Supplementary Data for deepBlockAlign: A tool for aligning RNA-seq profiles of read block patterns

David Langenberger *, Sachin Pundhir *, Claus T. Ekstrøm, Peter F. Stadler, Steve Hoffmann and Jan Gorodkin

*: Equal contribution

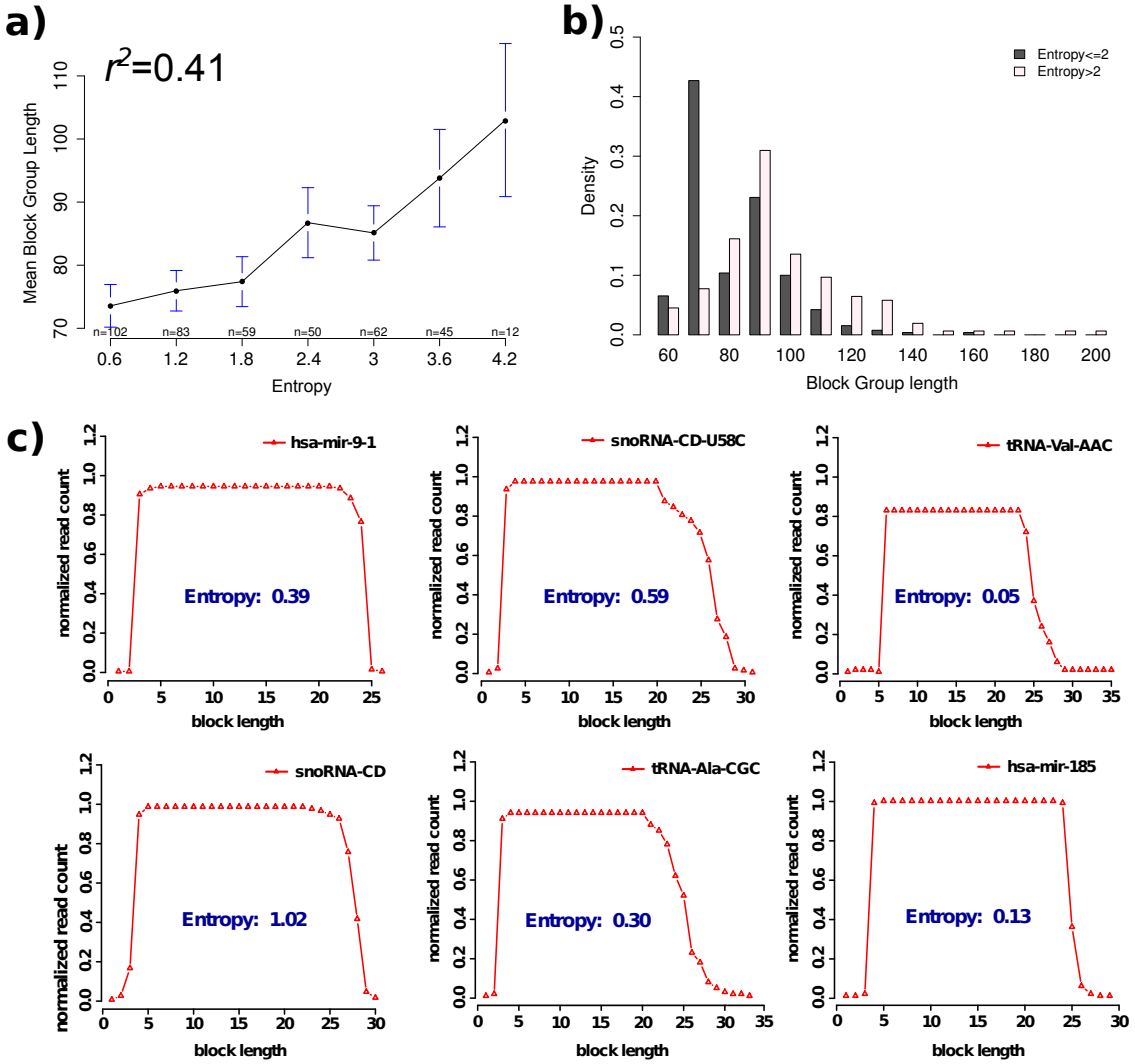

Figure 1: Significance of block length and shape in optimal alignment of block groups. a) For ∼10 nt (75 to 85 nt) increase in length, we observed 5-fold change in the entropy (0.6 to 3) of block groups. Furthermore, there is a moderate linear correlation (r^2=0.41) between the entropy and length of block groups. This suggests that beside length, other factors like functional annotation of block group determines the read processing pattern (shape) within a block group. b) Density distributions of length for block group with entropy <=2 and >2, respectively. Although distinct, the two distributions overlap with each other suggesting that we can observe low entropy within a block group despite having larger length and vice-versa. c) Similar read processing patterns (entropy) within read blocks despite having distinct functional annotations and length. This suggests i) similar read processing patterns of distict length do occur, thus length of block is an important parameter while comparing two read processing patterns ii) two read processing patterns can differ despite having same entropy.

