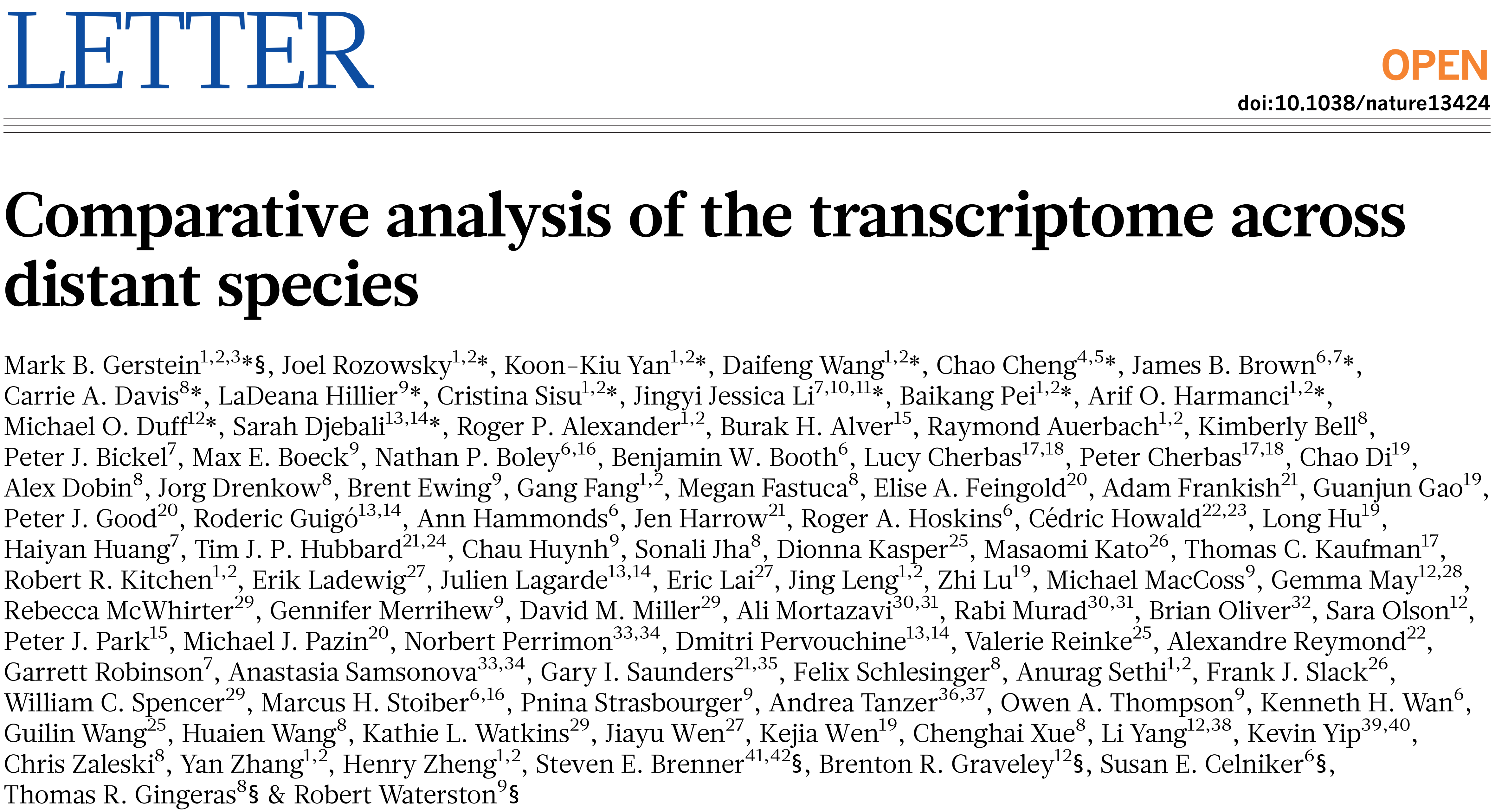


LETTER

OPEN

doi:10.1038/nature13424

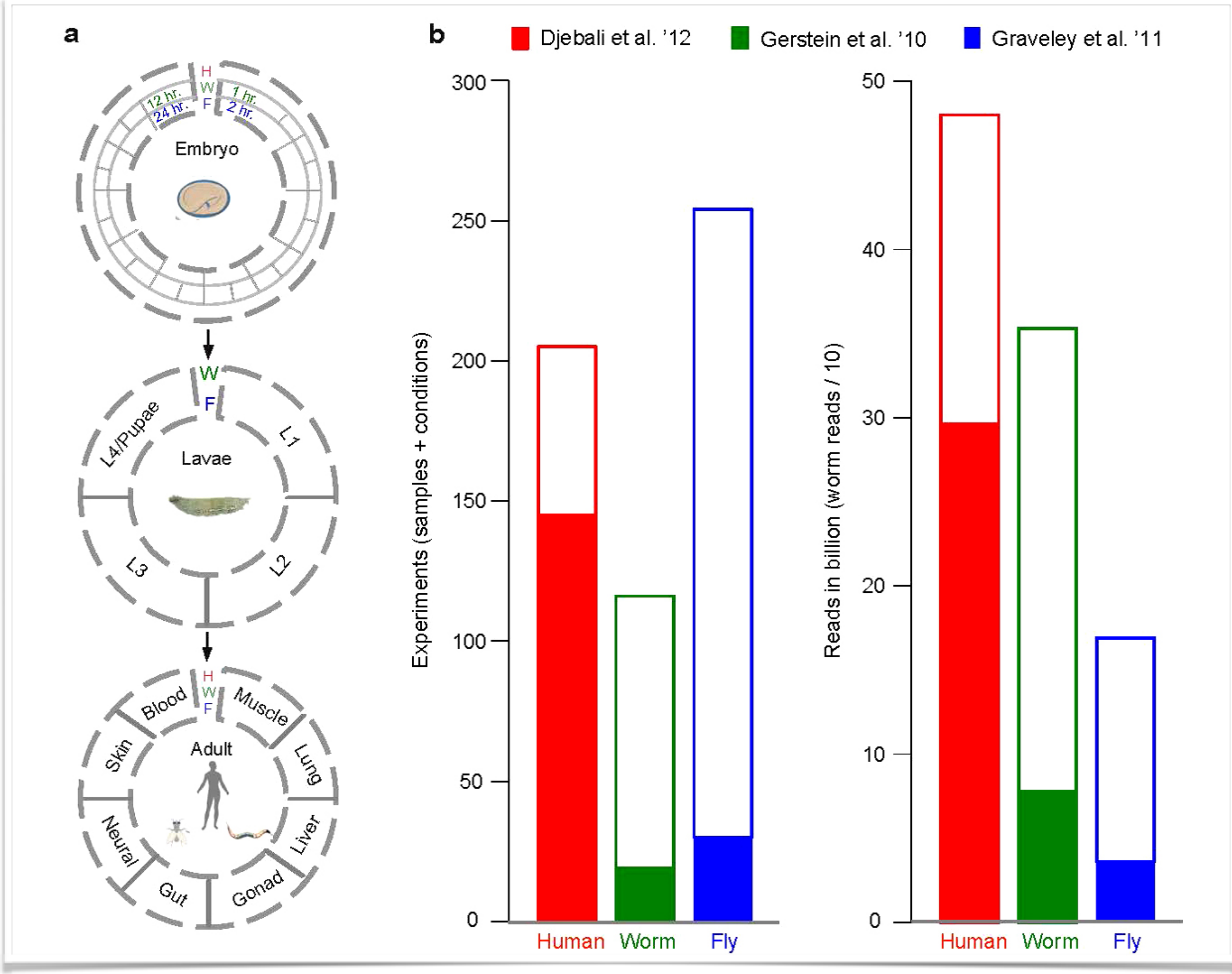
Comparative analysis of the transcriptome across distant species

Mark B. Gerstein[1,2,3]*§, Joel Rozowsky[1,2]*, Koon-Kiu Yan[1,2]*, Daifeng Wang[1,2]*, Chao Cheng[4,5]*, James B. Brown[6,7]*, Carrie A. Davis[8]*, LaDeana Hillier[9]*, Cristina Sisu[1,2]*, Jingyi Jessica Li[7,10,11]*, Baikang Pei[1,2]*, Arif O. Harmanci[1,2]*, Michael O. Duff[12]*, Sarah Djebali[13,14]*, Roger P. Alexander[1,2], Burak H. Alver[15], Raymond Auerbach[1,2], Kimberly Bell[8], Peter J. Bickel[7], Max E. Boeck[9], Nathan P. Boley[6,16], Benjamin W. Booth[6], Lucy Cherbas[17,18], Peter Cherbas[17,18], Chao Di[19], Alex Dobin[8], Jorg Drenkow[8], Brent Ewing[9], Gang Fang[1,2], Megan Fastuca[8], Elise A. Feingold[20], Adam Frankish[21], Guanjun Gao[19], Peter J. Good[20], Roderic Guigó[13,14], Ann Hammonds[6], Jen Harrow[21], Roger A. Hoskins[6], Cédric Howald[22,23], Long Hu[19], Haiyan Huang[7], Tim J. P. Hubbard[21,24], Chau Huynh[9], Sonali