Supplementary Materials

Poly(A)-ClickSeq: click-chemistry for next-generation 3´-end sequencing without RNA enrichment or fragmentation

Andrew Routh*[1,2], Ping Ji[1], Elizabeth Jaworski[1], Zheng Xia[3,4], Wei Li[4], Eric J. Wagner*[1,2]

1) Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, Texas, USA

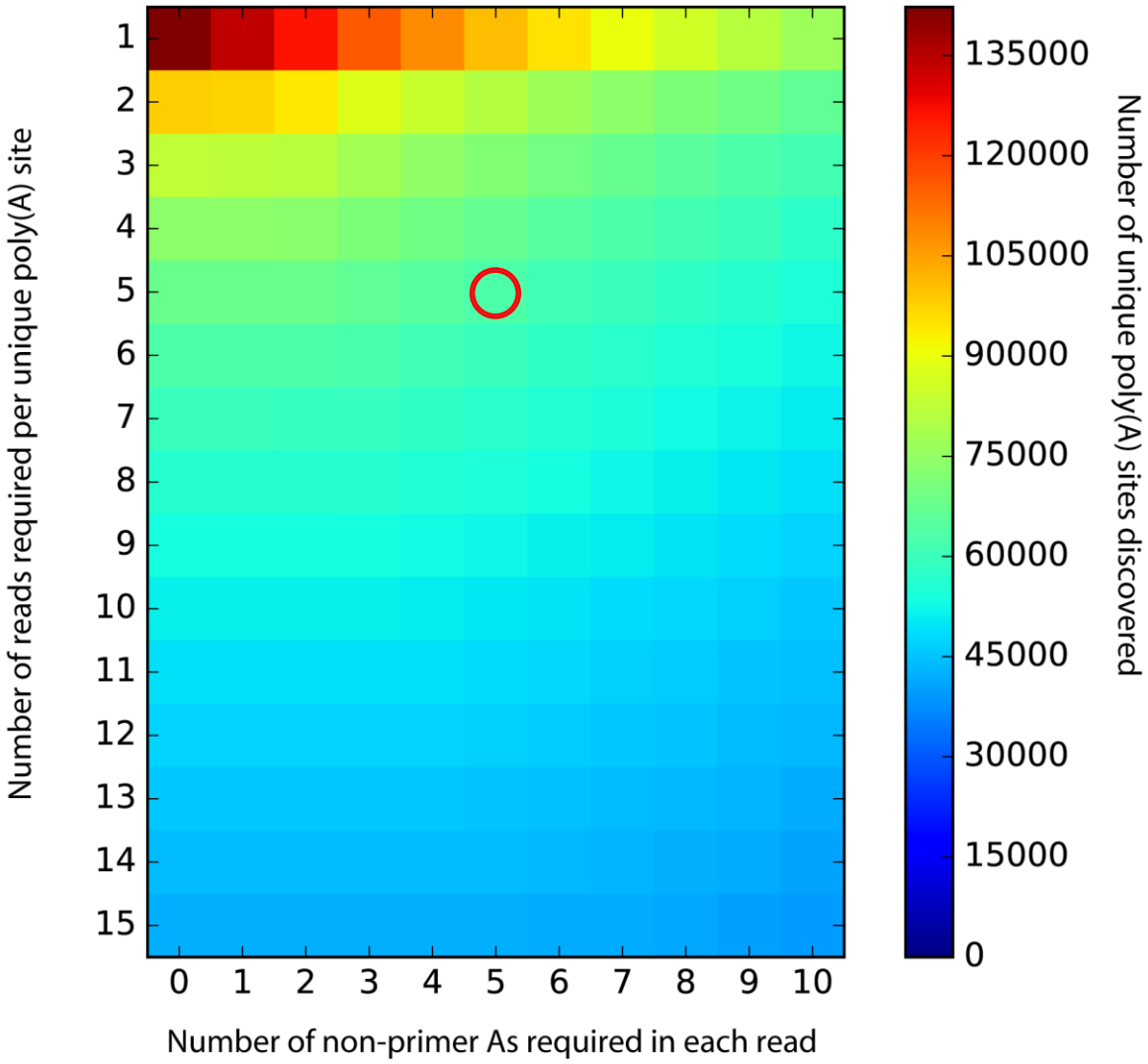
2) Sealy Centre for Structural Biology and Molecular Biophysics, The University of Texas Medical Branch, Galveston, Texas, USA.

