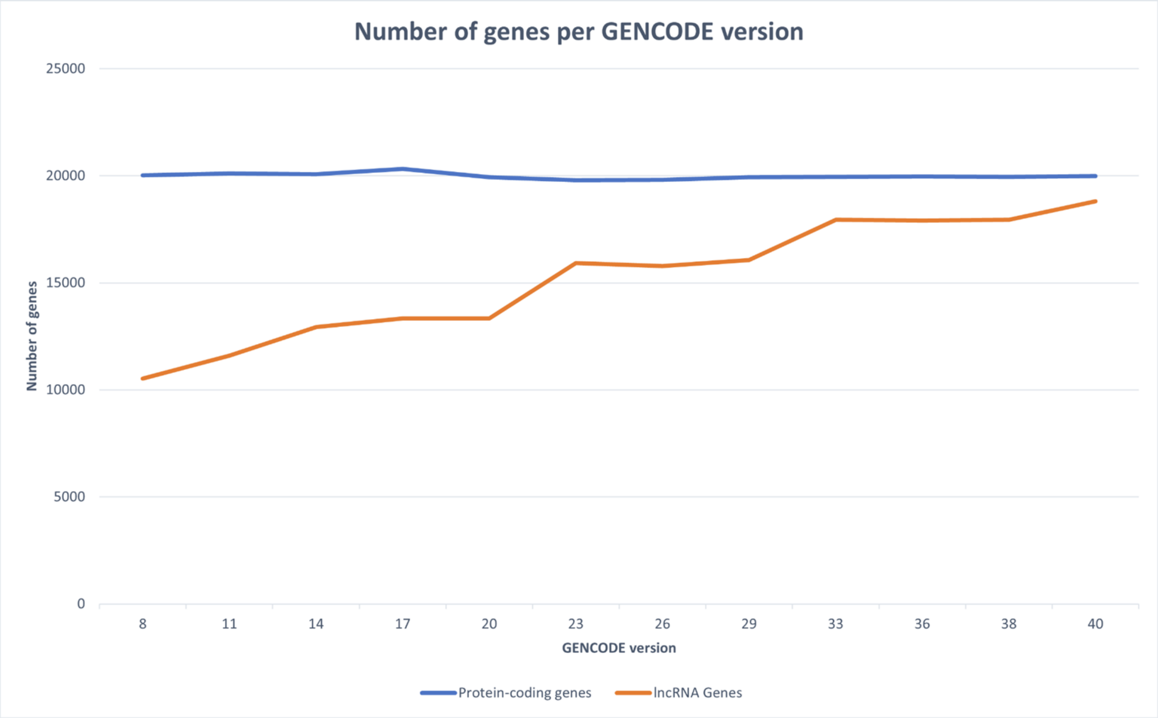
Genomic landscape of conserved RNA secondary structure signatures and their homologs

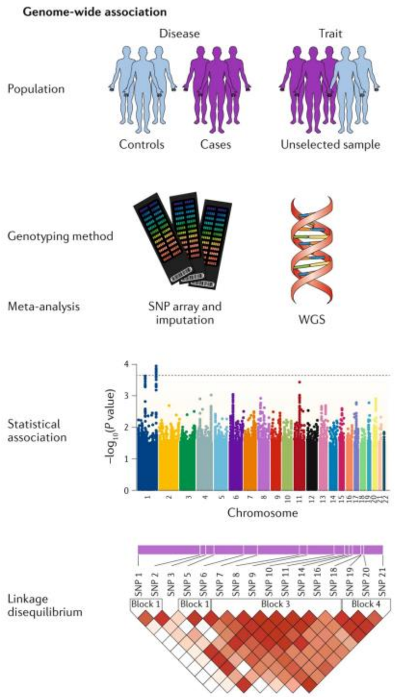
Vanda Gaonac’h-Lovejoy
Martin Smith Lab,
Benasque
Université de Montréal
August 2022

Increasing number lncRNAs with every new GENCODE version

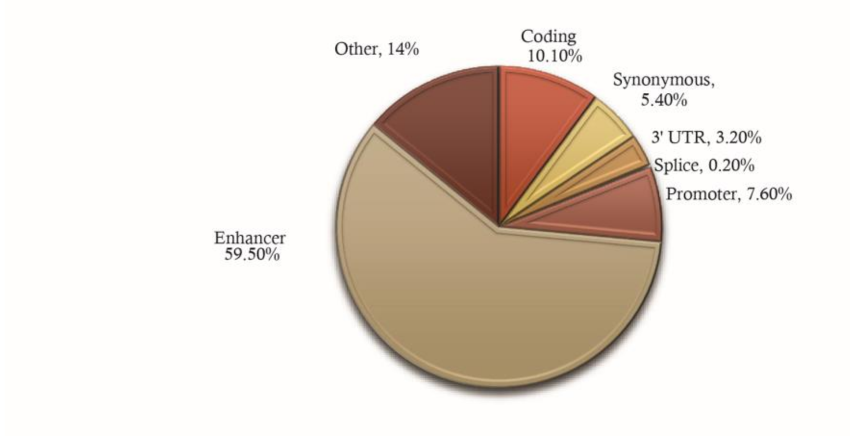


GENCODE

>90% of disease-associated mutation occur in non-coding genome

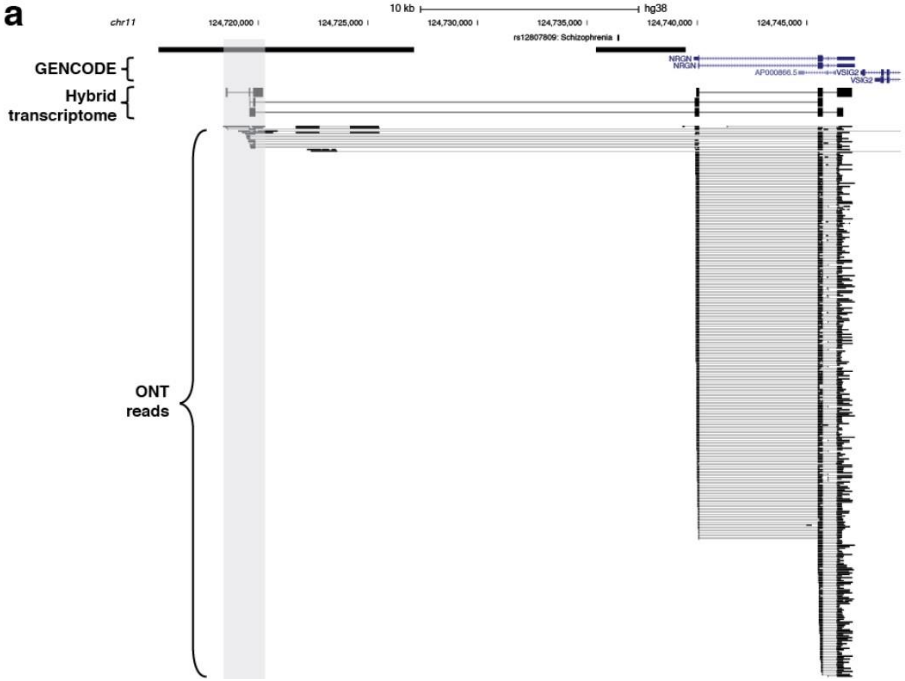


Vivian Tam *et al*. Nature rev Gen 2019.



