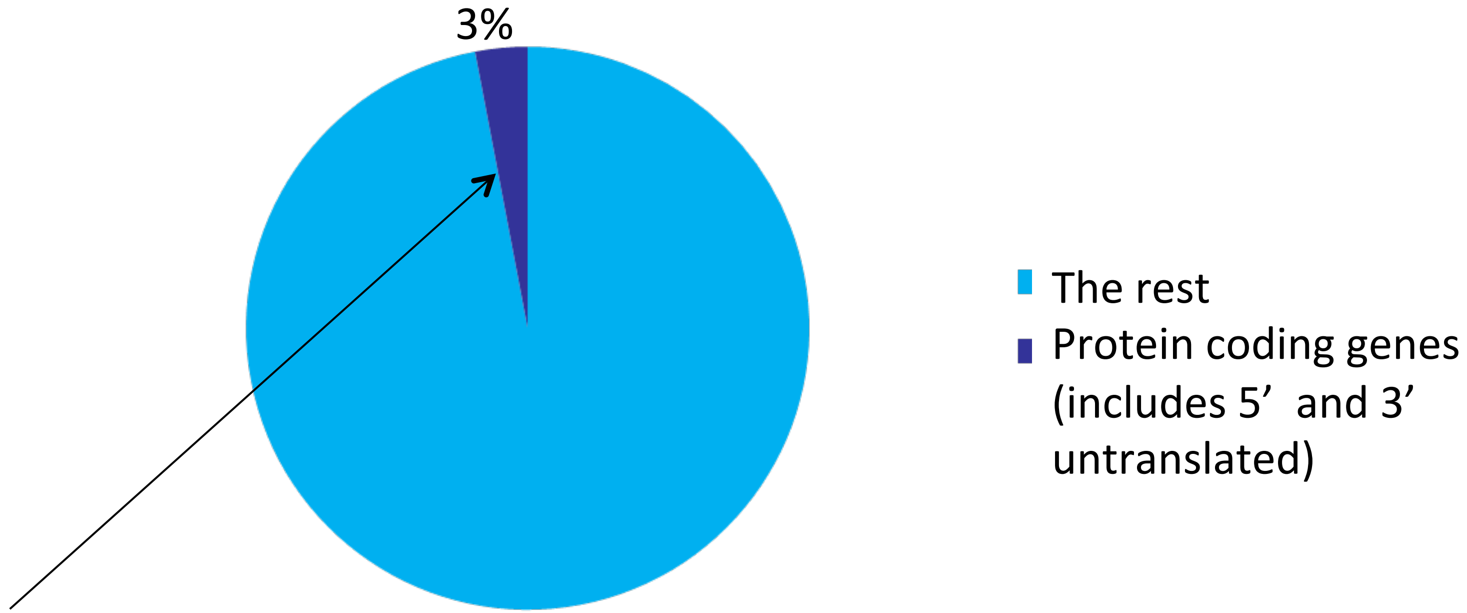
Predicting genes using RNA sequencing

9th October 2012

Predicting genes – finding the critical 3%

Human Genome of 3,300,000,000 bases

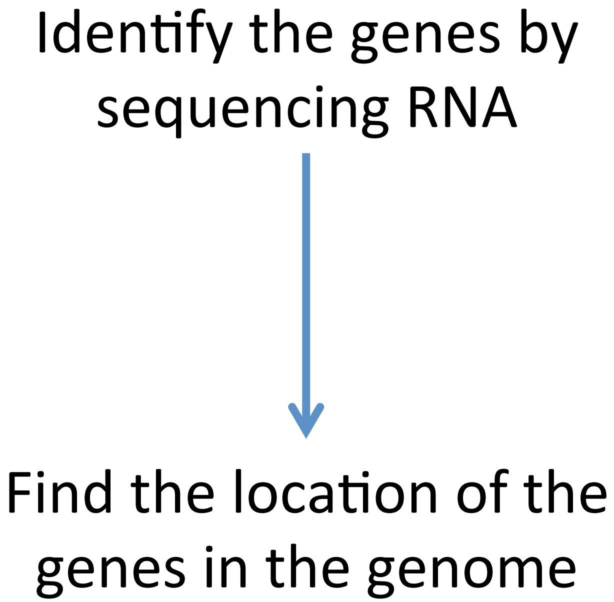


In Ensembl 68
22,058 protein coding genes

3% - like finding 7or 8 people in this crowd

Easier if you know who you are looking for…

Predicting genes using RNA sequencing



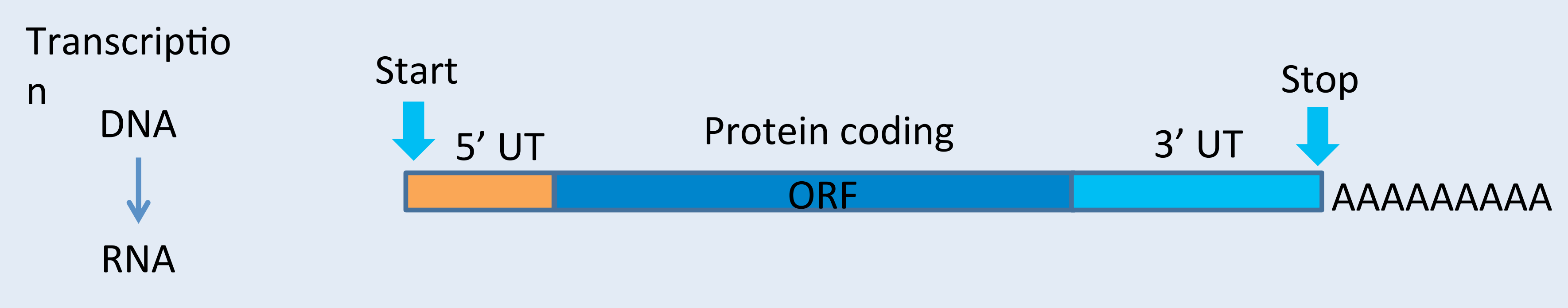
Identify the genes by sequencing RNA

Find the location of the genes in the genome

Predicting genes using RNA sequencing

1. Getting the RNA template
2. Sequencing the RNA
3. Predicting gene models
4. Producing a gene build
5. Using the gene build

Predicting genes – what are we looking for?



Transcription

DNA

RNA

Start

5’ UT

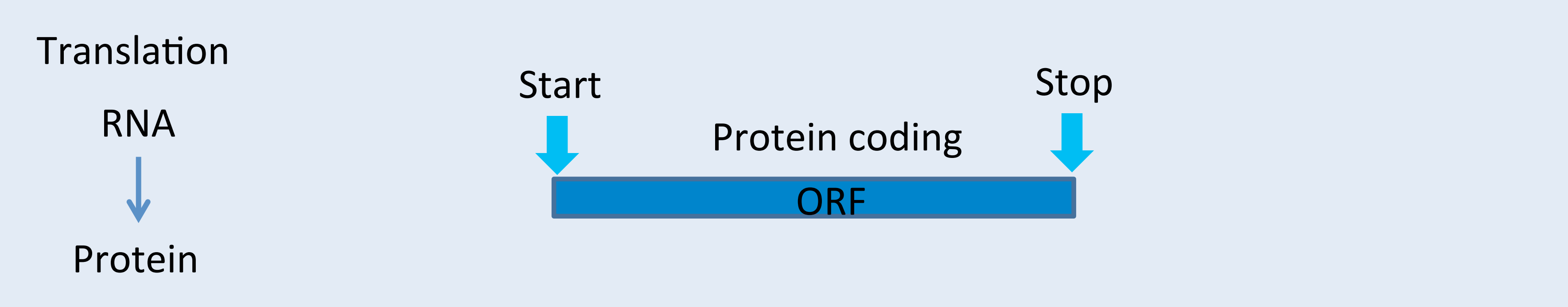
Protein coding

3’ UT

Stop

ORF

AAAAAAAAA



Translation

RNA

Protein

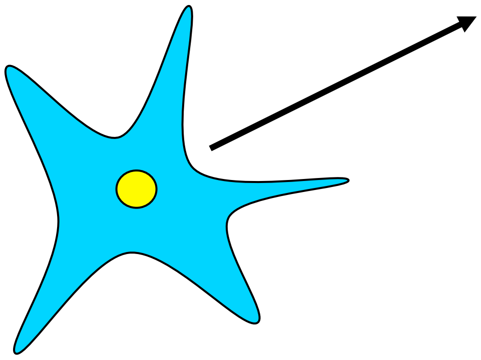
Start

Protein coding

Stop

ORF

Getting hold of the RNA



Whole cell gives total RNA

rRNA

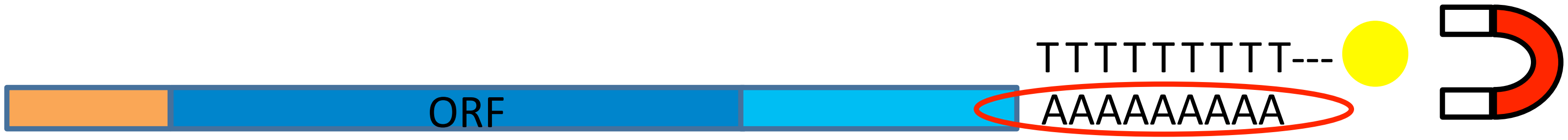
mRNA

snRNA

tRNA

snoRNA

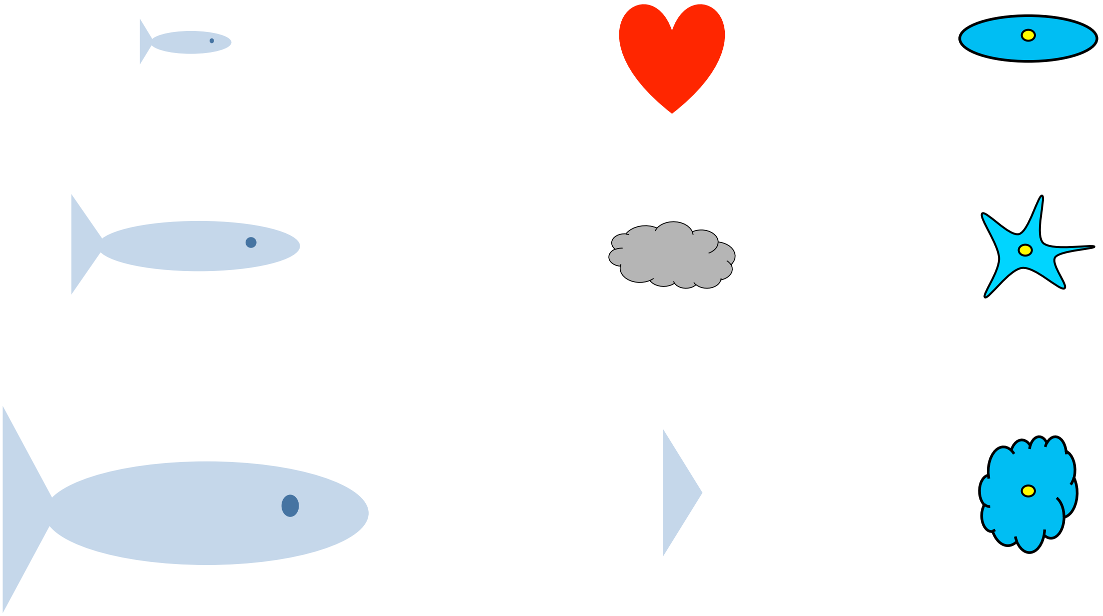
miRNA



Getting hold of the RNA

For each sample pick can pick …

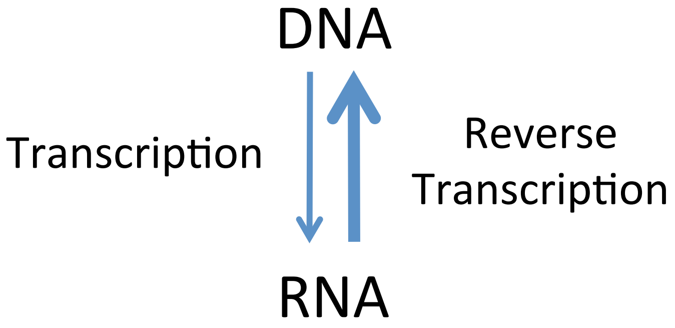
One Developmental Stage + One Tissue + One Cell type = One RNA sample



