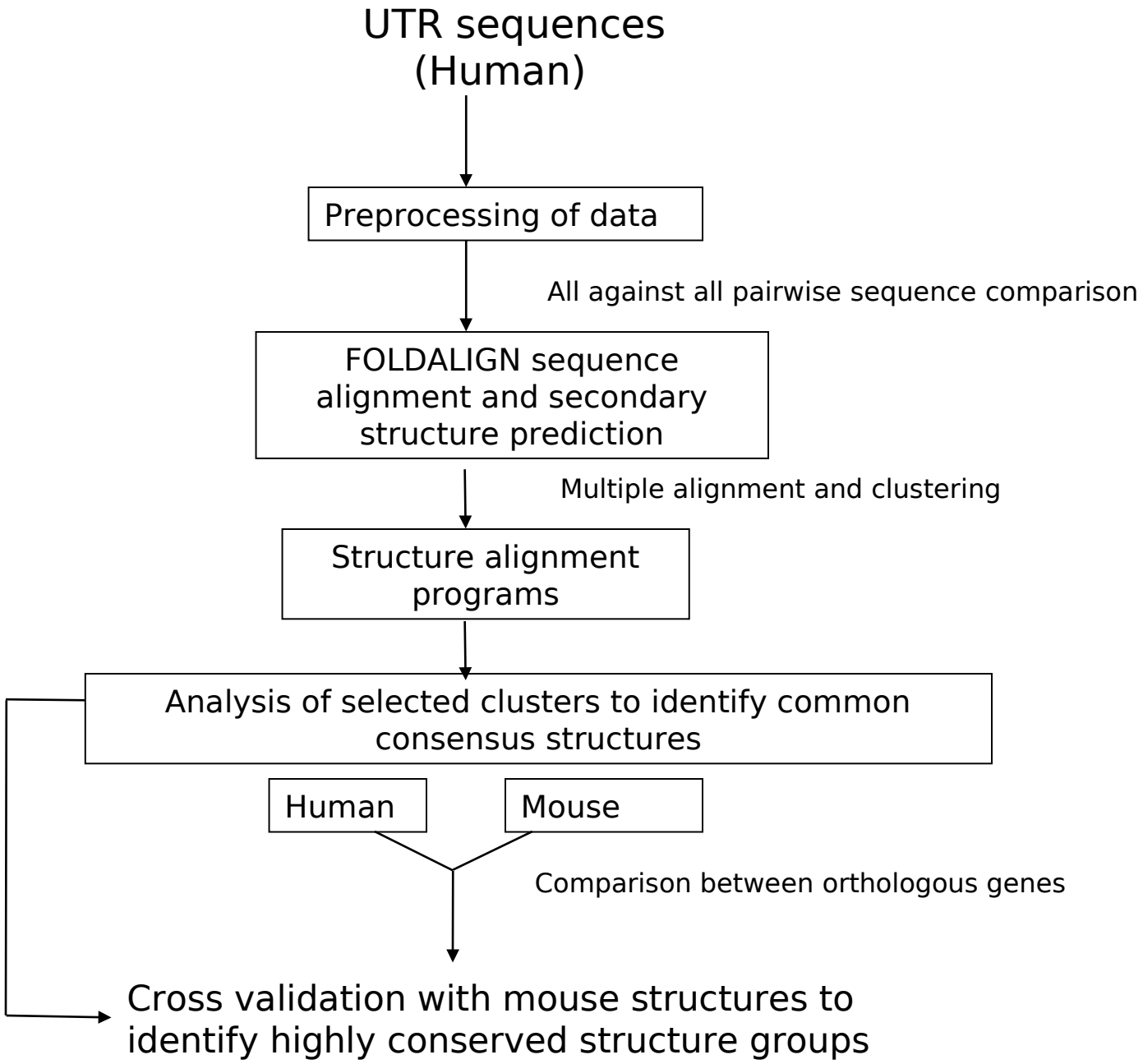
Sequence analysis of Human UTRs

Winterseminar 2012, Bled

Simranjeet Kaur
Centre for non-coding RNA in technology and Health

Outline

- An approach to uncover new conserved RNA structures in human UTRs
- Systematic study of combined sequence and structure similarity among human UTRs.


