Features of functional human genes

Helena Cooper & Paul Gardner

August 12, 2022

Main aim

- **What is sufficient evidence to call a transcript a non-coding RNA?**
  - Important for annotating genomes.
  - What should appear in databases such as Rfam and RNAcentral?
  - Expanded scope to include long ncRNAs and protein-coding exons.

Rfam RNAcentral

Cooper & Gardner (2020) Features of Functional Human Genes. *bioRxiv.*

- There are some polarizing opinions on what evidence is required to say something is a ncRNA, e.g.
  - "a few RNAseq reads in a single experiment is sufficient" (causal effect)
  - "must be expressed, KO impacts phenotype, and evolutionarily conserved..." (selected effect)
- Shouldn't a functional ncRNA be distinguishable from junk DNA?

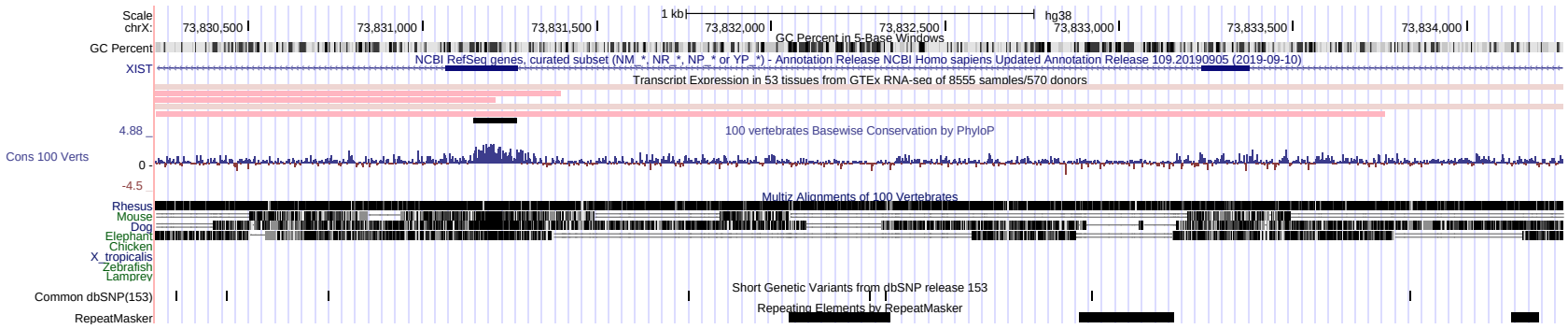
Vertebrate genomes & junk DNA

- Vary in length by an order of magnitude, e.g. bird genomes are $\approx$ 1Gb, while salamander genomes are $\approx$ 32Gb
  - Variation largely driven by decaying remnants of transposons
- $\approx$ 300Mb of sequence is conserved across the vertebrates
- Randomly generated sequences when inserted into genomes are also transcribed (and translated)
- Which suggests the number of functional elements should not scale with genome length

Ohno S (1972) So much'junk'DNA in our genome. In *Evolution of Genetic Systems, Brookhaven Symp. Biol.*

Our experiment

- Compare the strength of association between "known" **human** genes & control regions for a range of genomic features.
  - Positive controls: sampled 1,000 genes from each of the "ncRNA", and multiexonic "protein" and "lncRNA" HGNC classes
  - Negative controls: length-matched regions 20 Mb away to avoid linkage