Each sample has a different set of mRNAs each representing a genes –
the transcriptome of the sample

Preparing the sequence template – reverse transcription

Sequencing machines use DNA as a template so the RNA must be converted back into complementary DNA or cDNA



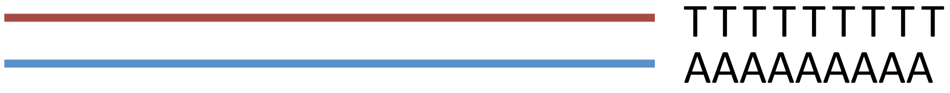
DNA

Transcription

Reverse Transcription

RNA

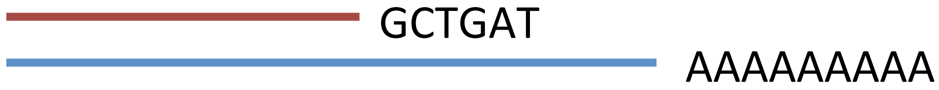
PolyT prime



TTTTTTTTT

AAAAAAAAA

Random prime



GCTGAT

AAAAAAAAA

Sequencing methods

Capillary sequencing

One sequence from one template about 800 bases long

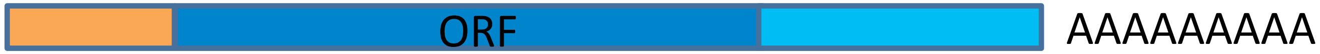
Template linked directly to sequence

Illumina sequencing or GAII or HiSeq or MiSeq

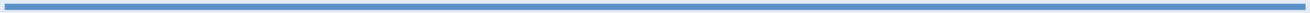
Millions of sequences from millions of templates all at the same time about 100 bases long

No template/sequence information

Sequencing methods

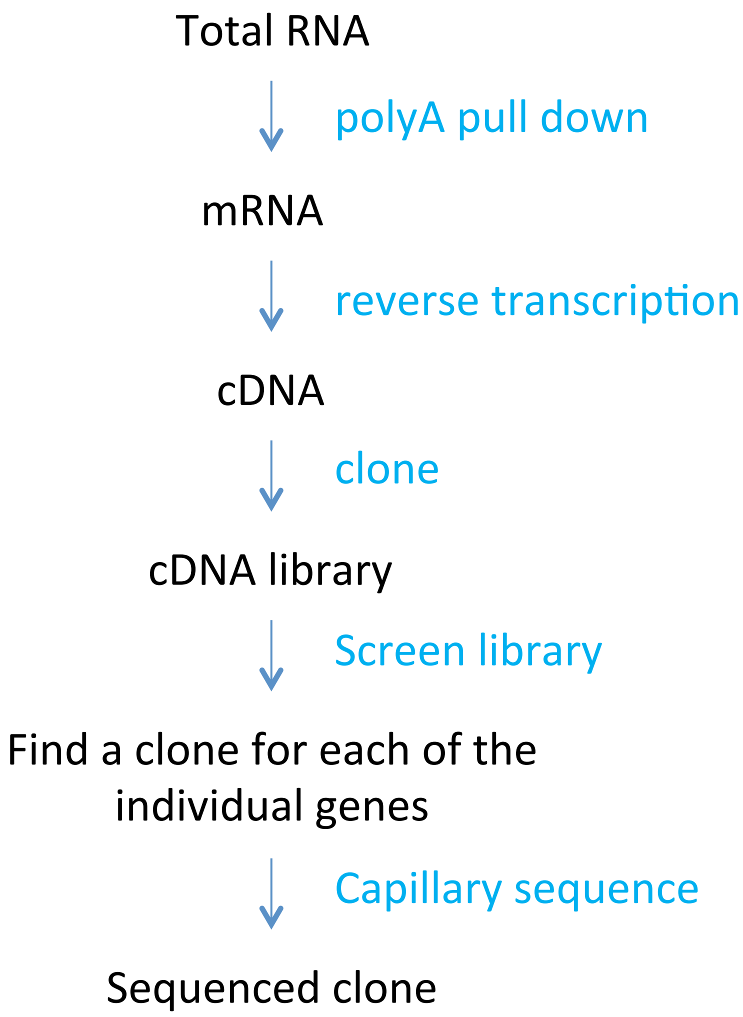


Traditional method
Sequence a single cDNA clone by capillary sequencing



NGS method
Sequence many small fragments of cDNA by Illumina sequencing and assemble into one sequence

Traditional RNA sequencing



Total RNA

↓ polyA pull down

mRNA

↓ reverse transcription

cDNA

↓ clone

cDNA library

↓ Screen library

Find a clone for each of the individual genes

↓ Capillary sequence

Sequenced clone

All tissues/cell types at all developmental stages

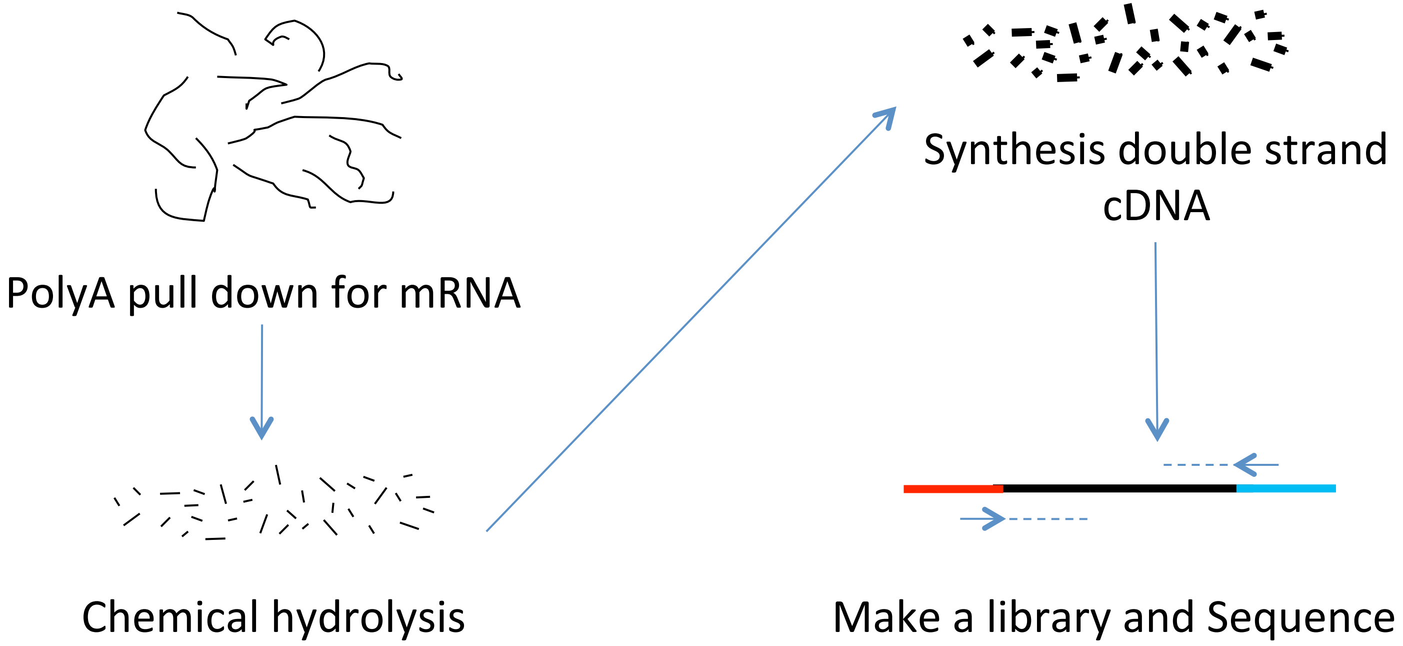
Thousands of clones stored in freezers

Screening many libraries to find rare clone

Whole thing costs loads!

