Comparative analysis of read processing patterns across 11 total RNA-seq datasets.

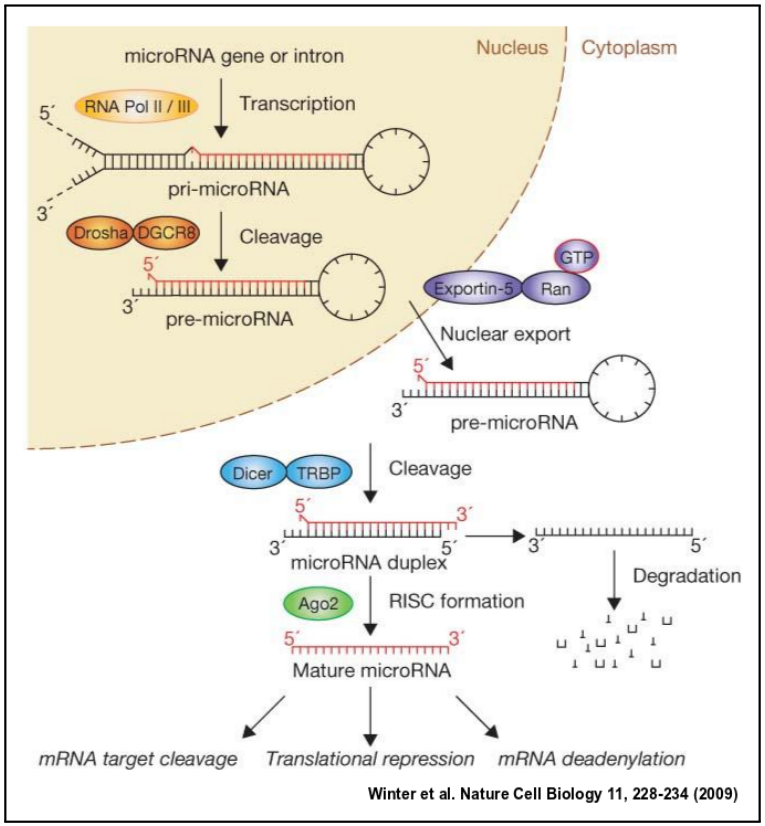
Sachin Pundhir

Center for Non-Coding RNA in Technology and Health,
LIFE, University of Copenhagen, Denmark

February 15, 2012

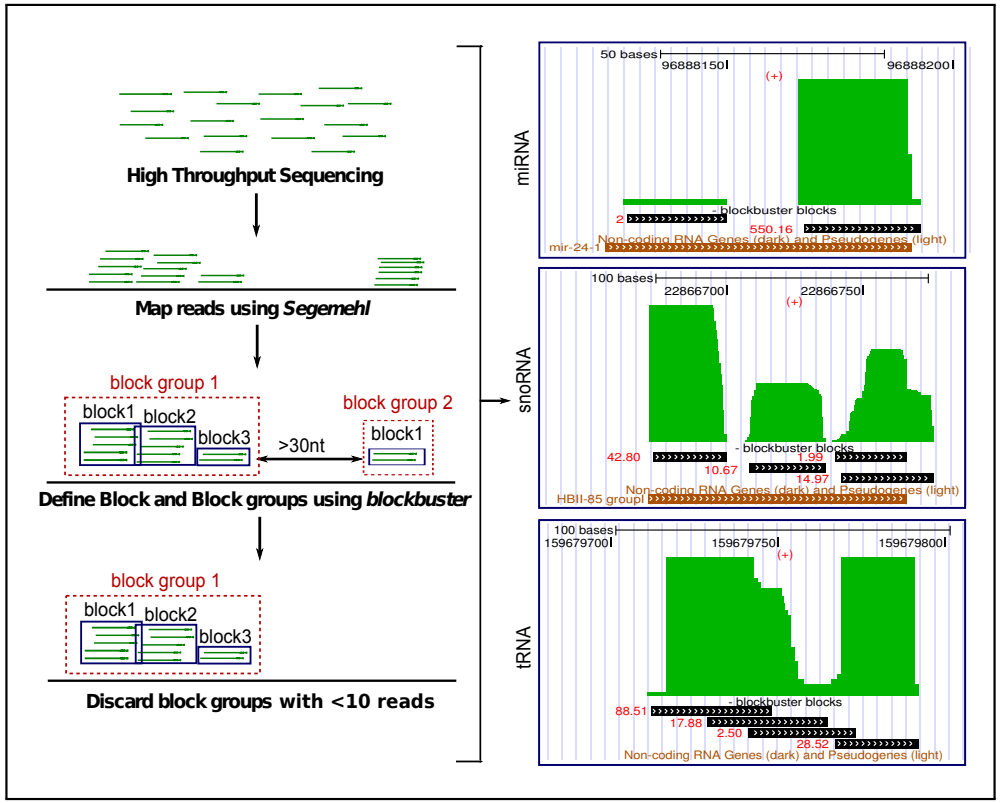
Read Processing Pattern?