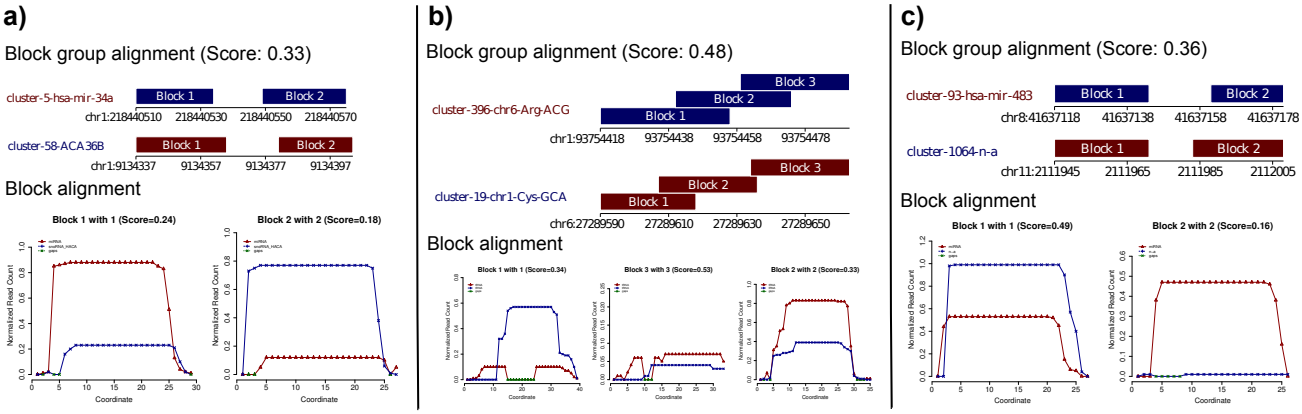

Figure 2: Three optimal alignments from deepBlockAlign suggesting the significance of block shape. In all the three examples (a, b and c), we observe a low alignment score despite having same number and distance between the blocks, primarily due to low block scores (different shapes and expression). In deepBlockAlign, we can tune the importance of block distance and shape by two parameters, distance weight and block weight, respectively. Since, for two block groups to share similar read processing, the relative position of blocks should be same, we have kept a higher distance weight of 6 as compared to block weight of 1.