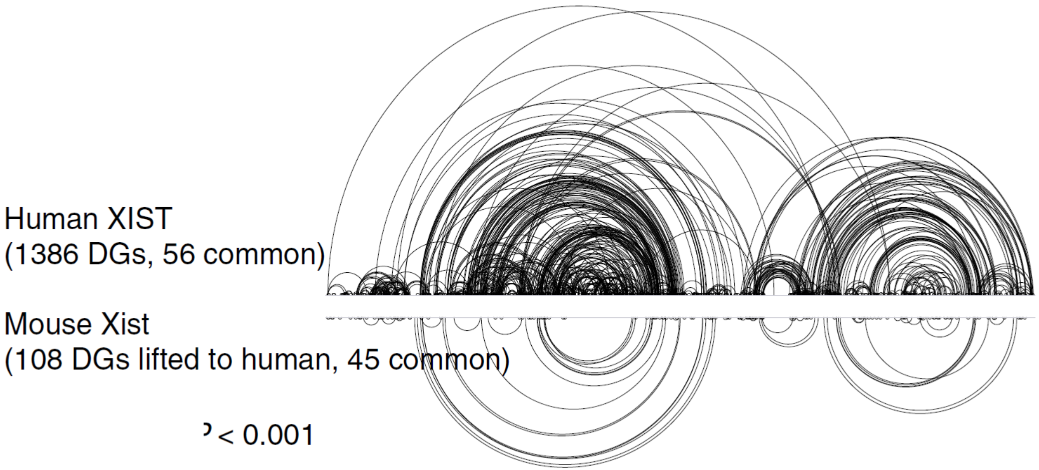
K Farh *et al*. Nature 2015.

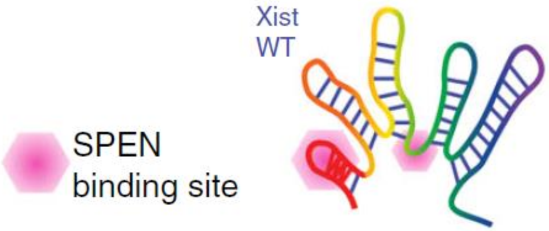
GWAS loci express lncRNAs



Simon A. Hardwick *et al*. Frontiers.2019.

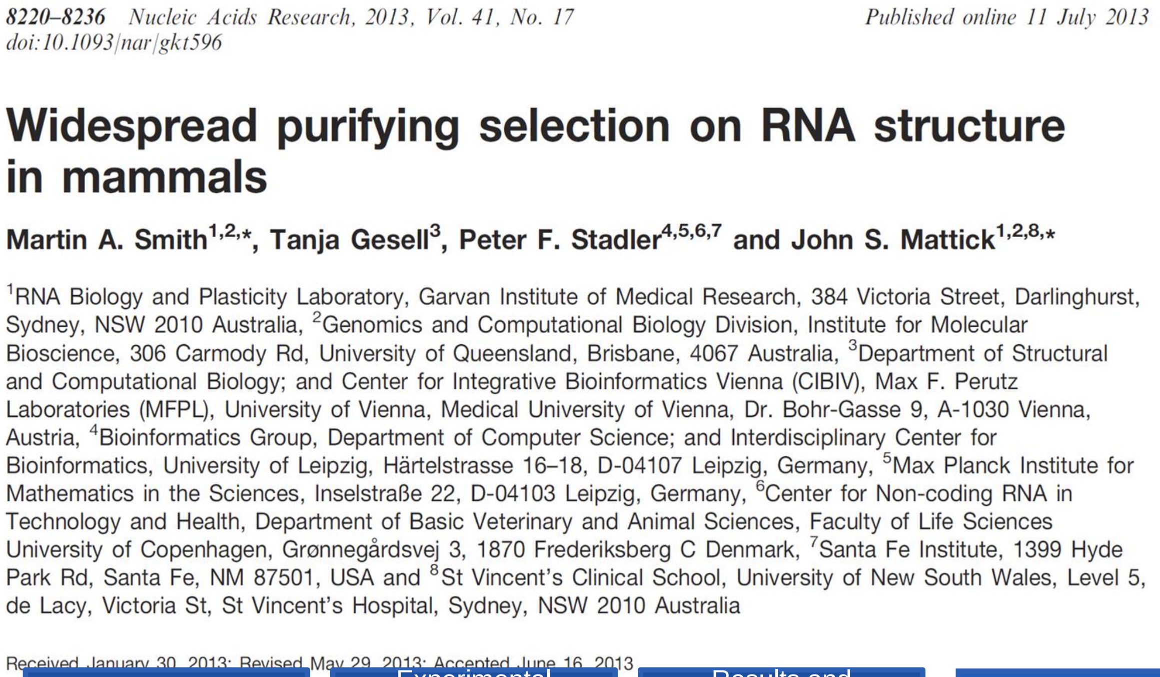
Xist is modular and conserved in evolution





Zhipeng Lu *et al*. Nature Commu.2020.