- Post-transcriptional processing of transcripts generate short-RNA sequence fragments.
- When mapped to the genome, they form distict patterns termed as 'Read Processing Patterns' eg. miR-miR* pattern from miRNA.



Read Processing Pattern: significance

- Conveys information about the structure of parent transcript and modality of processing.
- Study can lead us to identify commonality and diversity in read processing machanisms.



Eleven total RNA-seq Datasets

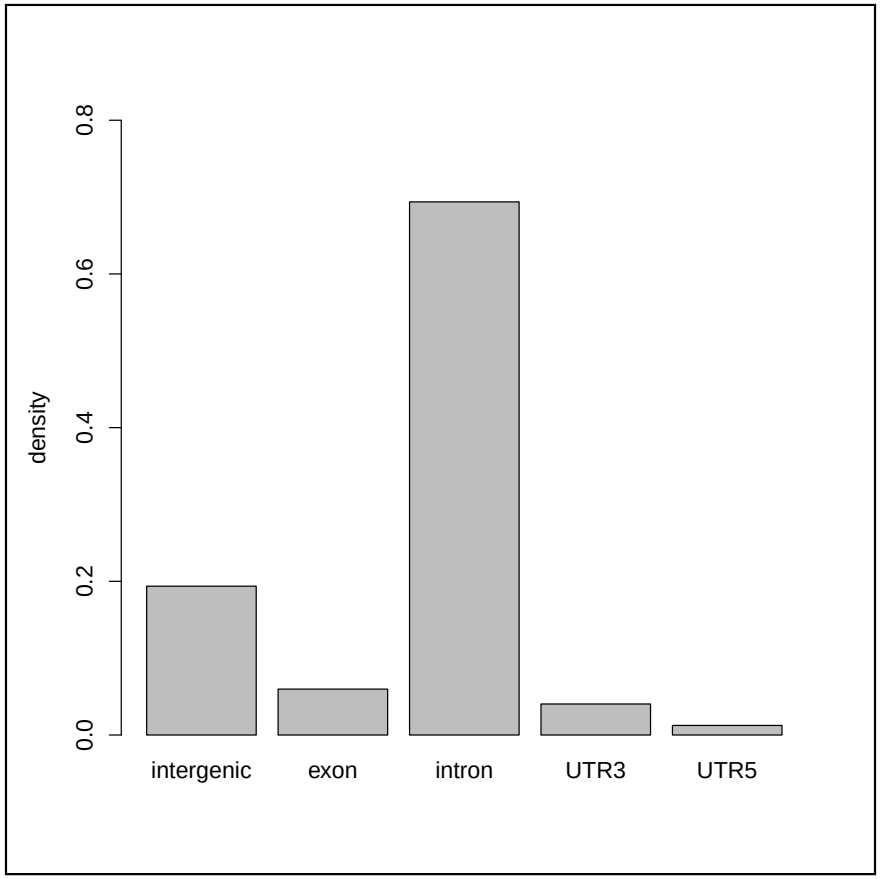
- prepared using ribosomal RNA-free total-RNA from 11 different human tissues.
- reads mapped to human genome (hg19) using *segemehl*.
- divided closely mapped reads (< 30 nt) into blocks and block groups using *blockbuster*.