Jha[8], Dionna Kasper[25], Masaomi Kato[26], Thomas C. Kaufman[17], Robert R. Kitchen[1,2], Erik Ladewig[27], Julien Lagarde[13,14], Eric Lai[27], Jing Leng[1,2], Zhi Lu[19], Michael MacCoss[9], Gemma May[12,28], Rebecca McWhirter[29], Gennifer Merrihew[9], David M. Miller[29], Ali Mortazavi[30,31], Rabi Murad[30,31], Brian Oliver[32], Sara Olson[12], Peter J. Park[15], Michael J. Pazin[20], Norbert Perrimon[33,34], Dmitri Pervouchine[13,14], Valerie Reinke[25], Alexandre Reymond[22], Garrett Robinson[7], Anastasia Samsonova[33,34], Gary I. Saunders[21,35], Felix Schlesinger[8], Anurag Sethi[1,2], Frank J. Slack[26], William C. Spencer[29], Marcus H. Stoiber[6,16], Pnina Strasbourger[9], Andrea Tanzer[36,37], Owen A. Thompson[9], Kenneth H. Wan[6], Guilin Wang[25], Huaien Wang[8], Kathie L. Watkins[29], Jiayu Wen[27], Kejia Wen[19], Chenghai Xue[8], Li Yang[12,38], Kevin Yip[39,40], Chris Zaleski[8], Yan Zhang[1,2], Henry Zheng[1,2], Steven E. Brenner[41,42]§, Brenton R. Graveley[12]§, Susan E. Celniker[6]§, Thomas R. Gingeras[8]§ & Robert Waterston[9]§

