V7 – Genomics data

Program for today:

- SNP frequencies in 1000 Genomes data
- Repeats in imprinted vs. biallelically expressed genes
- Non-canonical translation

It is necessary to filter / clean the gene sets so that

the research question being addressed can be answered in the best way.

Removing sequence redundancy

Let's assume we want to know whether the **amino acid composition** of certain protein sequences differs in one genomic region from the other regions.

For example, we want to know whether **transmembrane (TM) segments** of membrane proteins are more hydrophobic than the rest of the protein sequence

To check this, we could simply analyze all protein sequences from NCBI, predict the TM segments in them and compare the amino acid compositions.

However, this search would likely be **biased** by

- what proteins have been sequenced and which ones not, and
- by duplicated sequencing experiments.

→ It is very important to **remove sequence redundancy** before such analyses!

This can be done by software tools such as CDhit or BlastClust

BlastClust

blastclust -i infile -o outfile -p F -L .9 -b T -S 95

The sequences in "infile" will be clustered and the results will be written to "outfile".

The input sequences are identified as nucleotide (-p F); "-p T", or protein.

To register a pairwise match two sequences will need to be 95% identical (-S 95) over an area covering 90% of the length (-L .9) of each sequence (-b T) .

https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html

Refseq

The Reference Sequence (RefSeq) collection at NCBI provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

RefSeq transcript and protein records are generated in different ways:

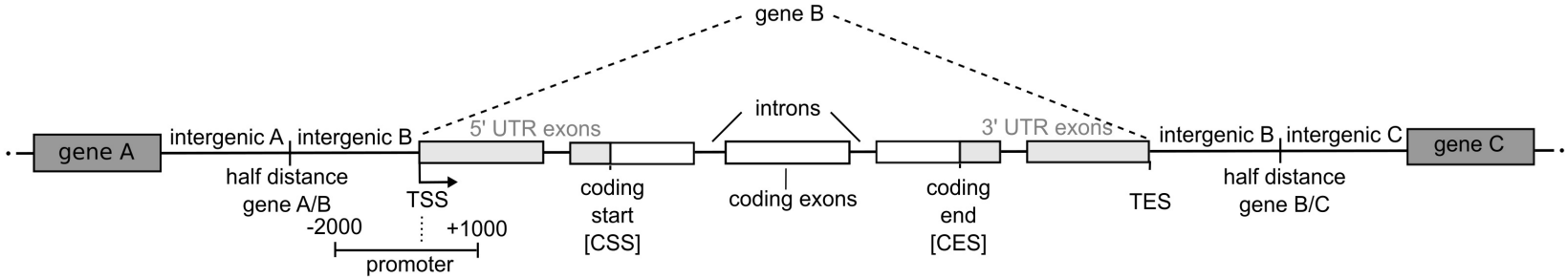
- Computation Eukaryotic Genome Annotation Pipeline
  Prokaryotic Genome Annotation Pipeline
- Manual curation
- Propagation from annotated genomes that are submitted to members of the International Nucleotide Sequence Database Collaboration (INSDC)

First research question:

Are the **Single Nucleotide Polymorphism (SNP) frequencies** in different genomic regions similar to eachother or not?

https://www.ncbi.nlm.nih.gov/refseq/about/

Definition of genomic regions



Every **gene** is located between two **intergenic regions**. Our definition for these is:

First intergenic region : interval between the transcription start site (TSS) of the considered gene and the mid-upstream position between this TSS and the transcription end site (TES) of the closest upstream gene.

Second intergenic region : defined analogously according to the TSS of the closest downstream gene.

Intragenic region of a gene : part between its TSS and its TES.

Gene **promoter** : region from 2000 bp upstream to 1000 bp downstream of the TSS.

Exons : intervals between the exon start positions and exon end positions (taken from UCSC genome browser).

5' UTRs : exonic segments between the TSS and the CSS

3' UTRs : exonic regions between the CES and the TES.

Introns : regions between the exonic gene parts.

Neininger & Helms, submitted

1000 Genomes project

The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data up to date.

The goal of the 1000 Genomes Project was to find most genetic variants with frequencies of at least 1% in the populations studied.

http://www.internationalgenome.org/

Identify SNPs in 1000 Genomes data

We used only the European super-population with 503 individuals and we focused on **autosomes** (chromosomes 1 – 22). Genes on sex chromosomes X and Y are ignored.

We keep autosomal SNPs with a minor allele frequency larger than zero → SNP exists

allele : variant form of a given gene

major allele : most common variant

minor allele: second-most common variant

We removed:

- genes starting with "SNO“ (small nuclear RNAs) or "MIR“ (microRNAs)
- genes with CDS start equal to the CDS end

Neininger & Helms, submitted

Problem: there exist many overlapping genes

Overlap between three human genes: *MUTH, FLJ13949*, and *TESK2*.

Dark boxes : coding sequence.

Light boxes : untranslated regions.

Table 1. Frequency of Different Types of Overlaps Between Protein-Coding Genes in Human and Mouse Genomes

	Human		Mouse	
	Overlapping genes	Genes with overlapping exons	Overlapping genes	Genes with overlapping exons
Total	774	542	578	455
Embedded	126 (16.28%)	15 (2.77%)	53 (9.17%)	7 (1.54%)
Tail to tail	414 (53.49%)	360 (66.42%)	314 (54.32%)	280 (61.54%)
Head to head	234 (30.23%)	167 (30.81%)	211 (36.51%)	168 (36.92%)
Involving coding sequence		299 (55.17%)		232 (50.99%)
Coding–coding overlap		57 (10.52%)		31 (96.81%)

Veeramachaneni et al.
Genome Res. (2004) 14: 280-286

Overlapping genes

One could speculate that overlapping genes would be more conserved between species than non-overlapping genes because a mutation in the overlapping region would cause changes in both genes.

Then, one would expect that evolutionary selection against these mutations is stronger.

However, Veeramachaneni *et al*. found that this is not the case.

Overlapping human and mouse genes were similarly conserved as non-overlapping genes.

Note that only a small fraction of the analyzed genes preserved exactly the same pattern of gene structure and overlap pattern in human and mouse.

Veeramachaneni et al.
Genome Res. (2004) 14: 280-286

How to deal with overlapping genes

In the case of overlapping genes, it is problematic to define the **genomic regions** because they have a different meaning for the 2 overlapping genes.

Therefore, we distinguished 2 cases:

(1) Overlaps where one gene is located inside another gene.

Such genes inside other genes were excluded from the SNP analysis.