Source	**SRA Id**	**# reads**[a]	**# block groups**		**read length (nt)**
			all	annotated[b]	
adipose	ERR015534	27,263,226	141,316	819	35
heart	ERR015536	26,650,067	131,035	481	50
kidney	ERR015538	26,883,350	215,779	616	50
lung	ERR015540	22,607,226	189,959	597	50
skeletal muscle	ERR015542	48,689,233	212,384	646	50
testes	ERR015544	48,158,986	387,580	684	50
colon	ERR015535	41,011,318	258,233	637	50
hypothalamus	ERR015537	30,936,056	145,375	892	35
liver	ERR015539	28,095,610	92,543	747	35
ovary	ERR015541	55,530,131	588,225	848	50
spleen	ERR015543	43,537,618	434,967	743	50

[a]reads obtained after quality filter, [b]block groups overlapping with annotated ncRNAs

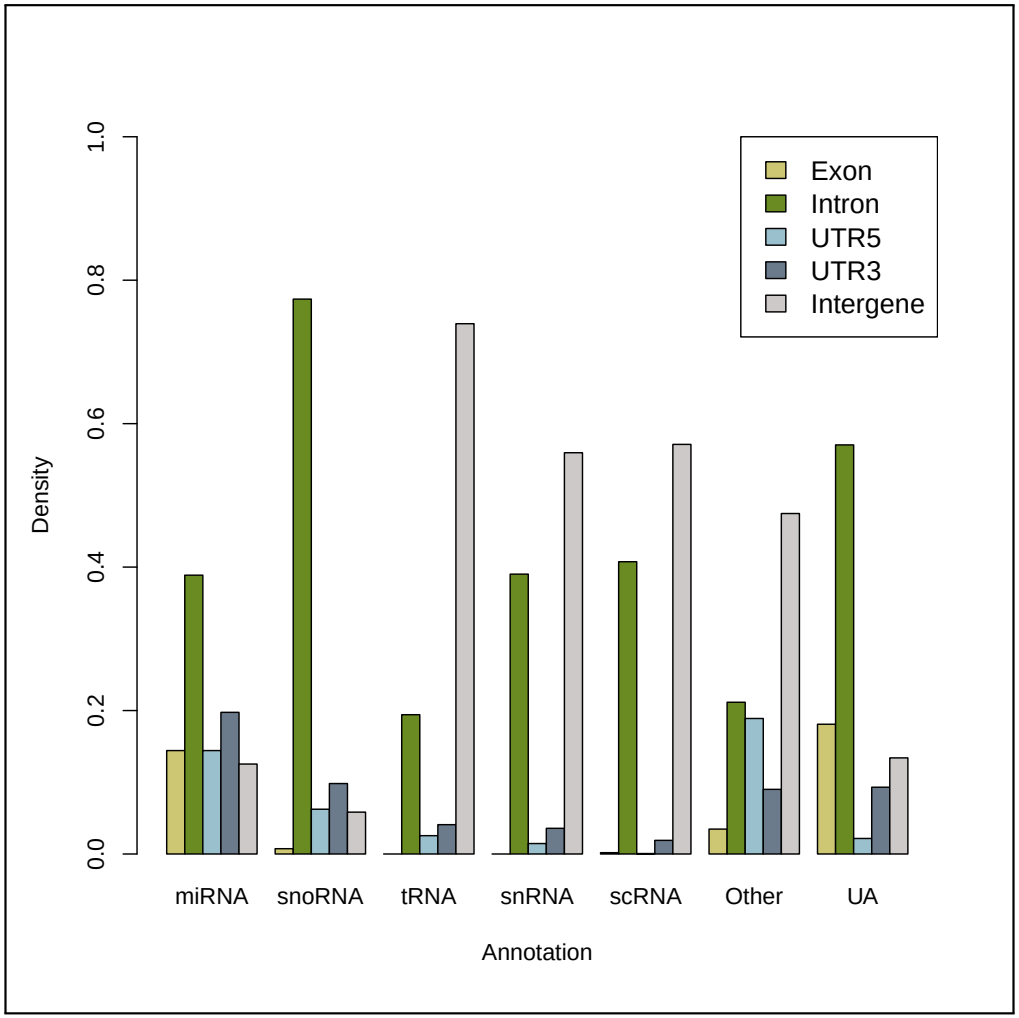
Density distribution of unique block groups from ovary, spleen and testes at various loci

- most unique block groups from ovary, spleen and testes map to introns.