Cristina Sisu
Gerstein Lab
11th Jan 2016

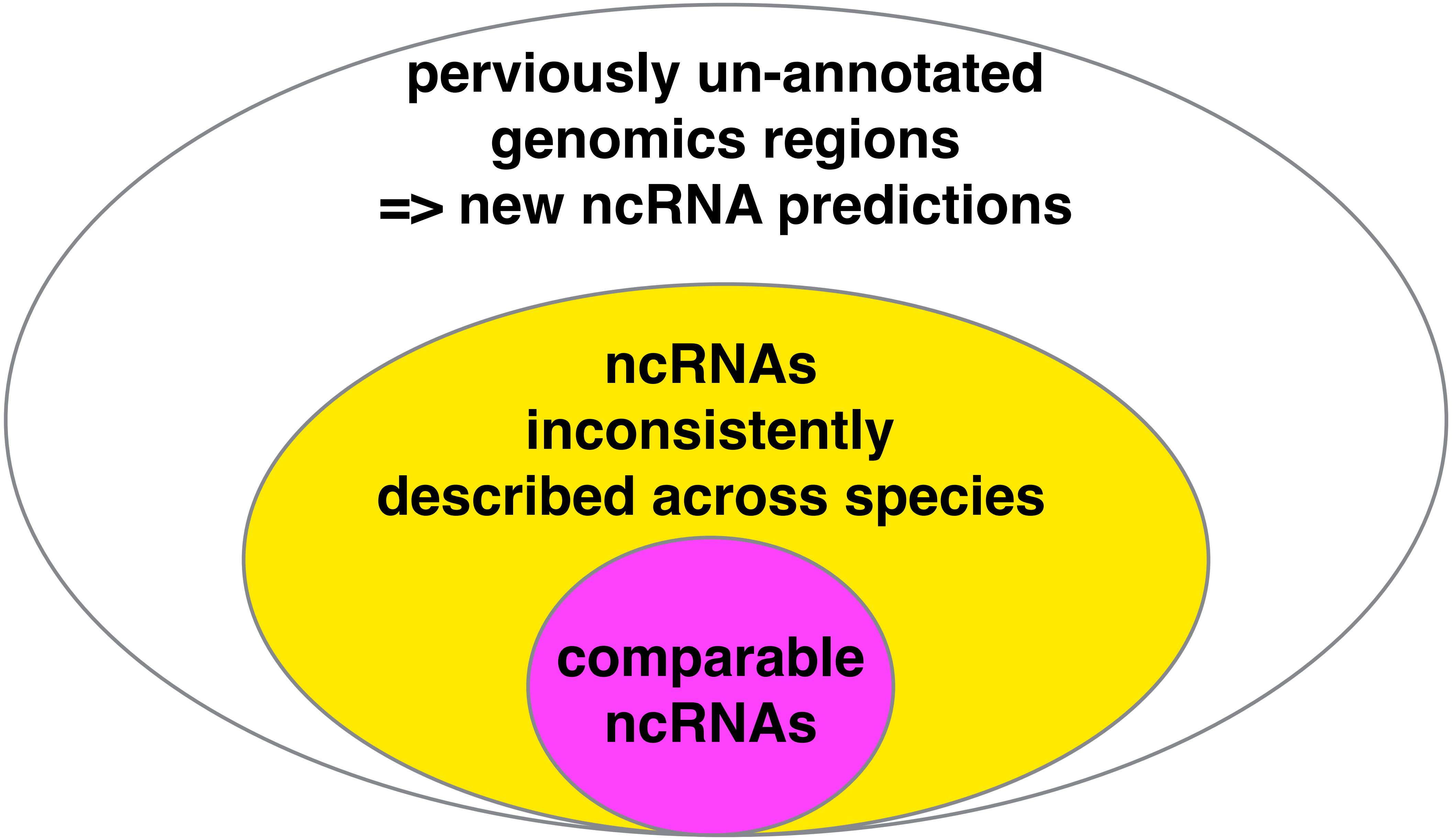
ENCODE/modENCODE RNA Resource

- 575 experiments
  - 380 New
- >65 billion reads
  - >50% New
  - Illumina GAIIx
  - HiSeq2000
- Experimental conditions
  - 116 worm
  - 254 fly
  - 41 cell lines human



ncRNA annotation

“only a subset of ncRNAs, are annotated consistently across organisms”