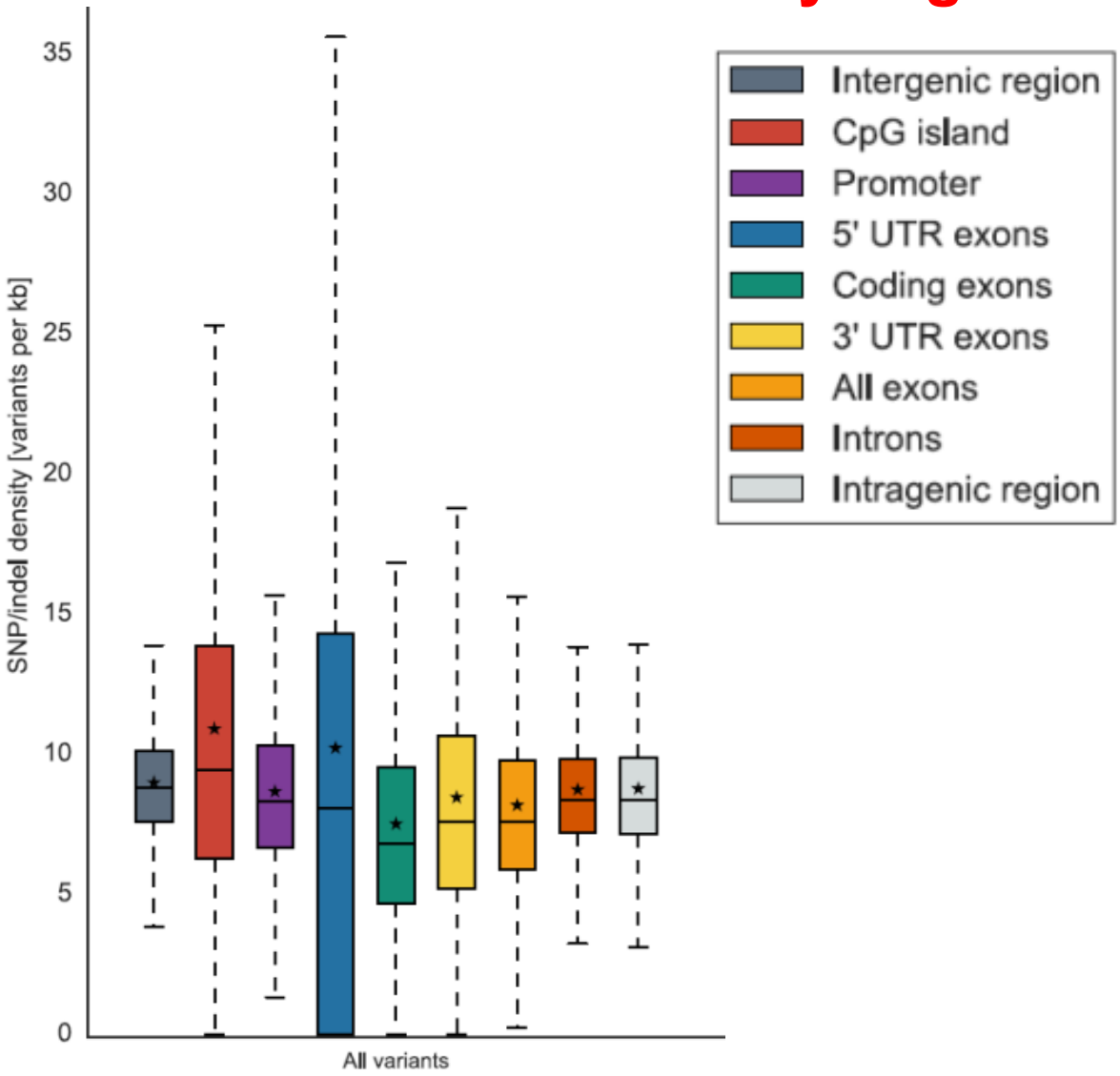
(2) staggered overlaps (genes overlap partially).

We collected all genes with staggered overlap. From each “bundle", only one gene was selected randomly to avoid overlapping genes.

In total, about 5% of all genes were removed due to overlaps.

Neininger & Helms, submitted

SNP density in genomic regions



Neininger & Helms, submitted

Number of SNP variants per kb for different genomic regions.

→ lowest SNP density in coding exons (green)

→ highest SNP density in CpG islands (due to frequent deamination of methylated cytosines into thymines)

Second-highest SNP density in intergenic regions (low evolutionary pressure)

Imprinted genes

Imprinted genes violate the usual rule of inheritance

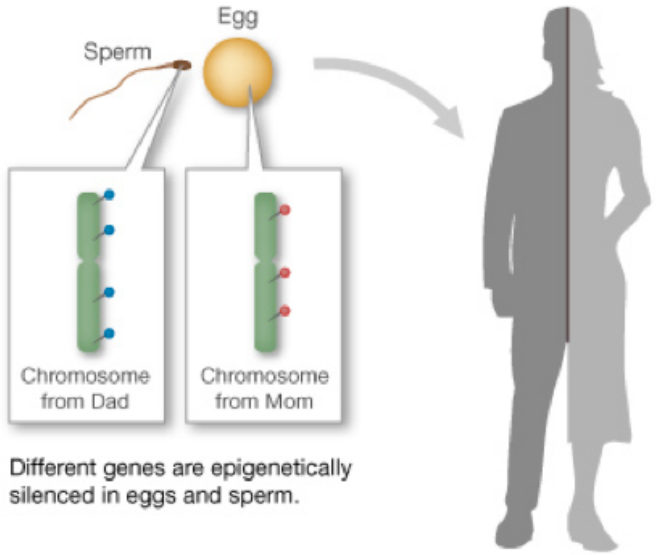
Bi-allelic genes :

1 gene copy (allele) encoding e.g. hemoglobin from dad

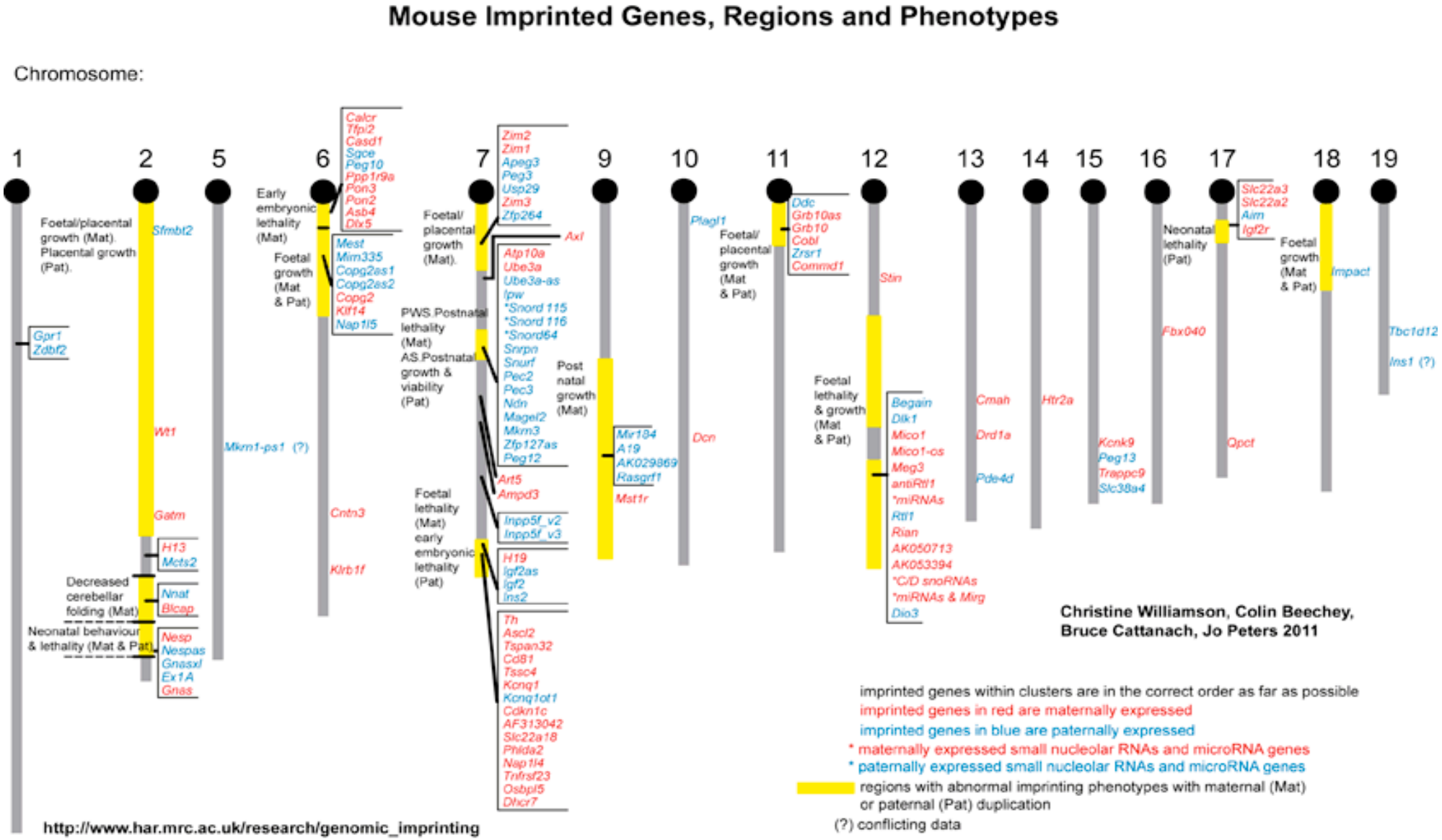
1 gene copy (allele) encoding e.g. hemoglobin from mom