Next generation RNA sequencing

RNAseq



PolyA pull down for mRNA

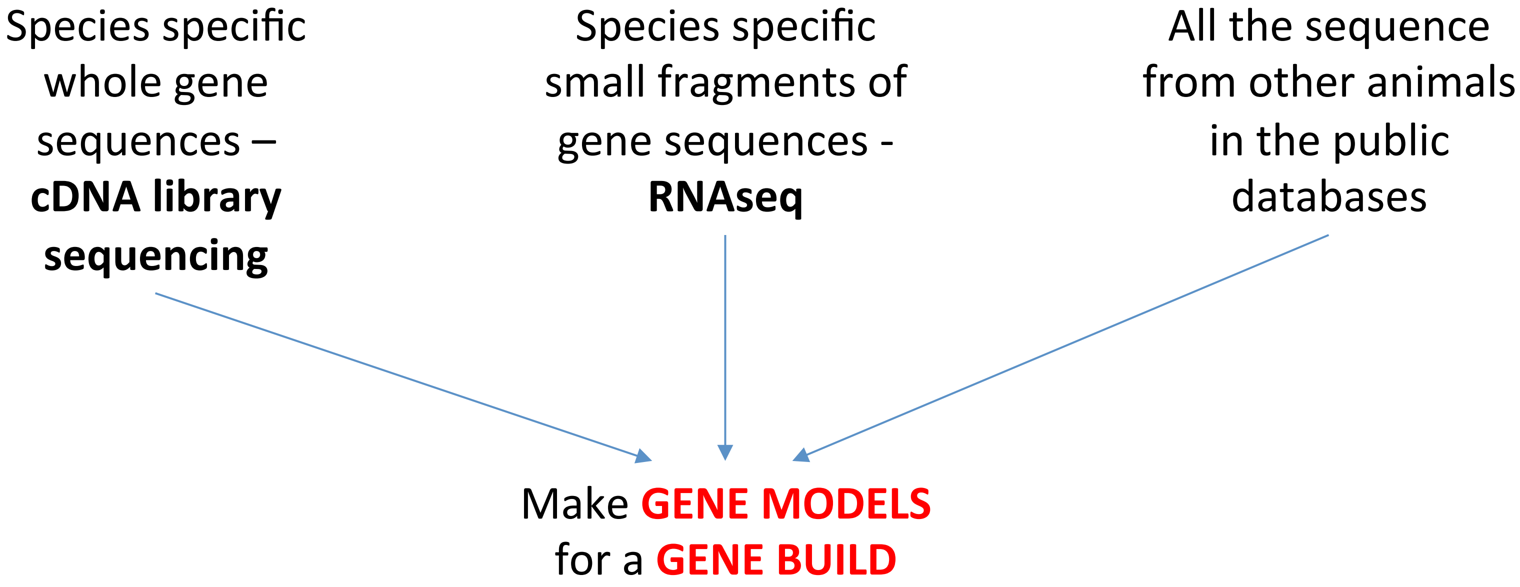
Chemical hydrolysis

Synthesis double strand cDNA

Make a library and Sequence

Note each fragment is sequenced from both ends to make a pair of reads

Annotation genes on genome reference sequence



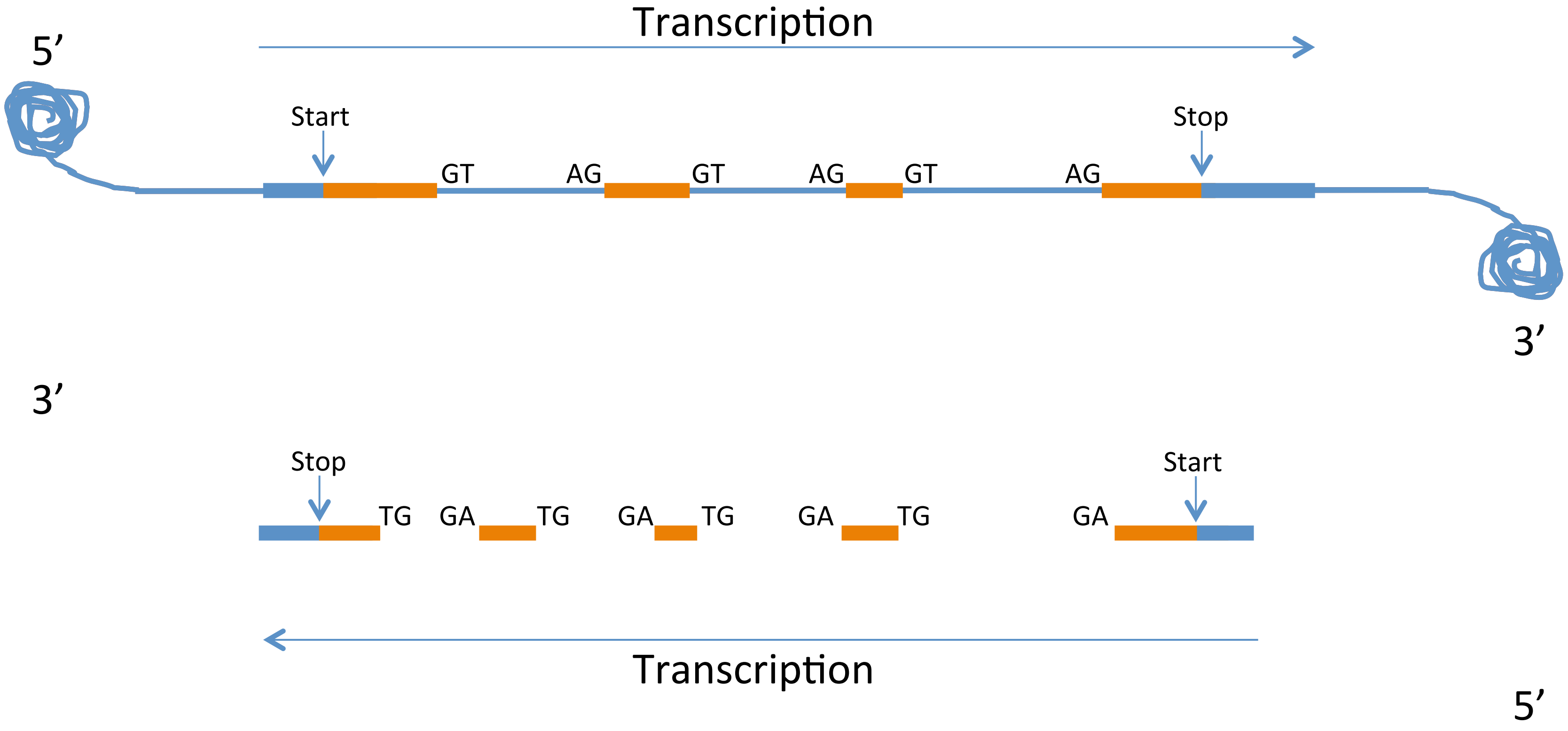
Species specific whole gene sequences – **cDNA library sequencing**

Species specific small fragments of gene sequences - **RNAseq**

All the sequence from other animals in the public databases

Make **GENE MODELS** for a **GENE BUILD**

Annotation gene models



Building genes from species-specific cDNAs

```
>part of human C22 reference sequence
gccgggctcatgctgggcacctccgggcaagctttgtgacttagaggtgtggggccactggtcaccctcggtggctcagaggctgtggctcc
atggctcatgagcgcctcctgtgtcccagacctgagaccatgcgtatgtcccgaggcggctgctccggtctcttgcggtgtggaaagggaac
gacctacgagggcggtgtccgagagcctgccttggccttctggccaggtcatatcgctcccggtcagtccgcaggccctctccttggaaccct
ggccccaccaccccaaccttgatggcgaactgagtgactgaccagcctcctgcccccaggcgtgacccacggagctggccagctccctg
gacctgctgcctaccctggcagccctggctggggccccactgcccaatgtcaccttggatggctttgacctcagcccccctgctgctgggcaca
ggcaaggtagggccggtgacccctgatcccagatccttggcccctgtcctggccttcccctggggtgagtgtgggcagtgcctgagagtctgt
gcctcagtgcctcctgcactgagtggcatccaagtggcgccacctctcaggttcctgggtgggcaagaagcggtgcacg
```

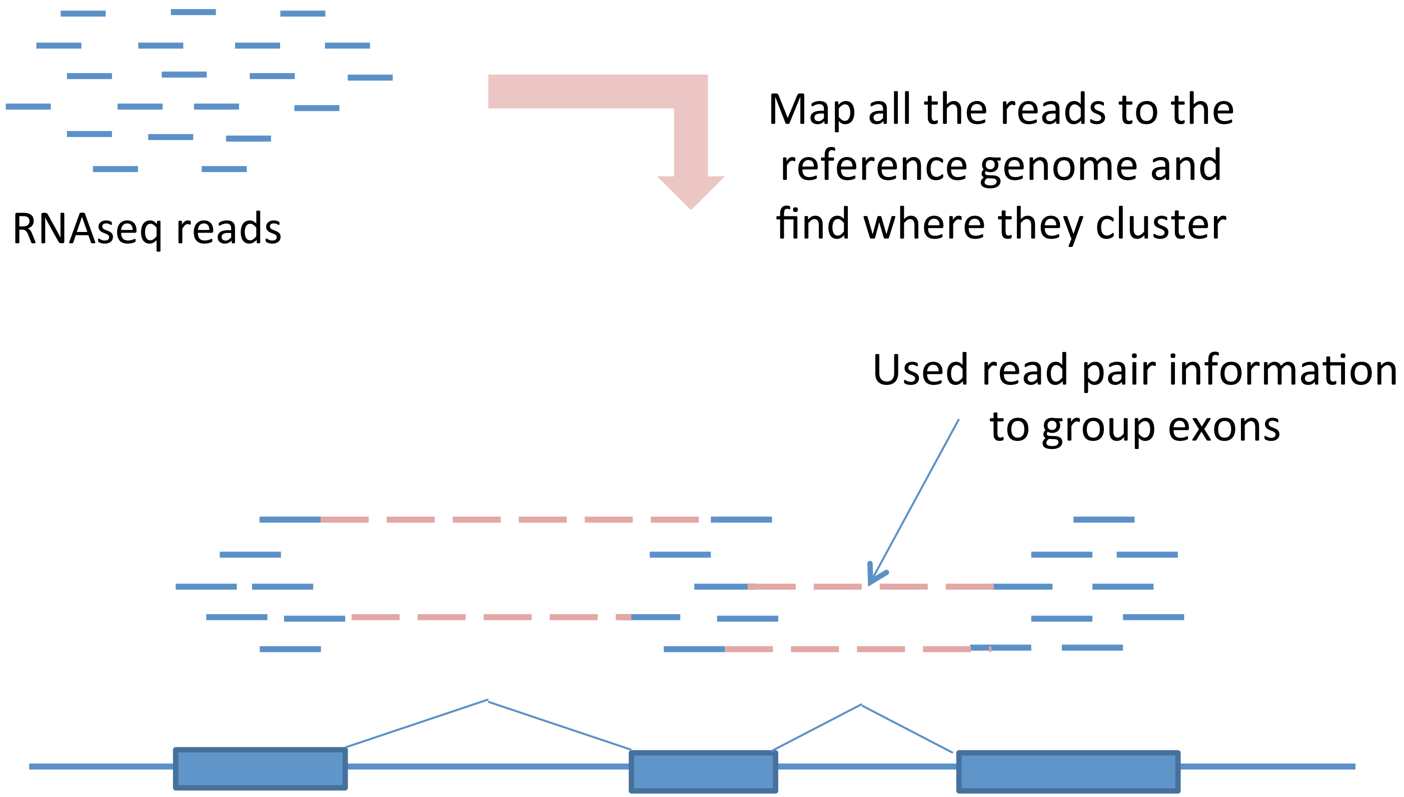
PLUS

```
>exons of the gene ARSA cDNA sequence
acctgagaccatgcgtatgtcccgaggcggctgctccggtctcttgcggtgtggaaagggaacgacctacgagggcggtgtccgagagcctgccttggcctt
ctggccaggtcatatcgctcccggcgtgacccgacgagctggccagctccctggacctgctgcctaccctggcagccctggctggggccccactgcccaatg
tcaccttggatggctttgacctcagcccccctgctgctgggcacaggcaag
```

EQUALS

```
>part of human C22 sequence annotated with two exons from the ARSA cDNA sequence
gccgggctcatgctgggcacctccgggcaagctttgtgacttagaggtgtggggccactggtcaccctcggtggctcagaggctgtggctcc
atggctcatgagcgcctcctgtgtcccagacctgagaccatgcgtatgtcccgaggcggctgctccggtctcttgcggtgtggaaagggaac
gacctacgagggcggtgtccgagagcctgccttggccttctggccaggtcatatcgctcccggtcagtccgcaggccctctccttggaaccct
ggccccaccaccccaaccttgatggcgaactgagtgactgaccagcctcctgcccccaggcgtgacccacggagctggccagctccctg
gacctgctgcctaccctggcagccctggctggggccccactgcccaatgtcaccttggatggctttgacctcagcccccctgctgctgggcaca
ggcaaggtagggccggtgacccctgatcccagatccttggcccctgtcctggccttcccctggggtgagtgtgggcagtgcctgagagtctgt
gcctcagtgcctcctgcactgagtggcatccaagtggcgccacctctcaggttcctgggtgggcaagaagcggtgcacg
```

Building genes from RNAseq



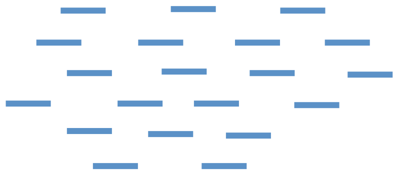
RNAseq reads

Map all the reads to the reference genome and find where they cluster

Used read pair information to group exons

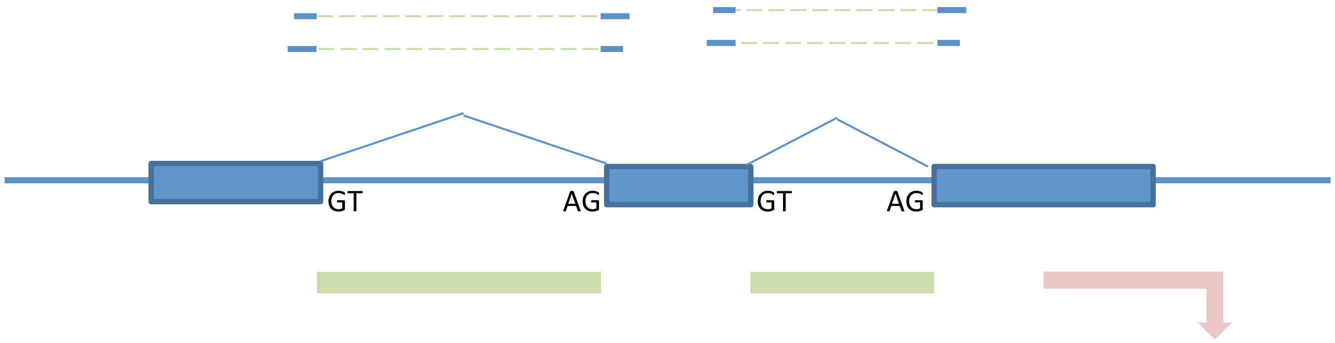
Approximate gene model

Building genes from RNAseq



Unmapped RNAseq reads

Map the rest of the reads to the approximate gene models allowing for read splicing