perviously un-annotated genomics regions
=> new ncRNA predictions

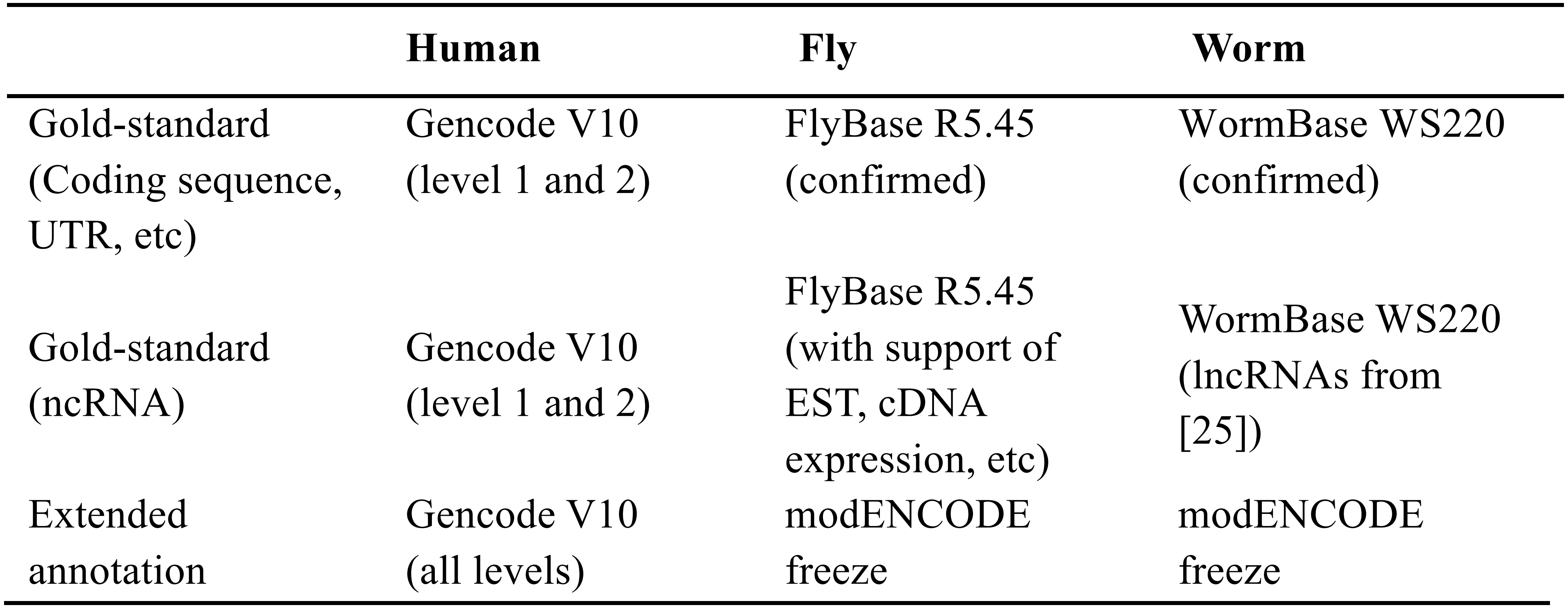
ncRNAs inconsistently described across species

comparable ncRNAs

comparable RNAs

5 biotypes for **gold-standard** annotations :

- **CDS**, **UTR**, **canonical ncRNAs** (include miRNA, tRNA, rRNA, snRNA, and snoRNA), **lncRNA**, and **ancestral repeats** for human & **unexpressed intergenic** regions for worm and fly



	Human	**Fly**	**Worm**
Gold-standard (Coding sequence, UTR, etc)	Gencode V10 (level 1 and 2)	FlyBase R5.45 (confirmed)	WormBase WS220 (confirmed)
Gold-standard (ncRNA)	Gencode V10 (level 1 and 2)	FlyBase R5.45 (with support of EST, cDNA expression, etc)	WormBase WS220 (lncRNAs from [25])
Extended annotation	Gencode V10 (all levels)	modENCODE freeze	modENCODE freeze

Annotation & refinement

- **lncRNA** subdivision:
  - lncRNA overlapping with > 1bp, know gold standard biotypes were reclassified with the respective labels
  - antisense lncRNA — have >50% sequence overlapping known CDS on the opposite strand
  - intronic lncRNA — fully embedded in protein coding gene introns on the same stand
  - ambiguous ncRNA — overlap know biotypes but do not fulfil criteria for reclassification

- **miRNA**:
  - pre-miRNA known annotations : 1,756 in human, 221 in worm and 235 fly.
  - pre-miRNA hairpins from miRBase v18
  - mirtrons : *Genome Res* 22, 1634-45, (2012)
    - pri-miRNA collected from RefSeq & literature

- **short ncRNA**:
  - extracted from the know annotations from GENCODE, FlyBase and WormBase