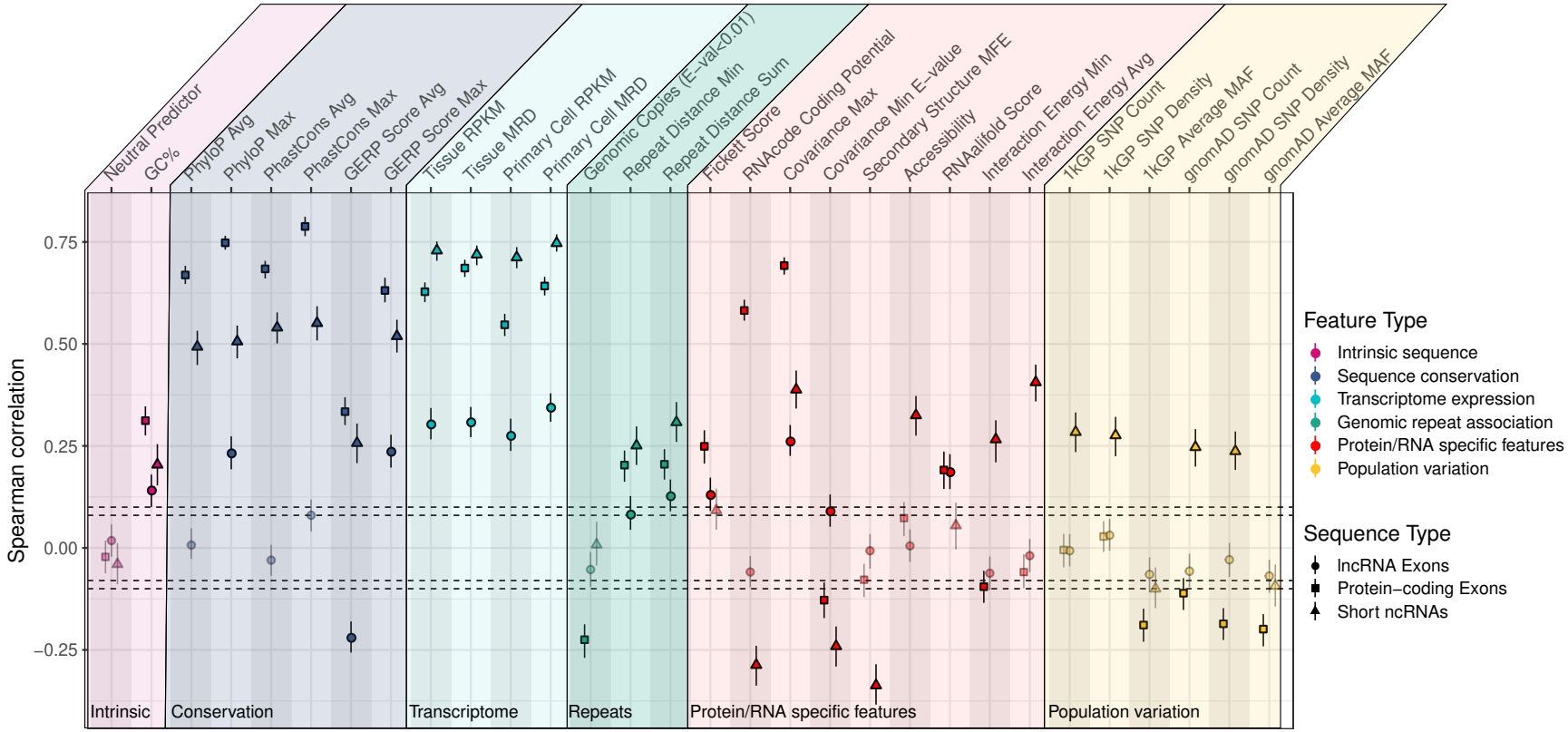
Selected genome features...

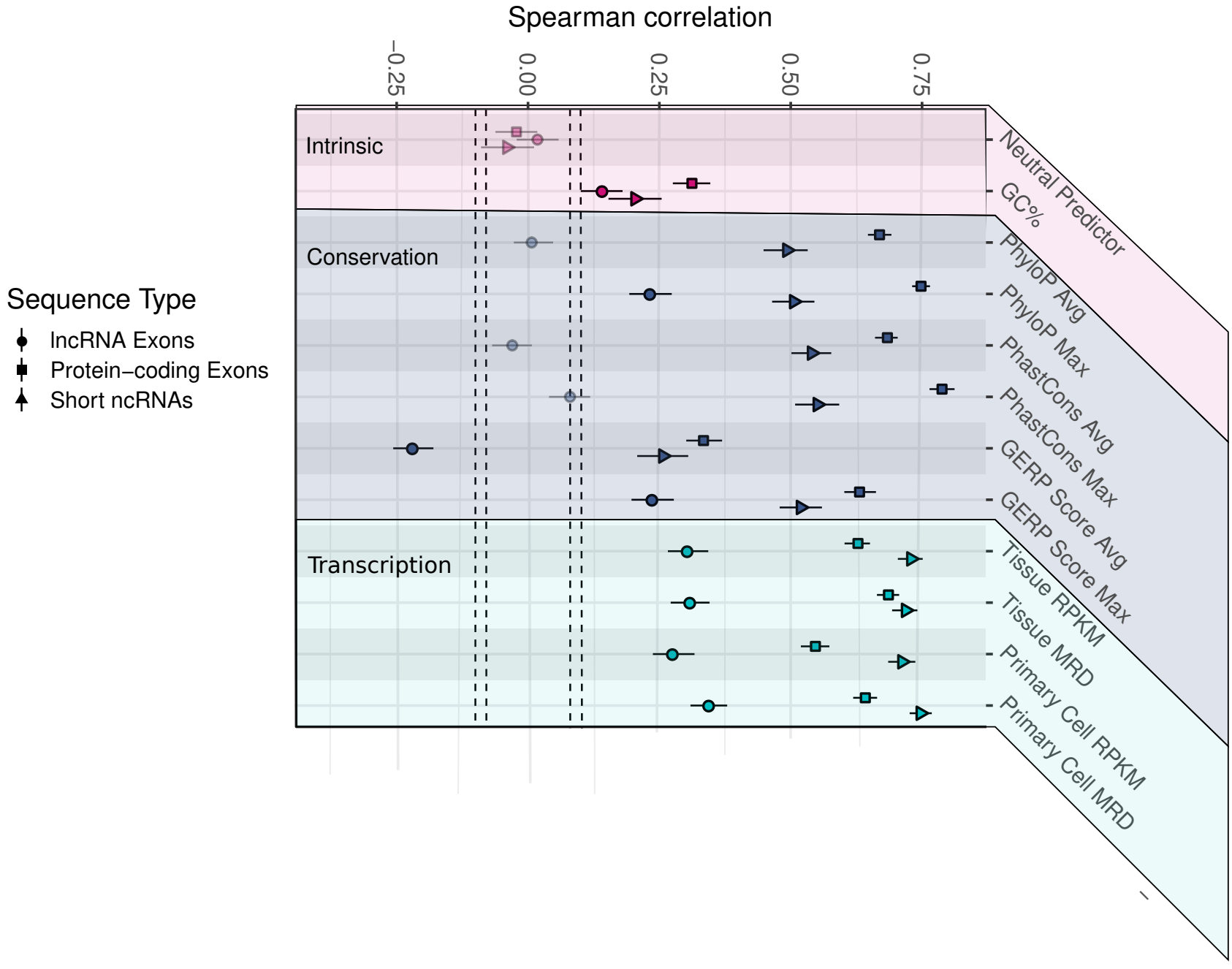
Inclusion criteria:

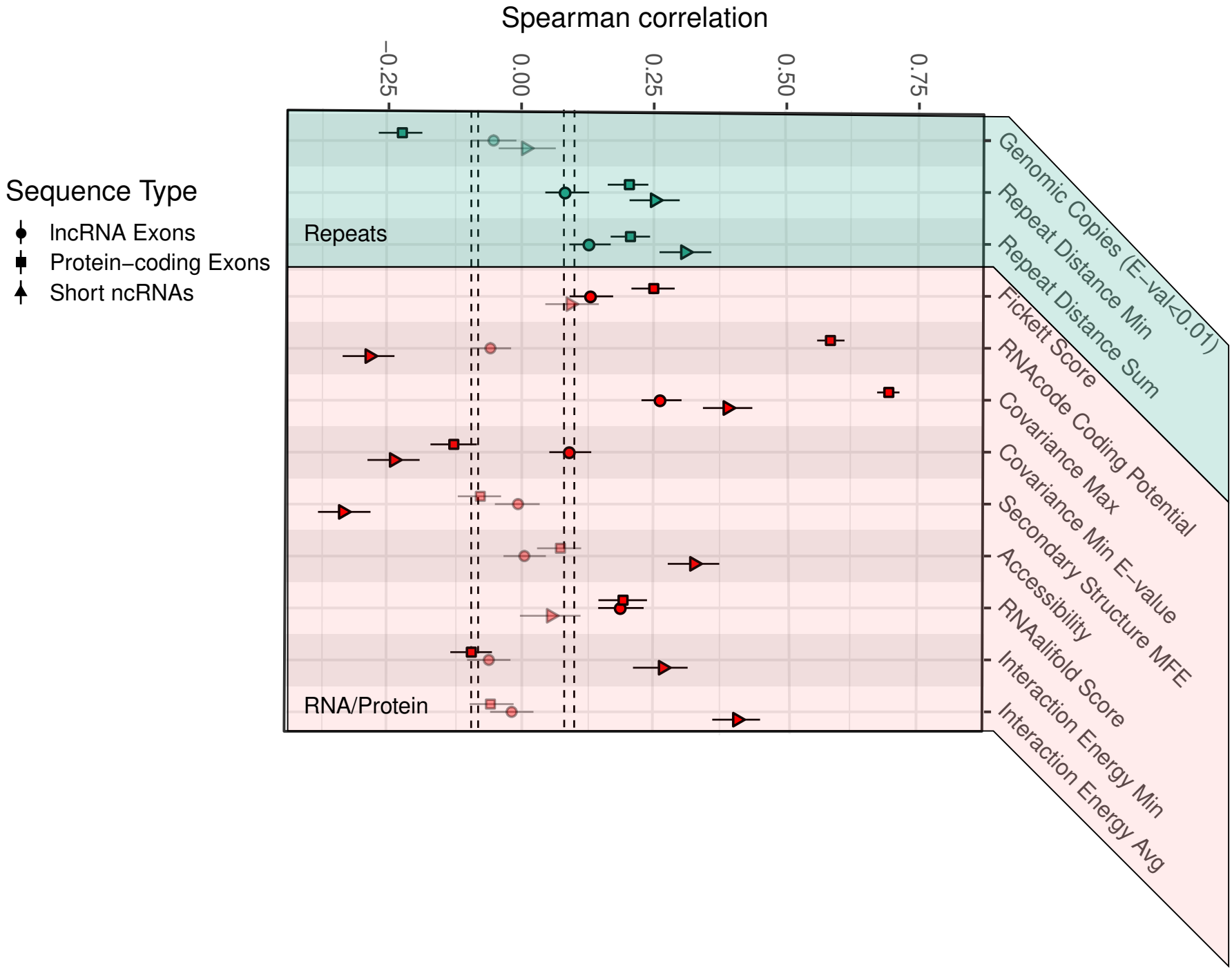
- Expected to relate to gene function
- Must be a genome-wide statistic
- Readily accessible for the GRCh38
- Non-redundant
- Selected:
  - Intrinsic features (G+C, start)
  - Conservation (PhastCons, PhyloP, GERP)
  - Population variation (1000g, gnomAD)
  - Transcription (ENCODE RNAseq)
  - Genome repeat (copy num., distance to Tn)
  - Protein/RNA specific features (coding, structure, interactions)