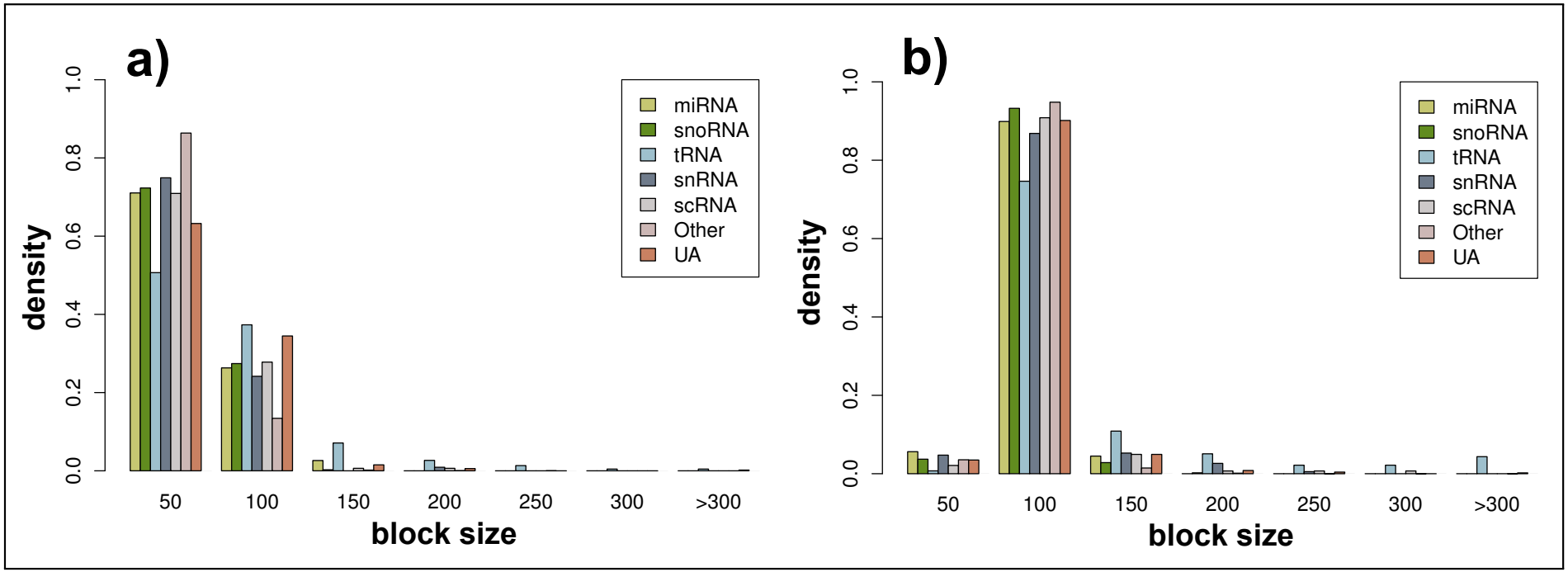
Distinct distribution of ncRNAs at various genomic loci

- miRNA and snoRNAs are preferentially expressed from introns.
- snRNA, scRNA and tRNA are transcribed mostly in intergenic regions.