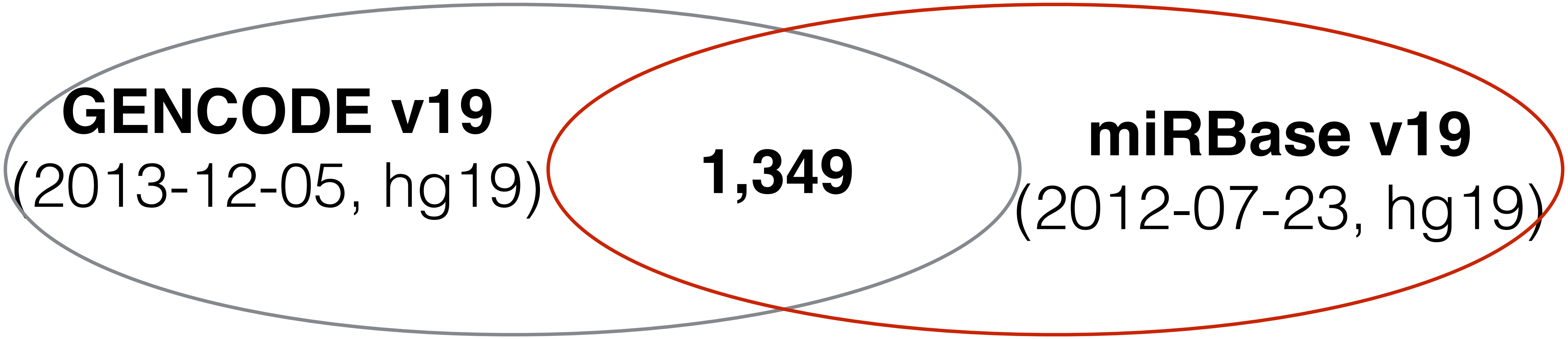
miRNA annotation in GENCODE

	V10	V19	V24
V10	1,756	1,558	1,474
V19		3,055	2,765
V24			4,093

- Common miRNA annotation in different GENCODE versions are determined by their transcript id’s.

- V10 is the version used in comparative genomics paper, while v19 is the latest version for hg19, and v24 is the current version for hg38.

GENCODE vs miRBase



GENCODE v19
(2013-12-05, hg19)

1,349

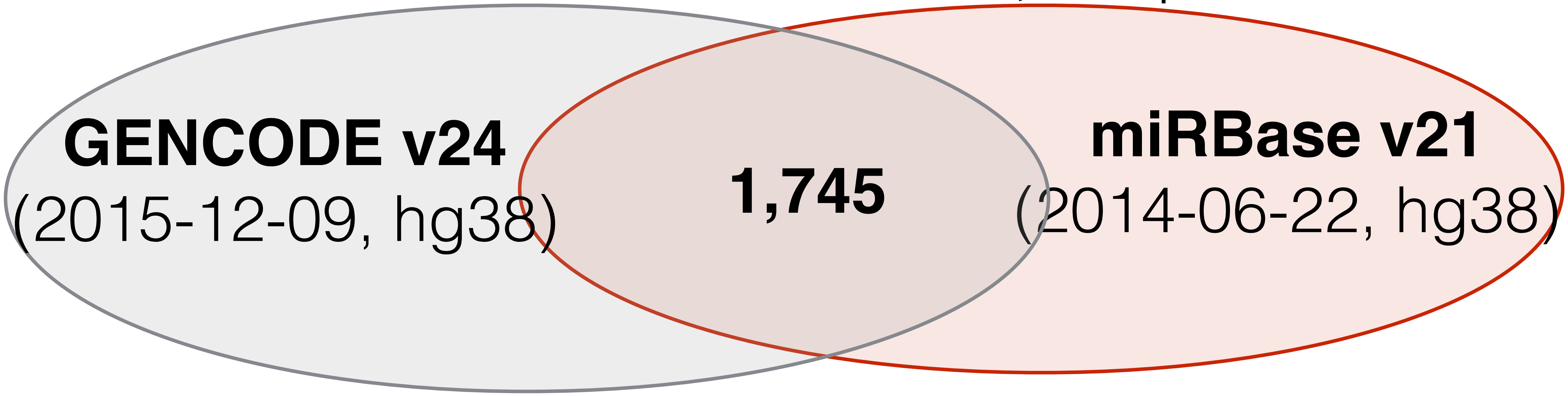
miRBase v19
(2012-07-23, hg19)

3,055 pre-miRNAs

1,595 pre-miRNAs

4,093 pre-miRNAs

1,881 pre-miRNAs



GENCODE v24
(2015-12-09, hg38)

1,745

miRBase v21
(2014-06-22, hg38)