3) Oregon Health and Science University, Portland, Oregon, USA

4) Department of Molecular and Cellular Biology, Baylor College of Medicine, Texas, USA

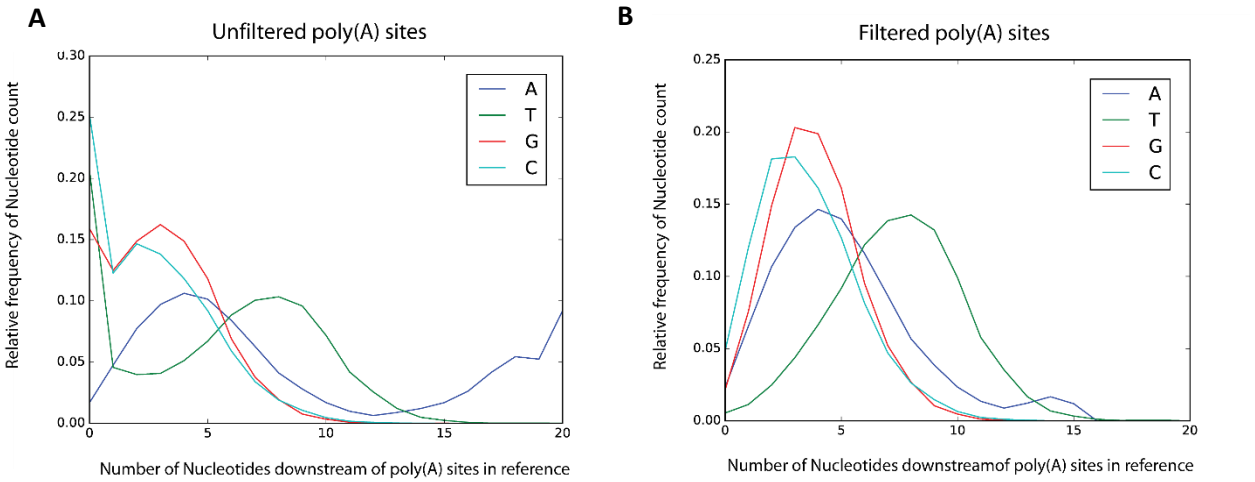
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555.

*Correspondence to: Andrew Routh: alrouth@utmb.edu and Eric Wagner: ejwagner@utmb.edu



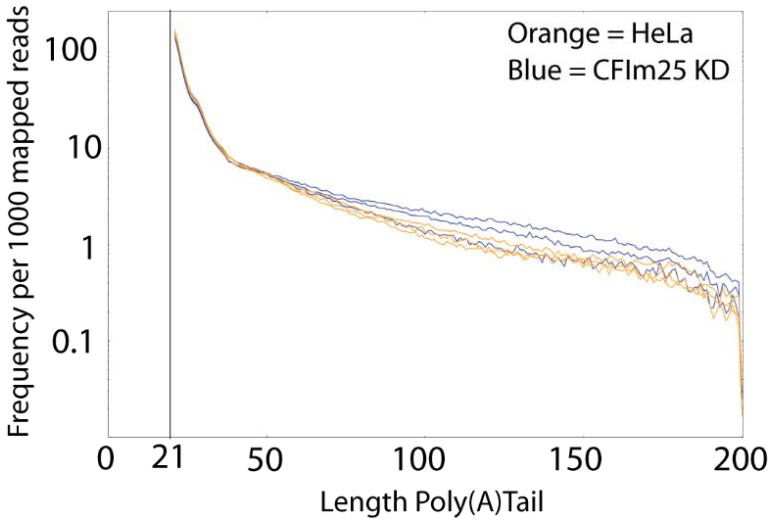
Supplementary Figure 1:

A heat-map illustrating the impact of quality filters on the number of poly(A) sites discovered. To accept a putative poly(A) site, the mapped reads can be required to contain a minimum number of non-primer derived 'A's (x-axis) and each poly(A) site can be required to be represented by a minimum number of reads (y-axis). The number of poly(A) sites found (denoted by the colour-bar) using a range of combination of these two filters are shown here. In our analyses we required as least 5 reads each with 5 or more non-primer derived A's (indicated by red circle) to confirm a poly(A) site.