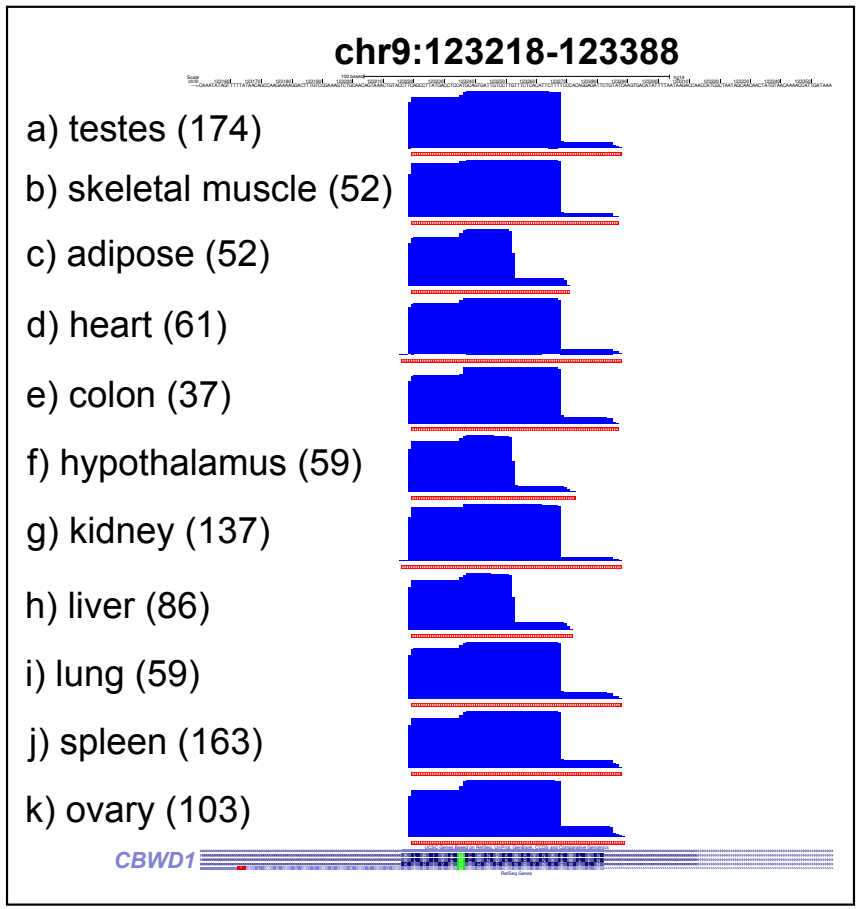
Factors effecting read processing patterns

- sequencing protocol and local sequence content (GC%) as suggested in an earlier study.
- read length:
  - read length (35 nt) - size of read processing patterns (50 nt).
  - read length (50 nt) - size of read processing patterns (100 nt).
- absolute expression of a transcript may also effect the read processing patterns.



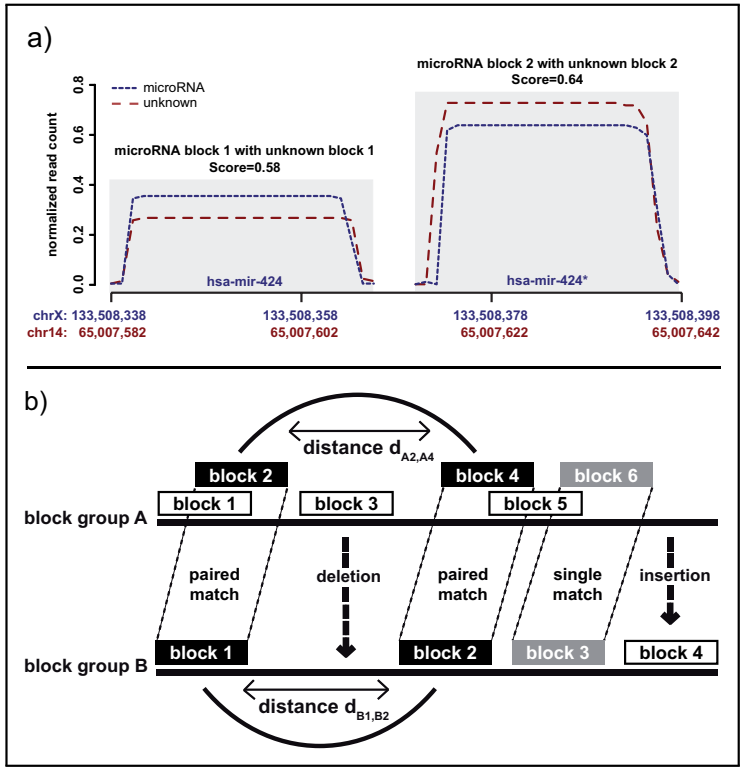
Read processing patterns are not sequencing artifact

- same sequencing protocol and comparison of similar loci across multiple experiments.
- no effect of read length and expression on read processing pattern.
- read processing patterns are representative of the underlying mechanism in which the host RNA-transcript is processed by the cellular machinery and not of a sequencing artifact.