Child: expresses equal amounts of the 2 types of hemoglobin

Mono-allelic (imprinted) genes : one allele silenced by **DNA methylation**



Imprinted genes cluster in the genome

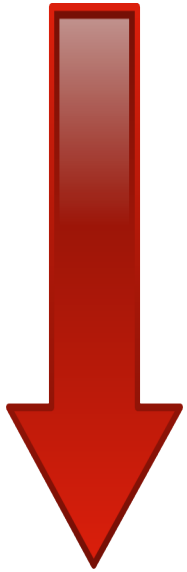


Parental conflict hypothesis = “battle of the sexes”

Paternally expressed genes

embryonic growth in placenta

Maternally expressed genes



embryonic growth in placenta

Aim of the study

Aim: distinguish general properties of imprinted genes from biallelically expressed (BE) genes.

Example features:

- Imprinted genes could be either more or less **conserved** during evolution than BE genes. Note: imprinting is found in mammals with placenta – also in plants

- Imprinted genes may have different **functions** than BE genes → V8

- Imprinted genes may have more or less CpG island promoters than BE genes

- ….

Hutter, Bieg, Helms & Paulsen,
BMC Genomics (2010) 11, 649

Preparation of data set

If several transcripts are known for one gene, we took the most 5' annotated transcriptional start site and the most 3' annotated transcriptional termination site and constructed the **longest possible transcript**.

Similarly, splice variants and overlapping exons were merged in a way so that the largest possible coding regions were constructed.

The genomic sequence that was assigned to a gene contained the transcribed sequence and intergenic regions upstream and downstream of the transcription unit.

For determining the intergenic region, the DNA sequence between two genes was cut into two halves, each half was assigned to the nearest gene.

Hutter, Bieg, Helms & Paulsen,
BMC Genomics (2010) 11, 649

Phast regions

As a set of sequences with high conservation in eutherian mammals, we used the UCSC phastCons28wayPlacMammal most conserved sequences (PCSs).

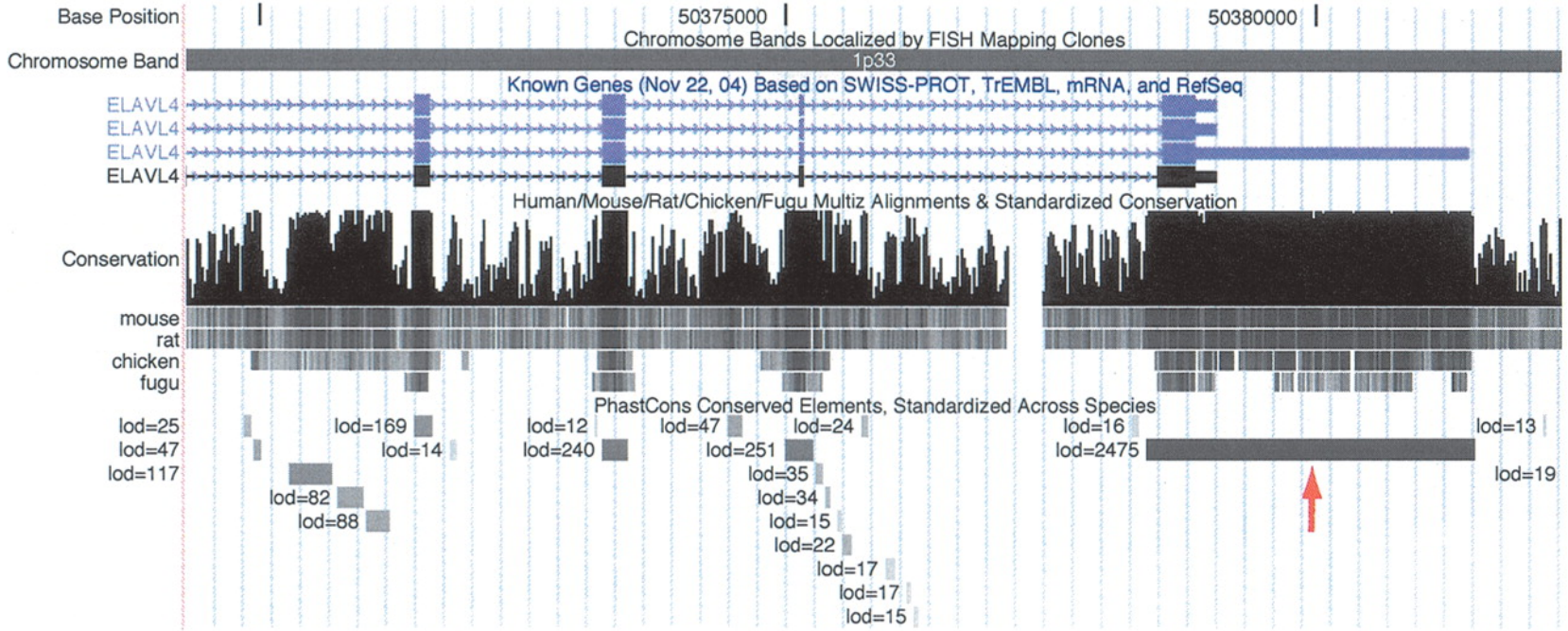
Such highly conserved regions were originally identified from a genome-wide multiple alignment of 29 vertebrate species by the Phast program and afterwards projected onto a reference genome.

The PCSs analyzed here are a subset of these regions showing conservation in 18 eutherian mammals.

We assigned the PCSs to the longest possible RefSeq transcripts based on the human genome March 2006 assembly (hg18).