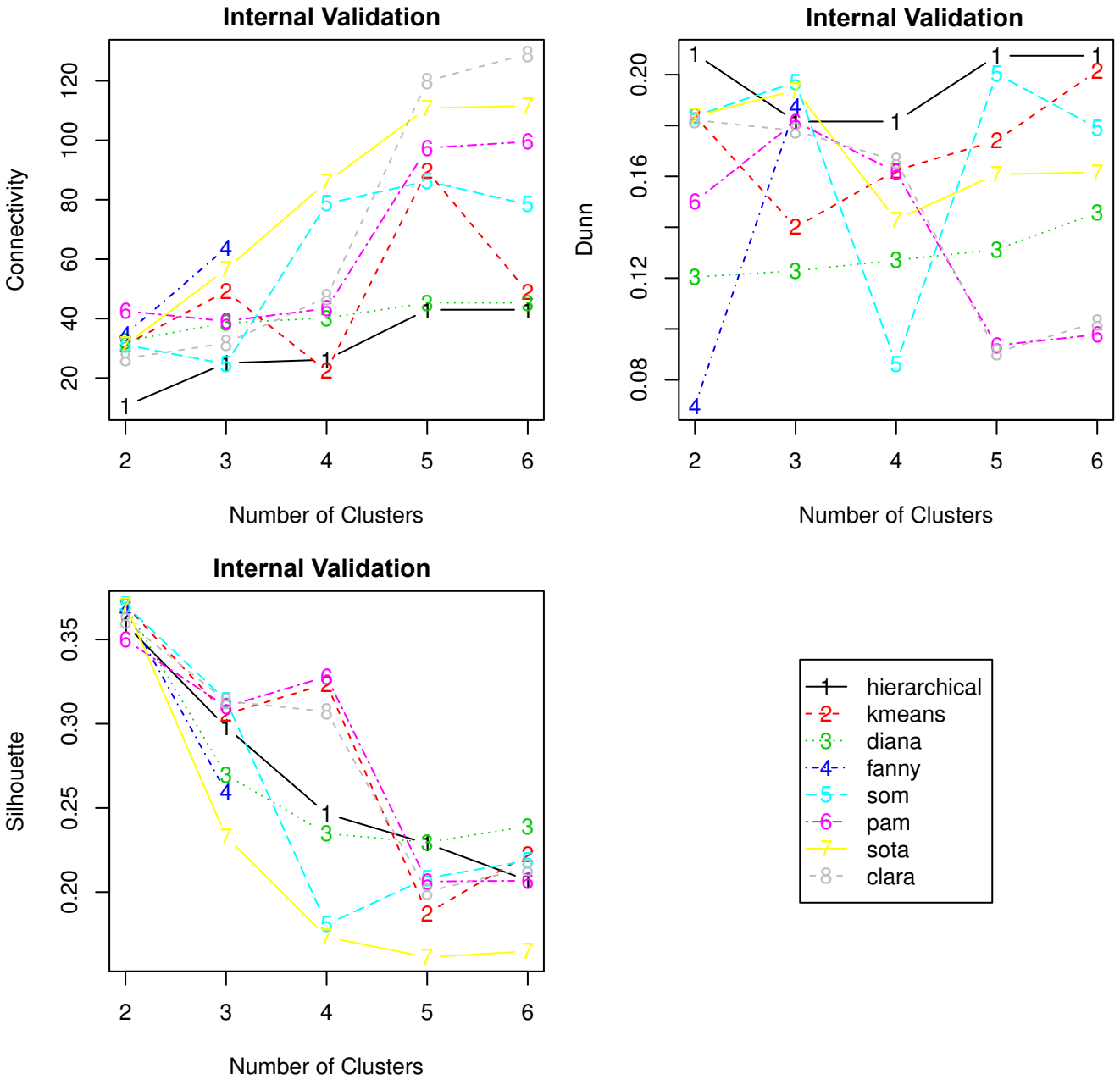

Figure 3: Plots of the connectivity measure, dunn index and silhouette width: eight clustering algorithms and varying number of clusters (1 to 6) were evaluated for unsupervised clustering of ncRNAs in human embryoid body cell dataset. While connectivity measure should be minimized, dunn index and silhouette width should be maximized for optimal performance. As observed hierarchical clustering with two clusters is the most optimal since it clearly has the lowest connectivity and highest Dunn index while at the same time, all clustering methods have virtually the same silhouette width. We therefore employed average linkage hierarchical clustering for subsequent analysis of our dataset.

Table 1: Block groups significantly clustering together. All the block groups significantly clustering together at a p-value of <0.1 were retrieved after the hierarchical clustering of 455 block groups from Human_eb dataset. [a]18 snoRNAs clustered within the miRNA cluster, [b]13 tRNAs clustered within the miRNA cluster, [c]6 unannotated block groups clustered within the miRNA cluster, [d]11 tRNAs clustered together with tRNAs, but having distinct anti-codons, suggesting a similar read processing pattern,[e]8 block group alignments comprising distinct ncRNA classes.