UTR sequences
(Human)
Preprocessing of data
All against all pairwise sequence comparison
FOLDALIGN sequence alignment and secondary structure prediction
Multiple alignment and clustering
Structure alignment programs
Analysis of selected clusters to identify common consensus structures
Human
Mouse
Comparison between orthologous genes
Cross validation with mouse structures to identify highly conserved structure groups

UTR sequences for human (assembly GRCh37) retrieved from Ensembl release 64 using BioMart

	5’ UTRs	**3’ UTRs**	**Total**
UTRs	78,662	71,427	
Genes	20,399	20,390	20,764
Average transcript per gene	3.85	3.5	
Identical UTRs	11,429	12,343	
UTR candidates	67,233	59,084	126,317

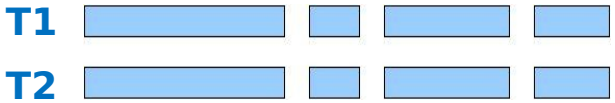
Genes having both 5’ and 3’ UTRs - 20,025

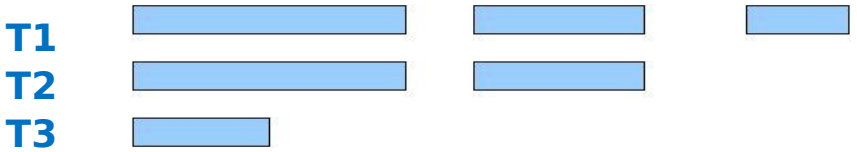
More insights into the human UTR data-set

	5’ UTRs	**3’ UTRs**	**Total**
UTRs after removing duplicates	66,665	58,869	125,534
Genes	19,451	19,573	
Average transcript per gene	3.43	3.01	
Genes with single transcripts	6,002	6,283	
Genes having both 5’ and 3’ UTRs			18,969
UTRs with single exons	35,932	30,733	
Max. length of UTRs	17,184 (average- 256)	17,159 (average- 950)	
Max length for UTRs with single exons	17,184 (average- 179)	14,960(average- 930)	

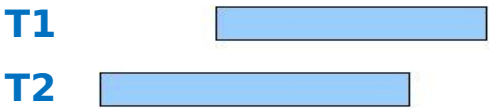
Removing cis redundancy from the dataset:-

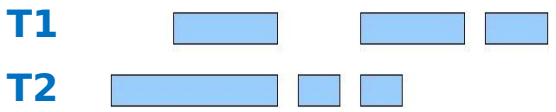
-Merged transcripts that were completely within other transcripts of a gene.





-Merged overlapping coordinates.





-New dataset (non-redundant UTRs for each gene) after merging overlapping and common exons:-

5'UTRs- 38,494

3'UTRs- 31,557

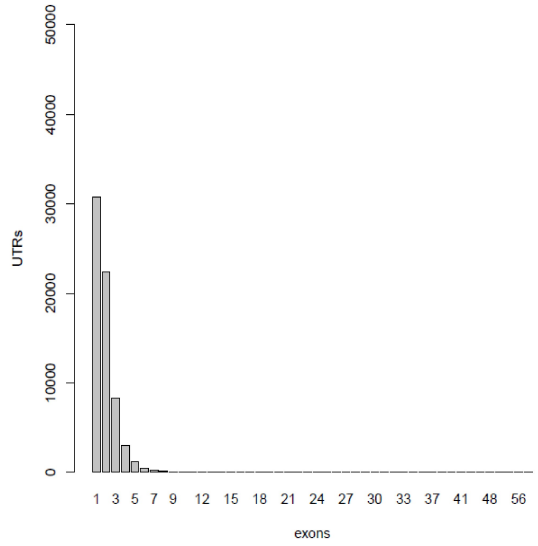
T1- Transcript 1

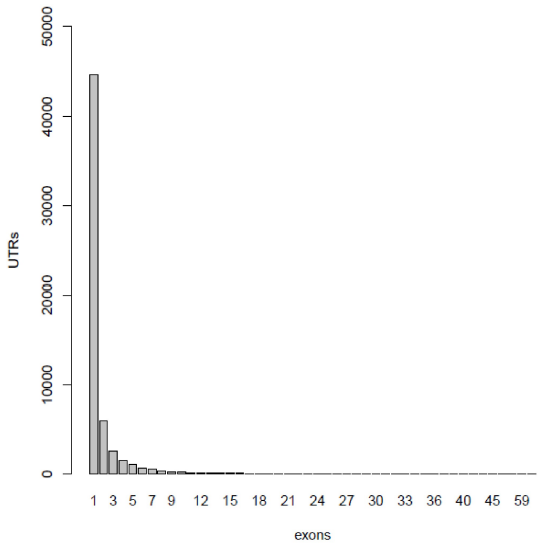
T2- Transcript 2

Distribution plots after merging overlapping coordinates (cis redundancy reduction)

