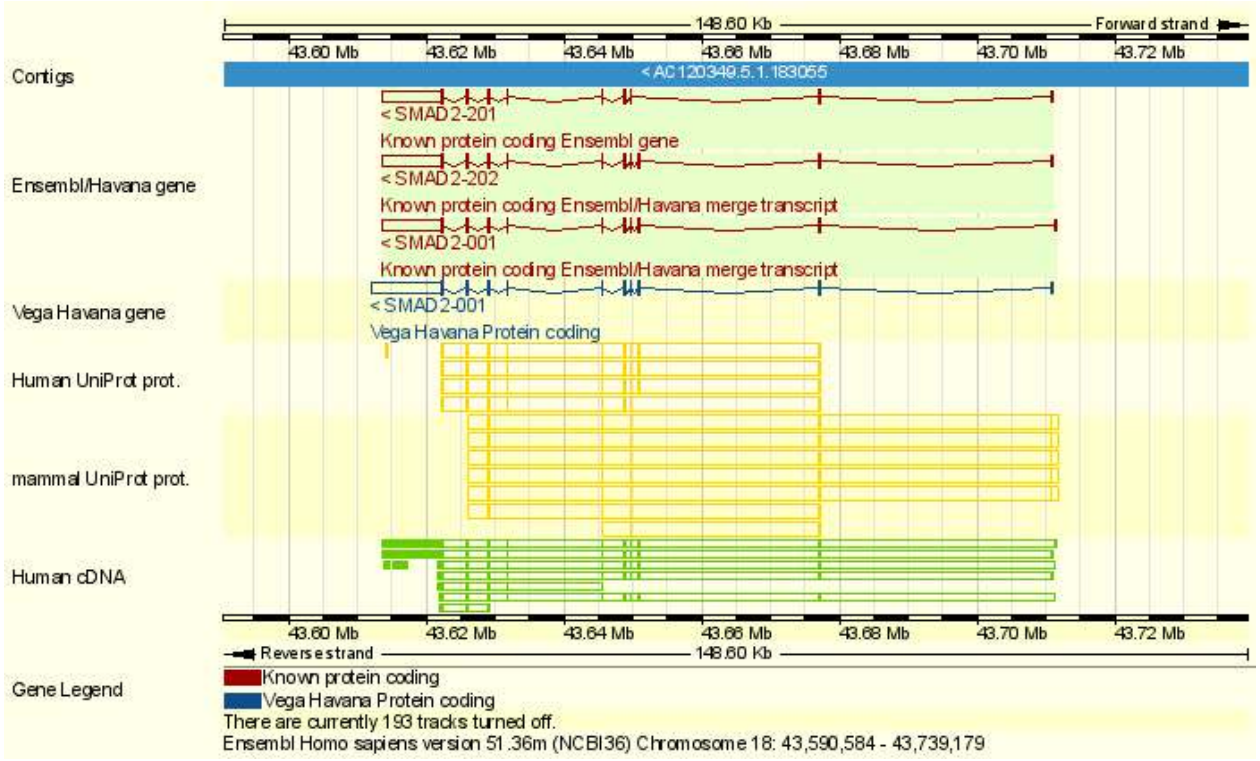



http://vega.sanger.ac.uk/

# Other Manually Curated Sets

- Havana
  - *Mus musculus* (finished chromosomes 2, 4, 11 and X)
  - *Danio rerio* (finished BAC clones)
- Manually curated gene sets imported into Ensembl
  - WormBase (*Caenorhabditis elegans*)
  - FlyBase (*Drosophila melanogaster*)
  - SGD (*Saccharomyces cerevisiae*)

# Havana Genes




Ensembl transcripts

Havana transcripts

Human proteins
(normal)

Human cDNAs
(half-height)

# Ensembl-Havana Merged Gene Set

**Automatically annotated Ensembl** transcript sets for human and mouse are merged into **manually curated Havana** transcript set.
Comparison provided for users