Hutter, Bieg, Helms & Paulsen,
BMC Genomics (2010) 11, 649

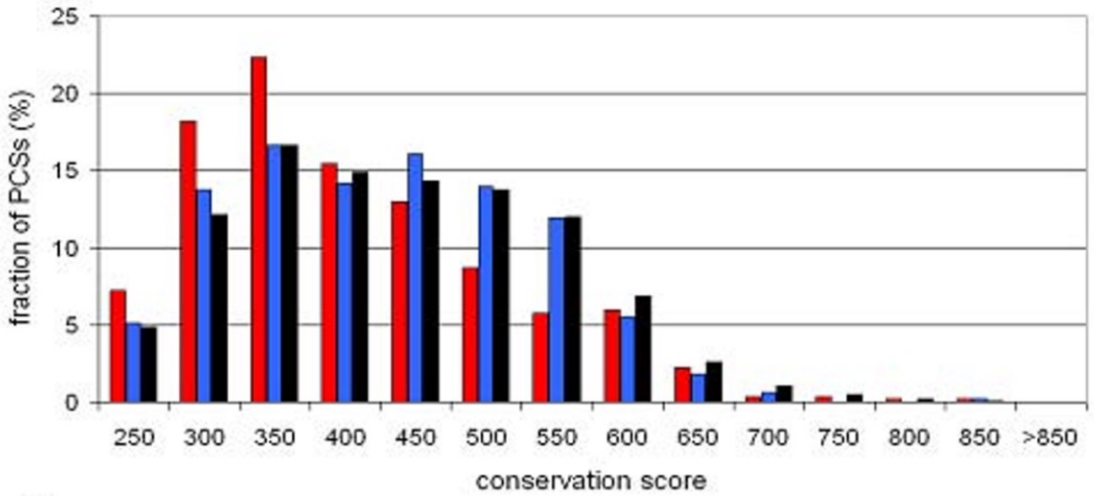
ELAVL4 is a Phast region

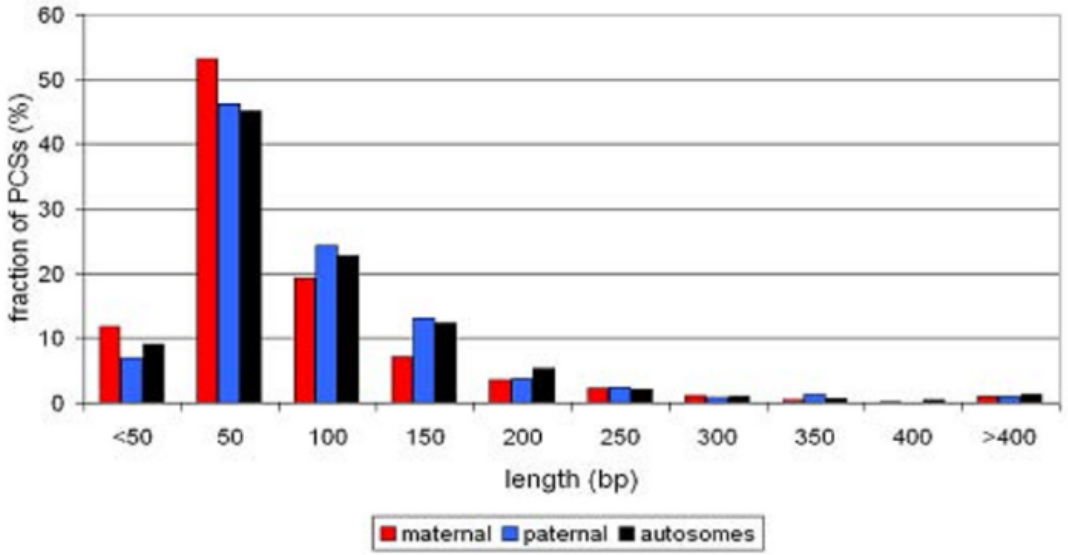


Extreme conservation at the 3′ end of the *ELAVL4* (*HuD*) gene, an RNA-binding gene associated with paraneoplastic encephalomyelitis sensory neuropathy and homologous to *Drosophila* genes with established roles in neurogenesis and sex determination. The 3117-bp conserved element that overlaps the 3′ UTR of this gene (red arrow) is the fifth highest scoring conserved element in the human genome. Several conserved elements in introns are also visible.

Siepel et al. Genome Res. (2005) 15: 1034-1050

Length and conservation of PCS sequences





Hutter, Bieg, Helms & Paulsen,
BMC Genomics (2010) 11, 649

(A) conservation scores and

(B) lengths of PCSs that overlap with coding exons.

PCSs of paternally expressed ones (blue bars) are similar to PCSs of autosomal genes (black bars).

In contrast, the PCSs of maternally expressed genes (red bars) are shorter (they are shifted to the left) and have lower conservation scores.

→ increased divergence of maternally expressed genes due to reduced selective pressure ??

Isoforms

Gene isoforms are mRNAs that are produced from the same locus but are different in their

- transcription start sites (TSSs),
- protein coding DNA sequences (CDSs) and/or
- untranslated regions (UTRs),

All this may potentially alter gene function.

www.wikipedia.org

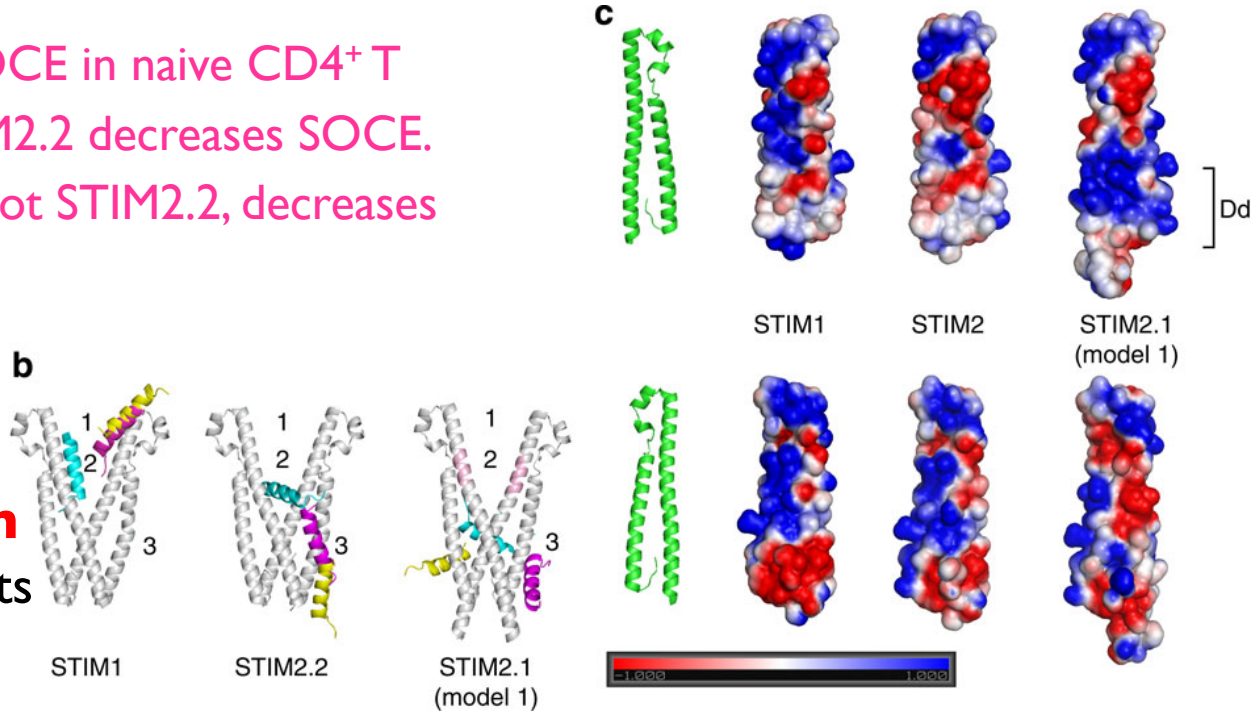
Alternative splicing may affect PP interactions: STIM2 splice variant

STIM proteins regulate store-operated calcium entry (SOCE) by sensing Ca^{2+} concentration in the ER and forming oligomers to trigger Ca^{2+} entry through plasma membrane-localized Orai1 channels.

Niemeyer and co-workers characterized a *STIM2* splice variant which retains an additional 8-AA exon within the region encoding the channel-activating domain.

STIM2.1 knockdown increases SOCE in naive $CD4^+$ T cells, whereas knockdown of STIM2.2 decreases SOCE. Overexpression of STIM2.1, but not STIM2.2, decreases SOCE.

STIM2.1 interaction with Orai1 is impaired and prevents Orai1 activation.



Miederer, ..., Lee, ..., Helms, Barbara Niemeyer
Nature Commun 6, 6899 (2015)

Alternative splicing

Alternative splicing (AS) of mRNA can generate a wide range of mature RNA transcripts.

It is estimated that AS of pre-mRNA occurs in 95% of multi-exon human genes.

There is abundant evidence for the expression of **multiple transcripts** in cells.

However, it is less clear whether these transcripts are expressed more or less equally across tissues or whether it would be biologically relevant to designate one transcript per gene as **dominant** and the rest as **alternative**.

Ezkurdia et al J Proteome Res. (2015) 14: 1880–1887.

Evidence from mRNA expression

Three contrasting large-scale expression studies came to different conclusions.

An EST-based study with 13 different tissues predicted that primary tissues generally had a single dominant transcript per gene.