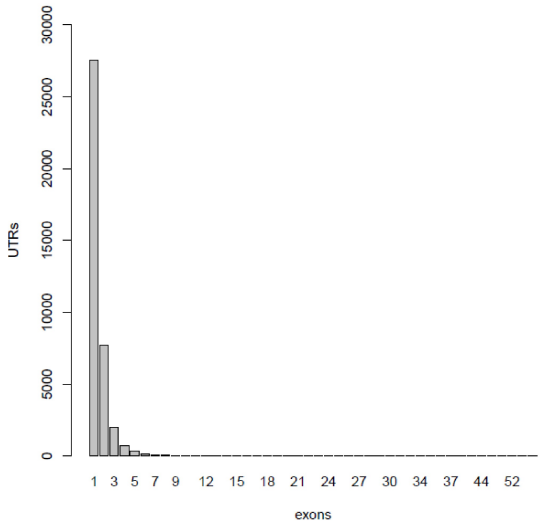
Exons distribution

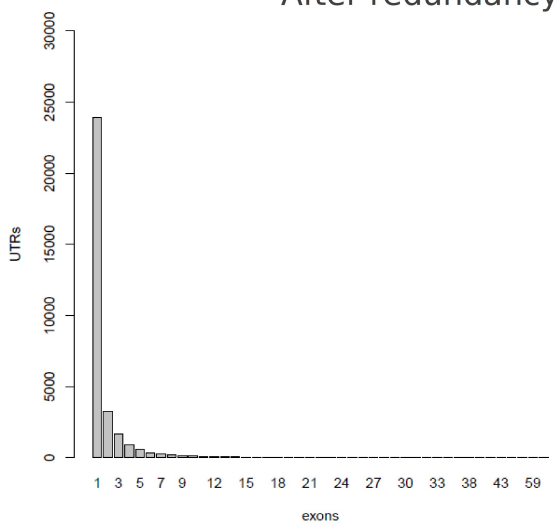
Before redundancy reduction





After redundancy reduction

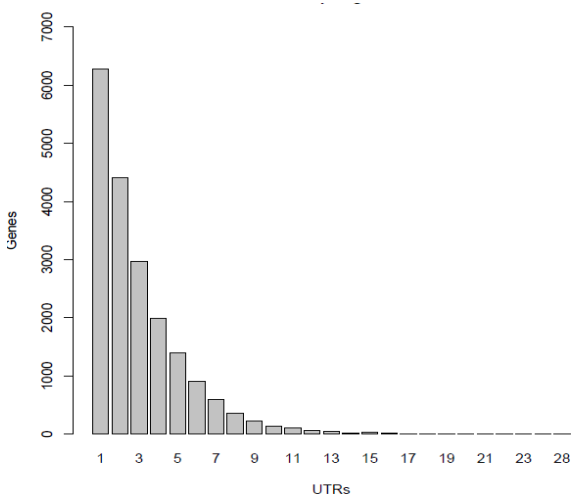




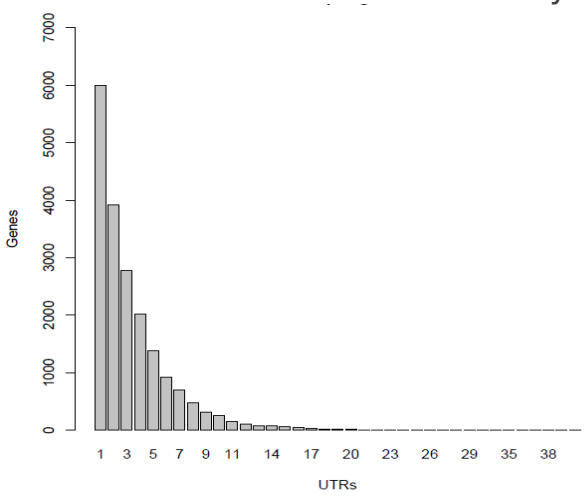
5’UTRs

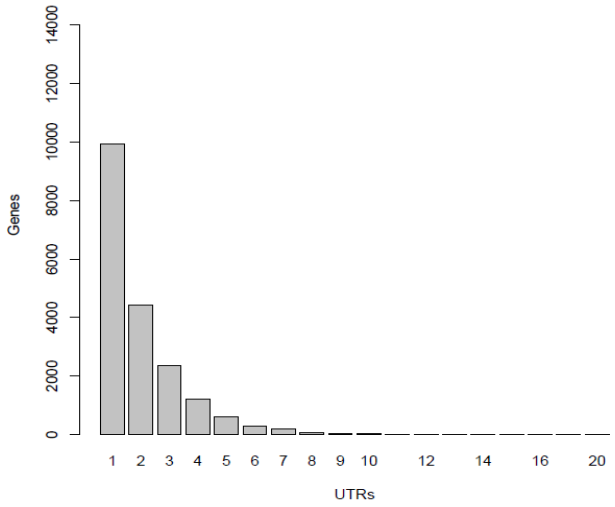
3’UTRs

UTRs distribution

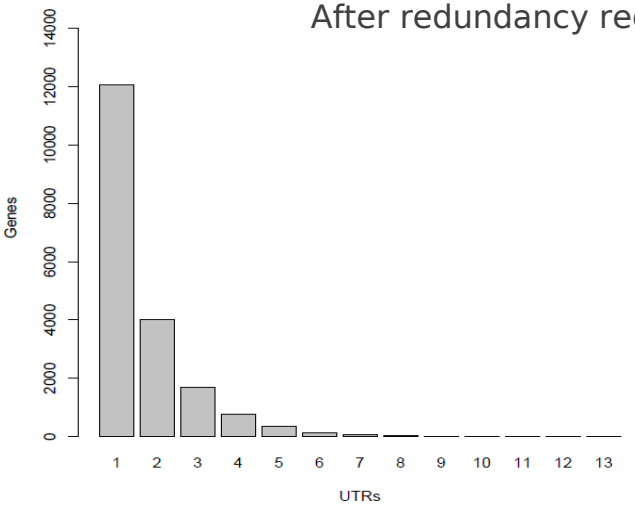


Before redundancy reduction





After redundancy reduction



5’UTRs

3’UTRs

Identifying repeat elements in UTRs