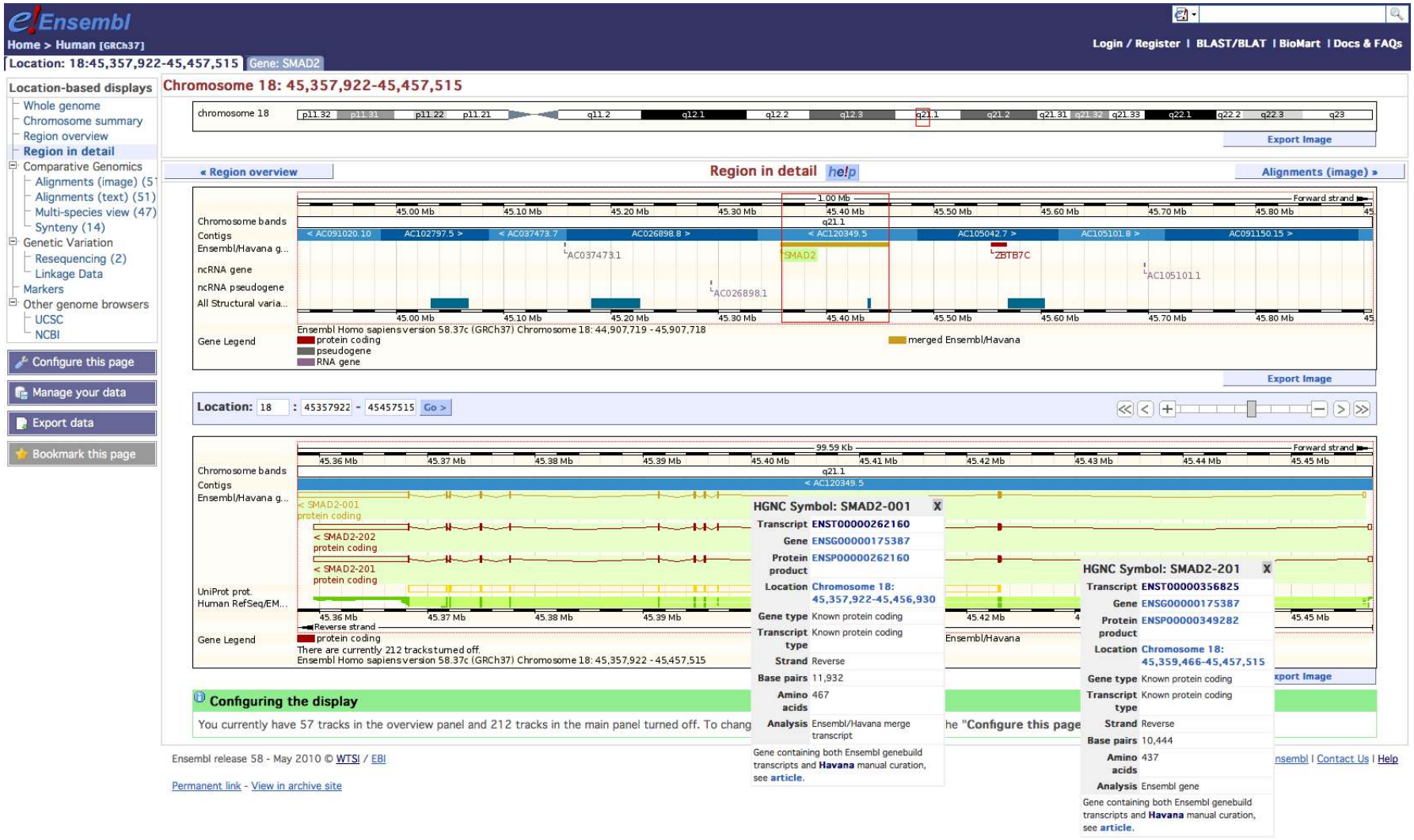
Transcripts:

- Ensembl/Havana: gold
- Ensembl: red / black
- Havana: blue

Genes:

- Ensembl/Havana: gold
- Ensembl: red / black
- Havana: blue

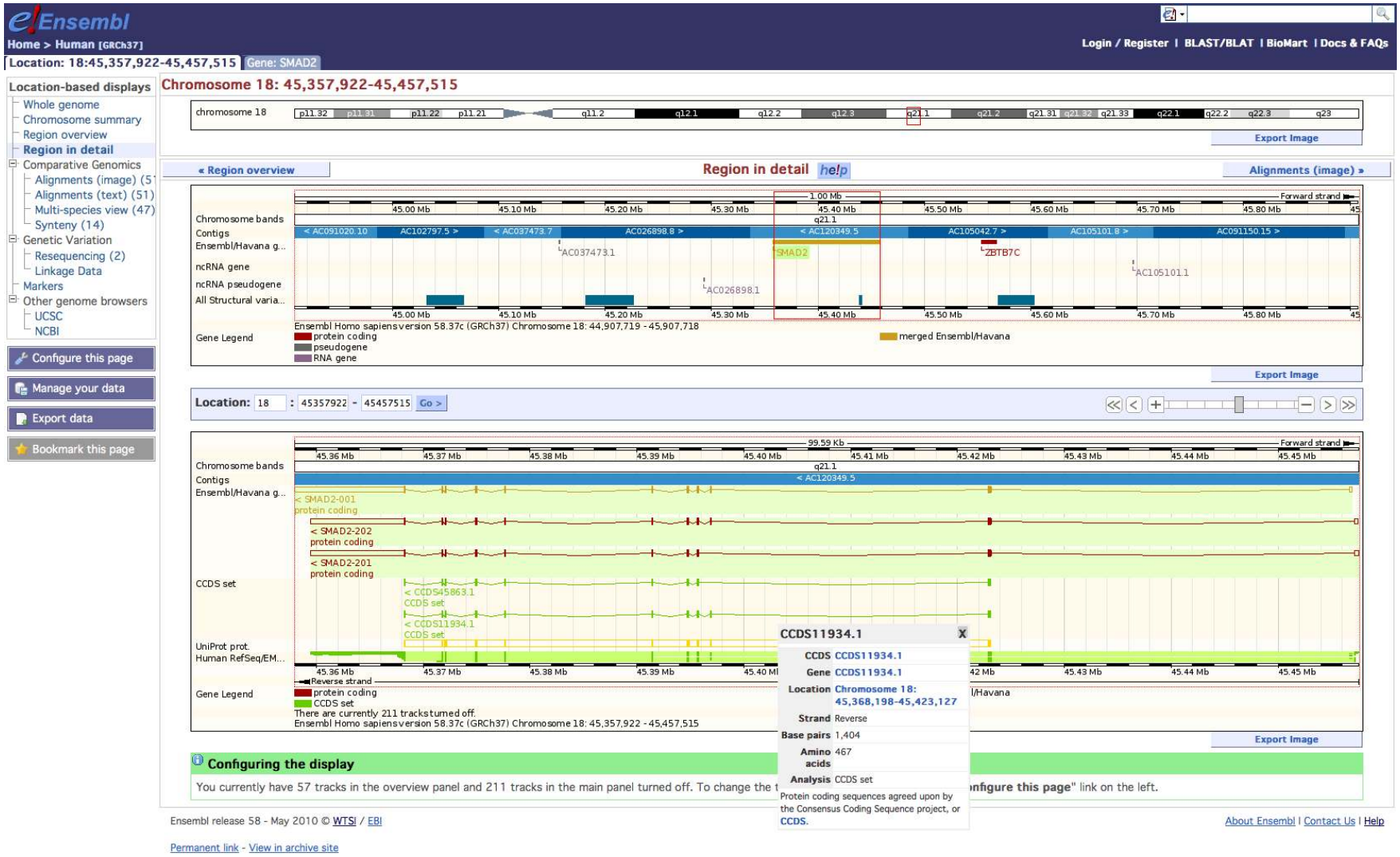
# Ensembl-Havana Merge in Location-View



# Consensus Coding Sequence (CCDS)

- Collaboration between NCBI, UCSC, Ensembl, Havana and UniProt to produce a set of stable, reliable, complete (ATG->stop) coding sequence (CDS) structures.
- Long term aim is to get to a single gene set for human and mouse
- The Ensembl annotation pipeline has been modified to retain these ‘blessed’ Consensus Coding Sequences