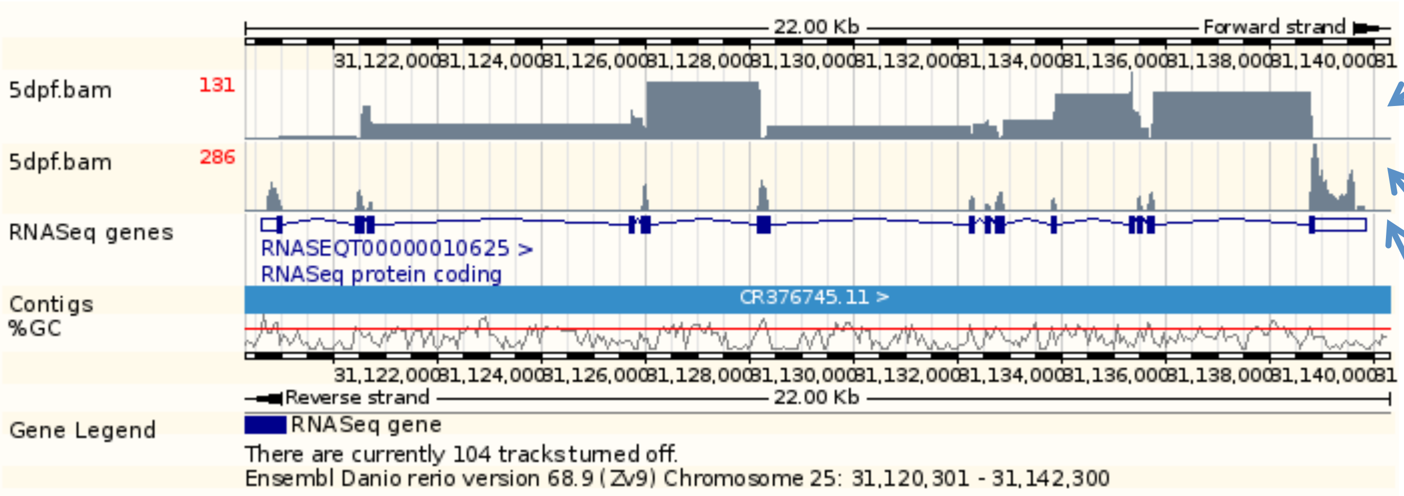


Intron driven annotation

Intron database

Building genes from RNAseq

Example of an RNAseq gene built from the short reads for apoptosis inhibitor 5 (api5)



There are currently 104 tracks turned off.

Ensembl Danio rerio version 68.9 (Zv9) Chromosome 25: 31,120,301 - 31,142,300

Sequence mapped to introns

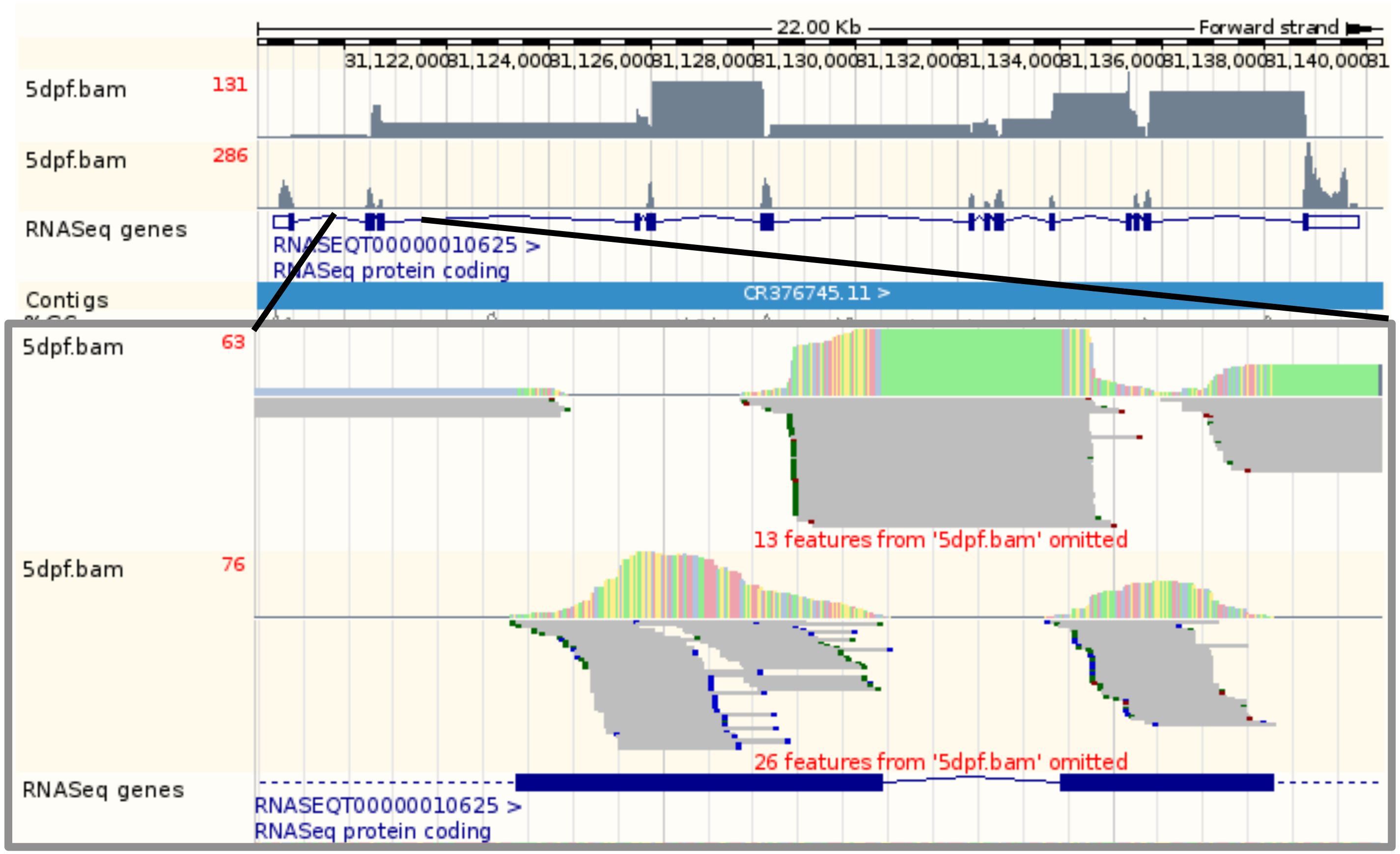
Sequence mapped to exons

RNAseq gene model

Zebrafish gene build http://www.ensembl.org/index.html release 68

Building genes from RNAseq

Example of an RNAseq gene built from the short reads for apoptosis inhibitor 5 (api5)



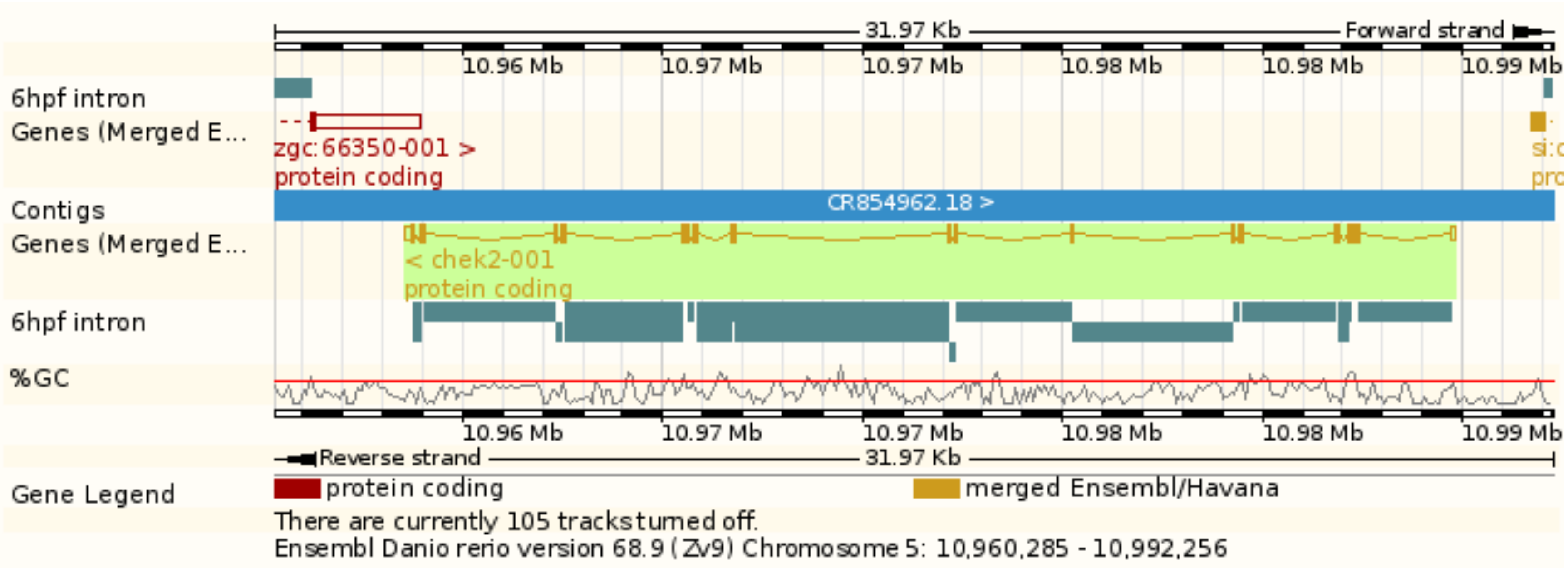
Sequence mapped to introns

Sequence mapped to exons

RNAseq gene model

Zebrafish gene build http://www.ensembl.org/index.html release 68

Introns show alternative transcripts



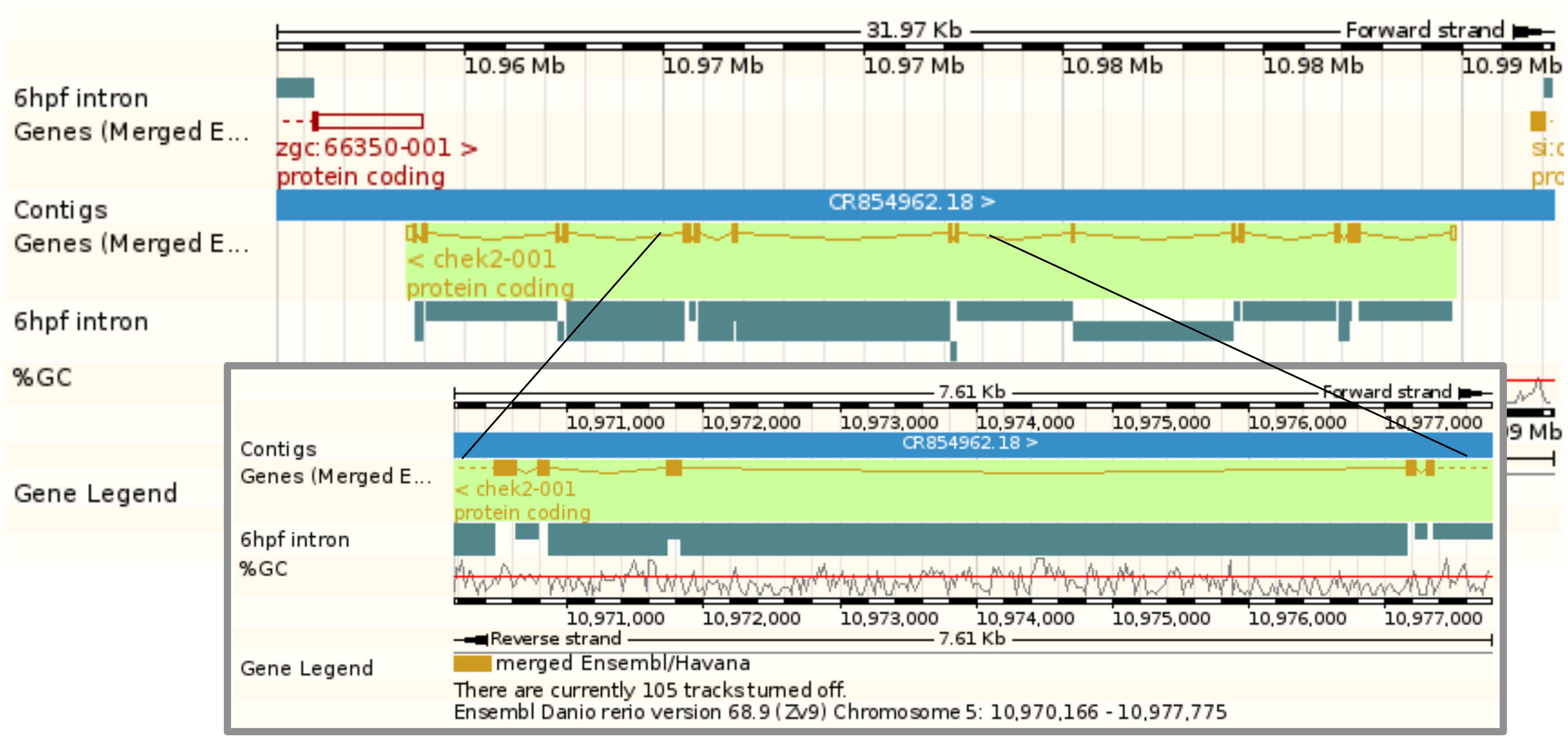
There are currently 105 tracks turned off.