Revisiting a previous study



8220–8236 Nucleic Acids Research, 2013, Vol. 41, No. 17
doi:10.1093/nar/gkt596

Published online 11 July 2013

Widespread purifying selection on RNA structure in mammals

Martin A. Smith[1,2,*], Tanja Gesell[3], Peter F. Stadler[4,5,6,7] and John S. Mattick[1,2,8,*]

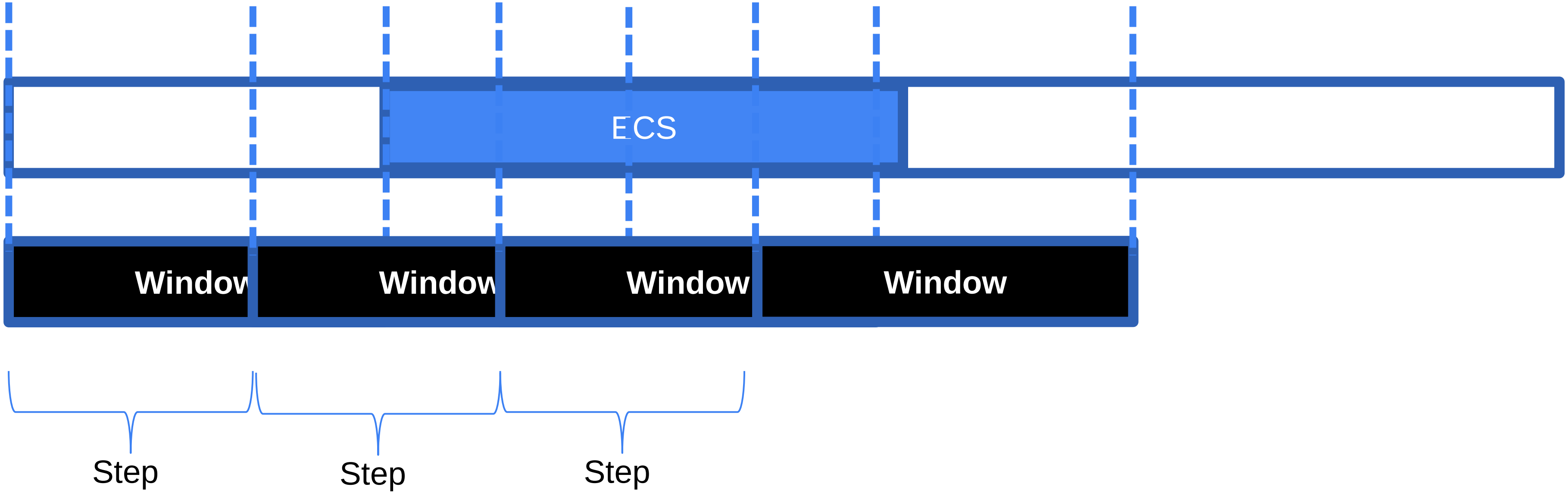
[1]RNA Biology and Plasticity Laboratory, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, NSW 2010 Australia, [2]Genomics and Computational Biology Division, Institute for Molecular Bioscience, 306 Carmody Rd, University of Queensland, Brisbane, 4067 Australia, [3]Department of Structural and Computational Biology; and Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), University of Vienna, Medical University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria, [4]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16–18, D-04107 Leipzig, Germany, [5]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany, [6]Center for Non-coding RNA in Technology and Health, Department of Basic Veterinary and Animal Sciences, Faculty of Life Sciences University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C Denmark, [7]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA and [8]St Vincent's Clinical School, University of New South Wales, Level 5, de Lacy, Victoria St, St Vincent's Hospital, Sydney, NSW 2010 Australia

Received January 30, 2013; Revised May 29, 2013; Accepted June 16, 2013

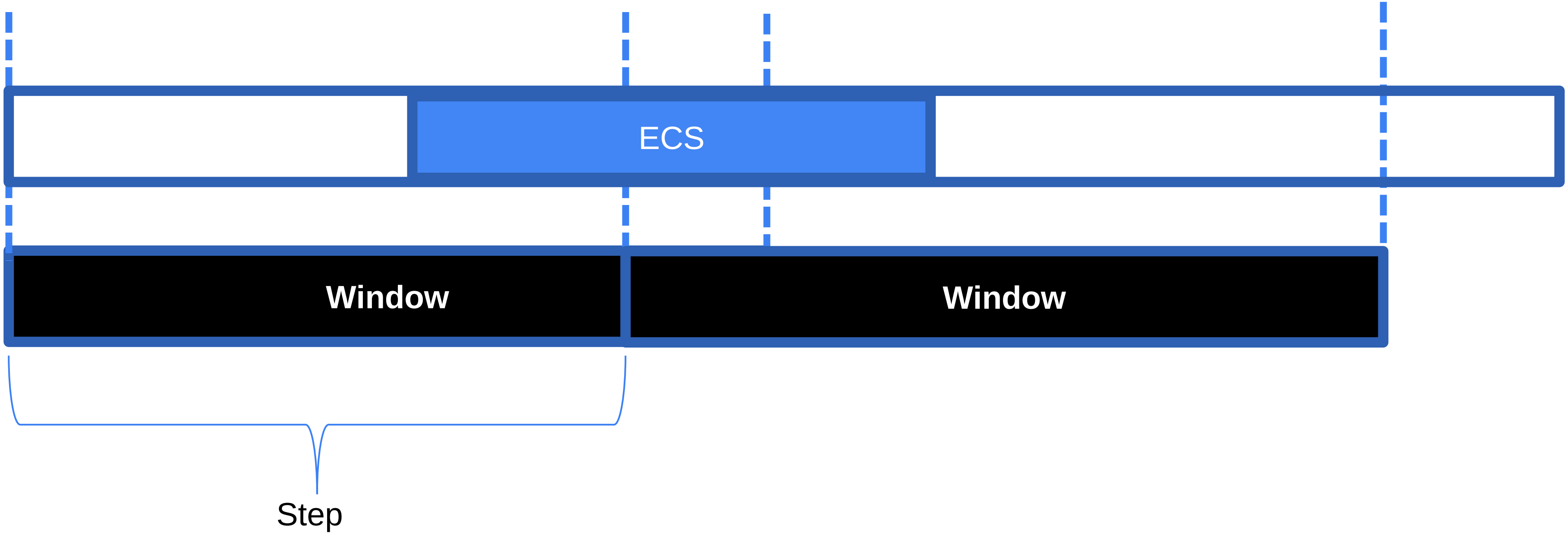
Research problem

- Increasing number of lncRNAs but no systematic approach for functional annotation
- Hypothesis: Comparative sequence analysis to identify, classify and map functional RNA structures
- Objective: Provide a rational framework for deciphering the structure functions of lncRNAs