# CCDS in Location-View




wellcome trust sanger institute

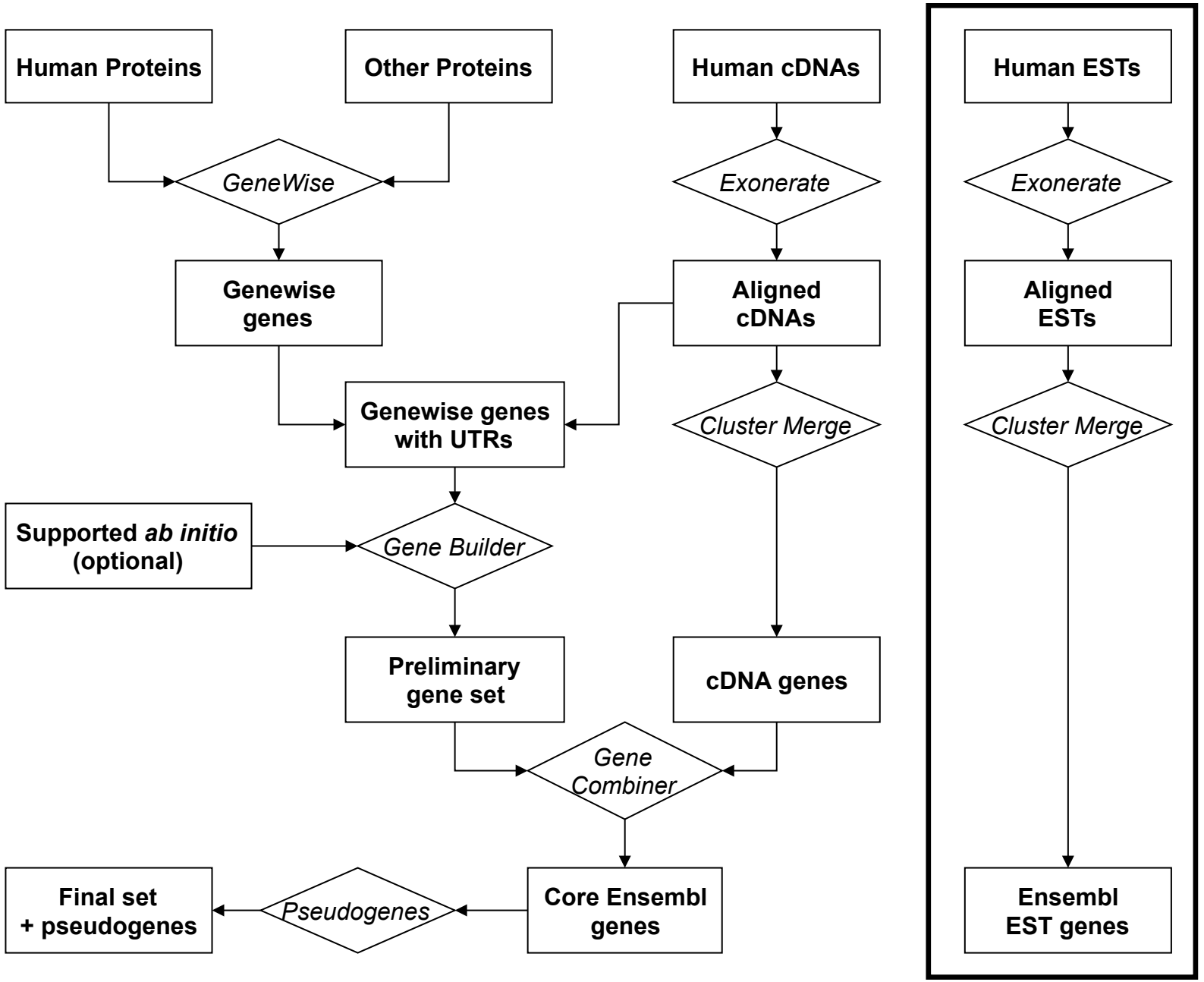
EMBL-EBI

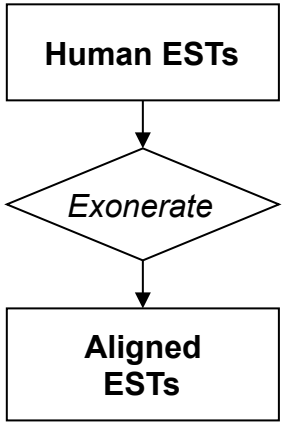
# Genes and Transcripts

- Automated annotation (Ensembl genes)
- Manual curation (Havana genes)
- Ensembl EST genes
- *Ab initio* predictions
- Non-coding RNA genes

# Expressed Sequence Tags

- Expressed Sequence Tags (ESTs) are **single** reads into cDNA library inserts
- High chance of sequencing mistakes
- EST libraries are regularly contaminated with genomic DNA
- Generally ~ 400 bp, so unlikely to cover a whole gene
- Therefore, less reliable but due to their large-scale, systematic nature a good source for rare alternative splice variants
- Ensembl provides an **EST Gene set** entirely **separate** from the **Ensembl Gene** set


Human Proteins
Other Proteins
Human cDNAs
Human ESTs
GeneWise
Exonerate
Exonerate
Genewise genes
Aligned cDNAs
Aligned ESTs
Genewise genes with UTRs
Cluster Merge
Cluster Merge
Supported ab initio (optional)
Gene Builder
Preliminary gene set
cDNA genes
Gene Combiner
Final set + pseudogenes
Pseudogenes
Core Ensembl genes
Ensembl EST genes