Ensembl Danio rerio version 68.9 (Zv9) Chromosome 5: 10,960,285 - 10,992,256

chek2 – checkpoint kinase 2

Zebrafish gene build http://www.ensembl.org/index.html release 68

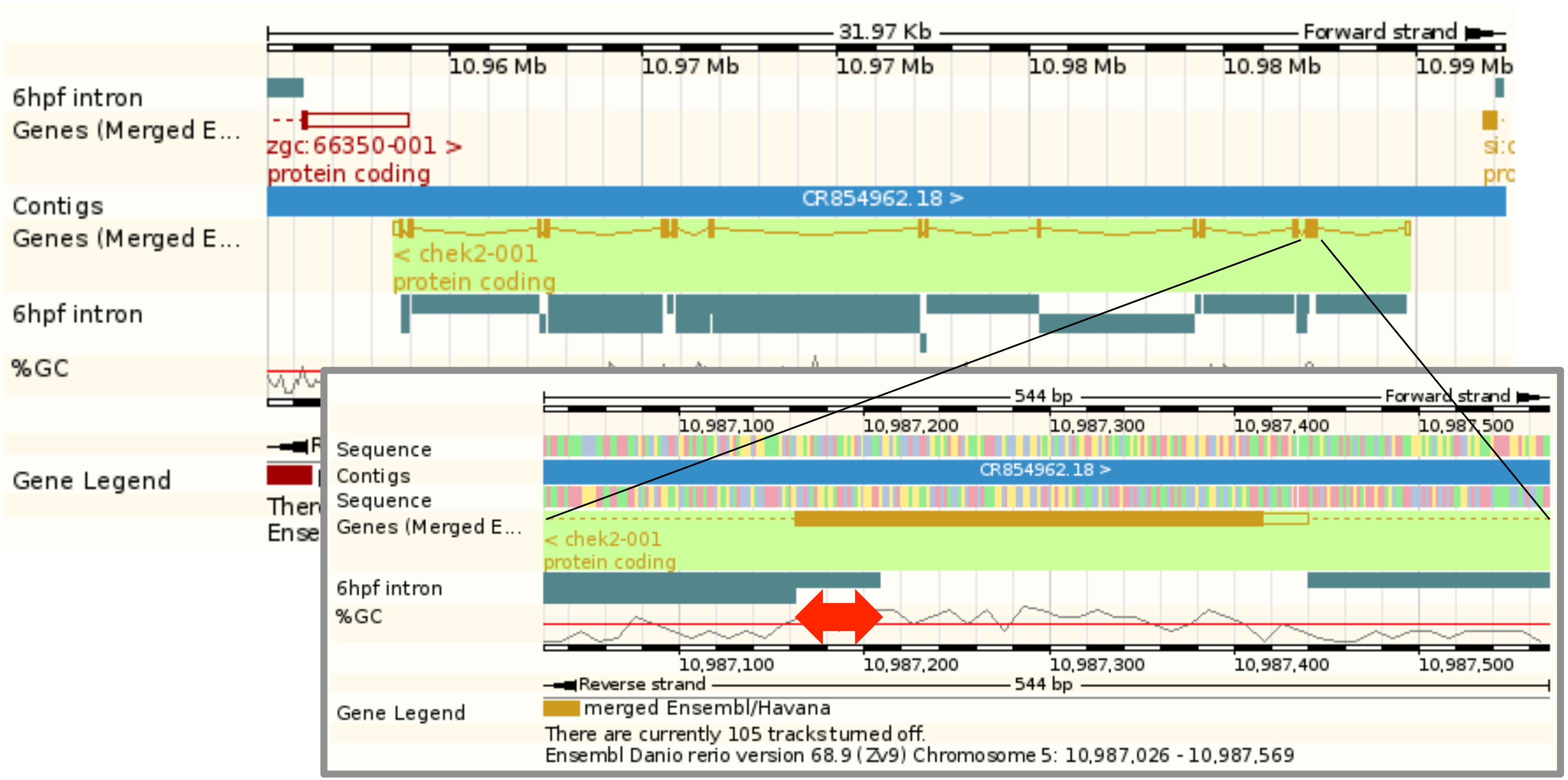
Introns show alternative transcripts



There are currently 105 tracks turned off.
Ensembl Danio rerio version 68.9 (Zv9) Chromosome 5: 10,970,166 - 10,977,775

Zebrafish gene build   http://www.ensembl.org/index.html release 68

Introns show alternative transcripts



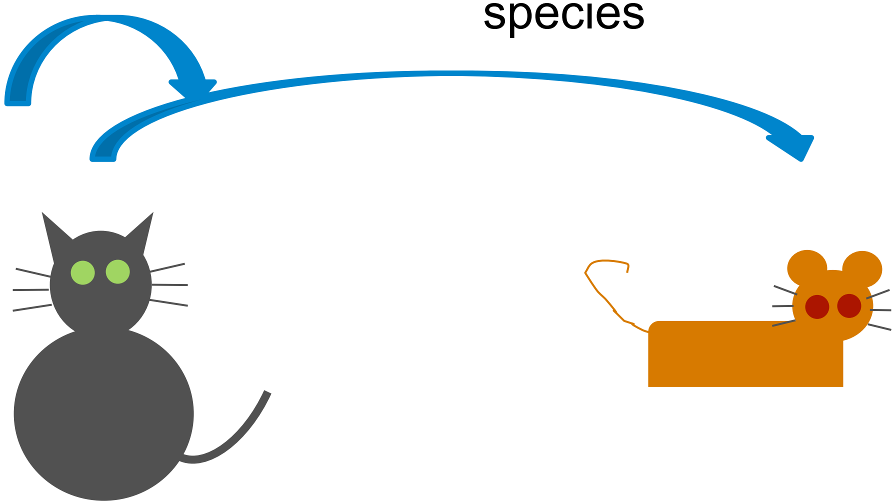
There are currently 105 tracks turned off.
Ensembl Danio rerio version 68.9 (Zv9) Chromosome 5: 10,987,026 - 10,987,569

Zebrafish gene build http://www.ensembl.org/index.html release 68

Building genes from homologues

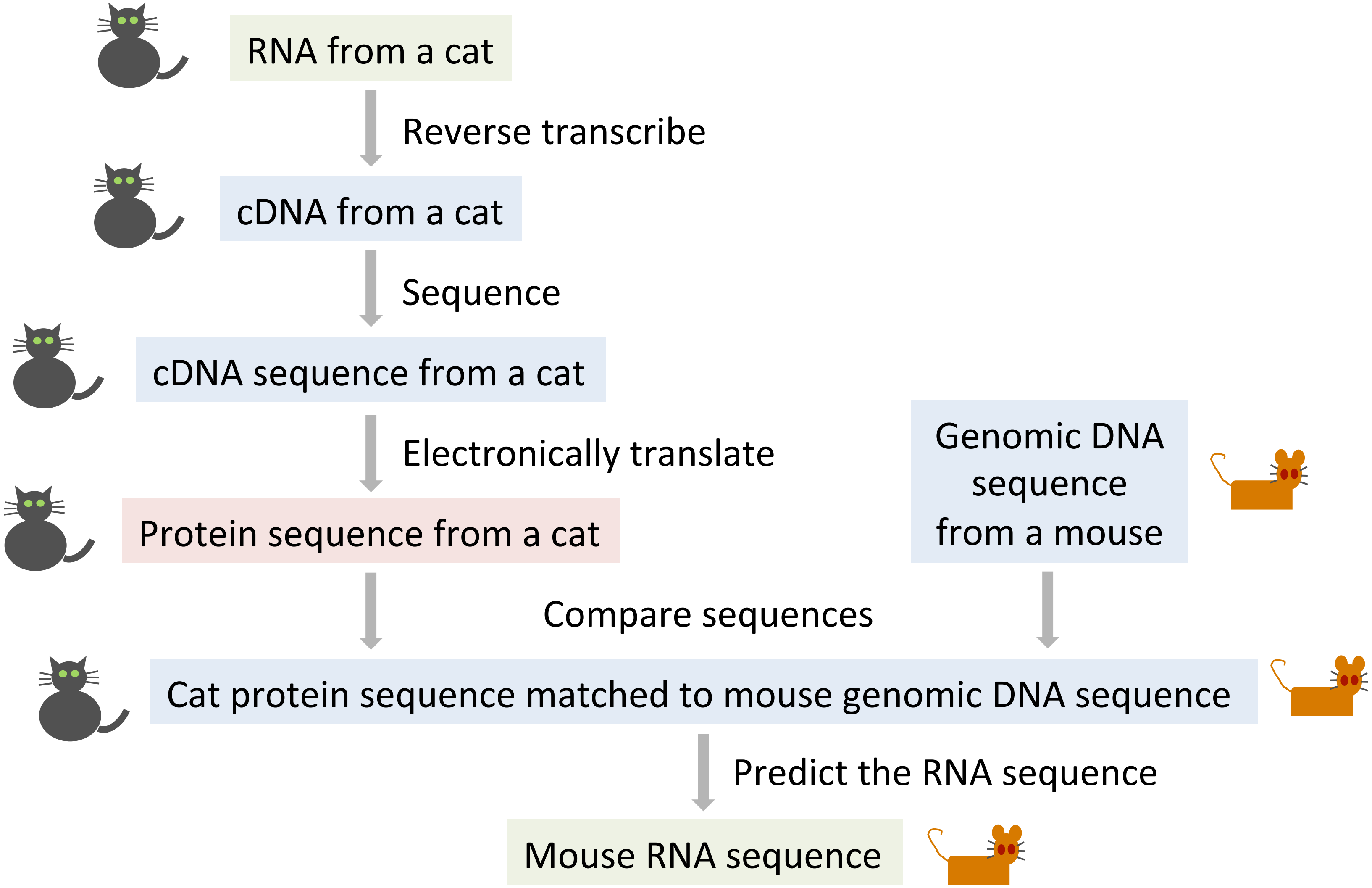
paralogue
Duplicate copies of a gene in the same species

orthologue
Divergent copies of the same gene in different species



If we sequence a cat gene and identified the protein then we can predict the mouse protein using the mouse genome