In contrast, a large-scale study using RNAseq found that > 75% of protein-coding genes had cell-line-specific dominant transcripts.

Those genes with the most splice variants had more dominant transcripts.

A second RNAseq study (Illumina Human BodyMap project) found that ca. 50% of the genes expressed in the 16 tissues studied had the same major transcript in all tissues, whereas another third of the genes had major transcripts that were tissue-dependent.

One curious result in this study was that the major transcript was noncoding in close to 20% of the protein-coding genes.

Ezkurdia et al J Proteome Res. (2015) 14: 1880–1887.

Detect isoforms in proteomic data

Here: re-analysis of 8 HT proteomics MS data sets.

We detected at least two peptides for 12 716 (63.9%) of the protein-coding genes but found alternative protein isoforms for just 246 genes (1.2%).

→ the vast majority of genes had peptide evidence for just **one protein isoform**.

The isoform with the highest number of peptides was the main proteomics isoform.

In this way, we could identify a unique main proteomics isoform for 5011 genes.

Ezkurdia et al J Proteome Res. (2015) 14: 1880–1887.

Comparison proteomics - RNAseq

CCDS variants are based on genomic evidence and are variants that are mutually agreed on by teams of manual annotators from NCBI, the Sanger Institute, EBI and UC Santa Cruz.

A total of 13 297 genes were annotated with a single CCDS variant. This unique manually curated variant agreed with the main proteomics isoform for 98.6% of the 3331 genes that we compared.

APPRIS annotates principal isoforms on the basis of conservation of structure and function and selected a main isoform for 15 172 of the coding genes.

We were able to compare the APPRIS principal isoforms and the main proteomics isoforms over 4186 genes. The main proteomics isoform agreed with the isoform with the most conserved protein features for 97.8% of these genes.

In contrast, the longest isoform coincided with the main proteomics isoform only for 89.6% of the genes.

Ezkurdia et al J Proteome Res. (2015) 14: 1880–1887.

Alternative translation: example TrpV6 channel protein

MUSCLE multiple sequence alignment of the translated 5′-UTR of TRPV6

Identical aa residues (compared with the human sequence) are *shaded*;

annotated N termini with the first Met^{+1} are in ***red***;

* : stop codon in frame

− : gap

Fecher-Trost et al. J. Biol. Chem. (2013) 288: 16629

The mammalian sequences upstream of the first AUG codon are conserved, but the one from rabbit contains an in-frame stop codon. In contrast, sequences from the other organisms contain several stop codons upstream of the annotated AUG and are not conserved. Sequence identity is highest among the 40 amino acids upstream of the first Met residue (position +1). This suggests that translation in mammals may start at a non-AUG

Alternative translation of human TRPV6

*A*lignment of 5′-UTR TRPV6 sequences including the AUG triplet encoding the first methionine (*red*, +1) of the human protein.

Red, putative initiation sites;

underlined, STOP-codon in frame.

Experiments in the Flockerzi group (Medical department, Homburg) showed that translation starts at Thr^{-40} .

Fecher-Trost et al. J. Biol. Chem. (2013) 288: 16629

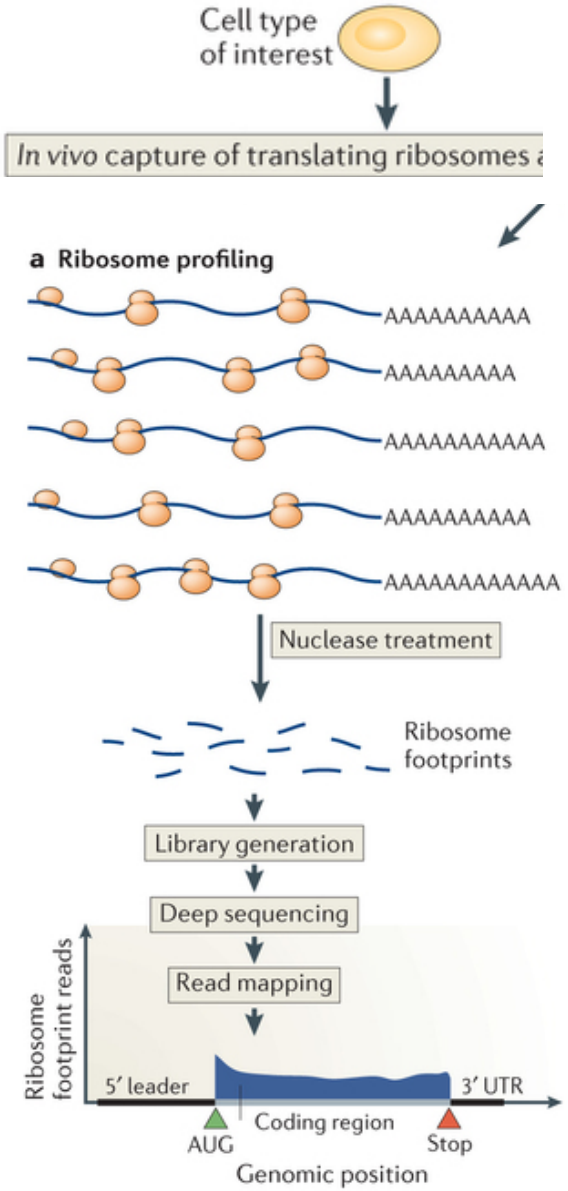
HT discovery of alternative translation: ribosome profiling

Ribosome-bound mRNAs are isolated by size.

Then they are treated with a nonspecific nuclease.

This results in protected mRNA fragments termed 'footprints'.

These ribosome footprints are isolated and converted to a library for deep sequencing.



Brar, Weissman, Nature Rev Mol Cell Biol 16, 651–664 (2015)

PreTIS: predict alternative translation initiation sites

```
  1 CGGUGAGGGU UCUCGGGCGG GGCCUGGGAC AGGCAGCUCC GGGGUCCGCG GUUUCACAUC
 61 GGAAACAAAA CAGCGGCUGG UCUGGAAGGA ACCUGAGCUA CGAGCCGCGG CGGCAGCGGG
121 GCGGCGGGGA AGCGUAUACC UAAUCUGGGA GCCUGCAAGU GACAACAGCC UUUGCGGUCC
181 UUAGACAGCU UGGCCUGGAG GAGAACACAU GAAAGAAAGA ACCUCAAGAG GCUUUGUUUU
241 CUGUGAAACA GUAUUUCUAU ACAGUUGCUC CAAUGACAGA GUUACCUGCA CCGUUGUCCU
301 ACUUCCAGAA UGCACAGAUG UCUGAGGACA ACCACCUGAG CAAUACUGUA CGUAGCCAGA
361 AUGACAAUAG AGAACGGCAG GAGCACAACG ACAGACGGAG CCUUGGCCAC CCUGAGCCAU
421 ...
```

Example mRNA sequence showing the categorization of true positive (TP) and true negative (TN) start sites.

Suppose that a ribosome profiling experiment detected the following start sites for a given mRNA sequence: CUG at position -78 and CUG at position -120 (blue colored codons). These start sites are then assumed to be TP start sites. In consequence, all near-cognate start sites not listed in the ribosome profiling dataset and upstream of the most downstream reported true start site were assumed to be TN (dark red colored codons).

Light red colored codons : start sites not considered as false starts in the analyses since they are located downstream of the most downstream reported true start site.

Grey colored downstream part : annotated CDS sequence

Italic (purple) upstream part : -99 upstream window needed to calculate some features.

All marked start sites (TP and TN) exhibit a surrounding window of ±99 nucleotides as well as a downstream in–frame stop codon. In total, this mRNA sequence would provide 2 **true start sites** and 9 **false start sites** out of 23 putative starts.