Repeat elements in UTRs identified and masked (excluding small RNAs) using RepeatMasker Version: 4/26/2011.

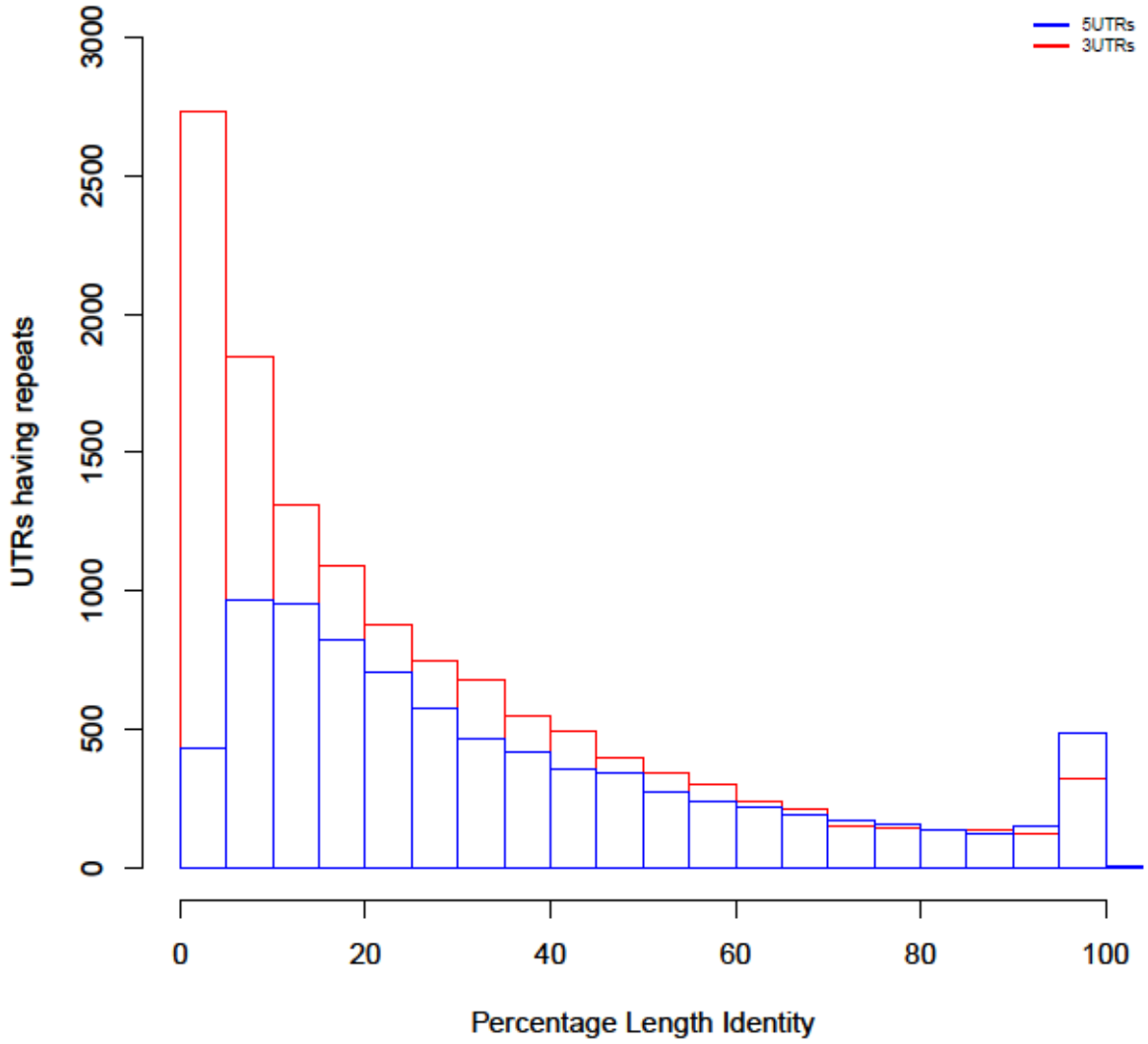
	5'UTRs	3'UTRs
Total UTR Sequences	38,494	31,557
Number of UTRs having repeats	8,191	12,824
Total Number of Repeats in UTRs	11,367	29,401
Average length of alignment with repeats in UTRs	25.00%	11.28%
Bases masked	1,106,390 bp (10.16%)	4,557,296 bp (13.04%)
UTRs having more than 80% length covered by repeats	903	721