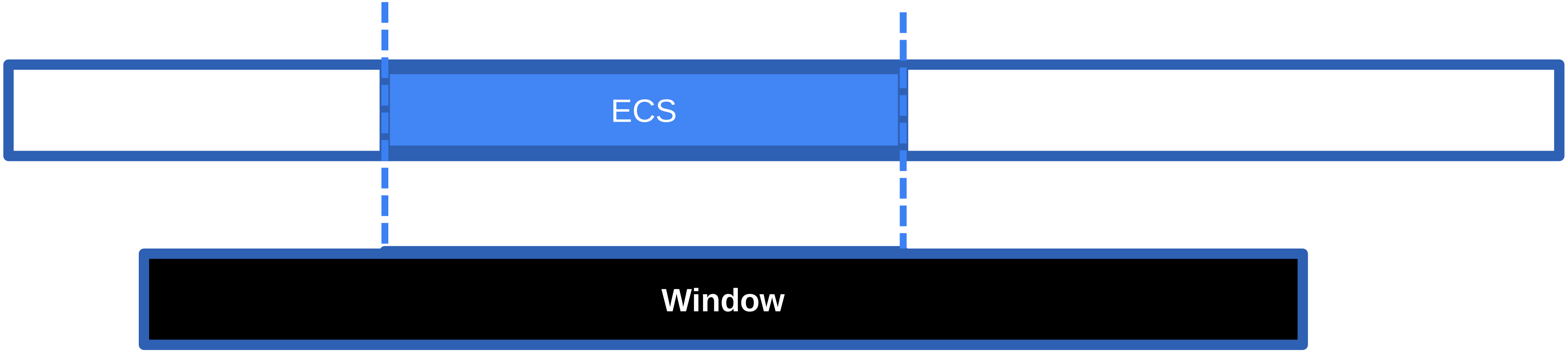
Previous study used fixed window length



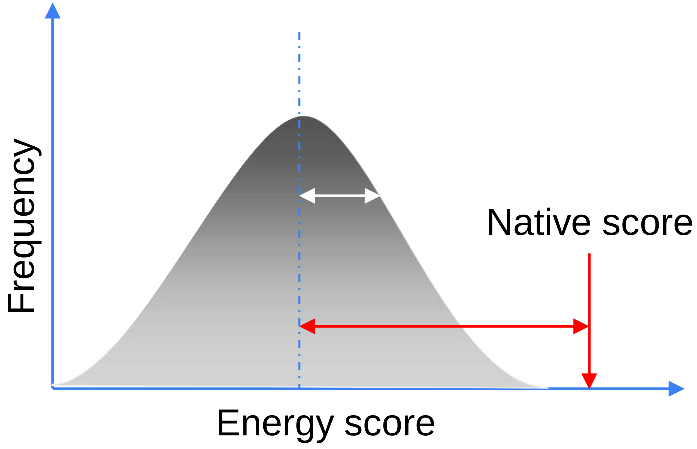
Added noise if window is too large



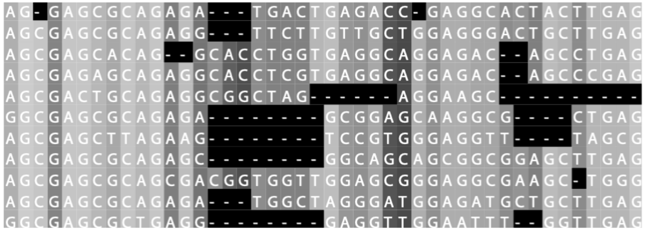
RNALalifold: Dynamic window approach



SISSIz: Detection of functional RNA structures



Gesell et al.Bioinformatics 2006
Gesell et al. BMC Bioinformatics 2008



```
AG-GAGCGCAGAGA---TGACTGAGACC-GAGGCACTACTTGAG
AGCGAGCGCAGAGG---TTCTTGTTGCTGGAGGGACTGCTTGAG
AGCGAGCACAG--GCACCTGGTGAGGCAGGAGAC--AGCCTGAG
AGCGAGAGCAGAGGCACCTCGTGAGGCAGGAGAC--AGCCCGAG
AGCGACTGCAGAGGCGGCTAG------AGGAAGC----------
GGCGAGCGCAGAGA--------GCGGAGCAAGGCG----CTGAG
AGCGAGCTTAGAAG--------TCCGTGGGAGGTT----TAGCG
AGCGAGCGCAGAGC--------GGCAGCAGCGGCGGAGCTTGAG
AGCGAGCGCAGCGACGGTGGTTGGAGCGGGAGGCGAAGC-TGGG
AGCGAGCGCAGAGA---TGGCTAGGGATGGAGATGCTGCTTGAG
GGCGAGCGCTGAGG--------GAGGTTGGAATTT--GGTTGAG
```

```
-55.76   ((((.........)))).(((((.......)))))
-0.86    .......((...........)).............
-8.56    (((................))).............
-1.38    ...............(.(.........))......
-12.76   ....((((.................))))......
-1.6     ...........(.(...............).)...
-8.63    .................(((.........)))...
```

This project

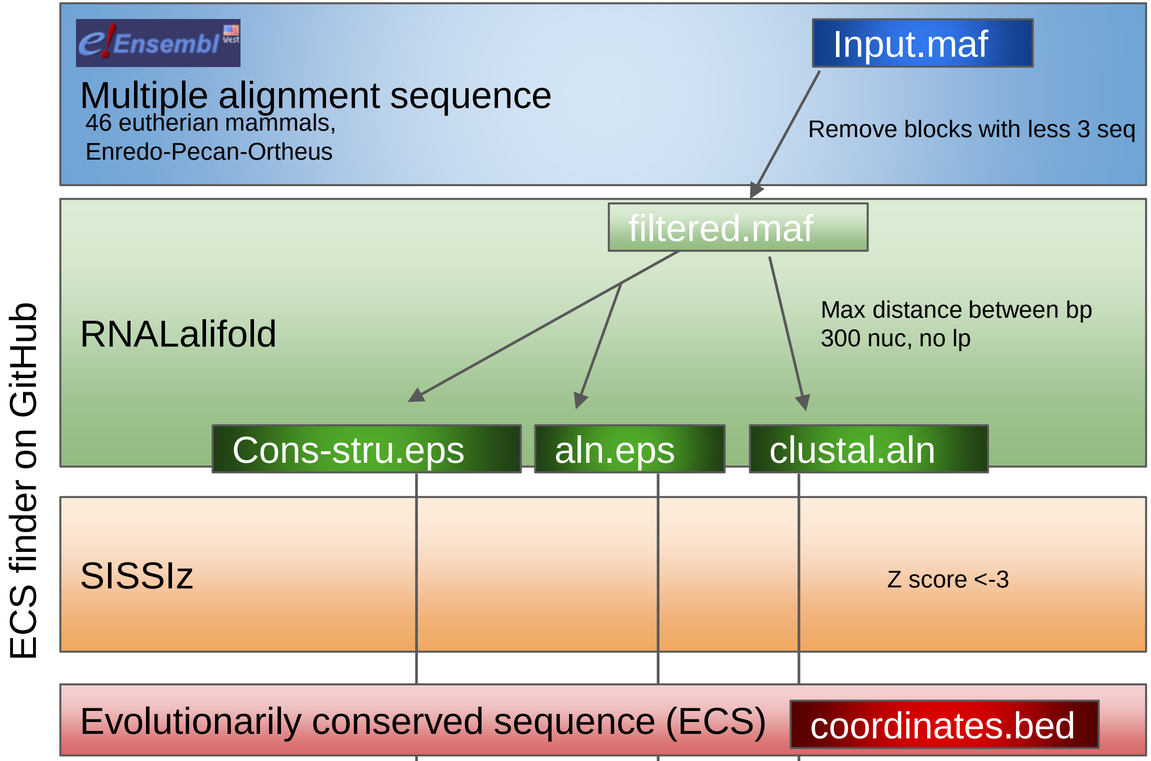
Deeper alignments:

- 46 mammals instead of 35
- Greater variability
- Likely to increase the specificity at the expense of loosing some sensitivity
- Harder to get a consensus structure

Dynamic window:

- RNALalifold instead of RNAalifold
- Locally more stable regions of interest
- Likely to increase sensitivity

Analytic pipeline



ECS finder on GitHub

Ensembl

Multiple alignment sequence

46 eutherian mammals,
Enredo-Pecan-Ortheus

Input.maf

Remove blocks with less 3 seq

filtered.maf

RNALalifold

Max distance between bp
300 nuc, no lp

Cons-stru.eps

aln.eps

clustal.aln

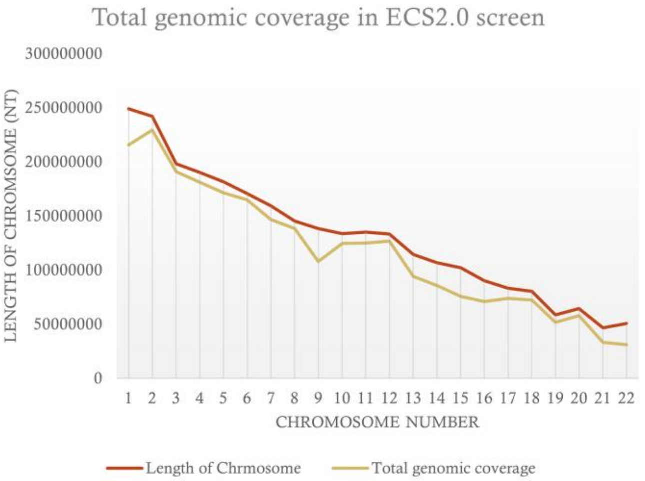
SISSIz

Z score <-3

Evolutionarily conserved sequence (ECS)

coordinates.bed

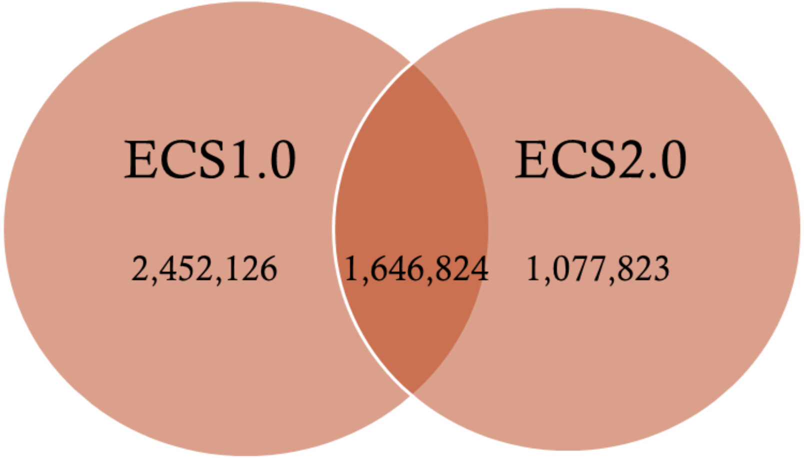
Detection of evolutionarily conserved RNA secondary structures (ECS)



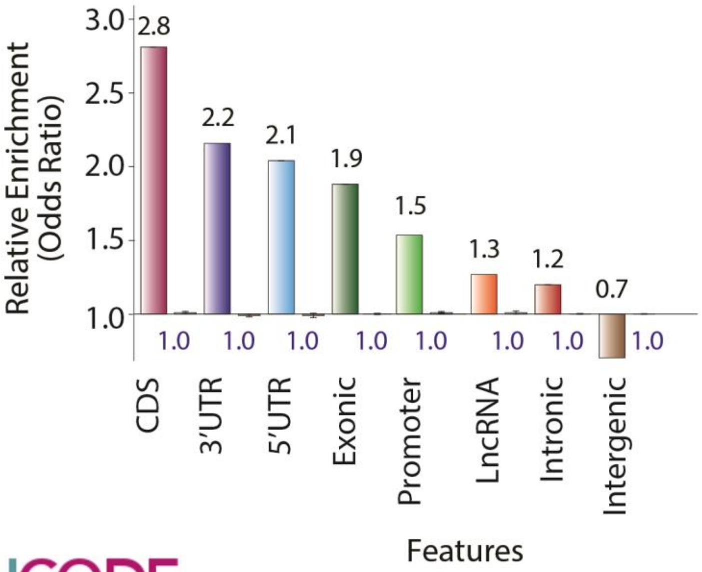
- 89% of the human genome sampled
- > 2 million evolutionarily conserved structures
- 6% genome is conserved at the secondary structure level
- Process completed in over 48,700 CPU hours (≈ 6 years)

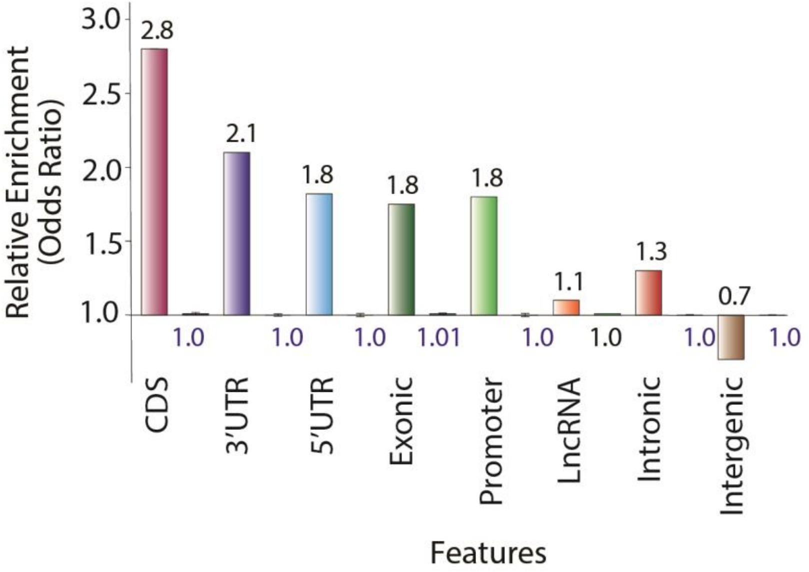
Revisited approach generated fewer predictions