Reuter et al Plos Comput Biol (2016) 12: e10005170

Data sets used for ML classifier

Cell line	Description	Genes	Start codons	TPs	TNs	Used for	Source
HEK293	Human embryonic kidney cells	3,566	AUG and near-cognate	4,482	49,520	Human prediction model	[3]
HEK293	Human embryonic kidney cells	391	AUG	332	447	Validation set	[5]
Mouse ES	Mouse embryonic stem cells	1,632	AUG and near-cognate	3,009	19,864	Mouse prediction model	[4]

Three different datasets were used in this study to establish a human and mouse prediction model and to cross-validate the regression models. The numbers indicate the filtered start sites used in the prediction approach.

doi:10.1371/journal.pcbi.1005170.t001

We only included curated mRNA sequences with available mRNA RefSeq identifier (starting with NM_).

Raw data is very unbalanced (number of TPs and TNs very different)

→ need to balance data sets (select random TN data points)

Reuter et al Plos Comput Biol (2016) 12: e10005170

Features used by PreTIS

Mean value and standard deviation of the 44 features that were used in the best human model.

PWM : probability weight matrix

$$\mathrm{PWM}_{(nt,i)} = log\left(\frac{PFM_{(nt,i)}}{bg_{nt}}\right)$$

Entries of position–frequency–matrix (PFM) : sum of occurrences of a nucleotide at position *i* divided by the total number of sequences contained in *S*.

Reuter et al Plos Comput Biol (2016) 12: e10005170

	Feature	True starts	False starts	P–value
1.	**5' UTR length**	414.41±270.48	675.41±545.35	$< 10^{-310}$
2.	**5' UTR conservation**	0.4±0.16	0.33±0.16	8.2×10^{-190}
3.	**PWM positive**	2.75±1.5	-0.14±2.82	5.5×10^{-173}
4.	K-mer: upstream AUG	0.22±0.57	0.59±0.9	5.1×10^{-144}
5.	**5' UTR: percentage A**	0.18±0.05	0.2±0.05	9.6×10^{-100}
6.	**Kozak sequence context**	2.67±1.07	2.3±1.11	9.2×10^{-95}
7.	**Translational efficiency of flanking sequence**	83.75±20.11	77.12±21.4	1.1×10^{-83}
8.	K-mer: position -12 is C	0.13±0.34	0.3±0.46	2.7×10^{-77}
9.	K-mer: upstream Asparagine	1.25±1.37	1.61±1.61	4.0×10^{-43}
10.	K-mer: downstream AUG	1.14±1.15	0.92±1.1	9.2×10^{-41}
11.	K-mer: upstream A	17.24±7.43	18.81±7.89	4.0×10^{-40}
12.	K-mer: in-frame upstream Alanine	3.69±2.6	3.16±2.29	4.0×10^{-37}
13.	K-mer: upstream Alanine	10.27±4.5	9.38±4.6	6.2×10^{-37}
14.	**5' UTR: percentage G**	0.32±0.06	0.31±0.05	7.1×10^{-37}
15.	**Codon conservation**	0.23±0.42	0.12±0.32	3.2×10^{-36}
16.	K-mer: position -3 is A	0.31±0.46	0.2±0.4	3.4×10^{-35}
17.	K-mer: upstream CCG	2.98±2.43	2.56±2.31	7.1×10^{-34}
18.	K-mer: downstream CCA	2.04±1.54	1.75±1.45	1.1×10^{-32}
19.	K-mer: position -12 is A	0.3±0.46	0.19±0.4	4.0×10^{-32}
20.	K-mer: in-frame upstream Methionine	0.07±0.29	0.2±0.48	3.3×10^{-31}
21.	K-mer: upstream Arginine	12.15±4.34	11.33±4.64	1.5×10^{-29}
22.	K-mer: upstream Histidine	1.7±1.52	1.97±1.65	2.2×10^{-27}
23.	K-mer: GCC	6.4±3.87	5.77±3.75	1.1×10^{-25}
24.	K-mer: position 4 is G	0.37±0.48	0.28±0.45	2.3×10^{-25}
25.	K-mer: upstream Threonine	3.56±2.08	3.91±2.19	4.9×10^{-25}
26.	K-mer: upstream CGG	3.14±2.51	2.77±2.41	3.2×10^{-24}
27.	K-mer: upstream C	30.4±8.98	28.96±9.04	1.0×10^{-23}
28.	K-mer: position -2 is G	0.23±0.42	0.32±0.47	1.2×10^{-23}
29.	K-mer: upstream Stop	2.3±1.71	2.66±2.0	1.4×10^{-23}
30.	K-mer: UAG	1.34±1.2	1.57±1.35	5.6×10^{-23}
31.	K-mer: upstream CAU	0.58±0.85	0.73±0.95	3.4×10^{-22}
32.	K-mer: upstream Serine	9.44±3.29	8.93±3.14	5.7×10^{-22}
33.	K-mer: downstream Glutamine	3.57±2.01	3.26±1.88	2.4×10^{-21}
34.	K-mer: AGG	4.29±2.51	4.7±2.69	2.1×10^{-20}
35.	K-mer: AGC	4.4±2.43	4.02±2.19	2.1×10^{-20}
36.	K-mer: downstream ACC	1.45±1.26	1.27±1.17	2.0×10^{-19}
37.	K-mer: UAA	1.22±1.42	1.51±1.76	6.2×10^{-19}
38.	K-mer: downstream Proline	9.3±5.63	8.56±5.47	3.5×10^{-18}
39.	K-mer: upstream CAA	0.75±0.92	0.91±1.06	1.3×10^{-17}
40.	K-mer: in-frame upstream Histidine	0.54±0.77	0.67±0.87	1.7×10^{-17}
41.	K-mer: upstream GAU	0.63±0.85	0.77±0.96	2.1×10^{-16}
42.	K-mer: in-frame upstream GCC	1.21±1.4	1.02±1.22	6.7×10^{-16}
43.	K-mer: in-frame upstream GCG	1.14±1.42	0.97±1.27	6.2×10^{-14}
44.	**PWM negative**	1.94±1.34	1.59±1.09	1.6×10^{-08}