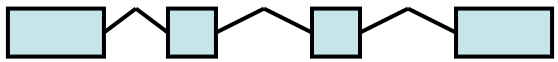

Human ESTs
Exonerate
Aligned ESTs

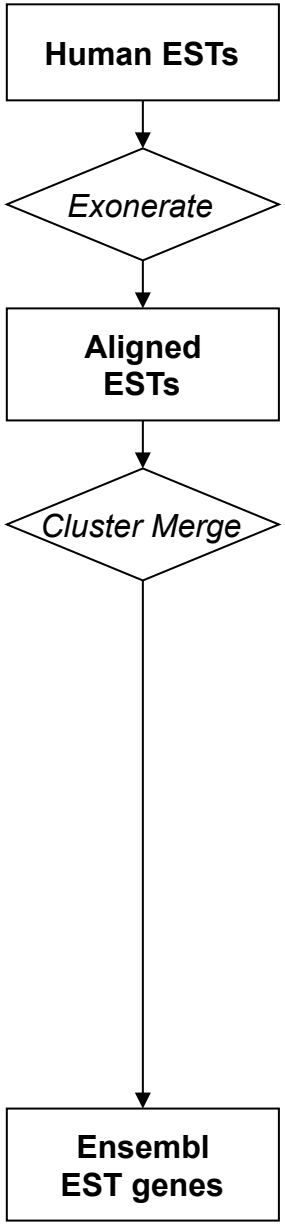
# Initial EST Analysis

**Pre-filter species-specific ESTs**
- Remove short sequences < 100 bp
- Remove ESTs with too many Ns

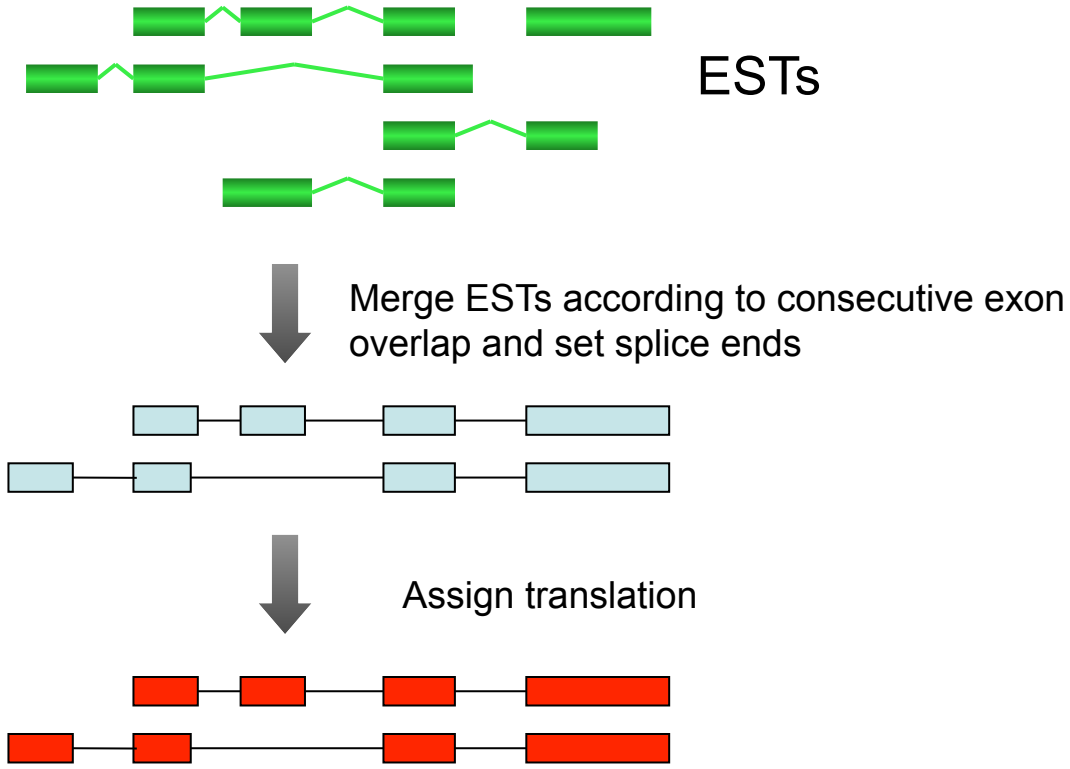
**Map ESTs**
- Align to the genome with Exonerate
- Record alignment coverage, identity and genomic location

**Filter on identity and depth**
- 5.5 million human ESTs in dbEST
- Only about 1/3 is mapped




Human ESTs
Exonerate
Aligned ESTs
Cluster Merge
Ensembl EST genes

# Alternative Splicing Forms




Alternative transcripts with translation and UTRs

# EST Genes and ESTs


Human ESTs
(stacked unlimited)

Human cDNAs
(half-height)

Human proteins
(normal)

EST transcripts

Ensembl transcripts

**Tracks are limited to 7 at any one point under the "Normal" and "Half-height" track display options.**
**The full data set is accessible via the "Stacked unlimited" configuration option.**