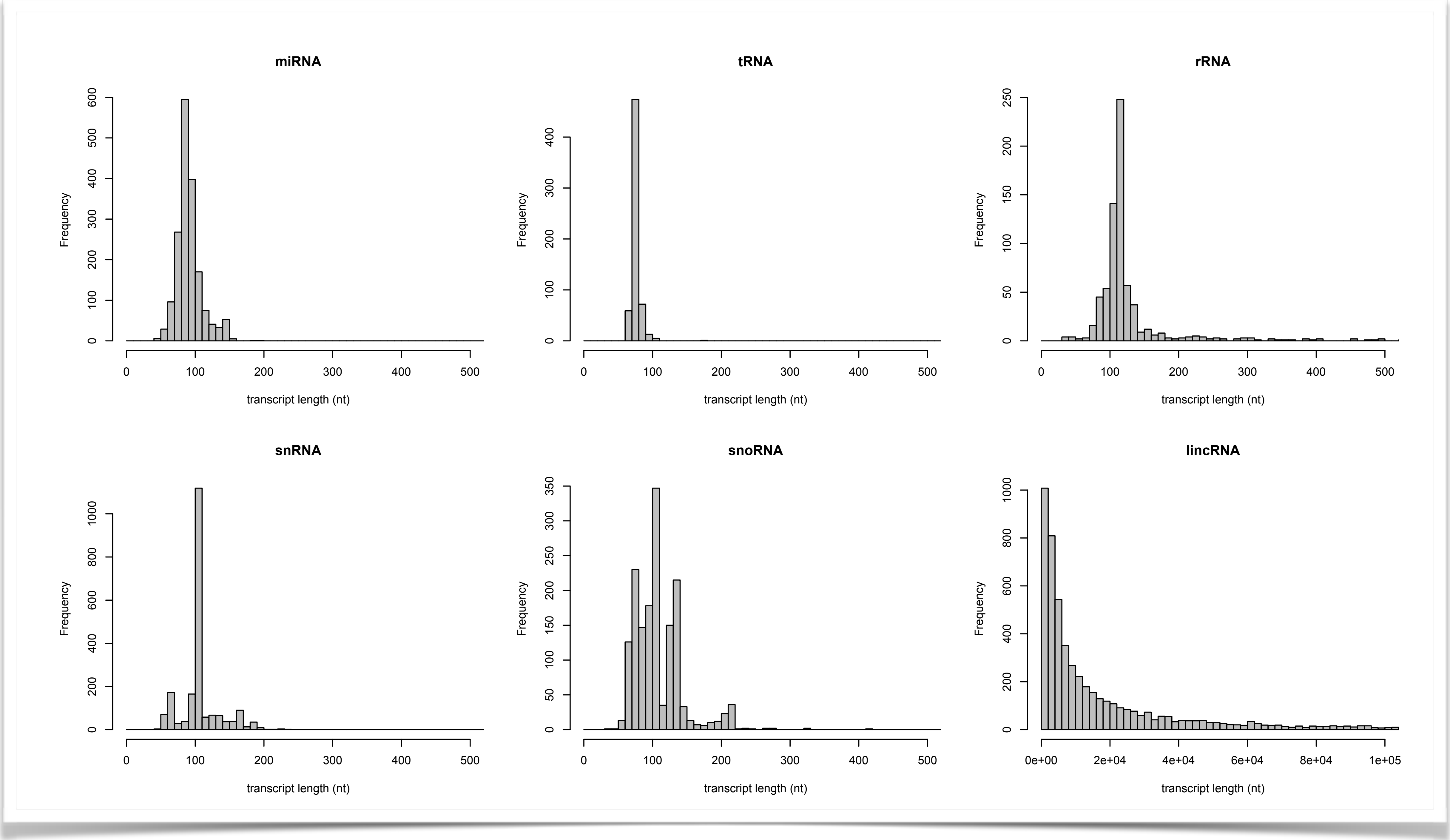
non-comparable ncRNA

- ribosomal RNA
    - inconsistency of cross species annotations

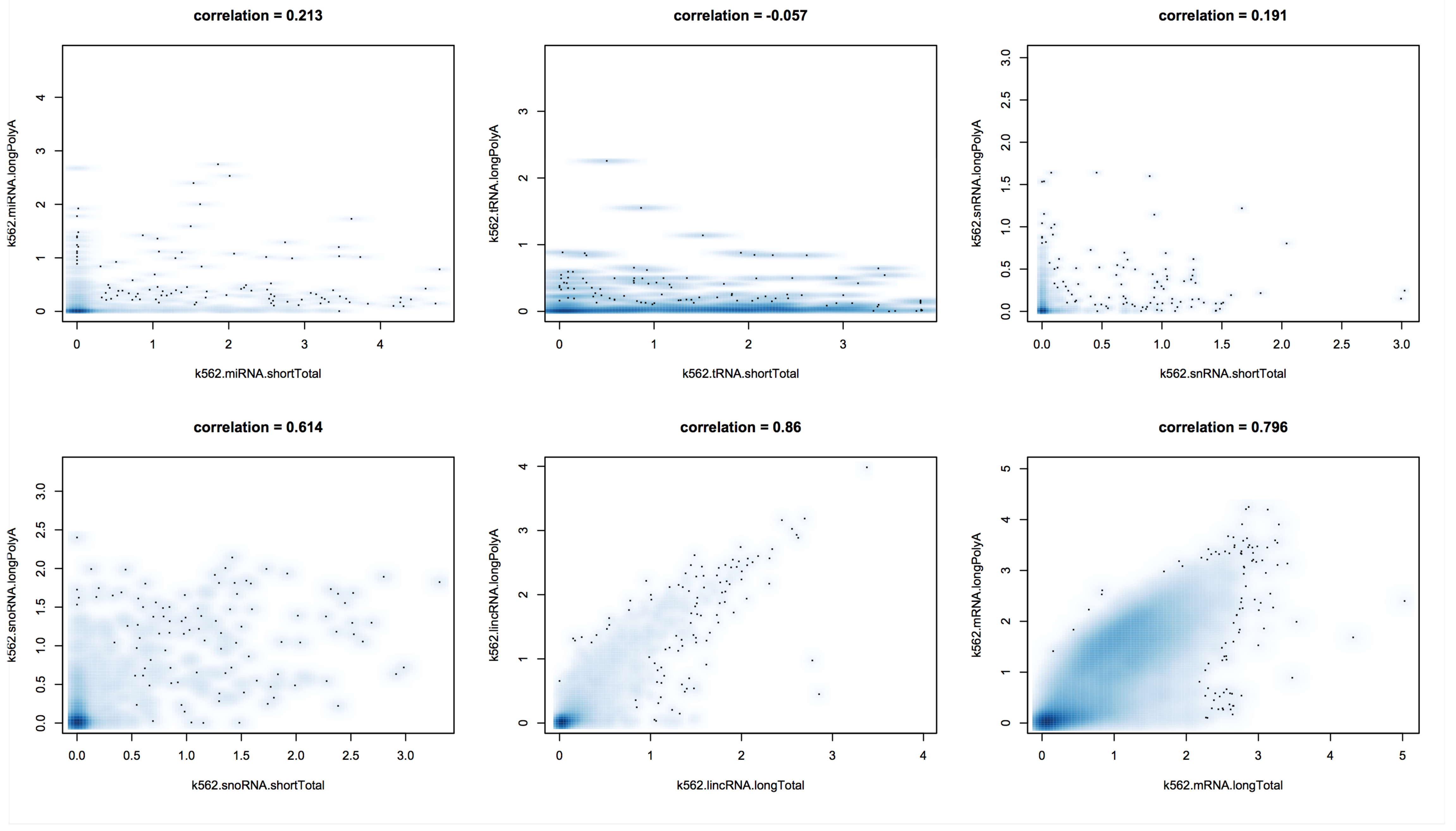
- piwi-interacting RNA
    - cross species annotation using human piRNA clusters
    - refinement of previous annotation with respect to ~100M human testis small RNA reads
    - 88 human loci, 27 fly loci, and 35329 worm loci as piRNA clusters