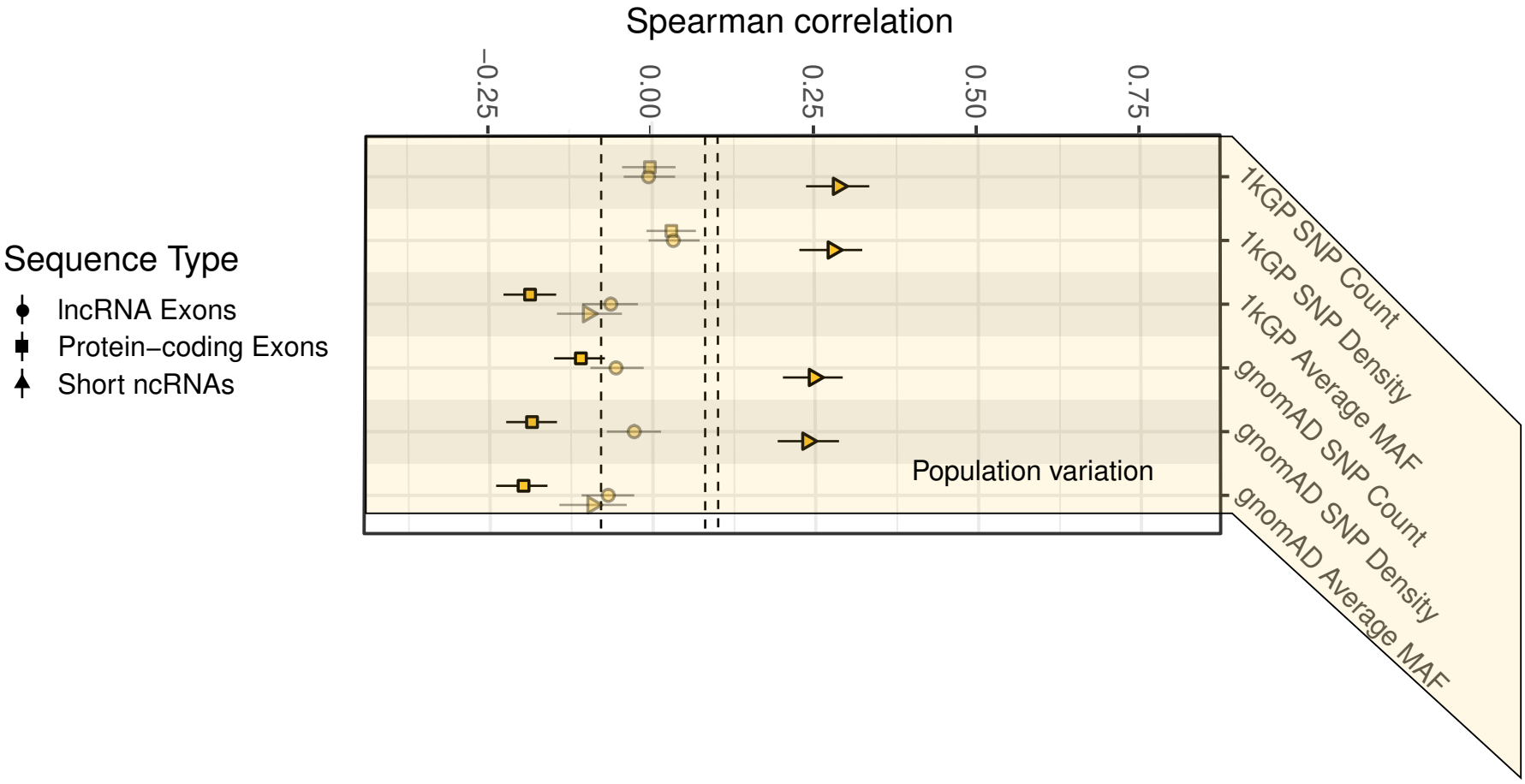
Feature Name (Figure)	Feature Name (CSV)
Intrinsic sequence	
GC%	GC_percentage
Neutral Predictor	Start
Sequence conservation	
PhastCons Max	MaxPhastCons
PhastCons Avg	MeanPhastCons
PhyloP Max	MaxPhyloP
PhyloP Avg	MeanPhyloP
GERP Score Max	mammals_max_gerp
GERP Scrore Avg	mammals_mean_gerp
Transcriptome Expression	
Tissue RPKM	RPKM_tissue
Tissue MRD	MRD_tissue
Primary Cell RPKM	RPKM_primary_cell
Primary Cell MRD	MRD_primary_cell
Genomic repeat association	
Genomic Copies (E-val<0.01)	Genome_copy_number
Repeat Distance Min	Dfam_min_distance
Repeat Distance Sum	Dfam_sum_distance
Protein/RNA specific features	
Protein-coding signals:	
Fickett Score	Fickett_score
RNAcode Coding Potential	RNAcode_score
RNA structure:	
Covariance Max	Max_covariance
Covariance Min E-value	Min_covariance_Eval
Secondary Structure MFE	MFE
Accessibility	Accessibility
RNAalifold Score	RNAalifold_score
RNA:RNA interactions:	
Interaction Energy Min	InteractionMIN
Interaction Energy Avg	InteractionAVE
Population variation	
1kGP SNP Count	1000G_SNPs
1kGP SNP Density	1000G_SNPsDensity
1kGP Average MAF	aveMAF
gnomAD SNP Count	gnomAD_SNP_count
gnomAD SNP Density	gnomAD_SNP_density
gnomAD Average MAF	gnomAD_avg_MAF