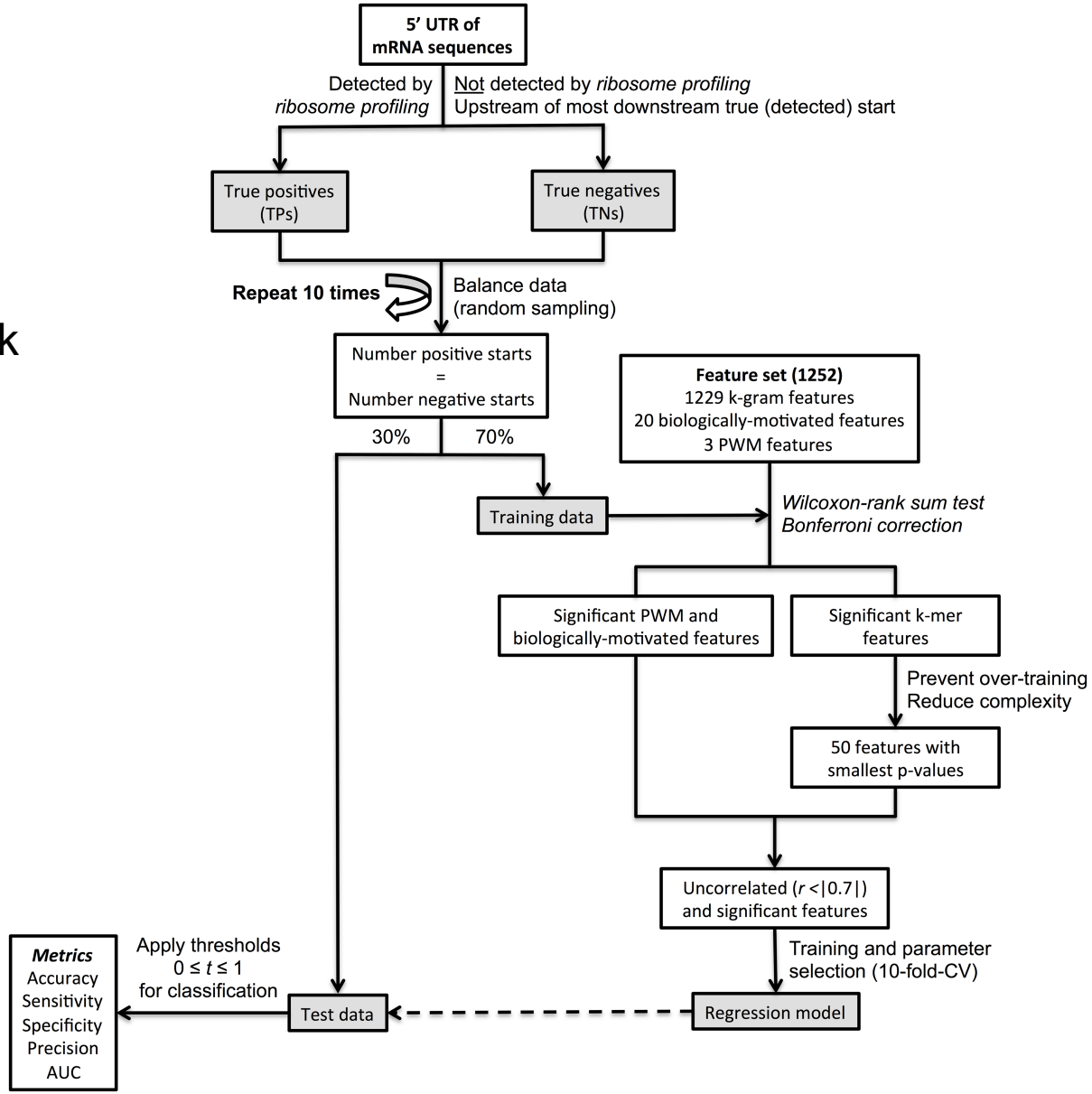
Mean value and standard deviation of the 44 features that were used in the best human model (biologically-motivated and PWM features are shown in bold). All 4,482 true and 49,520 false start sites were considered for this analysis. All listed features showed significant differences between true and false start sites (P–values < 1.6×10^{-8}). Note that due to numerical reasons, very small p–values (< 10^{-310}) are represented as 0.0 in python programming language (*scipy version 0.17.0*). The PWM–scores are based on the test data (compare to Fig 4).

doi:10.1371/journal.pcbi.1005170.t003

Flow-chart of regression approach

Data balancing was repeated ten times to investigate model robustness.

Significant features were identified by the Wilcoxon-rank sum test.



Reuter et al Plos Comput Biol V7(2016) 12: e10005170

Evaluation

	Accuracy	Specificity	Sensitivity	Precision	AUC	Threshold
HEK293						
Linear SVR	0.80±0.01	0.80±0.01	0.81±0.01	0.80±0.01	0.80±0.01	0.62±0.01
RBF SVR	0.82±0.01	0.81±0.01	0.83±0.02	0.82±0.01	0.82±0.01	0.55±0.02
Polynomial SVR	0.80±0.01	0.80±0.01	0.81±0.02	0.80±0.01	0.80±0.01	0.59±0.02
Linear Regression	0.80±0.01	0.80±0.01	0.81±0.01	0.80±0.01	0.80±0.01	0.55±0.01
Mouse ES						
Linear SVR	0.75±0.01	0.75±0.01	0.76±0.01	0.75±0.01	0.76±0.01	0.65±0.03
RBF SVR	0.76±0.01	0.76±0.01	0.76±0.02	0.76±0.01	0.76±0.01	0.58±0.03
Polynomial SVR	0.75±0.02	0.75±0.01	0.76±0.02	0.75±0.02	0.75±0.02	0.62±0.03
Linear Regression	0.76±0.01	0.75±0.01	0.76±0.01	0.75±0.01	0.76±0.01	0.55±0.01

The prediction was repeated 10 times to evaluate the model robustness. Shown are the average performance measures.

doi:10.1371/journal.pcbi.1005170.t002

All human models perform very similarly with accuracies of about 80%

while the average performance of the mouse model is lower with average accuracies of about 76%,

Reuter et al Plos Comput Biol (2016) 12: e10005170

PWM_positive scores

Frequency distribution of $PWM_{positive}$ scores for the test samples of the best performing run 2.

The PWM was established using the true start sites in the training data of run 2. The difference between TPs and TNs was found to be highly significant ($p = 5.5 \times 10^{-173}$, Wilcoxon–rank sum test).

Reuter et al Plos Comput Biol (2016) 12: e10005170

Is model transferable to other species?

Performance of the best human HEK293 model applied to the mouse ES dataset

→ model is reasonably transferable,

suggests universal translation code

Unbalanced datasets

	Mouse ES		**Mouse ES**	
Threshold	$t = 0.54$		$t = 0.52$	
	TP	TN	TP	TN
Predicted positive	2,161	4,569	2,273	5,072
Predicted negative	848	15,295	736	14,792
Total	3,009	19,864	3,009	19,864
Accuracy	0.76		0.75	
Sensitivity	0.72		0.76	
Specificity	0.77		0.74	
Precision	0.32		0.31	