Building genes from homologues



RNA from a cat

Reverse transcribe

cDNA from a cat

Sequence

cDNA sequence from a cat

Electronically translate

Protein sequence from a cat

Genomic DNA sequence from a mouse

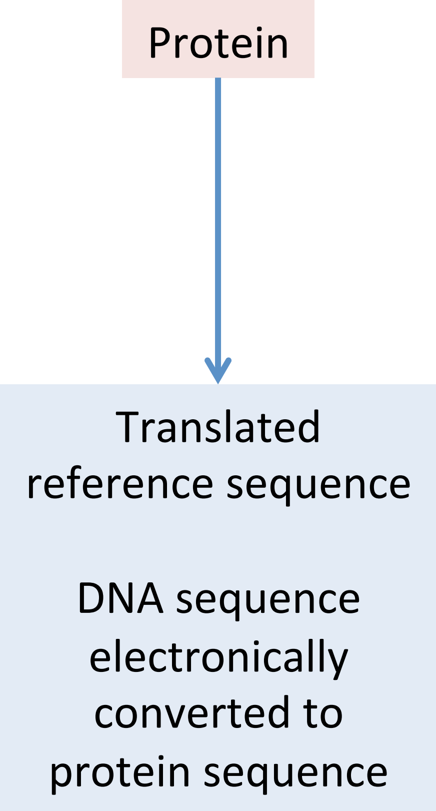
Compare sequences

Cat protein sequence matched to mouse genomic DNA sequence

Predict the RNA sequence

Mouse RNA sequence

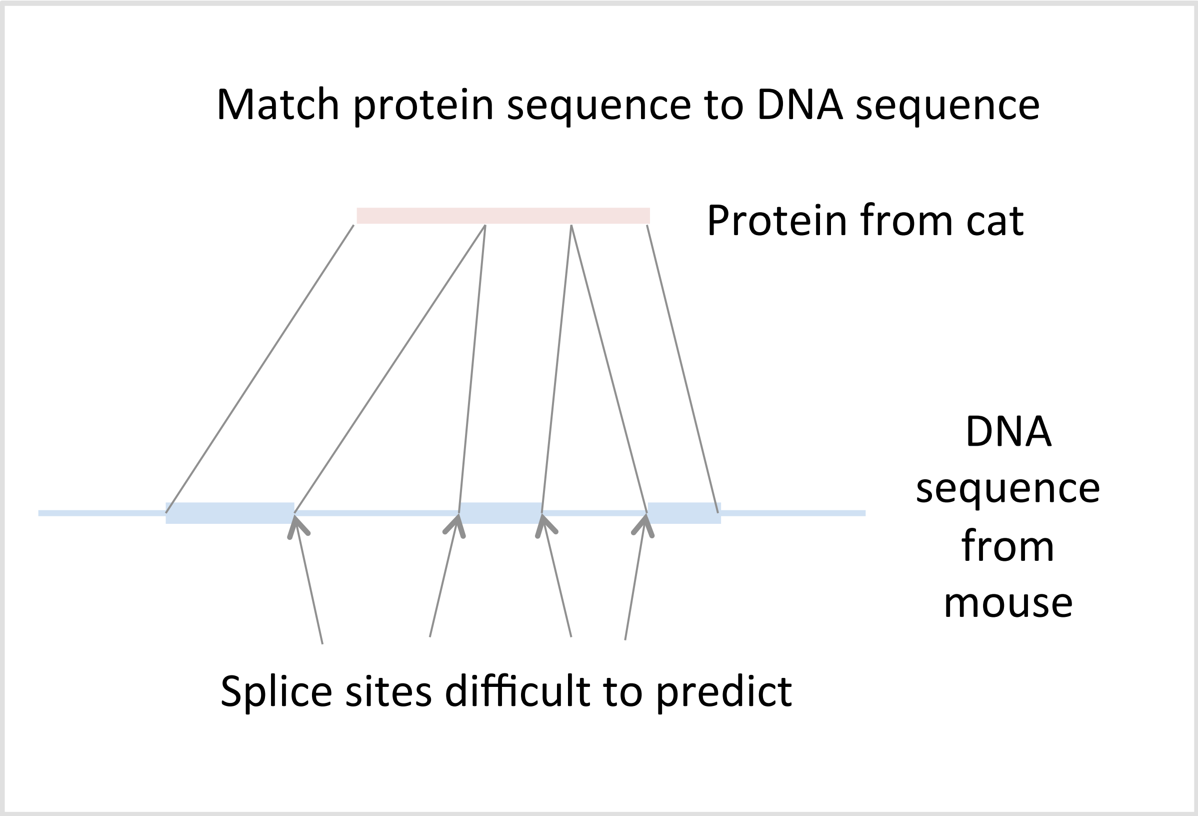
Building genes from homologues



Protein

Translated reference sequence

DNA sequence electronically converted to protein sequence



Match protein sequence to DNA sequence

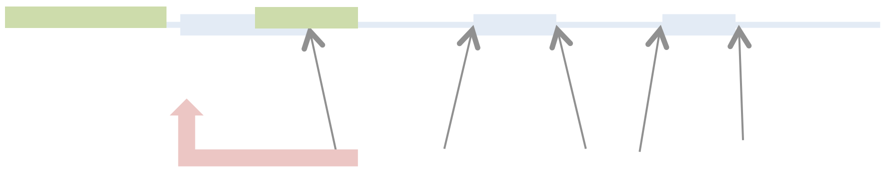
Protein from cat

DNA sequence from mouse

Splice sites difficult to predict

Building genes from homologues

Intron driven annotation



DNA sequence from mouse

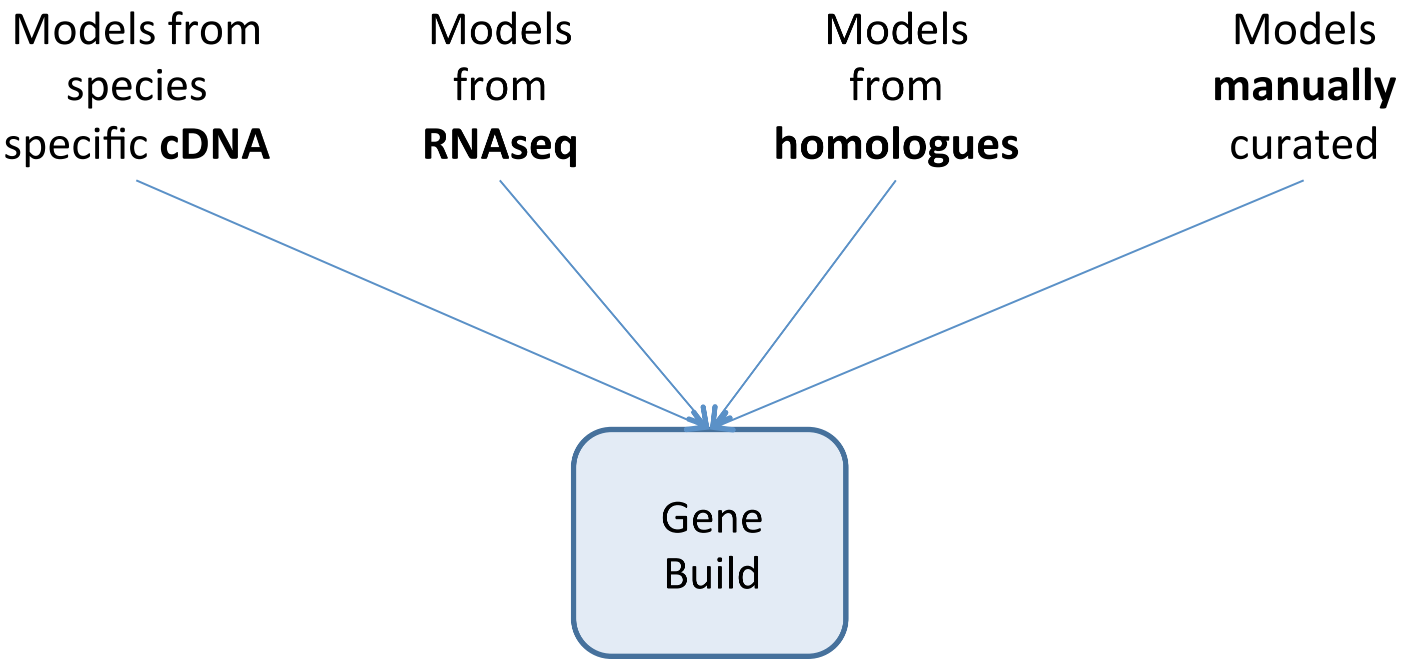
Define the exact splice sites using the mouse RNAseq **intron database**

Building genes manually

VEGA gene annotation - Manually curated

- Look at all the evidence for each transcript
- Annotate all the alternative splices for each gene
- Solve problems which aren’t possible in an automated system

Bring all the gene model together



Models from species specific **cDNA**

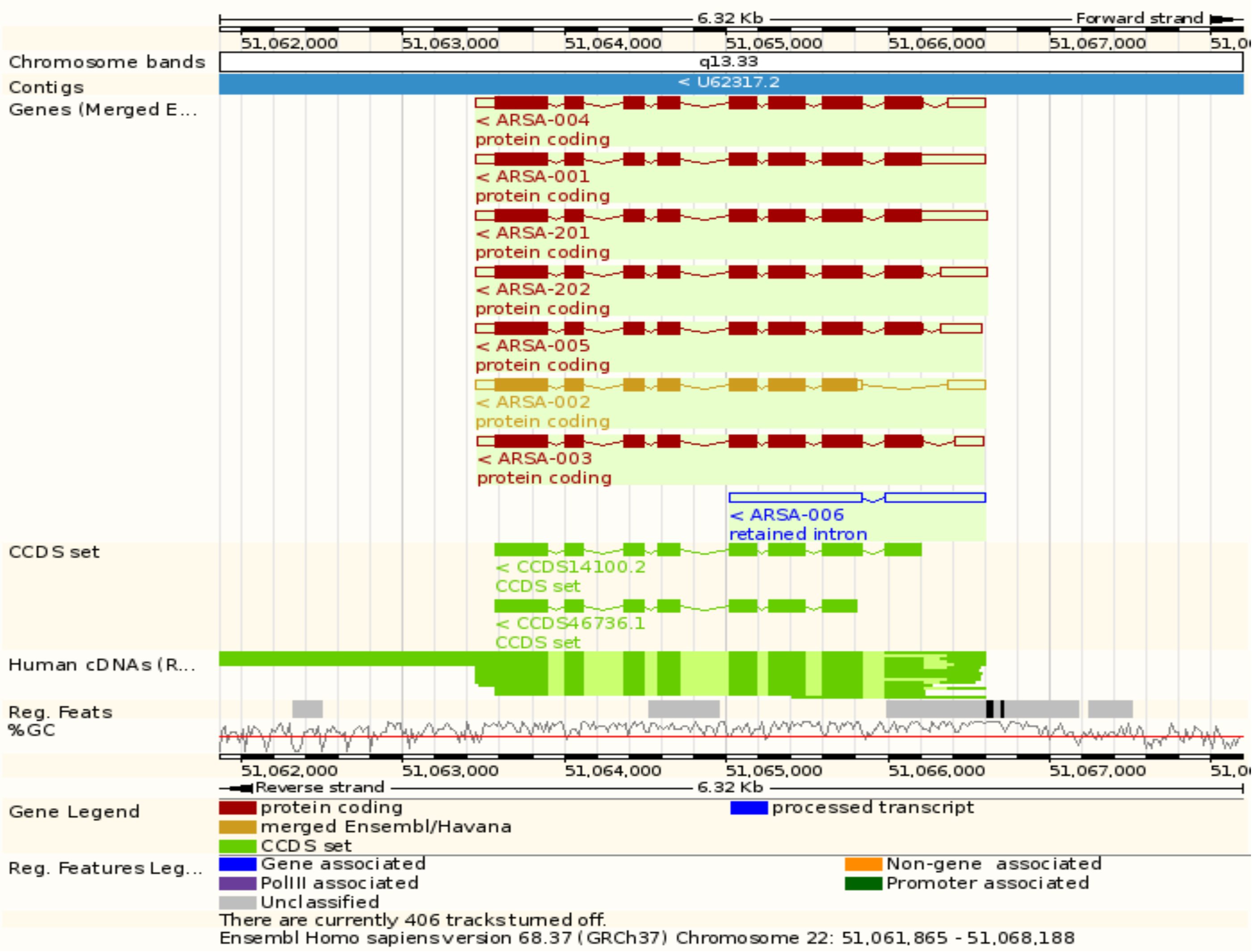
Models from **RNAseq**

Models from **homologues**

Models **manually** curated

Gene Build

Displaying the gene build



There are currently 406 tracks turned off.
Ensembl Homo sapiens version 68.37 (GRCh37) Chromosome 22: 51,061,865 - 51,068,188