Read Processing Pattern: alignment and comparison

- In an earlier study, we developed a tool named deepBlockAlign (http://rth.dk/dba) for the alignment of two read processing patterns.
- deepBlockAlign gives a score between 0 and 1, suggesting the similarity between the two processing patterns.



a) Block alignment

- Each position i in block $\vec{X}$ is represented as the normalized difference between total reads and start reads, such that $x_i = (x_{1i} - x_{2i})/N_X$.
- Next, a Needleman-Wunsch like algorithm is employed to determine optimal alignment between two block profiles, $\vec{X}$ and $\vec{Y}$.
- In other words, we compute the block scores based on the comparison of *block size*, *read count* and *entropy*.

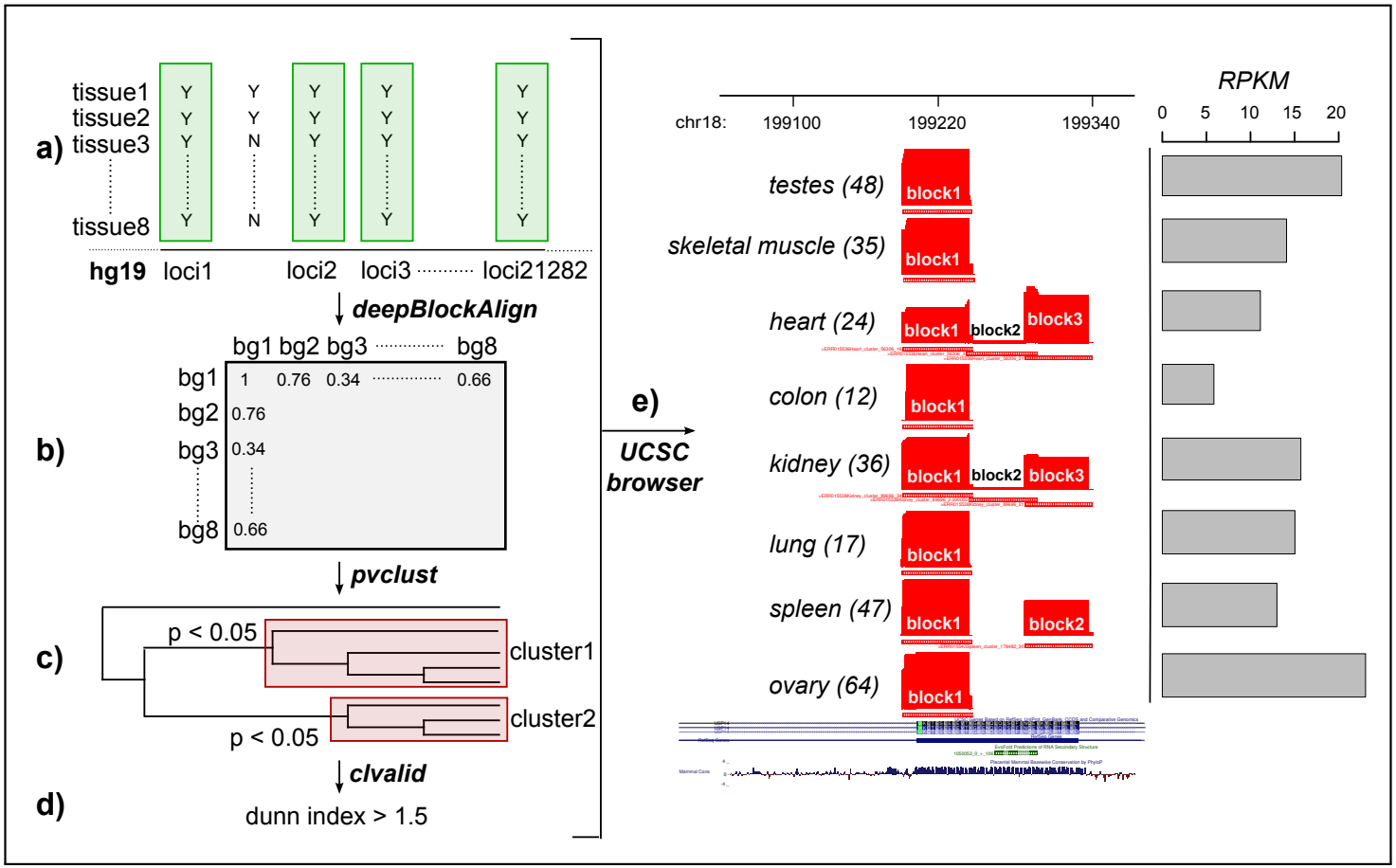
b) Block group alignment

- A variant of the Sankoff (1985) algorithm is used to compute optimal alignment between block groups.
- Here, a similarity measure is used that combines the *block score* and *block distance*.