- others: mitochondrial RNAs, piRNAs, rRNAs, Y RNAs, and misc_RNA

Length distribution in coding and non coding annotated human RNAs

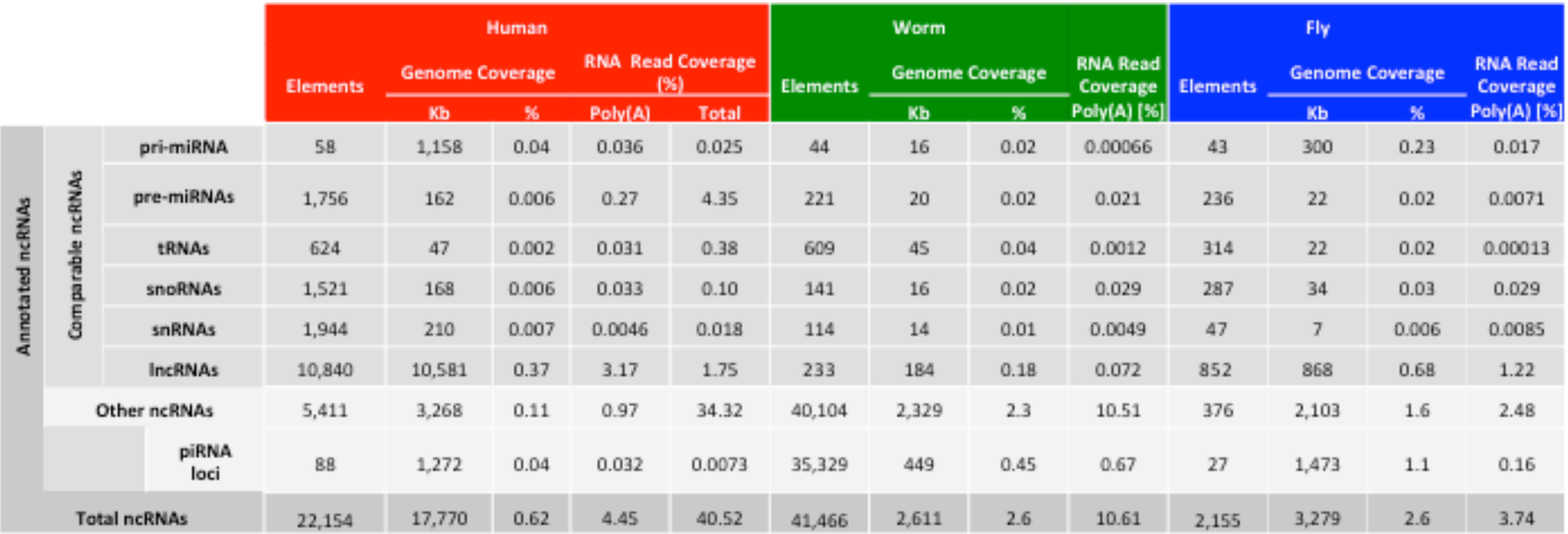


Effect of poly(A) RNA purification, during sample-prep, on ability to detect coding and non-coding RNAs