Block group	Coordinate	Hit	Hit Coordinate	Score	p-value
scRNA[e]	chr11:93092461-93092530(+)	miRNA	chr9:96888128-96888193(+)	0.95	<0.05
snoRNA (U28)[a]	chr11:62378662-62378725(-)	miRNA	chr9:130046836-130046905(-)	0.94	<0.05
snoRNA (E3)[a]	chr3:187987781-187987842(+)	miRNA	chr1:154656768-154656831(-)	0.94	<0.05
snoRNA (U43)[a]	chr22:38045005-38045065(-)	miRNA	chr19:56888333-56888394(+)	0.93	<0.05
snoRNA (HBI-100)[a]	chr1:174204156-174204220(-)	miRNA	chr1:203684070-203684135(-)	0.93	<0.05
snoRNA (ACA45)[a]	chr15:81221811-81221879(+)	miRNA	chr8:41637117-41637181(-)	0.92	<0.05
snoRNA (U58A)[a]	chr18:45271651-45271716(-)	miRNA	chr1:197094639-197094701(-)	0.91	<0.05
Unannotated[c]	chr14:65007582-65007644(-)	miRNA	chrX:133508338-133508398(-)	0.90	<0.05
snoRNA (U8)[a]	chr17:8017495-8017584(-)	miRNA	chr3:161605235-161605325(+)	0.88	<0.05
snoRNA (SNORD119)[e]	chr20:2391597-2391681(-)	tRNA_Ser (GCT)	chr6:27373749-27373843(+)	0.87	<0.05
snoRNA (U27)[a]	chr11:62379063-62379132(-)	miRNA	chr1:197094639-197094701(-)	0.87	<0.05
Unannotated[c]	chr8:41637118-41637181(+)	miRNA	chr8:41637117-41637181(-)	0.87	<0.05
snoRNA (ACA36B)[a]	chr1:218440510-218440575(-)	miRNA	chr8:141811860-141811929(-)	0.86	<0.05
tRNA_Trp (CCA)[d]	chr12:97422159-97422239(+)	tRNA_Arg (CCG)	chr16:3140674-3140755(+)	0.85	<0.05
Unannotated[c]	chr6:95213558-95213629(-)	miRNA	chr11:43559534-43559599(+)	0.85	<0.05
Unannotated[c]	chr19:54175315-54175376(+)	miRNA	chr6:33283610-33283675(+)	0.84	<0.05
scRNA[e]	chr1:154720515-154720589(-)	miRNA	chr2:136139431-136139516(+)	0.84	<0.05
tRNA_Arg (ACG)[b]	chr6:27746322-27746384(-)	miRNA	chrX:85045299-85045382(-)	0.84	<0.05
rRNA (5.8s)[e]	chr19:23979061-23979157(-)	snRNA	chr15:94090123-94090225(+)	0.84	<0.05
snoRNA (U52)[a]	chr6:31912831-31912895(+)	miRNA	chr11:2111945-2112008(-)	0.83	<0.05
tRNA_Gly (CCC)[d]	chr2:70329618-70329697(-)	tRNA_Ala (TGC)	chr6:28719199-28719275(+)	0.83	<0.05
tRNA_Lys (CTT)[d]	chr16:3147404-3147479(-)	tRNA_Pro (TGG)	chr16:3178093-3178171(+)	0.82	<0.05
tRNA_Ala (CGC)[d]	chr6:26661704-26661789(+)	tRNA_Pro (TGG)	chr14:20222012-20222088(+)	0.82	<0.05
snoRNA (ACA61)[a]	chr1:28778862-28778939(-)	miRNA	chrX:73354946-73355028(-)	0.82	<0.05
snoRNA (U44)[a]	chr1:172101715-172101789(-)	miRNA	chrX:73354946-73355028(-)	0.81	<0.05
tRNA_Ala (AGC)[b]	chr2:27127582-27127662(+)	snoRNA (U49A)	chr17:16284074-16284145(+)	0.81	<0.05
scRNA [e]	chr8:124126138-124126233(+)	snoRNA (HBII-436)	chr15:22778233-22778320(+)	0.80	<0.05
tRNA_Arg (ACG)[b]	chr6:26645716-26645804(+)	miRNA	chr19:58931907-58931990(+)	0.80	<0.05