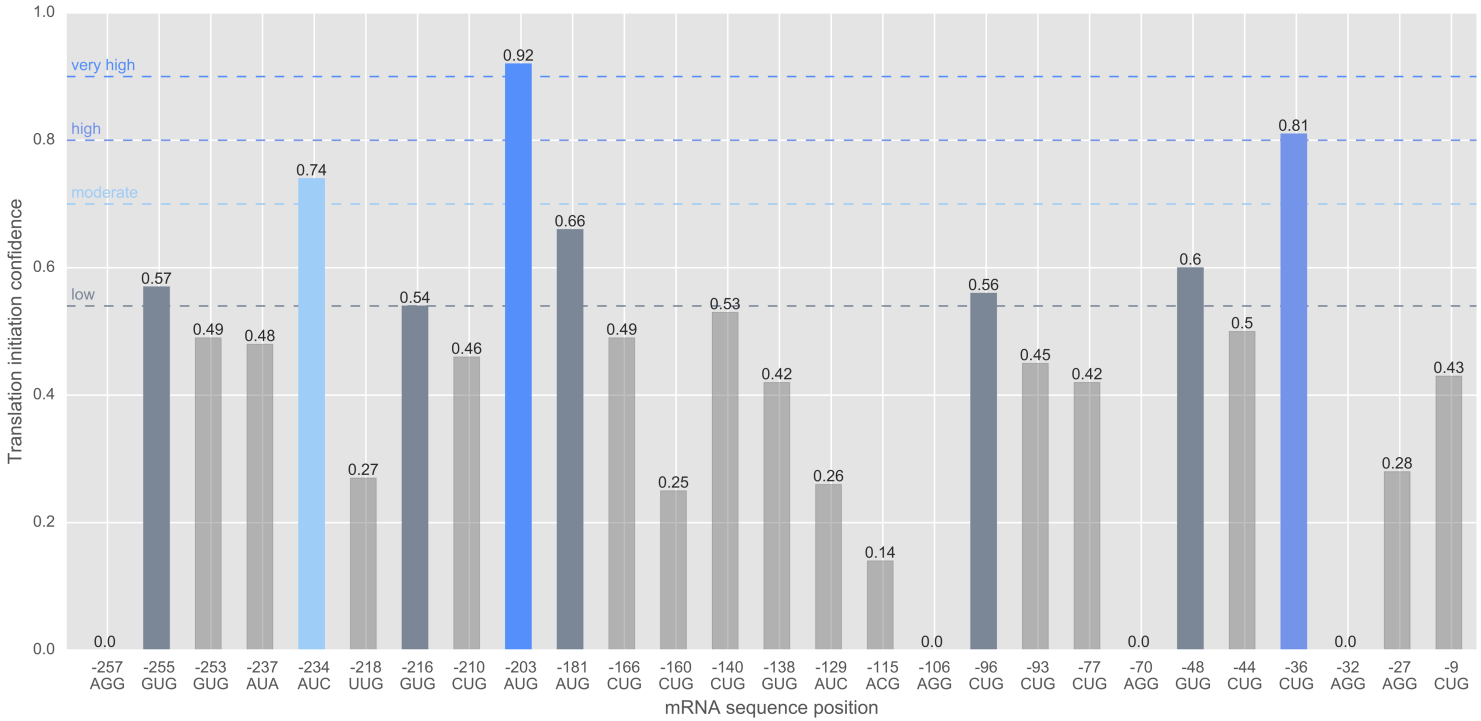
Balanced datasets

	Mouse ES		**Mouse ES**	
Threshold	$t = 0.54$		$t = 0.52$	
	TP	TN	TP	TN
Predicted positive	2,161	689	2,273	763
Predicted negative	848	2,320	736	2,246
Total	3,009	3,009	3,009	3,009
Accuracy	0.74		0.75	
Sensitivity	0.72		0.76	
Specificity	0.77		0.75	
Precision	0.76		0.75	

doi:10.1371/journal.pcbi.1005170.t004

Reuter et al Plos Comput Biol (2016) 12: e10005170

Alternative start codons of human gene GIMAP5



Predicted start sites were subdivided into 4 confidence groups and highlighted by different colors and dashed lines: very high (hot/best candidates with $c \geq 0.9$), high ($0.8 \leq c < 0.9$), moderate ($0.7 \leq c < 0.8$) and low ($t = 0.54 \leq c < 0.7$) initiation confidence c.

For this gene, we found one hot candidate with a very high confidence value of 0.92 of being a true start site (AUG at position -203).

Reuter et al Plos Comput Biol (2016) 12: e10005170

Virtual SNP analysis of gene GIMAP5

Mutation matrix showing the impact of the flanking sequence context of 4 putative start sites of gene *GIMAP5* on the predicted initiation confidence.

In each case, only one nucleotide is mutated with respect to the reference sequence (top line). Grey : start was predicted as true translational start (predicted initiation confidence > 0.54). white : start was classified as false start.

Mutations at the start sites itself were not considered. The numbers reflect the predicted initiation confidence values

(A) *CUG at position -36*

SNPs	-15 U	-14 C	-13 A	-12 G	-11 U	-10 G	-9 A	-8 C	-7 U	-6 G	-5 C	-4 C	-3 A	-2 C	-1 C	**1 C**	**2 U**	**3 G**	4 G	5 A	6 G	7 G	8 A	9 C	10 A	11 G	12 G	13 G
A	**0.80**	**0.80**		**0.80**	**0.83**	**0.82**		**0.73**	**0.84**	**0.82**	**0.81**	**0.84**		**0.85**	**0.83**				**0.80**		**0.82**	**0.86**		**0.83**		**0.86**	**0.89**	**0.85**
C	**0.81**		**0.80**	**0.64**	**0.83**	**0.81**	**0.75**		**0.80**	**0.82**			**0.67**			■			**0.78**	**0.81**	**0.82**	**0.86**	**0.79**		**0.80**	**0.82**	**0.81**	**0.83**
G	**0.80**	**0.79**	**0.79**		**0.77**		**0.78**	**0.78**	**0.78**		**0.76**	**0.80**	**0.74**	**0.80**	**0.80**			■		**0.73**			**0.77**	**0.79**	**0.73**			
U		**0.76**	**0.78**	**0.83**		**0.81**	**0.82**	**0.80**		**0.84**	**0.83**	**0.81**	**0.70**	**0.83**	**0.80**		■		**0.74**	**0.77**	**0.86**	**0.86**	**0.81**	**0.80**	**0.77**	**0.85**	**0.83**	**0.84**

(B) *CUG at position -44*

SNPs	-15 C	-14 C	-13 A	-12 G	-11 A	-10 G	-9 C	-8 C	-7 U	-6 C	-5 A	-4 G	-3 U	-2 G	-1 A	**1 C**	**2 U**	**3 G**	4 C	5 C	6 A	7 C	8 C	9 C	10 U	11 G	12 G	13 A
A	0.49	0.49		**0.57**		0.49	**0.55**	0.49	0.49	0.51		0.46	**0.66**	**0.61**					**0.54**	0.52		0.52	0.50	**0.54**	0.51	**0.54**	**0.56**	
C			0.50	0.34	0.49	0.47			0.48		0.50	0.52	0.50	**0.58**	0.48	■					0.48				0.48	**0.54**	0.52	0.47
G	0.49	0.48	0.49		0.42		0.51	0.46	0.45	0.51	0.45		**0.57**		0.46			■	**0.60**	0.44	0.49	0.47	0.45	0.47	0.45			0.48
U	0.51	0.49	0.48	0.52	0.47	0.48	**0.56**	0.49		0.49	0.51	0.49		**0.55**	0.45		■		0.50	0.46	0.51	0.50	0.50	0.50		**0.56**	0.53	0.50

(C) *AUA at position -237*

SNPs	-15 U	-14 G	-13 G	-12 G	-11 G	-10 G	-9 A	-8 C	-7 A	-6 C	-5 A	-4 C	-3 U	-2 C	-1 C	**1 A**	**2 U**	**3 A**	4 A	5 U	6 C	7 U	8 C	9 U	10 A	11 C	12 U	13 U
A	0.48	0.49	0.50	**0.56**	**0.54**	0.49		0.48		0.50		0.50	**0.63**	0.51	0.50	■		■		0.53	0.48	0.48	0.49	0.50		0.48	0.50	0.48
C	0.48	0.51	0.50	0.33	0.52	0.46	0.45		0.46		0.50		0.46						0.44	**0.54**		0.47		0.48	0.45		0.47	0.46
G	0.46						0.44	0.44	0.44	0.52	0.45	0.46	**0.55**	0.40	0.46				0.50	0.47	0.49	0.43	0.44	0.45	0.43	0.42	0.43	0.45
U		0.50	0.48	0.52	0.52	0.48	0.48	0.47	0.49	0.50	0.52	0.45		0.46	0.45		■		0.45		0.51		0.49		0.47	0.49		

(D) *CUG at position -160*

SNPs	-15 C	-14 C	-13 U	-12 C	-11 C	-10 U	-9 U	-8 A	-7 A	-6 C	-5 U	-4 G	-3 C	-2 G	-1 U	**1 C**	**2 U**	**3 G**	4 C	5 U	6 C	7 A	8 A	9 C	10 C	11 U	12 C	13 C
A	0.23	0.24	0.25	0.47	0.26	0.26	0.25			0.20	0.25	0.31	0.46	0.33	0.30				0.29	0.31	0.24			0.25	0.26	0.26	0.30	0.28
C			0.25			0.24	0.20	0.26	0.24		0.24	0.30		0.31	0.28	■				0.29		0.24	0.24			0.23		
G	0.23	0.23	0.24	0.40	0.21	0.25	0.22	0.21	0.22	0.28	0.20		0.34		0.26			■	0.32	0.23	0.25	0.20	0.20	0.22	0.21	0.20	0.24	0.24
U	0.25	0.25		0.44	0.25			0.25	0.27	0.26		0.27	0.25	0.30			■		0.25		0.27	0.24	0.23	0.24	0.25		0.27	0.27

Take home messages

- You may want to remove sequence redundancy

- Check for overlapping genes

- Which isoform is relevant?

  There are substantial differences between what is expressed at the transcript level and what is expressed at the protein level.

  CCDS and APPRIS appear good resources.

- Which translated variant is relevant? May want to try PreTIS

Reuter et al Plos Comput Biol (2016) 12: e10005170