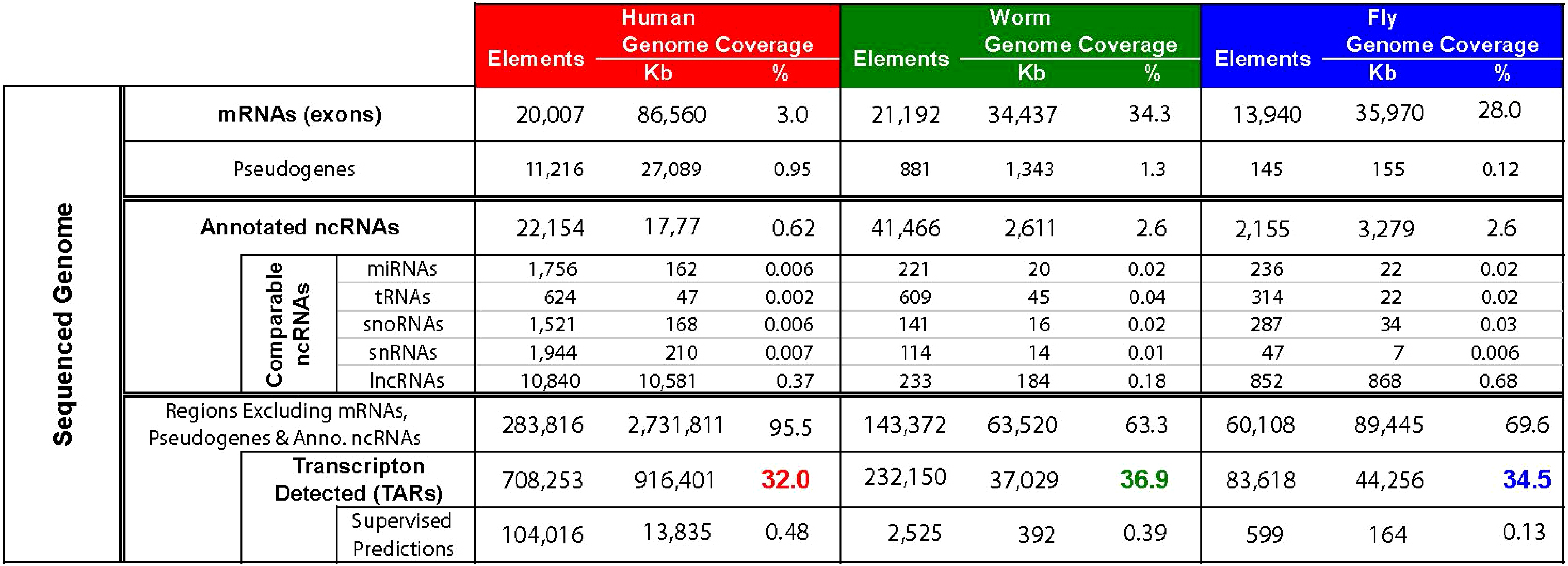
Poly(A) RNA-seq data perform poorly compared to short-total RNA-seq at detecting miRNAs, tRNAs, snRNAs and snoRNAs, but are much better able to detect lincRNAs and mRNAs.

Summary of Annotated ncRNAs



			Human					Worm				Fly			
			Elements	Genome Coverage		RNA Read Coverage (%)		Elements	Genome Coverage		RNA Read Coverage	Elements	Genome Coverage		RNA Read Coverage
				Kb	%	Poly(A)	Total		Kb	%	Poly(A) [%]		Kb	%	Poly(A) [%]
Annotated ncRNAs	Comparable ncRNAs	pri-miRNA	58	1,158	0.04	0.036	0.025	44	16	0.02	0.00066	43	300	0.23	0.017
		pre-miRNAs	1,756	162	0.006	0.27	4.35	221	20	0.02	0.021	236	22	0.02	0.0071
		tRNAs	624	47	0.002	0.031	0.38	609	45	0.04	0.0012	314	22	0.02	0.00013
		snoRNAs	1,521	168	0.006	0.033	0.10	141	16	0.02	0.029	287	34	0.03	0.029
		snRNAs	1,944	210	0.007	0.0046	0.018	114	14	0.01	0.0049	47	7	0.006	0.0085
		lncRNAs	10,840	10,581	0.37	3.17	1.75	233	184	0.18	0.072	852	868	0.68	1.22
	Other ncRNAs		5,411	3,268	0.11	0.97	34.32	40,104	2,329	2.3	10.51	376	2,103	1.6	2.48
		piRNA loci	88	1,272	0.04	0.032	0.0073	35,329	449	0.45	0.67	27	1,473	1.1	0.16
Total ncRNAs			22,154	17,770	0.62	4.45	40.52	41,466	2,611	2.6	10.61	2,155	3,279	2.6	3.74

Summary of Annotation + Novel Transcription



			Human			Worm			Fly		
			Elements	Genome Coverage		Elements	Genome Coverage		Elements	Genome Coverage	
				Kb	%		Kb	%		Kb	%
Sequenced Genome	**mRNAs (exons)**		20,007	86,560	3.0	21,192	34,437	34.3	13,940	35,970	28.0
	Pseudogenes		11,216	27,089	0.95	881	1,343	1.3	145	155	0.12
	Annotated ncRNAs		22,154	17,77	0.62	41,466	2,611	2.6	2,155	3,279	2.6
	Comparable ncRNAs	miRNAs	1,756	162	0.006	221	20	0.02	236	22	0.02
		tRNAs	624	47	0.002	609	45	0.04	314	22	0.02
		snoRNAs	1,521	168	0.006	141	16	0.02	287	34	0.03
		snRNAs	1,944	210	0.007	114	14	0.01	47	7	0.006
		lncRNAs	10,840	10,581	0.37	233	184	0.18	852	868	0.68
	Regions Excluding mRNAs, Pseudogenes & Anno. ncRNAs		283,816	2,731,811	95.5	143,372	63,520	63.3	60,108	89,445	69.6
	Transcripton Detected (TARs)		708,253	916,401	**32.0**	232,150	37,029	**36.9**	83,618	44,256	**34.5**
		Supervised Predictions	104,016	13,835	0.48	2,525	392	0.39	599	164	0.13