Langenberger D, Pundhir S, Ekstrm C, Stadler P, Hoffmann S, Gorodkin J. (2012) deepBlockAlign: A tool for aligning RNA-seq profiles of read block patterns. Bioinformatics 28(1):17-24.

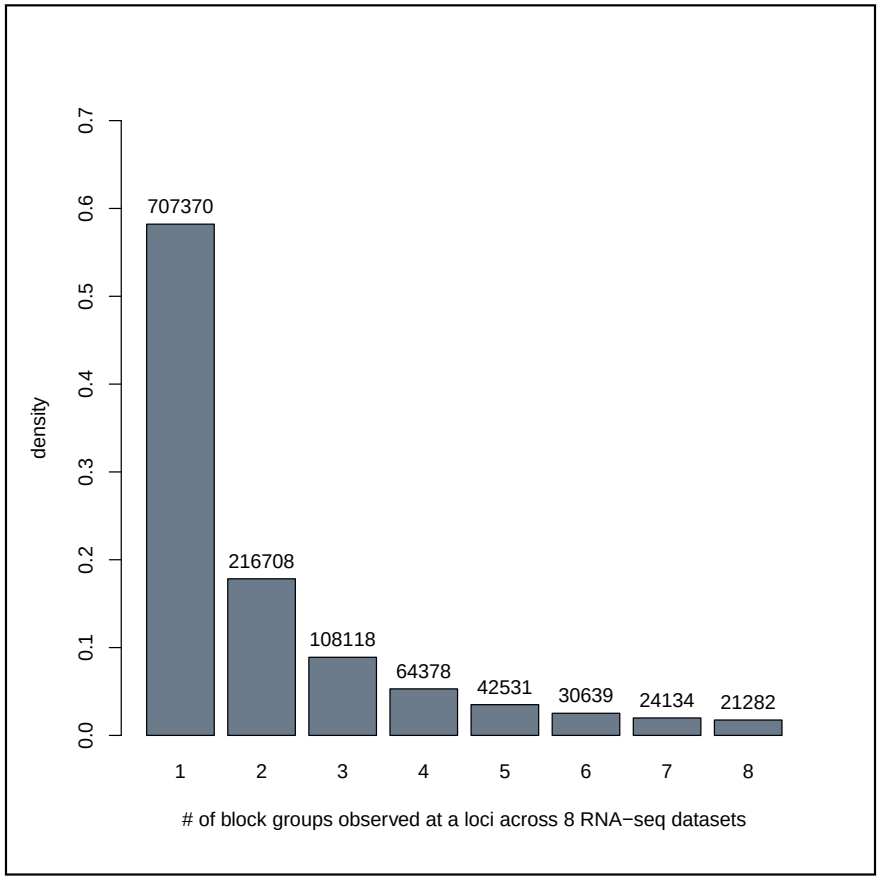
Pipeline: differential read processing patterns

- Pipeline to identify differentially processed transcripts from a locus across multiple RNA-seq experiments.