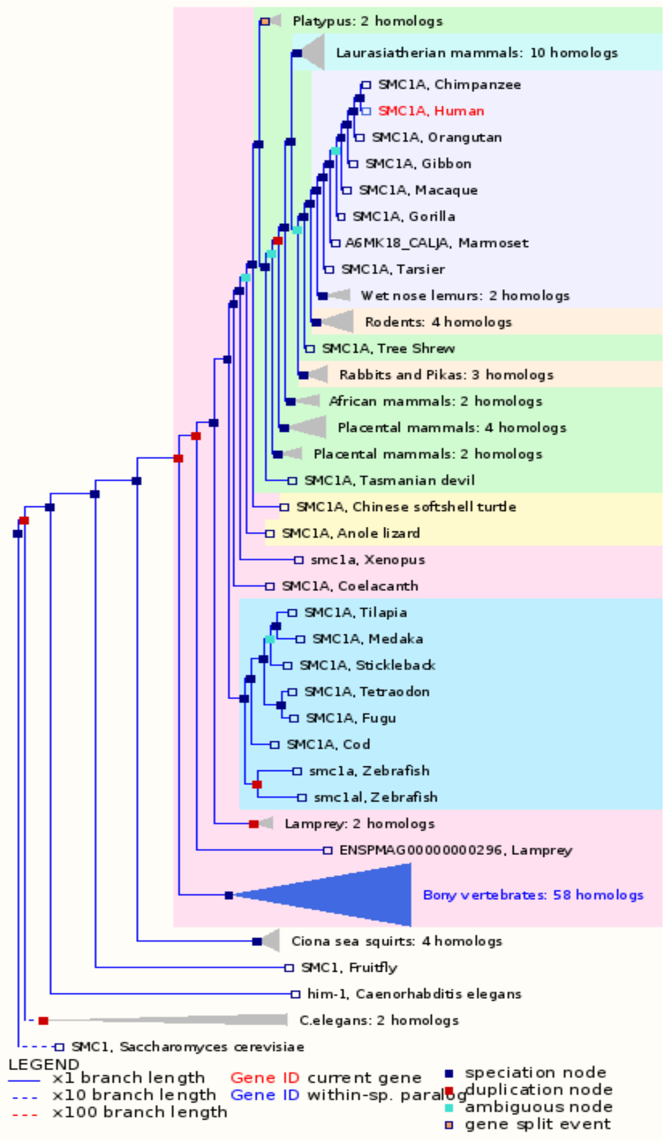
Human gene build http://www.ensembl.org/index.html release 68

Comparing gene builds from other organisms

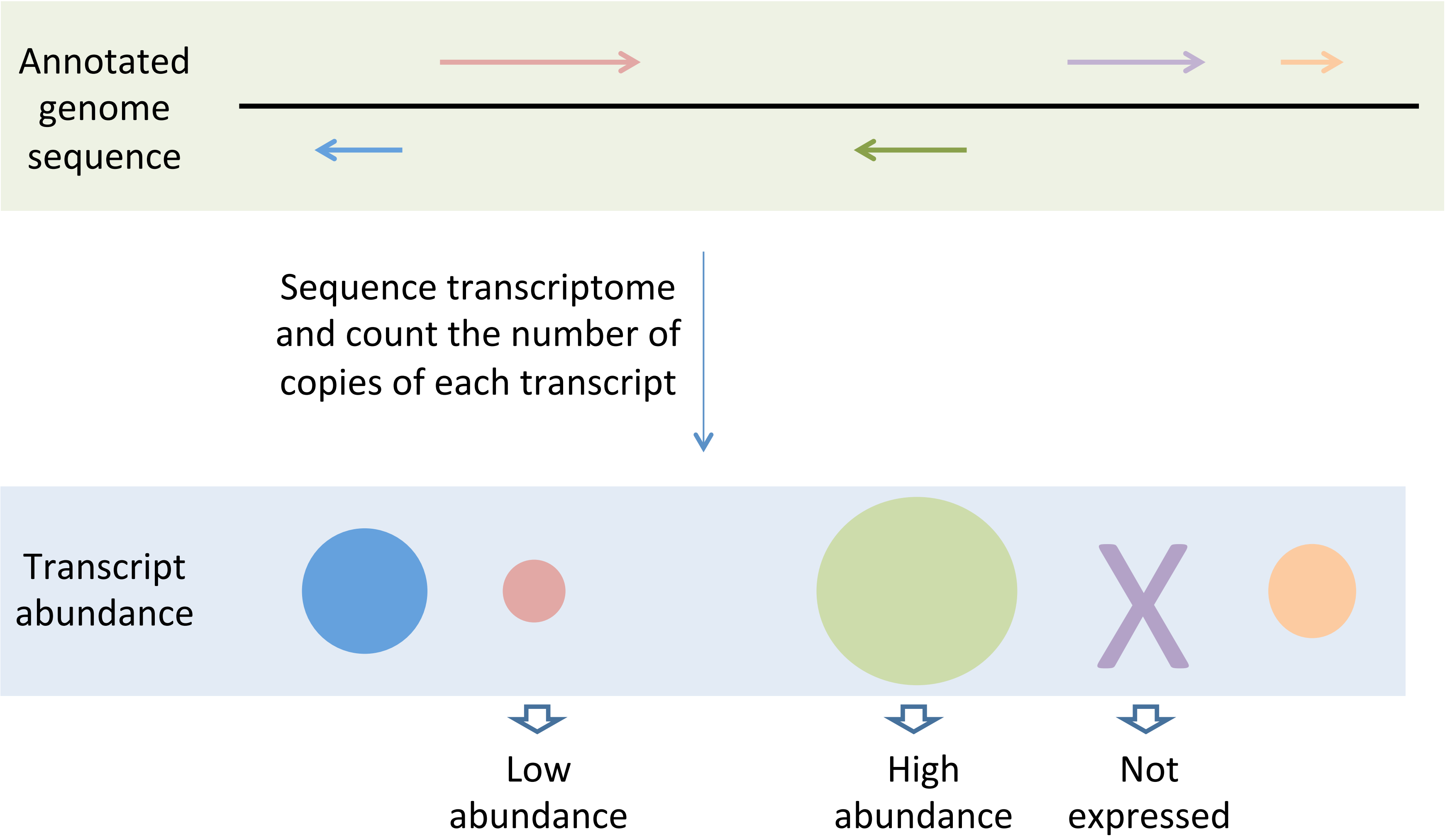
SMC1A is a protein important in cell division.

By matching gene builds across species we find candidate orthologues in other organisms. The position on the Gene Tree shows the degree of relationship