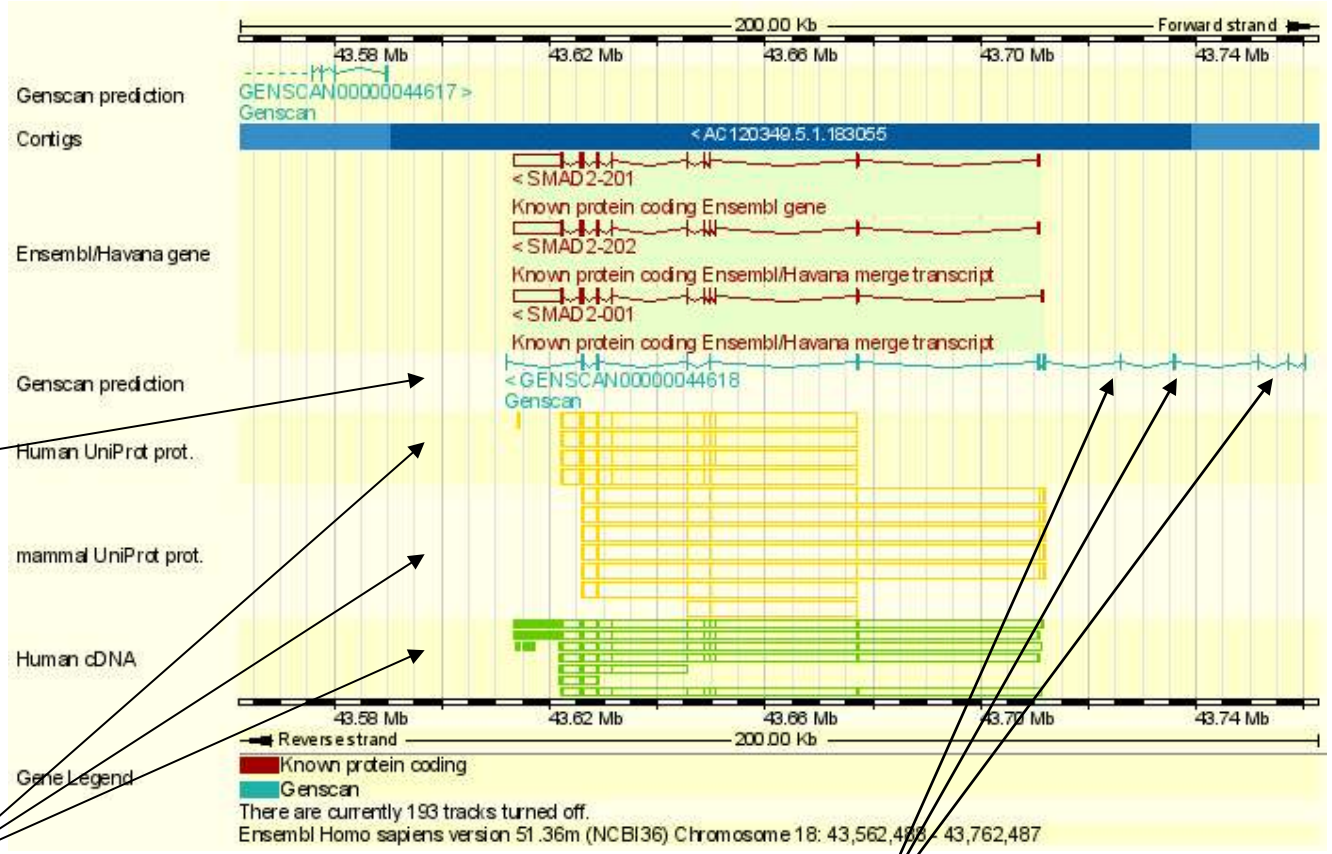
# Genes and Transcripts

- Automated annotation (Ensembl genes)
- Manual curation (Havana genes)
- Ensembl EST genes
- *Ab initio* predictions
- Non-coding RNA genes

# *Ab initio* Predictors

- Algorithm predicts translatable transcript structures solely on the basis of genome sequences
- **No** validation with biological expression information
- GENSCAN for vertebrate genomes
- SNAP better for invertebrates
- NB: Both programs are **over-predicting** transcript structures

# *Ab initio* Transcripts




GENSCAN prediction

Protein and mRNA evidence supporting GENSCAN exons

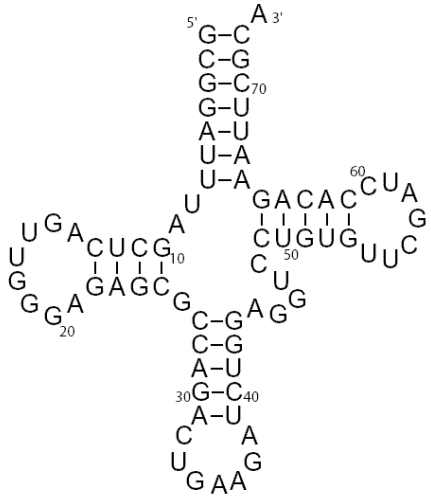
Unsupported GENSCAN exons

# Genes and Transcripts

- Automated annotation (Ensembl genes)
- Manual curation (Havana genes)
- Ensembl EST genes
- *Ab initio* predictions
- Non-coding RNA genes

# Non-coding RNAs (ncRNAs)

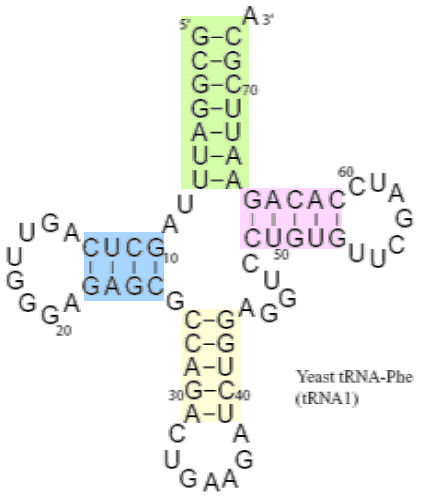
- Functional RNAs
- Families share conserved secondary structure
- Low sequence identity
- Ribosome (rRNA)
- Small nucleolar RNAs (snoRNAs)
- Spliceosome, small nuclear ribonucleoproteins (snRNPs)
- Small nuclear RNAs (snRNAs)
  - U1, U2, U4, U5 and U6
  - U11, U12, U4atac and U6atac
- tRNAs