Feature correlation with function...

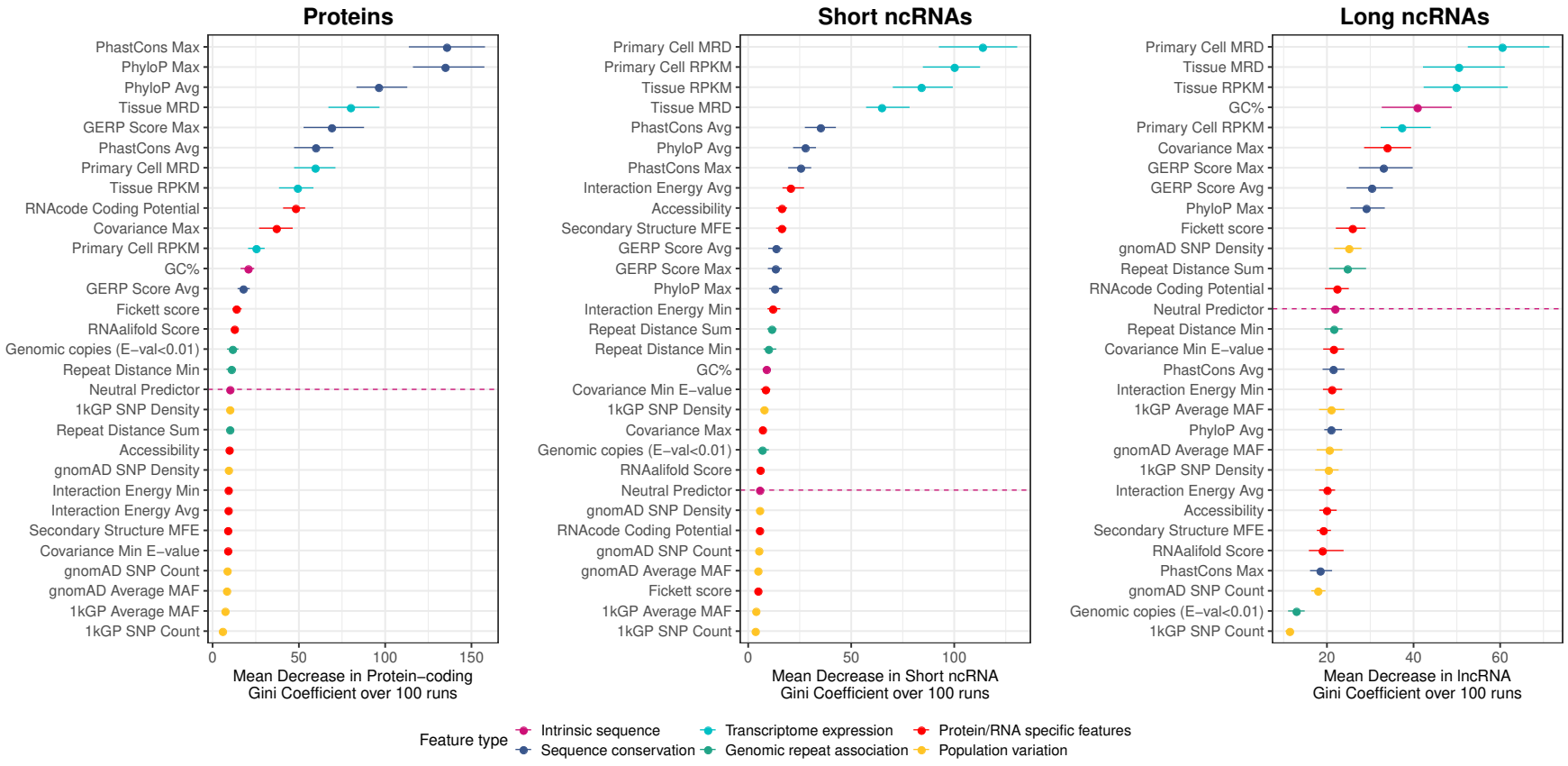



Spearman correlation
−0.25
0.00
0.25
0.50
0.75
Intrinsic
Conservation
Transcription
Neutral Predictor
GC%
PhyloP Avg
PhyloP Max
PhastCons Avg
PhastCons Max
GERP Score Avg
GERP Score Max
Tissue RPKM
Tissue MRD
Primary Cell RPKM
Primary Cell MRD
Sequence Type
lncRNA Exons
Protein−coding Exons
Short ncRNAs