tRNA_Ser (AGA)[b]	chr17:8070649-8070737(-)	miRNA	chr19:58931907-58931990(+)	0.79	<0.05
tRNA_Ser (AGA)[b]	chr6:26435789-26435881(+)	miRNA	chr9:85774505-85774590(-)	0.78	<0.05
snRNA[e]	chr6:89830005-89830081(-)	miRNA	chr19:56888333-56888394(+)	0.78	<0.05
tRNA_Trp (CCA)[d]	chr6:26439643-26439713(-)	tRNA_Arg (ACG)	chr6:27746322-27746384(-)	0.77	<0.05
tRNA_Trp (CCA)[b]	chr6:26439643-26439713(-)	miRNA	chrX:85045299-85045382(-)	0.76	<0.05
snoRNA (U20)[a]	chr2:232029399-232029478(-)	miRNA	chr19:58982726-58982807(+)	0.76	<0.05
tRNA_Ala (CGC)[d]	chr2:156965499-156965601(+)	tRNA_Cys (GCA)	chr7:148659128-148659208(+)	0.76	<0.05
snoRNA (U25)[a]	chr11:62379604-62379680(-)	miRNA	chrX:73423850-73423926(-)	0.76	<0.05
tRNA_Ala (AGC)[b]	chr8:67188978-67189082(+)	miRNA	chr13:49521267-49521325(-)	0.76	<0.1
tRNA_His (GTG)[d]	chr9:14423935-14424010(-)	tRNA_Gly (CCC)	chr16:626728-626807(-)	0.76	<0.05
snoRNA (HBII-85-27)[a]	chr15:22897814-22897906(+)	miRNA	chr13:49521267-49521325(-)	0.75	<0.1
tRNA_Leu (CAA)[d]	chr6:28971964-28972089(-)	tRNA_Tyr (GTA)	chr6:26683770-26683879(+)	0.74	<0.05
tRNA_Cys (GCA)[b]	chr1:93754418-93754491(-)	snoRNA	chr1:45016649-45016727(+)	0.73	<0.05
tRNA_Gln (CTG)[d]	chr1:145971672-145971766(+)	snoRNA (U61)	chrX:135789009-135789099(-)	0.71	<0.05
tRNA_Cys (GCA)[d]	chr17:34563487-34563590(-)	tRNA_His (GTG)	chr15:43278056-43278168(-)	0.71	<0.05
snoRNA (HBII-85-29)[a]	chr15:22902761-22902843(+)	miRNA	chr13:90801301-90801387(+)	0.69	<0.05
Unannotated[c]	chr3:105362210-105362273(-)	snoRNA	chr6:133179636-133179690(+)	0.69	<0.08
rRNA (28S)[e]	chr6:120625141-120625200(+)	tRNA_Cys (GCA)	chr3:133430650-133430707(-)	0.68	<0.05
scRNA[e]	chr7:148291340-148291443(+)	snoRNA (U50)	chr6:86443719-86443806(-)	0.67	<0.05
snoRNA (U38B)[a]	chr1:45016649-45016727(+)	miRNA	chr19:58881529-58881612	0.67	<0.05
Unannotated[c]	chr22:19595430-19595490(+)	snRNA	chr3:73242903-73242966(+)	0.64	<0.1
tRNA_Arg (ACG)[b]	chr6:27290942-27291027(+)	miRNA	chr3:49033062-49033151(-)	0.63	<0.05
snoRNA (HBII-251)[a]	chr1:31213599-31213671(-)	miRNA	chr7:127635165-127635232(+)	0.61	<0.05
tRNA_Lys (TTT)[d]	chr11:121935859-121935925(+)	tRNA_Gln (CTG)	chr6:27595287-27595352(+)	0.57	<0.05
tRNA_Lys (TTT)[b]	chr11:121935859-121935925(+)	snRNA	chr3:73242903-73242966(+)	0.56	<0.05
tRNA_Gln (CTG)[b]	chr6:27595287-27595352(+)	snRNA	chr3:73242903-73242966(+)	0.55	<0.05
tRNA_Gln (CTG)[b]	chr17:7963795-7963860(+)	snRNA	chr3:73242903-73242966(+)	0.55	<0.05
tRNA_Lys (TTT)[b]	chr16:72069735-72069798(-)	snRNA	chr3:73242903-73242966(+)	0.50	<0.1