Types of Repeats found in human UTRs

	5'UTRs			3'UTRs		
	Number of elements	**Length occupied (bps)**	**Percentage of sequence**	**Number of elements**	**Length occupied (bps)**	**Percentage of sequence**
SINEs ALUs MIRs	2,813 1,694 1,109	403,185 276,006 125,883	3.70% 2.53% 1.16%	10,456 7,047 3,307	2,126,539 1,685,870 429,529	6.08% 4.82% 1.23%
LINEs LINE1 LINE2 L3/CR1	1,258 496 663 80	202,448 98,278 92,827 9,306	1.86% 0.90% 0.85% 0.09%	4,345 2,161 1,764 320	1,034,935 686,766 285,734 48,813	2.96% 1.96% 0.82% 0.14%
LTR elements ERVL ERVL-MaLRs ERV_classI ERV_classII	129,709 25,925 38,230 59,013 5,009	129,709 25,925 38,230 59,013 5,009	1.19% 0.24% 0.35% 0.54% 0.05%	1,799 390 888 445 43	557,845 120,350 260,735 148,838 21,954	1.60% 0.34% 0.75% 0.43% 0.06%
DNA elements hAT-Charlie TcMar-Tigger	419 257 95	59,802 33,403 17,669	0.55% 0.31% 0.16%	2,511 1,445 544	436,966 229,544 126,319	1.25% 0.66% 0.36%
Unclassified	17	5,076	0.05%	53	17,862	0.05%
Total Interspersed repeats		800,220	7.35%		4,174,147	11.94%
Small RNAs*	18	1,547	0.01%	108	10,799	0.03%
Satellites	28	5,472	0.05%	19	5,517	0.02%
Simple repeats	2,144	109,978	1.01%	3,485	161,512	0.46%
Low complexity	3,749	189,173	1.74%	5,449	207,974	0.59%