Spearman correlation
−0.25
0.00
0.25
0.50
0.75
Sequence Type
lncRNA Exons
Protein−coding Exons
Short ncRNAs
Repeats
RNA/Protein
Genomic Copies (E−val<0.01)
Repeat Distance Min
Repeat Distance Sum
Fickett Score
RNAcode Coding Potential
Covariance Max
Covariance Min E−value
Secondary Structure MFE
Accessibility
RNAalifold Score
Interaction Energy Min
Interaction Energy Avg


Spearman correlation
−0.25
0.00
0.25
0.50
0.75
Sequence Type
lncRNA Exons
Protein−coding Exons
Short ncRNAs
Population variation
1kGP SNP Count
1kGP SNP Density
1kGP Average MAF
gnomAD SNP Count
gnomAD SNP Density
gnomAD Average MAF

Random forest result...



Conclusions

- Conservation **and** transcription is useful for identifying genes
- Covariation is surprisingly high in protein-coding alignments
- RNA structure and interactions important for short ncRNAs
- SNP data is not useful for determining function, MANY false positives in short ncRNAs
- It is difficult to distinguish many lncRNAs from neighbouring intergenic regions of the genome

BRIEF REPORT

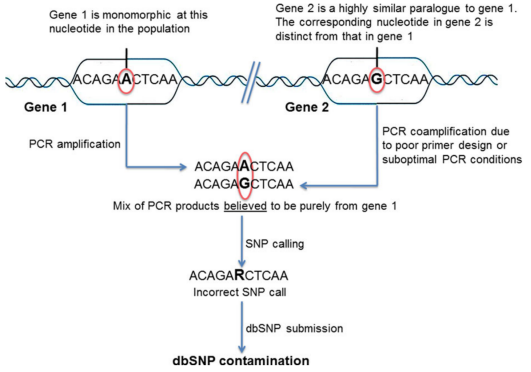
Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Single Nucleotide Differences (SNDs) Continue to Contaminate the dbSNP Database With Consequences for Human Genomics and Health

Jonathan W. Arthur,[1]* Florence S.G. Cheung,[2] and Juergen K.V. Reichardt[3]

[1]Children's Medical Research Institute, University of Sydney, Westmead, New South Wales, Australia; [2]Department of Microbiology, Immunology Programme Centre of Life Sciences, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [3]School of Pharmacy and Molecular Sciences, James Cook University, Townsville, Queensland, Australia



Reviewers...

R1: The authors ultimately conclude that evolutionary conservation and transcription should be "taken into consideration" when differentiating between functional sequences and noise: however, **this is a principle that biologists have long applied.**

R2: The study adds value to the current debate on the functionality of lncRNAs and makes a number of other interesting observations such as the covariation patterns in coding sequences or the excess of SNPs in small RNAs.

R3: ...we are far from knowing the full set of non-protein coding genes... The study is well designed and carefully executed. The manuscript is concise and clearly written...

R4: **I have major concerns about this manuscript.** While the title and abstract suggest that the authors seek to explore, challenge, and ultimately more precisely define notions of "functionality", no meaningful analysis along these lines is performed...