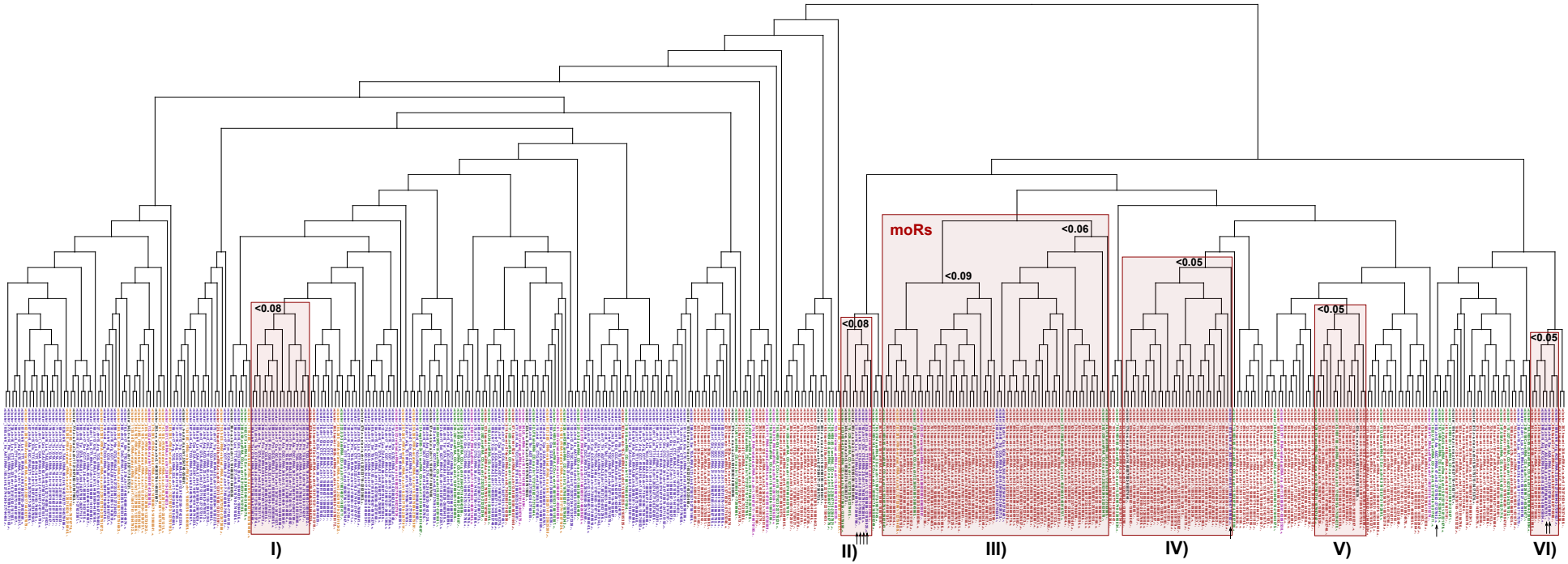

Figure 4: Hierarchical clustering of 455 block groups based on alignment score from deepBlockAlign. (a) A tree visualizing the clustering. microRNA loci (red) are well separated from tRNA genes (blue). Within the microRNA cluster, microRNA-offset RNAs (moRs) can be found in one sub-cluster (IV), illustrating the different read pattern, caused by the additional blocks flanking the mature microRNA regions. Some significant clusters having tRNAs, snoRNAs or unannotated block groups clustering together with microRNAs (II, III, V and VI). tRNAs that are reported to generate products with miRNA-like features (Lee et al., 2009; Haussecker et al., 2010; Burroughs et al., 2011; Cole et al., 2009) are highlighted with arrows. A cluster having tRNAs with different anti-codons but highly similar expression pattern (I).