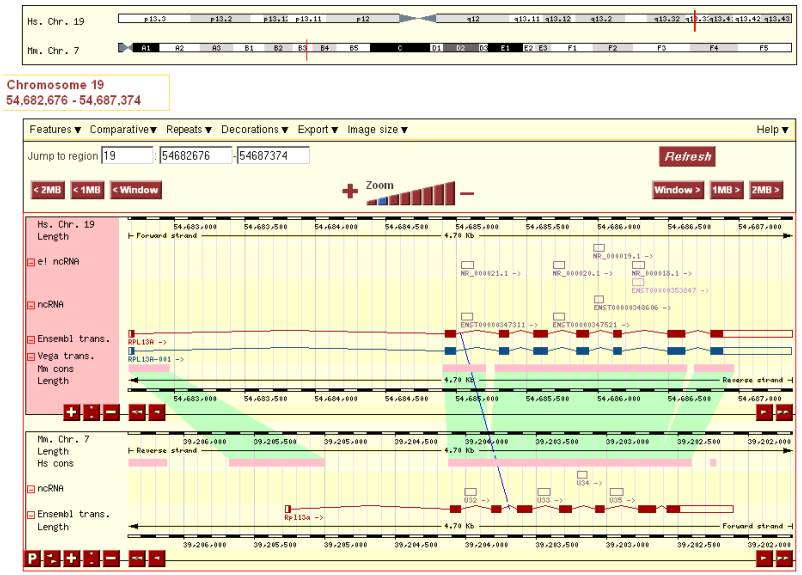
# RFAM - RNA family database

- Collection of RNA families
  - Hand-made multiple-sequence alignments
  - Consensus secondary structures
  - Covariance Models (CMs)
    - Simultaneously model RNA sequence and secondary structure
- Scan models over subset of EMBL to build family alignments

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

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



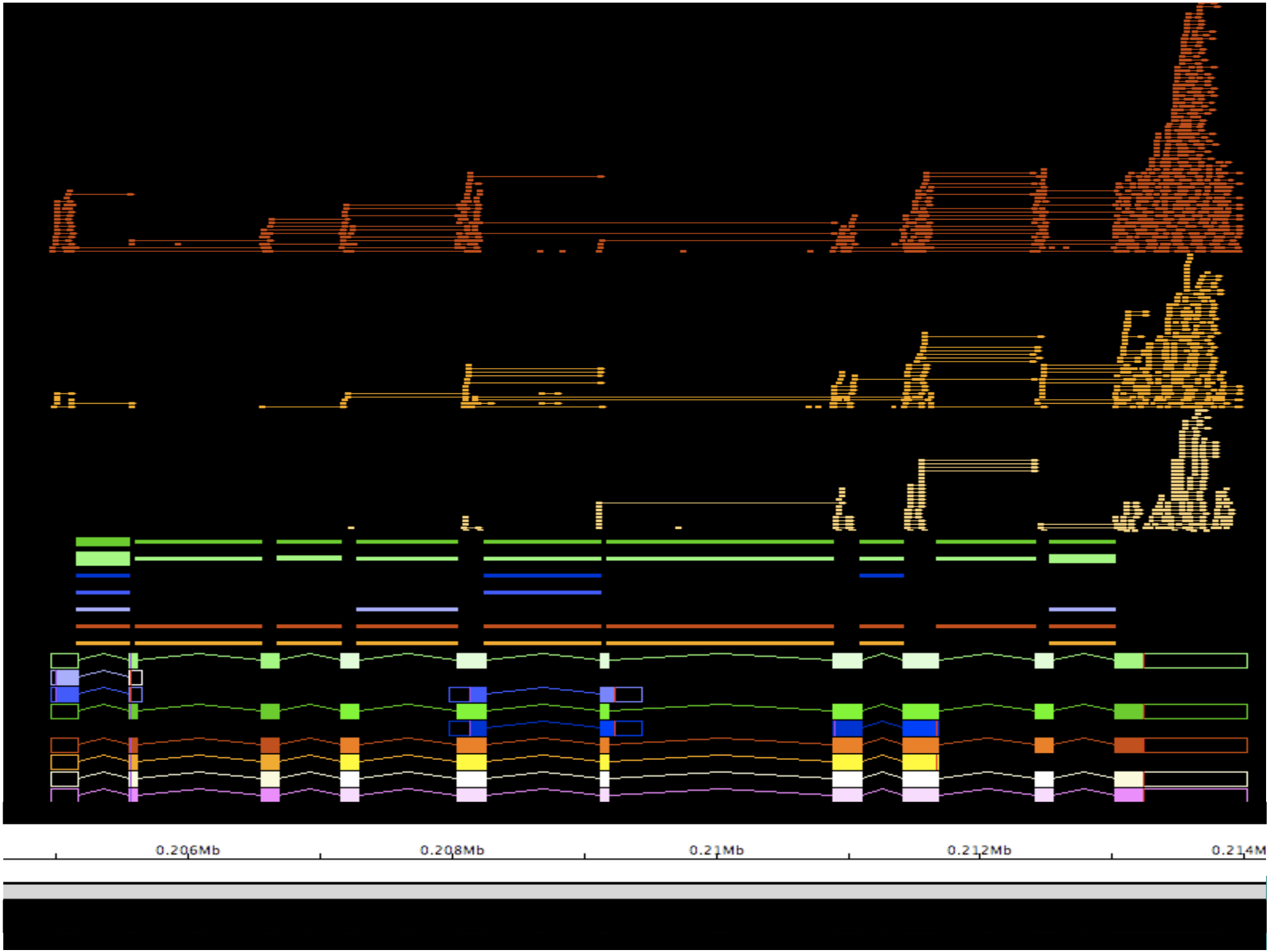
# microRNA (miRNA)