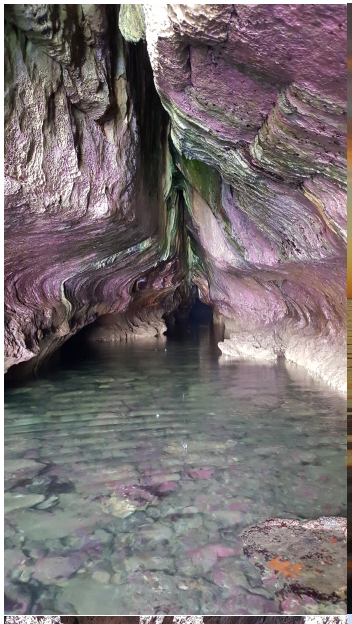
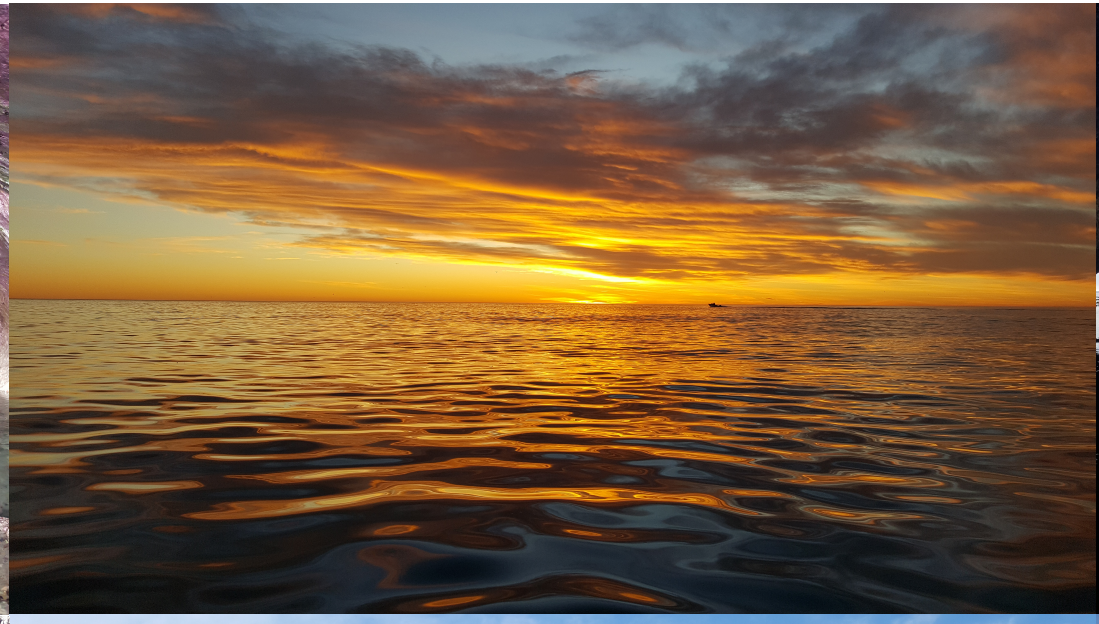
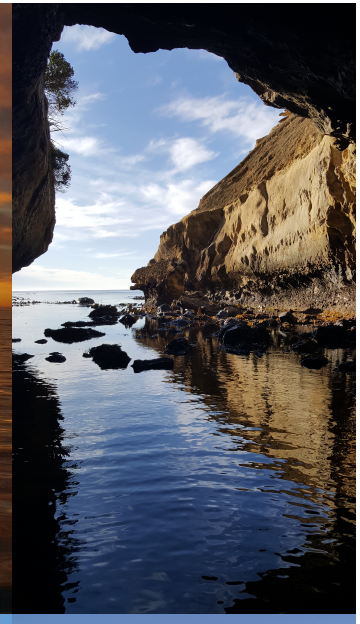
R5: The analysis is thorough and very nicely described. Such detailed description and comprehensive analysis ... is sure to be appreciated by many readers.

Does the bar need raising on the lncRNAs?