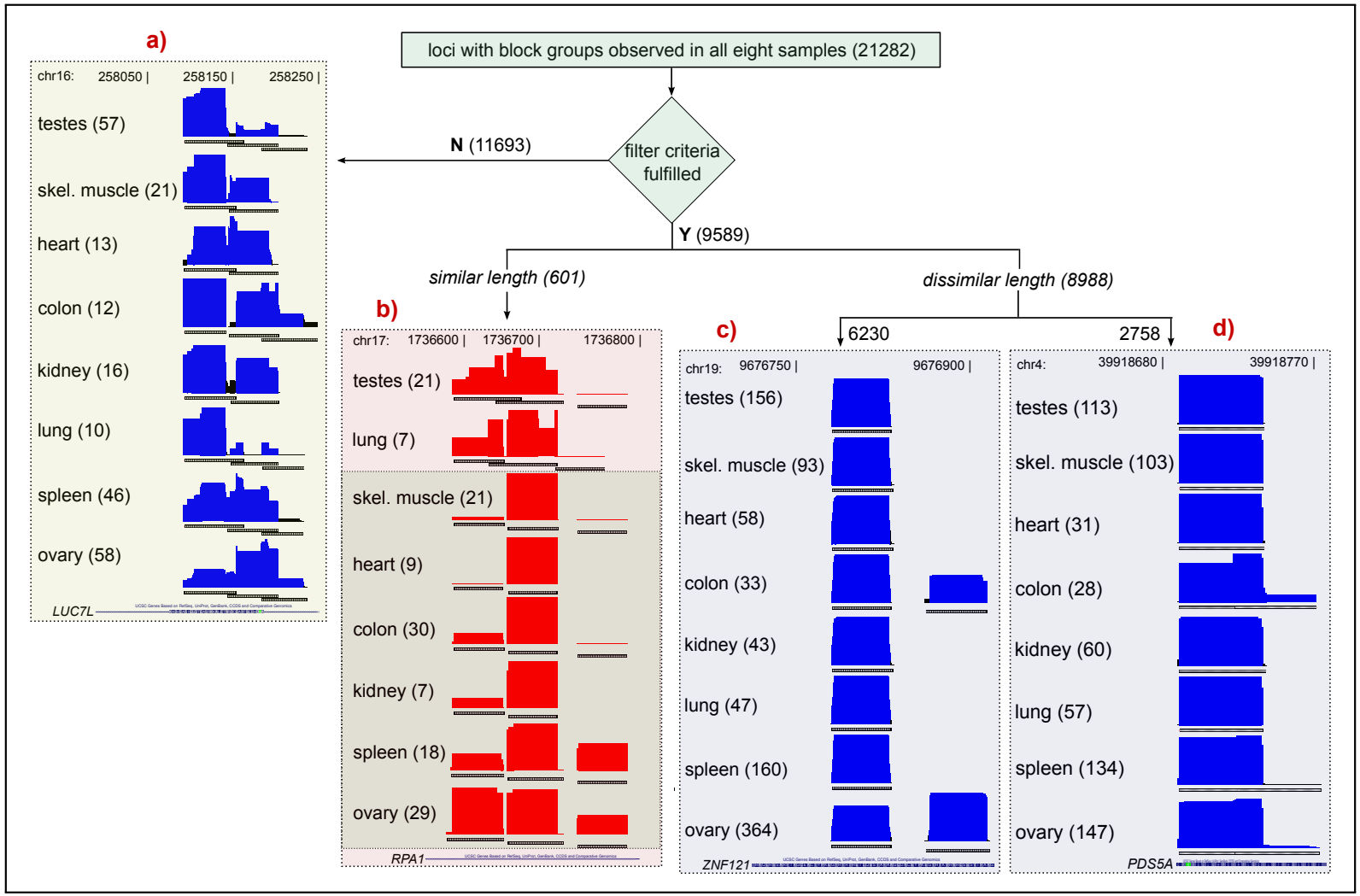


Results

- Total 1,285,5087 loci in human genome were identified having a block group in atleast one dataset.
- 21,282 loci were having block groups in all eight datasets.



Results: sub-classification of read processing patterns



filter criteria: a) $>= 2$ well-defined clusters of block groups and dunn Index $>= 1.5$. or; b) one well-defined cluster with all except one block group.

Conclusions

- ncRNAs are preferentially expressed from different genomic loci.
- read processing patterns are representative of the underlying mechanism in which the host RNA-transcript is processed by the cellular machinery and not of a sequencing artifact.
- Comparison of read processing patterns is useful to identify
  - common processing for specific and/or across different ncRNA classes.
  - novel processing patterns.
  - difference in processing across samples.
- read processing patterns can be sub-classified based on length and constituent blocks.
- no correlation between expression of transcript (RPKM) and its processing pattern.

Outlook

- Better measure to identify distinct sub-clusters from hierarchical cluster.
- Biological implications of loci posessing distinct read processing patterns.
- Statistical measure to identify differential read processing patterns.
- Analysis for preferetial location of different read processing patterns in various part or class of transcripts.

Acknowledgements

- Jan Gorodkin, RTH, University of Copenhagen
- All colleagues, RTH, University of Copenhagen
- LIFE, University of Copenhagen

- Thank You -