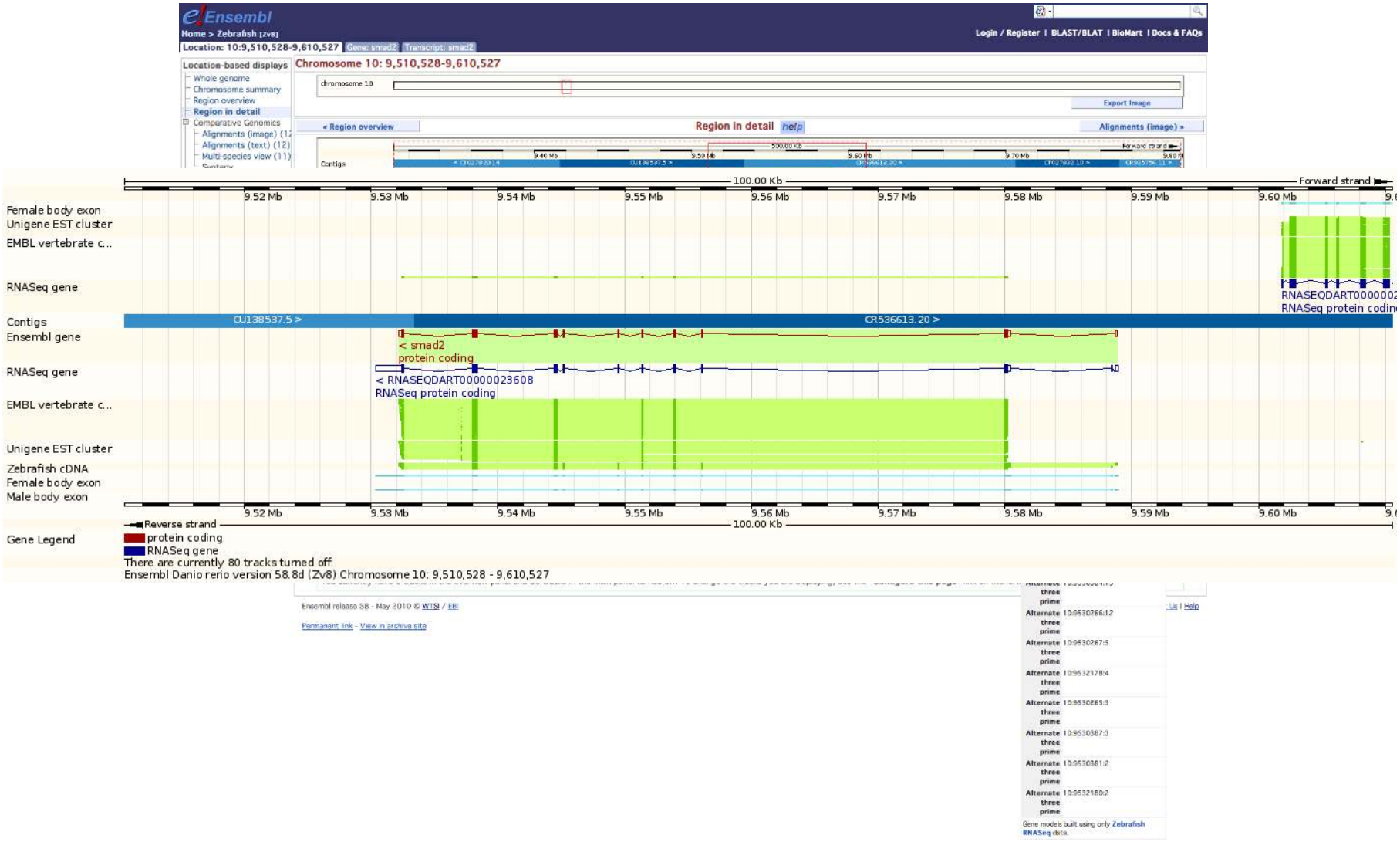

- Highly conserved across species
- Detectable with BLAST
- Precursor stem loop sequence ~ 70nt
- Mature miRNA ~ 21nt
- BLAST genomic against miRBase precursors
- RNAfold (Vienna RNA) used to test for stem loop
- Mature sequence identified
- (only 2 nucleotide changes tolerated)

# Future Strategy

- Next generation sequencing
- Massively parallel
- Very cheap in comparison to Sanger sequencing
- But rather short reads
- For mRNA sequencing: RNA-Seq
- An entirely new annotation strategy is needed
- Currently in development and evaluation phase


0.206Mb
0.208Mb
0.21Mb
0.212Mb
0.214M

# *Danio rerio* RNA-Seq genes (smad2)




# QUESTIONS

# ANSWERS