Table 1. Summary of 24 lncRNA annotation resources reviewed in this study

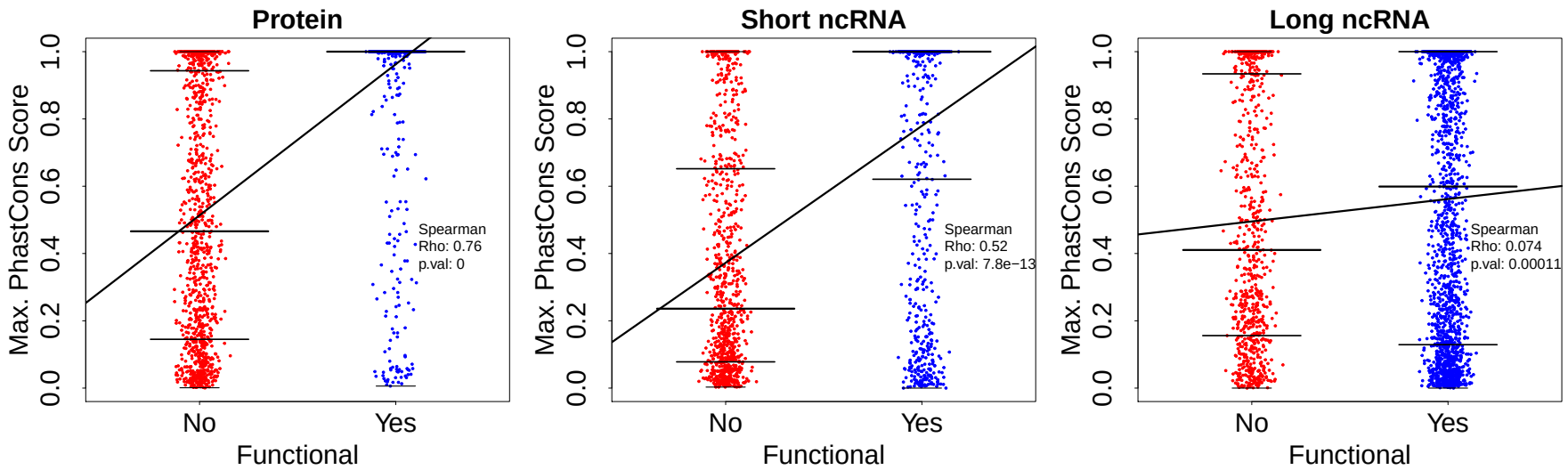
Source	Data type	Tissue/cell	Samples	lncRNA genes	Research scope	Method	Read type	Exon number	Length	Expression	Coding potential	Epigenetic signals	Ref.
CABILI	RNA-seq	24 tissues and cell types	24	8195	lincRNA	*ab initio* assembly	Paired end and single	≥2	>200 bp	≥3 reads per base	PhyloCSF<100, without hit in Pfam		[2]
KELLEY	RNA-seq	28 tissues and cell lines	70	9164	lincRNA	*ab initio* assembly	Paired end and single	≥2	>200 bp	≥1 FPKM	PhyloCSF<100		[17]
KRETZ	RNA-seq	keratinocytes	3	654	lincRNA	*ab initio* assembly	Paired end	≥2	>200 bp	>5 RPKM			[15]
DING	RNA-seq	Breast cancer tissues	25	344	lincRNA	*ab initio* assembly	Paired end			>10 read			[29]
KHALIL	ChIP-seq	6 cell lines	12	2510	lincRNA	*ab initio* assembly	Single		>5 Kb			H3K4me3 and H3K36me3	[25]
WHITE	RNA-seq	Lung cancer tissues	728	4067	lincRNA	*de novo* assembly	Paired end		>200 bp		GeneID		[16]
HE	RNA-seq	Prefrontal cortex	38	1888	lincRNA	*de novo* assembly	Single	≥2	>200 bp	>1 read	PhyloCSF < 100, ORF < 100, without hit in Pfam		[20]
HANGAUER	RNA-seq	23 tissues	127	3968	lincRNA	*de novo* assembly	Paired end	≥2	>200 bp	>1 RPKM	ORF < 100		[18]
IYER	RNA-seq	18 organs	7256	52 238	lncRNA	*ab initio* assembly	Paired end and single		>200 bp		Pfam/CPAT		[9]
TRIMARCHI	RNA-seq and ChIP-seq	T-ALL cell lines and primary leukemia samples	14	1984	lncRNA	*ab initio* assembly	Paired end	≥2	>200 bp	≥3 reads	PhyloCSF < 100	H3K4me3, H3K4me1 and H3K27ac	[14]
MORAN	RNA-seq and ChIP-seq	Islets and beta-cells	15	1128	lncRNA	*ab initio* assembly	Paired end		>200 bp	>0.5 RPKM	ORF < 130, without hit in Pfam	H3K4me3	[13]
SIGOVA1	RNA-seq and ChIP-seq	hESC	3	3983	lncRNA	*ab initio* assembly	Paired end		>100 bp	>0.07 FPKM	CPC < 0	H3K4me3	[26]
SIGOVA2	RNA-seq and ChIP-seq	Human endoderm cell	3	3544	lncRNA	*ab initio* assembly	Paired end		>100 bp	>0.07 FPKM	CPC < 0	H3K4me3	[26]
BELL	RNA-seq	Coronary artery smooth muscle cell	3	31	lncRNA	*ab initio* assembly	Single	≥2	>200 bp	>0.7 RPKM	PhyloCSF < 100, without hit in Pfam		[27]
YANG	RNA-seq	Failing LV samples	16	113	lncRNA	*ab initio* assembly	Paired end	≥2		>0.5 RPKM			[21]
NE	RNA-seq	Monocytes	8	2523	lncRNA	*ab initio* assembly	Paired end	≥2	>200 bp				[23]
PARALKAR	RNA-seq	Erythroblasts	15	594	lncRNA	*ab initio* assembly	Paired end	≥2	>200 bp	≥3 read	BlastX, HMMER, PhyloCSF, GetORF		[19]
SOWALSKY	RNA-seq	Castration-resistant prostate cancer (CRPC) tissues	8	2965	lncRNA	*ab initio* assembly	Paired end	≥2	>200 bp				[24]
YAN	RNA-Seq	Preimplantation embryos and hESCs	124	2121	lncRNA	*de novo* assembly	Single	≥2	>1 kb	>1 read	CPC < 0		[28]
NECSULEA1	RNA-Seq	8 organs	185	14 677	lncRNA	*de novo* assembly	Single	≥2	>200 bp	>10 reads			[22]
NECSULEA2	RNA-Seq	2 organs	53	12 667	lncRNA	*de novo* assembly	Single	≥2	>200 bp	>10 reads			[22]
GENCODE	Manually collected			13 869	lncRNA								[1]
LNCipedia	Intergrative database			17 385	lncRNA								[10]
NONCODE	Intergrative database			54 818	lncRNA								[11]

Xu et al. (2017) A comprehensive overview of lncRNA annotation resources. *Briefings in bioinformatics.*


Relating genome features to function...

- Spearman correlation coefficients
- Random forest feature importance



SNPs in ncRNAs...