- 60% of the hits had been identified in our 2013 study
- Revisited approach generated fewer predictions but likely to be more accurate



ECS are enriched in various functional motifs





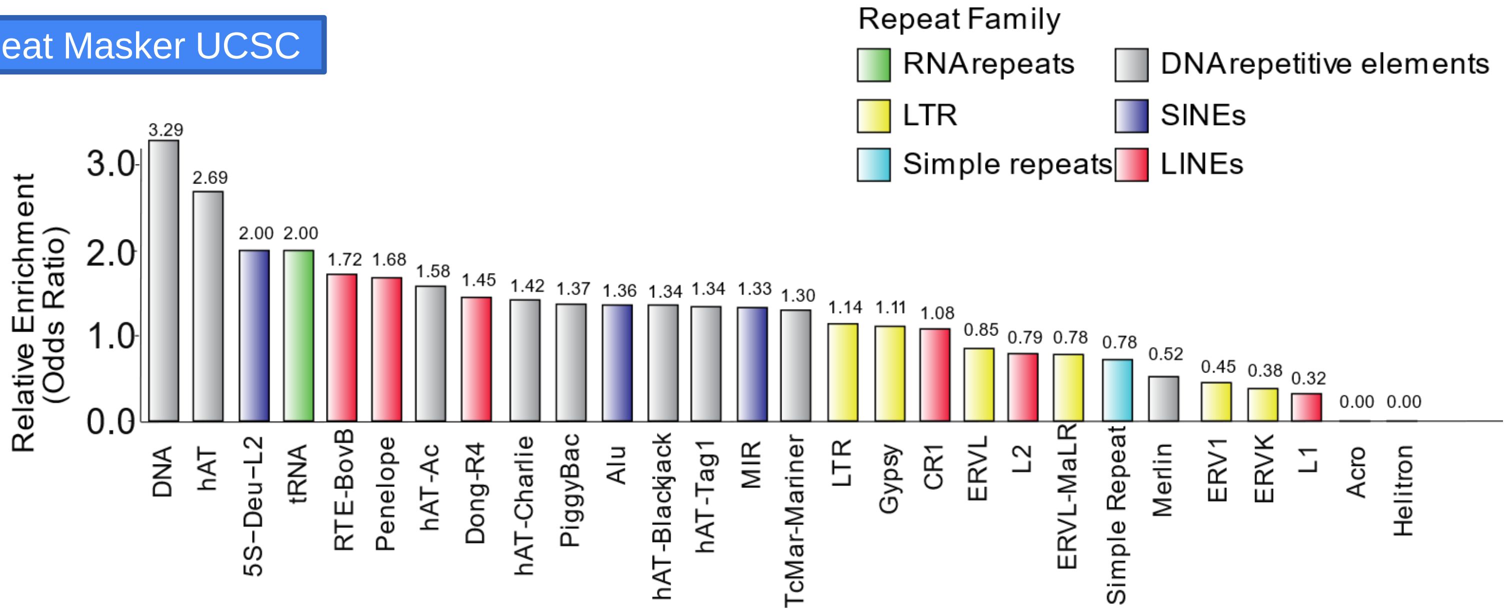
GENCODE

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

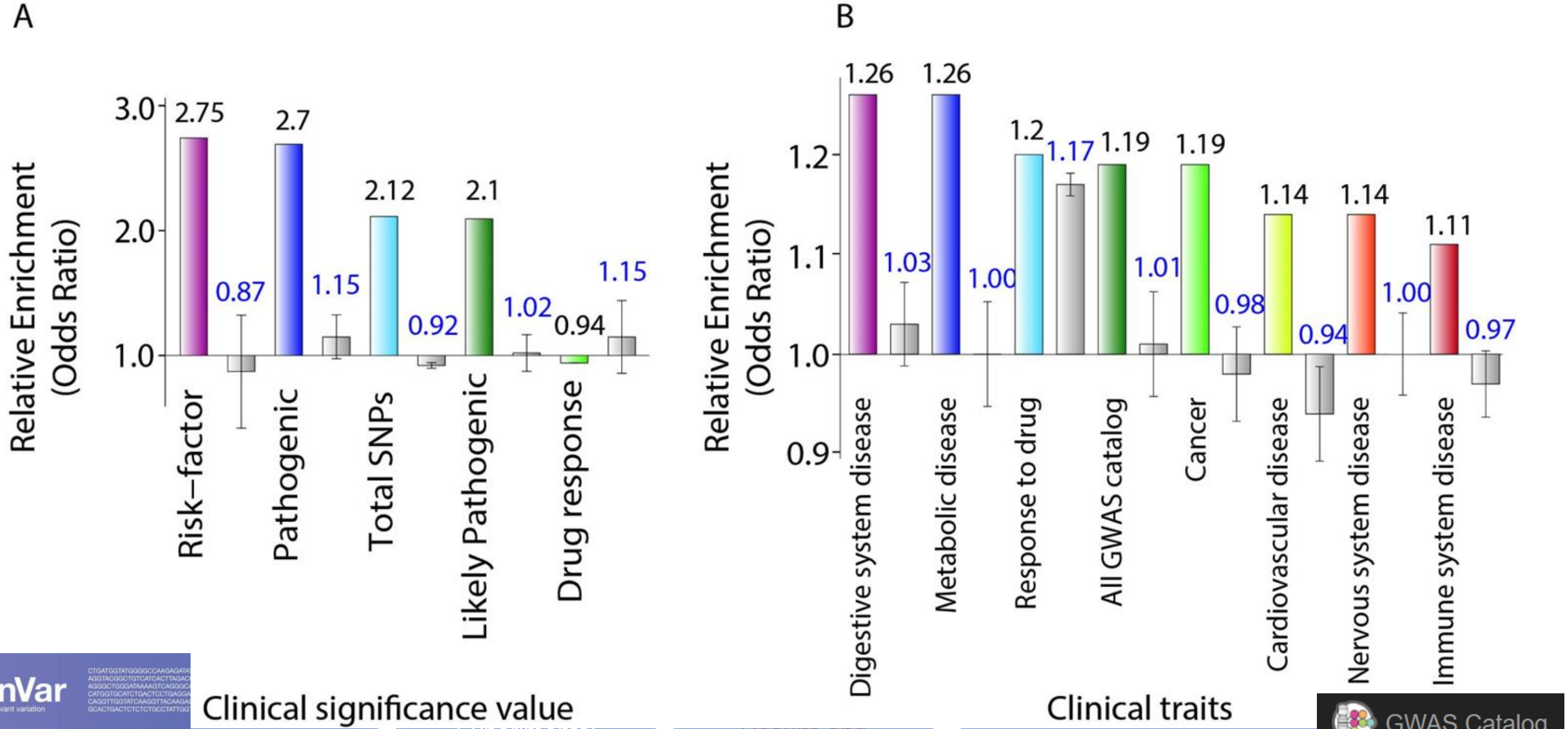
CHESS database

ECS are enriched in various transposable elements