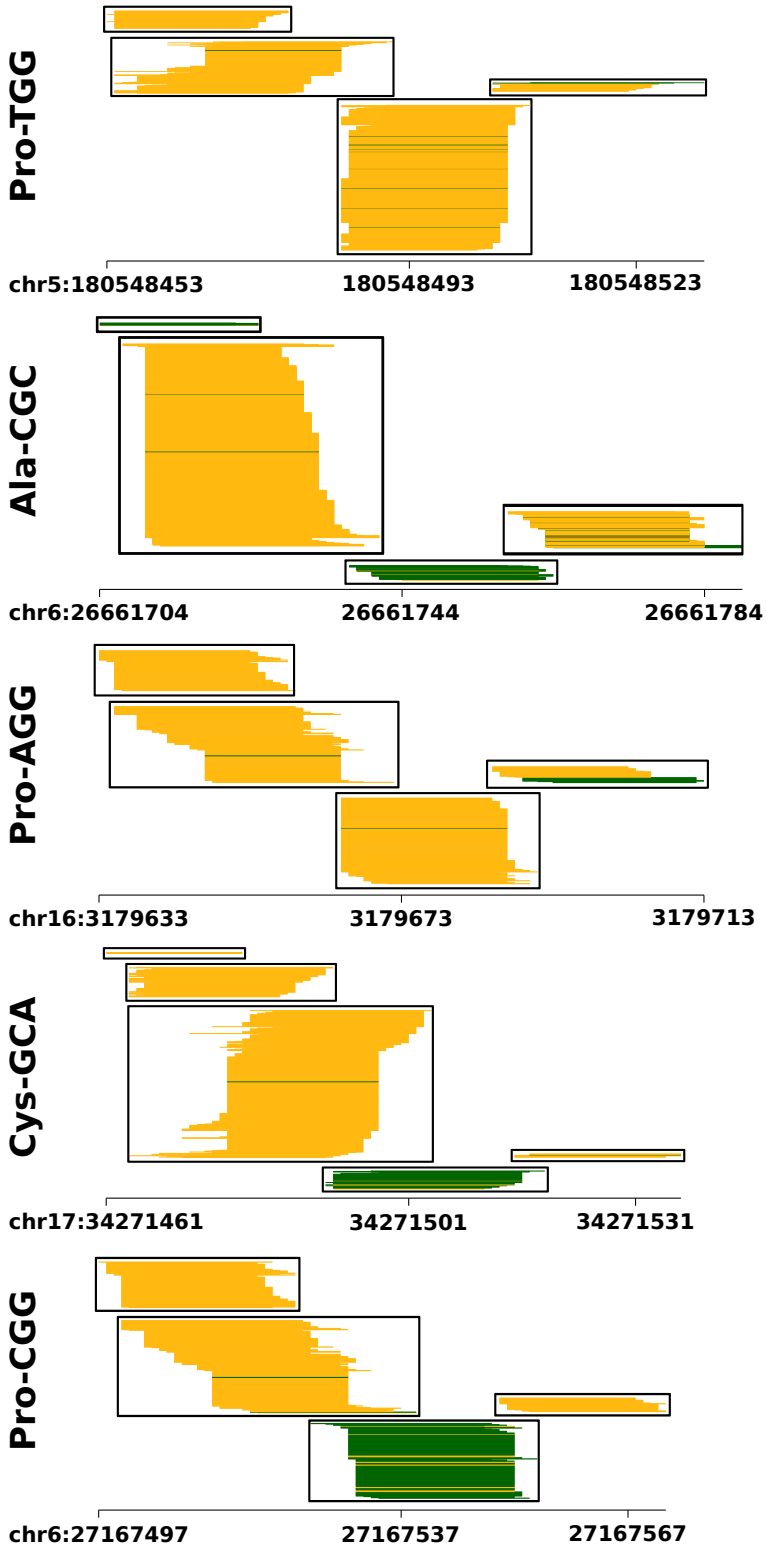

Figure 5: Similar read processing pattern of five tRNAs having different anticodons (Pro-TGG, Ala-CGC, Pro-AGG, Cys-GCA and Pro-CGG). On hierarchial clustering, these tRNAs significantly clustered together (p-value < 0.08).

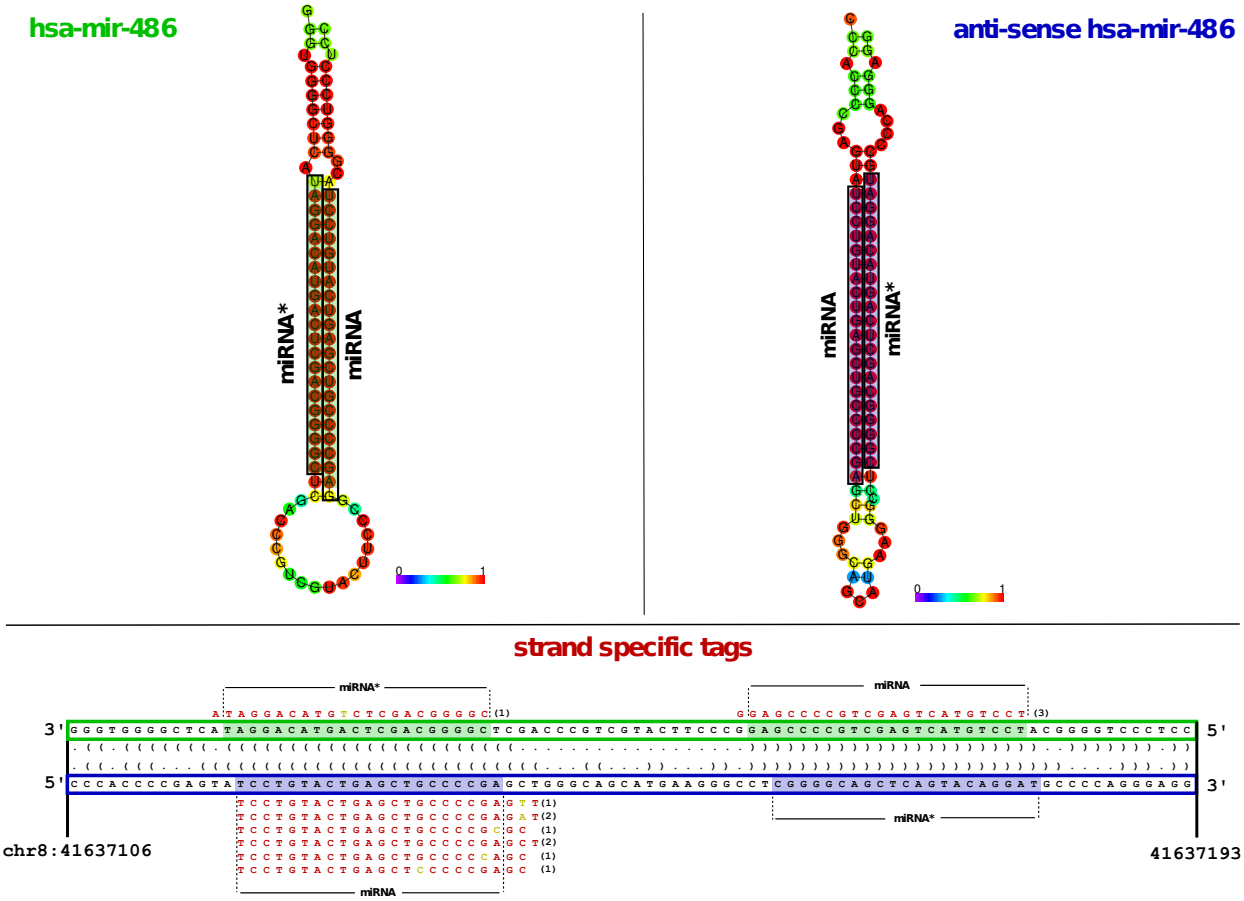

Figure 6: Putative anti-sense miRNA to hsa-mir-486. Strand-specific tags are highlighted in red along with their frequency enclosed in brackets. The sense and anti-sense miRNAs are marked with green and blue colors, respectively.