Gene Tree in Ensembl 68
ENSGT00580000081569



Using the annotation – transcriptome profiling



Annotated genome sequence

Sequence transcriptome and count the number of copies of each transcript

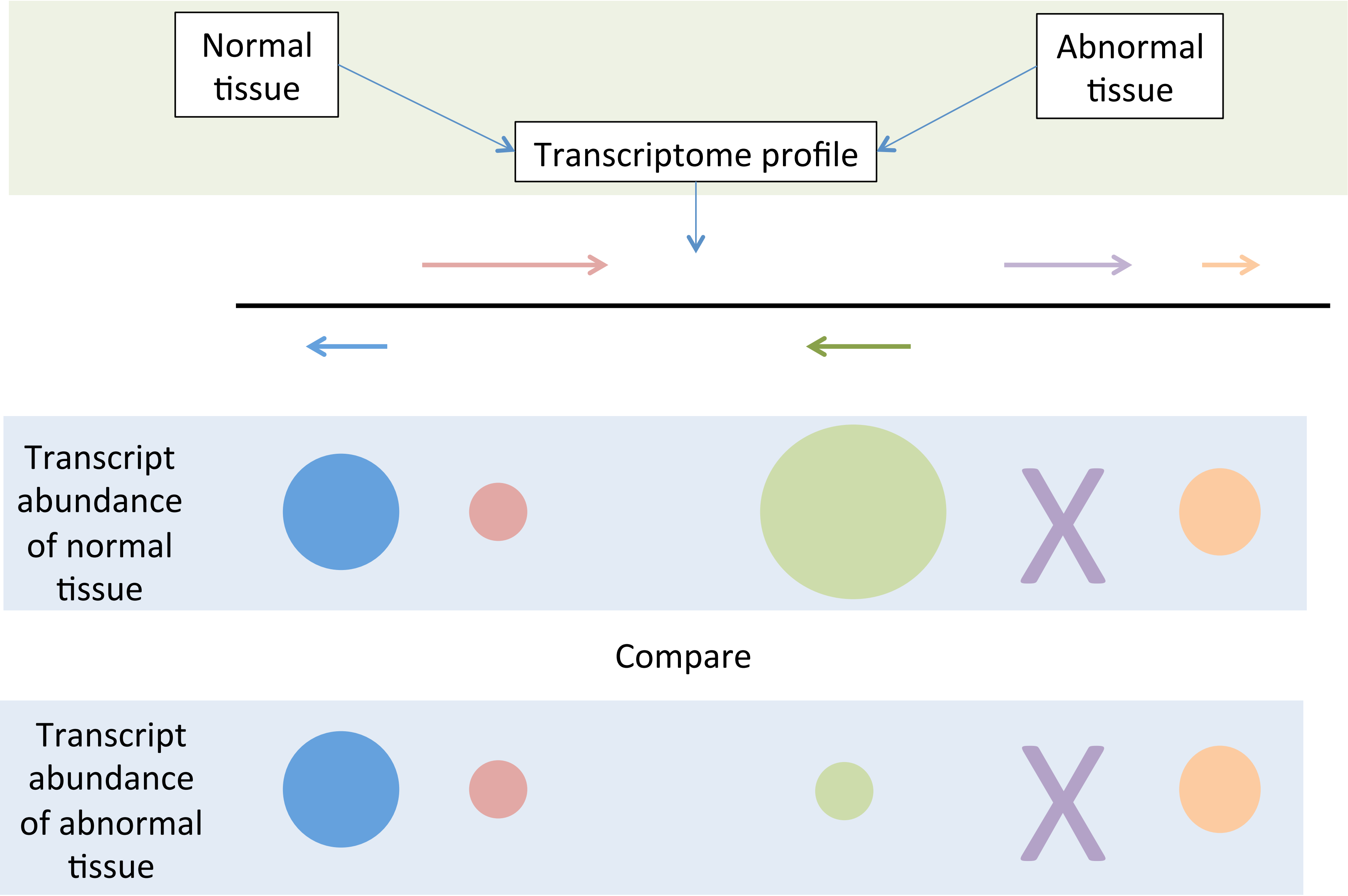
Transcript abundance

Low abundance

High abundance

Not expressed

Using the annotation – transcriptome comparison



Normal tissue

Abnormal tissue

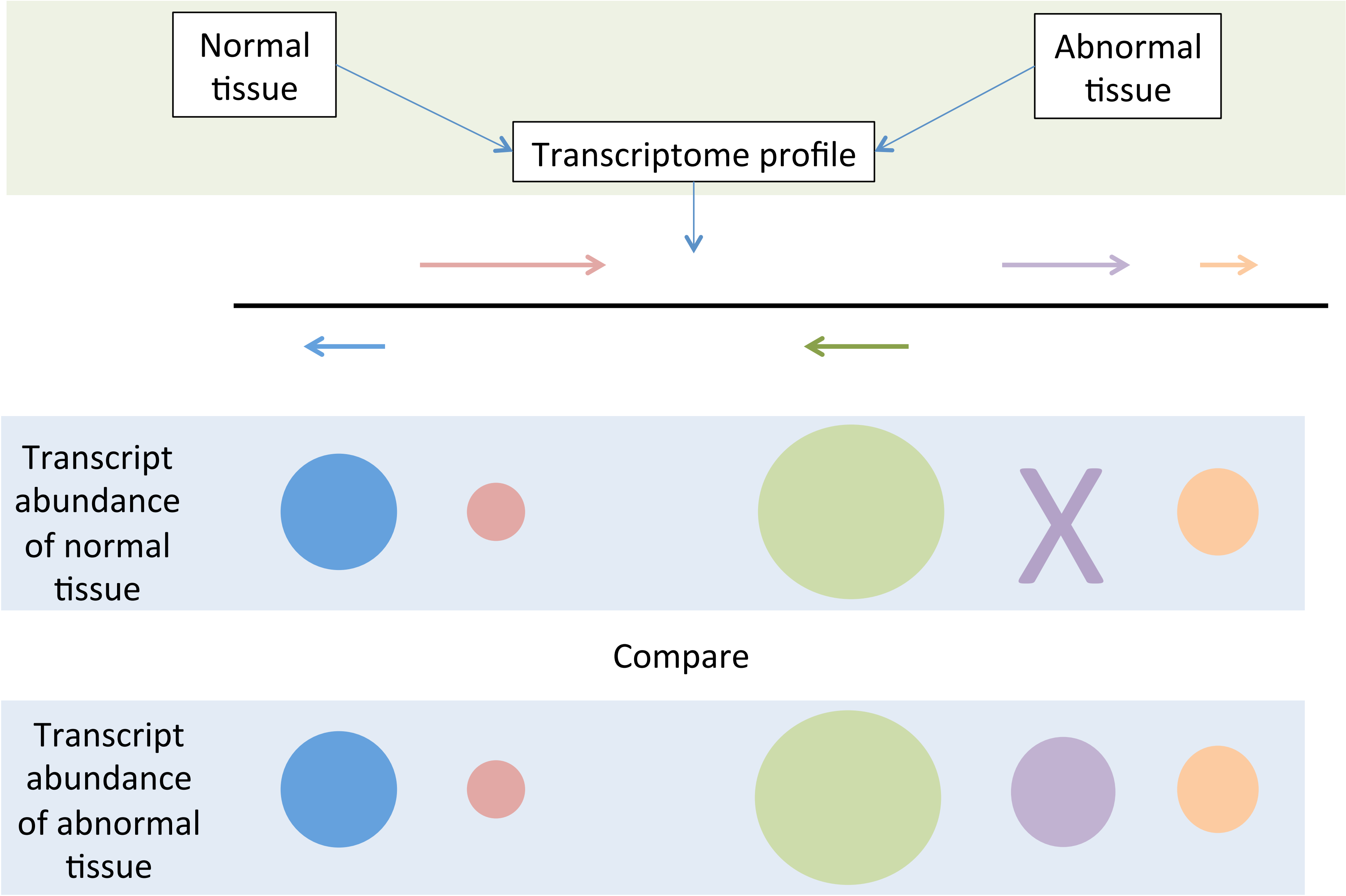
Transcriptome profile

Transcript abundance of normal tissue

Compare

Transcript abundance of abnormal tissue

Using the annotation – transcriptome comparison



Normal tissue

Abnormal tissue

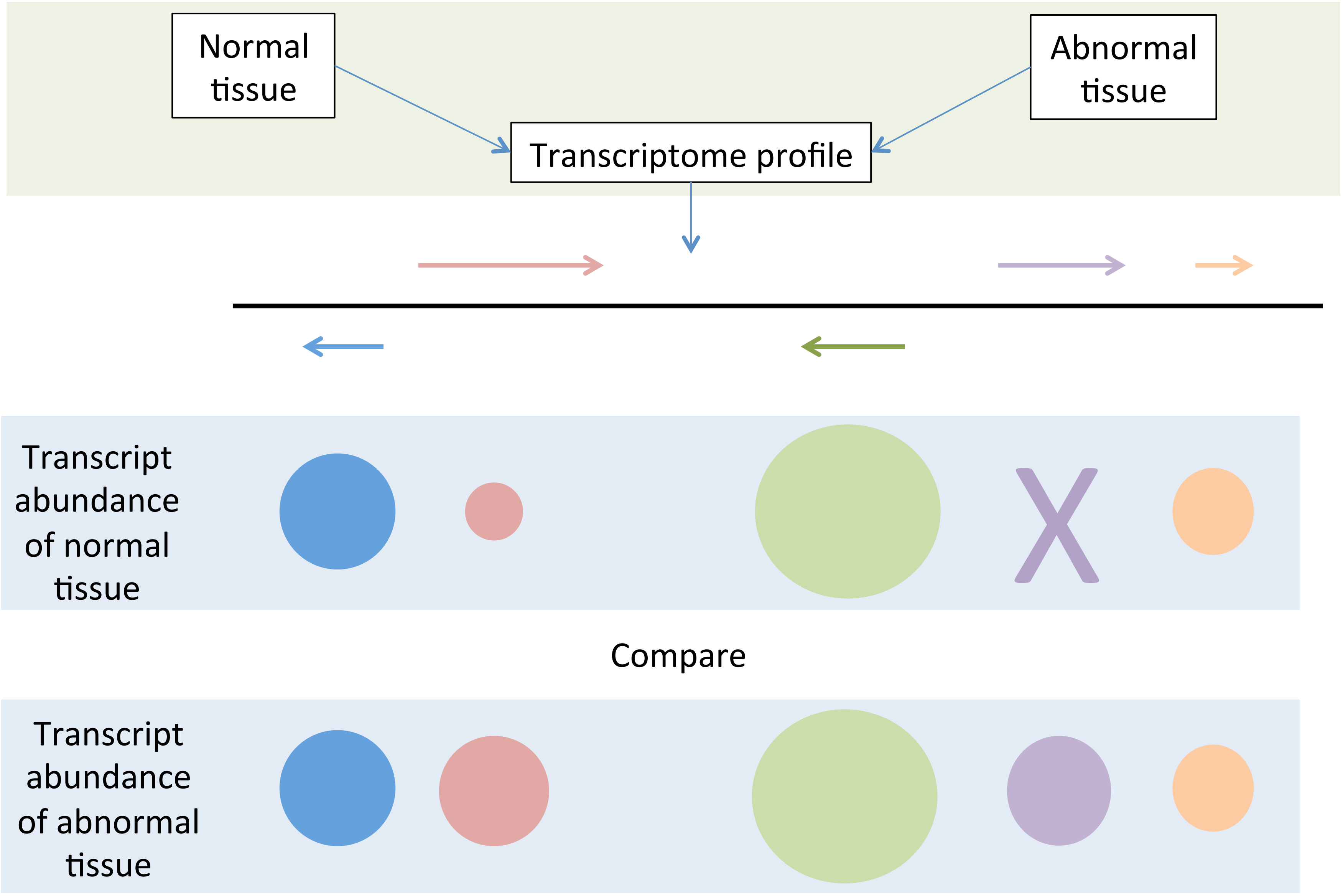
Transcriptome profile

Transcript abundance of normal tissue

Compare

Transcript abundance of abnormal tissue

Using the annotation – transcriptome comparison



Normal tissue

Abnormal tissue

Transcriptome profile

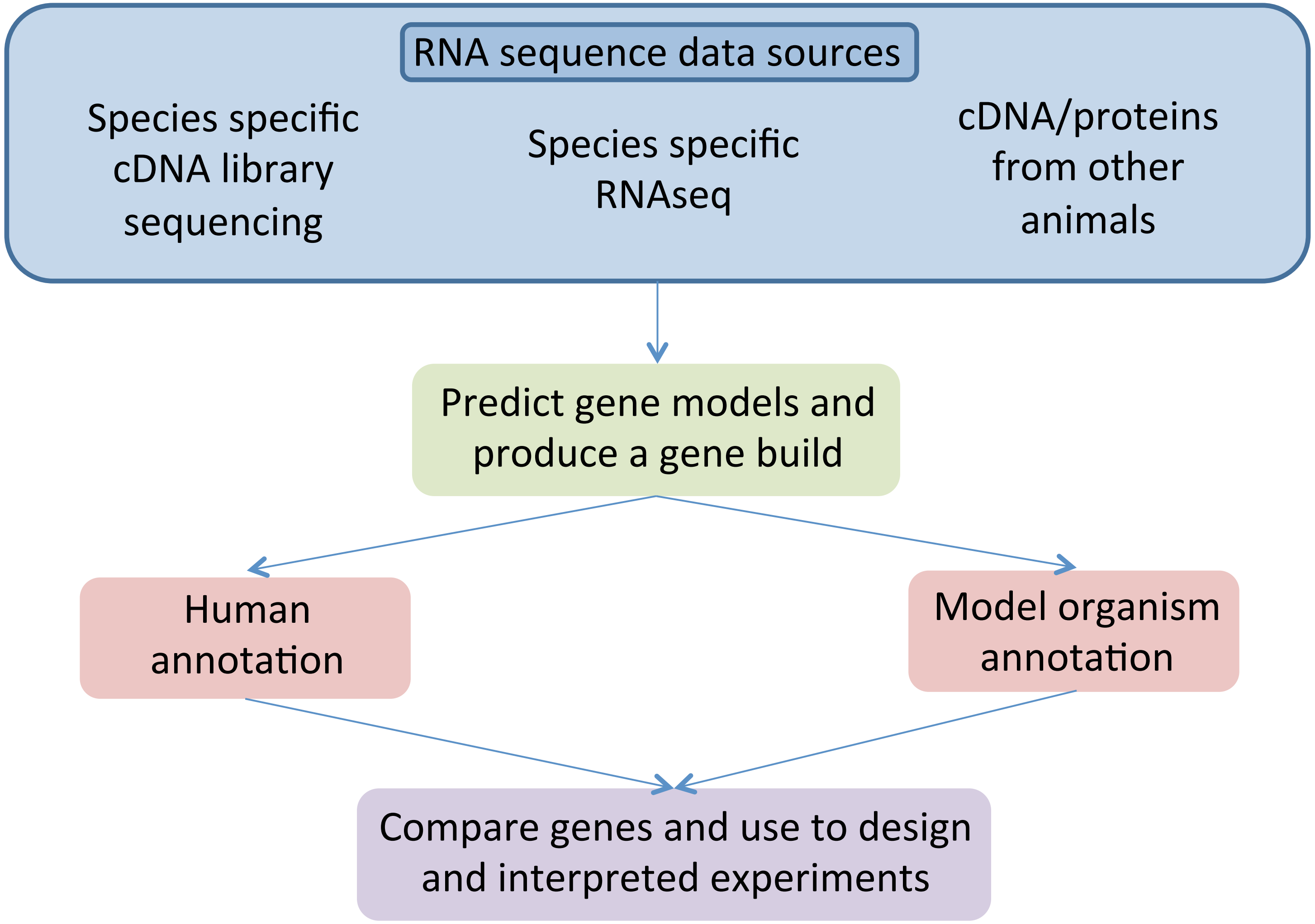
Transcript abundance of normal tissue

Compare

Transcript abundance of abnormal tissue

Summary

Identifying the 3% of the genome which contains the protein coding genes



RNA sequence data sources

Species specific cDNA library sequencing

Species specific RNAseq

cDNA/proteins from other animals

Predict gene models and produce a gene build

Human annotation

Model organism annotation

Compare genes and use to design and interpreted experiments