Repeat Masker UCSC



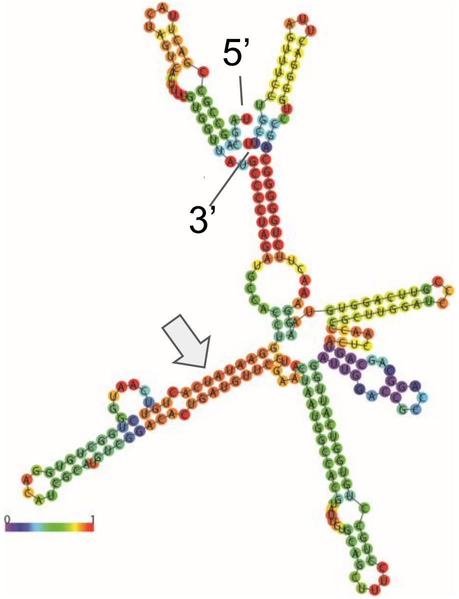
Non-coding ECSs are enriched in disease-associated SNPs



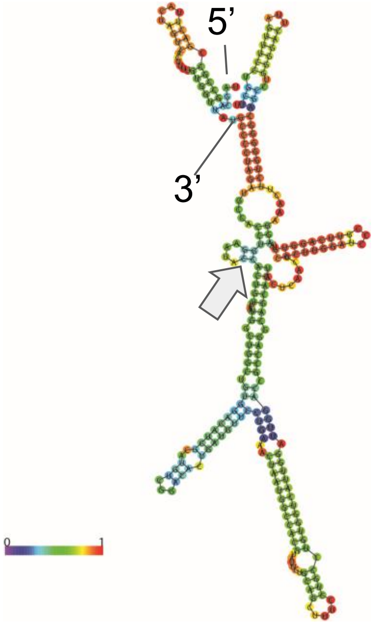
ClinVar
Clinically relevant variation

GWAS Catalog

Identified 23 pathogenic-associated SNPs that have riboSNitch potential

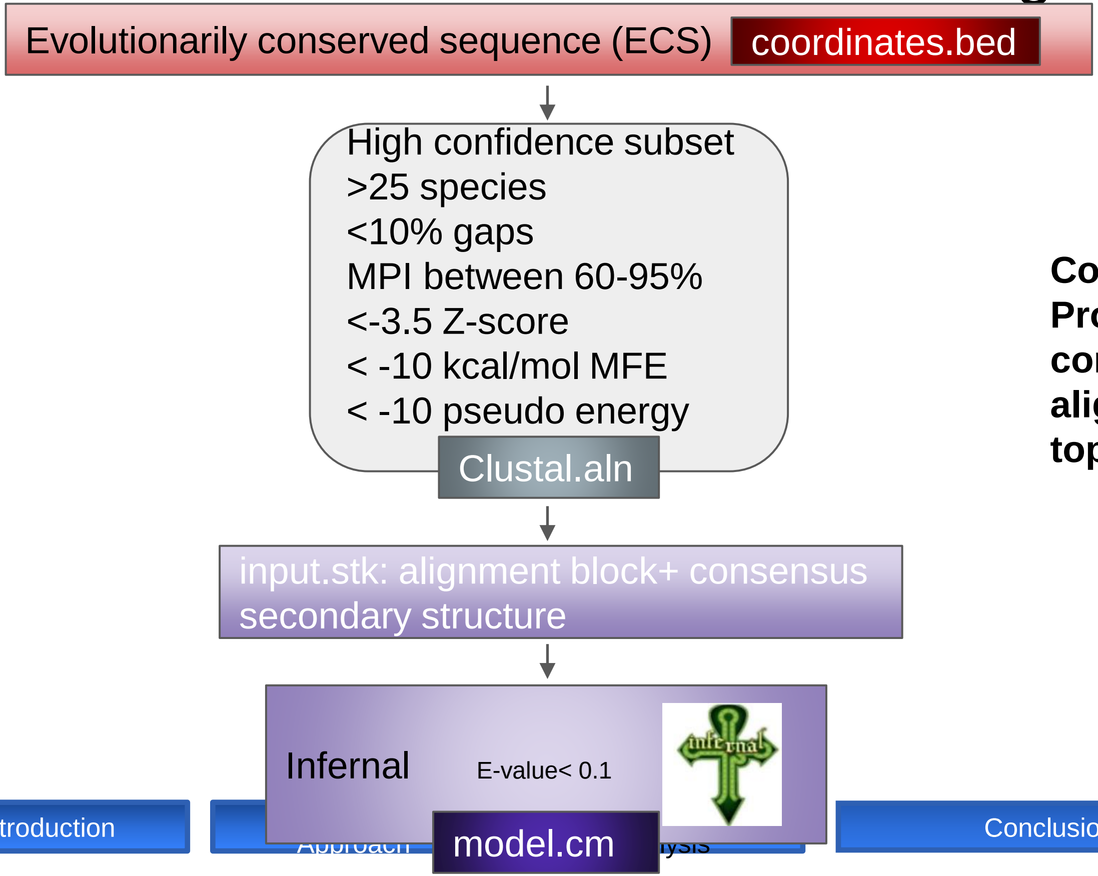


Wildtype



Mutant

Do these structures occur elsewhere in the genome ?