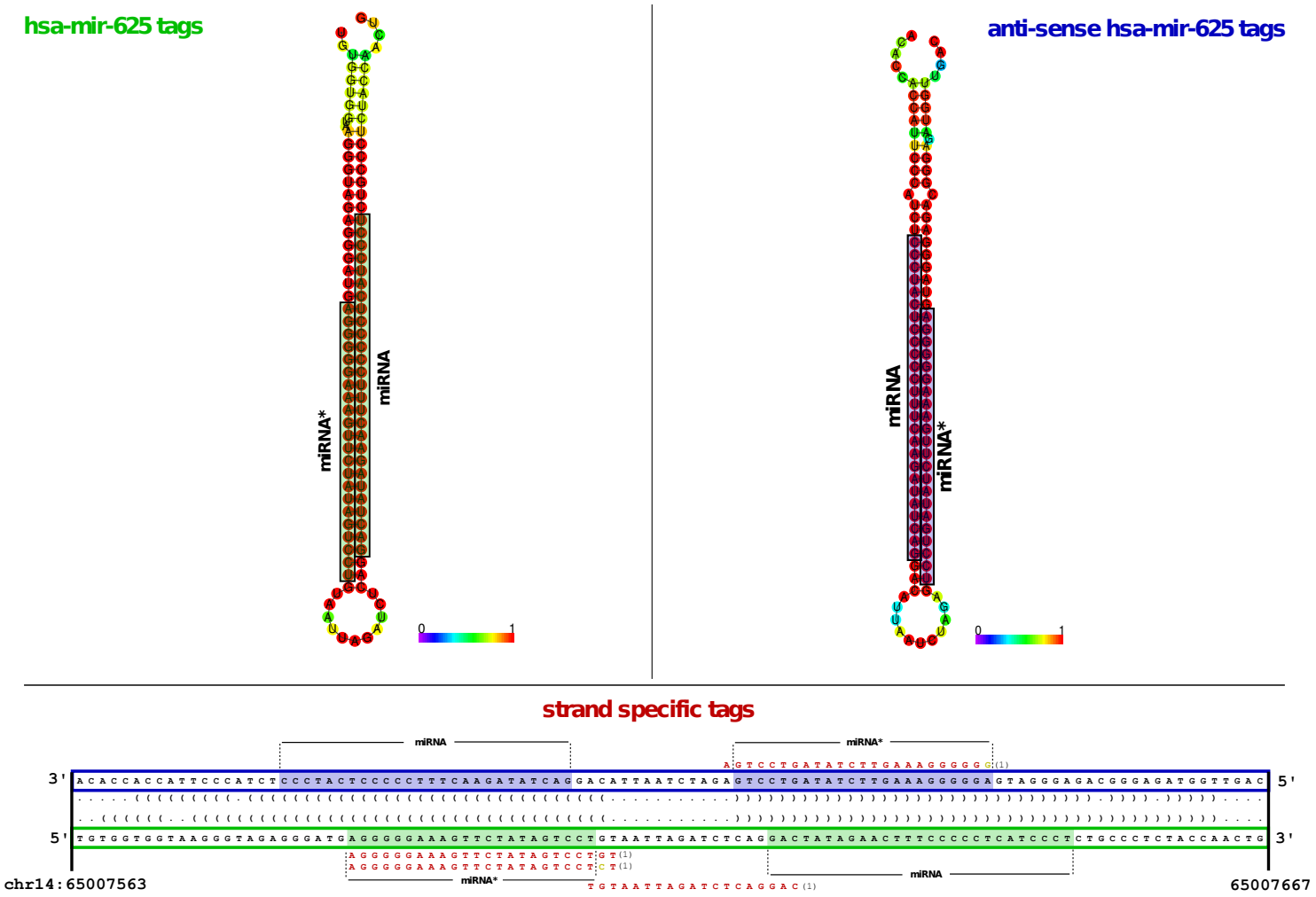

Figure 7: Putative anti-sense miRNA to hsa-mir-625. Strand-specific tags are highlighted in red along with their frequency enclosed in brackets. The sense and anti-sense miRNAs are marked with green and blue colors, respectively.