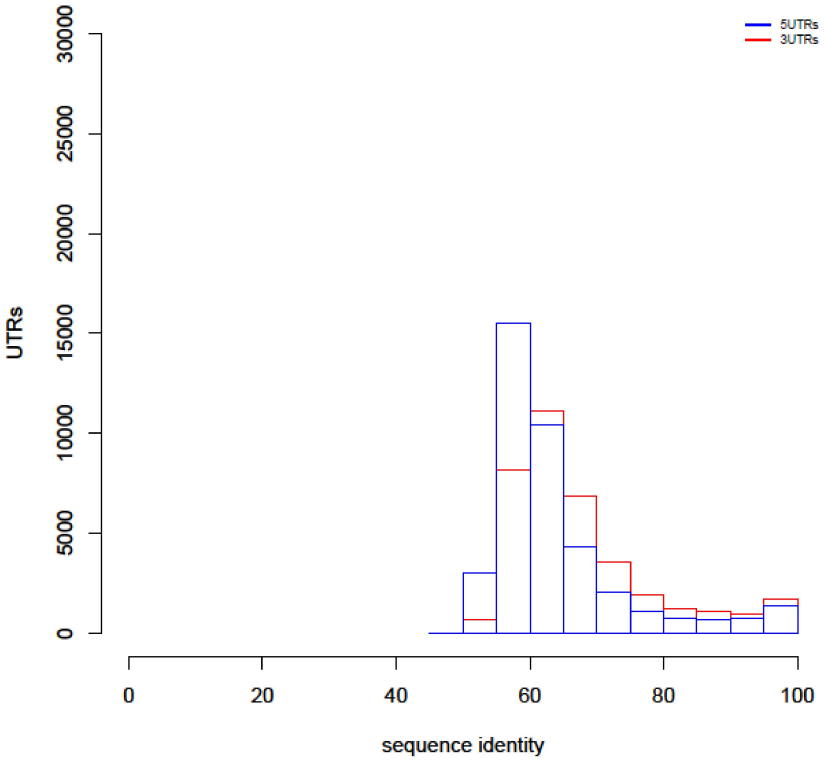
*Small RNAs includes snRNA, srpRNA, scRNA, rRNA, and tRNAs

Percentage length covered by repeats in UTRs

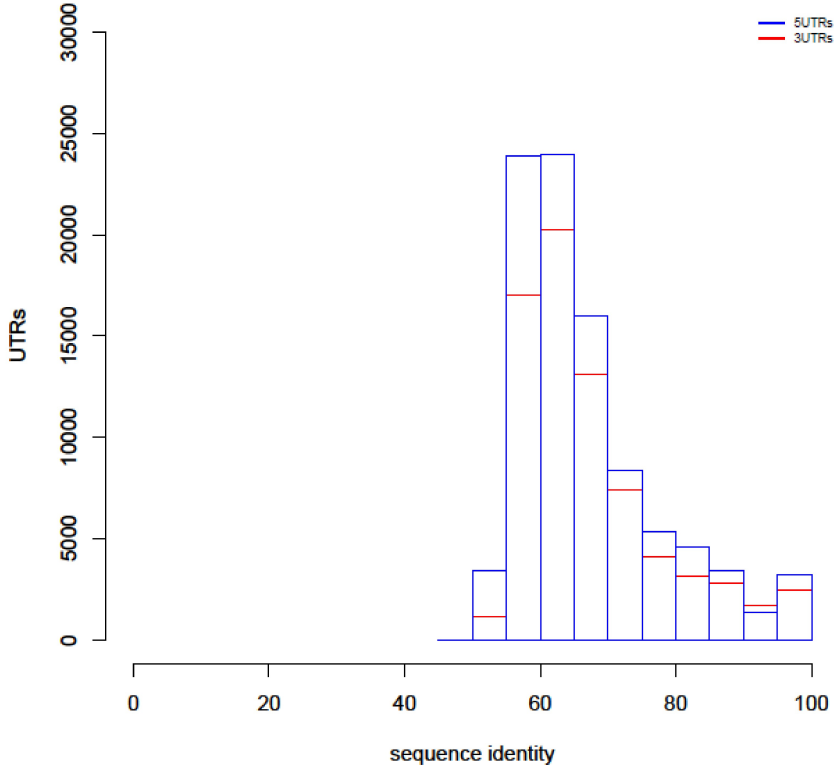


Computing pairwise alignments of UTRs (using ssearch33)

Repeat masked UTRs



Unmasked UTRs



Mouse UTRs (Ensembl release 65, assembly NCBIM37)

	5'UTRs	3'UTRs
Total UTRs	45,601	43,441
Genes	19,921	20,147
Average Length	231	1080
Maximum length	13,770	23,858
Average transcript per gene	2.28	2.15
UTRs after removing redundancy	29,536	25,204

Next steps:-

Removing redundancy from the UTR sequences

Mapping GO terms with UTRs and clustering of genes with overlapping GO terms.

Mapping all known regulatory elements with human and mouse UTRs.

Running structural alignments using Foldalign

Clustering of results based on Foldalign scores

Comparison with orthologous genes (Mouse UTRs)

Acknowledgements

My sincere thanks to

Jakob Hull Havgaard

Jan Gorodkin

THANK YOU