Evolutionarily conserved sequence (ECS) coordinates.bed

High confidence subset
>25 species
<10% gaps
MPI between 60-95%
<-3.5 Z-score
< -10 kcal/mol MFE
< -10 pseudo energy

Clustal.aln

input.stk: alignment block+ consensus secondary structure

Infernal E-value< 0.1

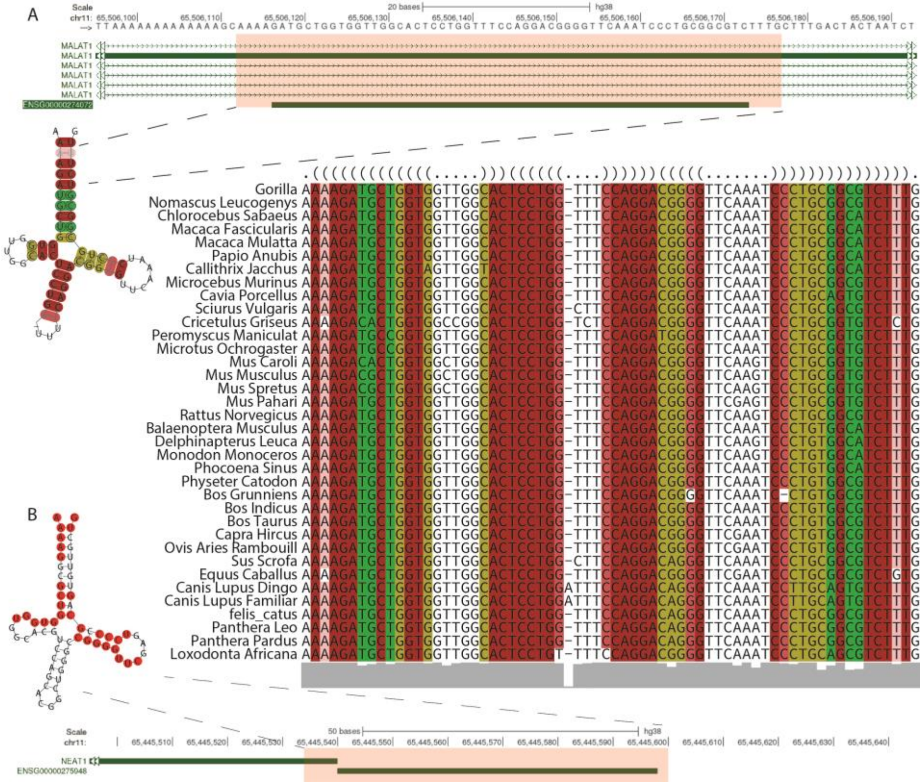
model.cm

Covariance models: Probabilistic models combining sequence alignment and structure topology

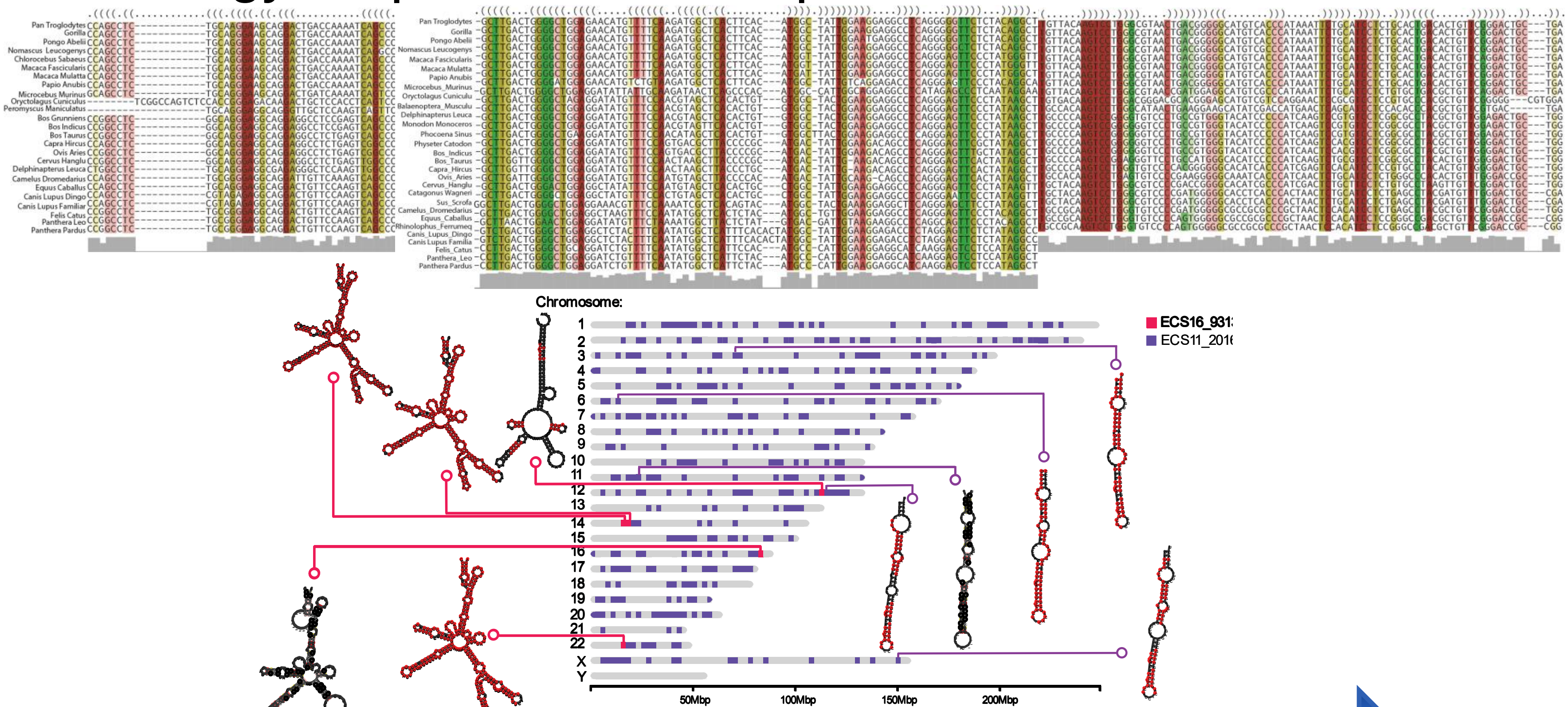
Identified 809,432 homologs from a subset of 23,818 ECS motifs

ECS

Homolog



Homology map from a non-repeat ECS model



Take home message

- ECSs are enriched in single nucleotide variants associated with various diseases and overlap over a thousand different splice sites associated with pathogenic diseases
- Some ECS have hundreds of homologs containing repetitive elements
- We can generate a network map of conserved structures and their homologs throughout the human genome

Acknowledgements

Fonds de recherche
Santé
Québec

Dr Martin Smith

Dr Bastien Paré

Mélanie Sagniez

Léa Kaufmann

Kristina Atanasova

Shawn Simpson

Yanis Bencheikh

Yuxin Zhou

Nicolas Roy

Jonathan Therrien