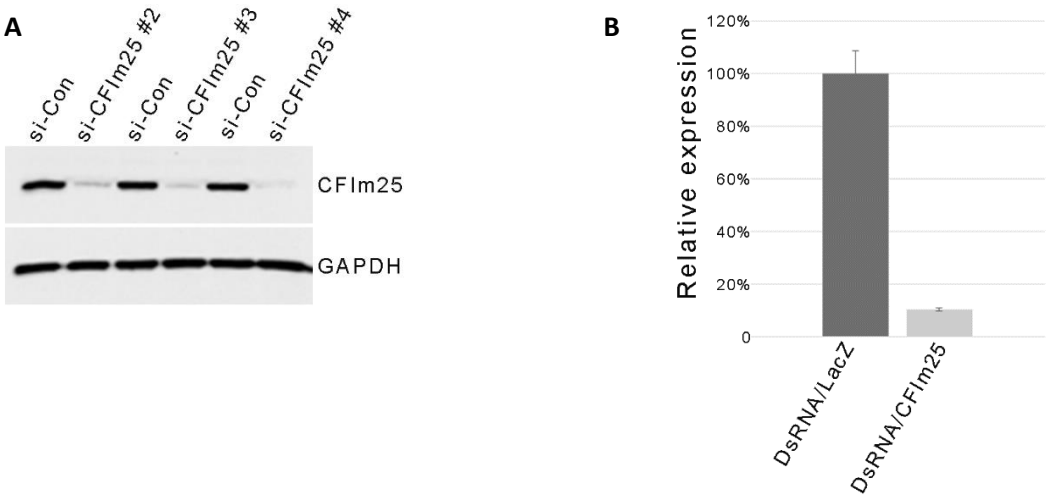
Supplementary Figure 2:

Nucleotides in the reference genomes upstream of Poly(A) sites were found be enriched for A's. This indicates that these may be derived from internally primed sequences, rather than from genuine poly(A) tails. The number of nucleotides found upstream of each poly(A) sites were counted for **A)** all poly(A) sites, and **B)** after removing poly(A) sites containing 15 or more A's within 20 nucleotides of the upstream reference sequence. The x-axis indicates the nucleotide count, and the y-axis gives the relative frequency with which each nucleotide count was found.



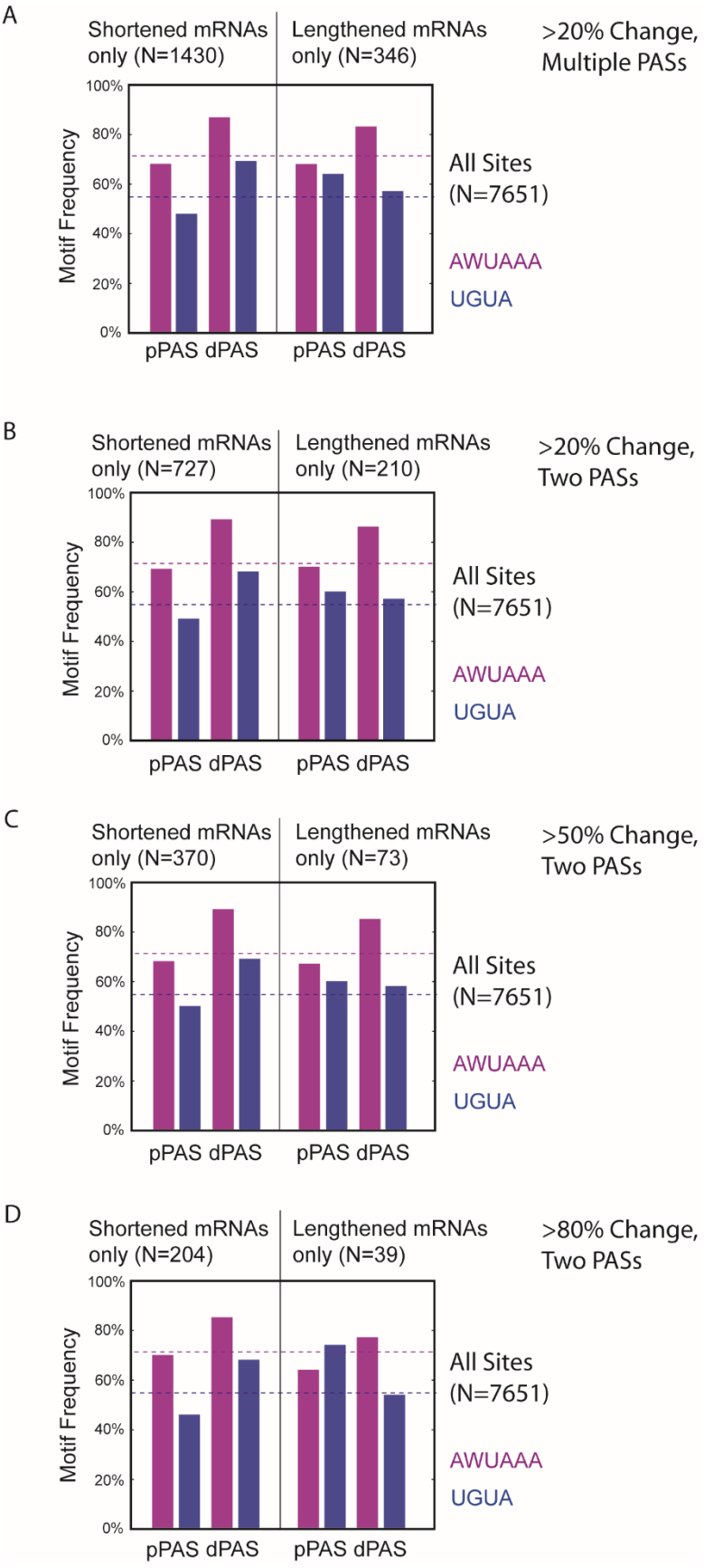
Supplementary Figure 3:

Poly(A) tail lengths are inferred from the read data. The number of 'A's remaining at the end of each read after adaptor trimming is determined and appended onto the read name. After mapping, the distribution of poly(A) tail length at each mapped loci can be determined. The distribution of tail lengths over all the mapped reads for each of the HiSeq Poly(A)ClickSeq datasets of HeLa total cellular RNA (orange) and CFIm25 KD total cellular RNA (blue) is shown. A range of poly(A) tail lengths from 21 up to 200 nts is found.

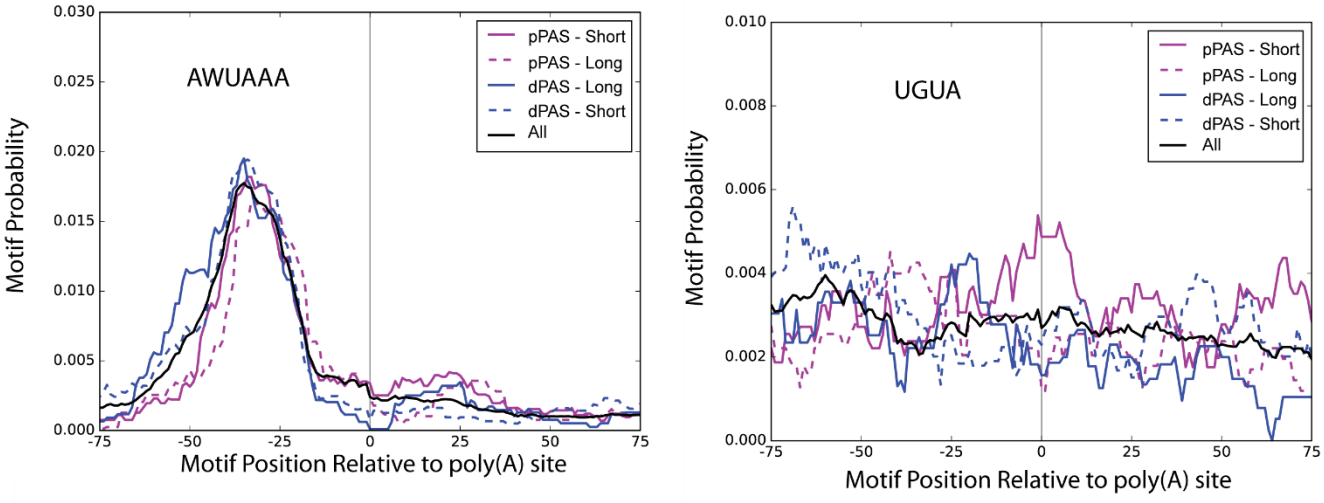


Supplementary Figure 4:

CF25Im was knocked down in HeLa cells by siRNAs. **A)** Western blotting shows <95% depletion compared to control cell. CF25d was knocked down in S2 cells by dsRNA. **B)** Real-time quantitative PCR of total cellular RNA shows <90% depletion relative to control dsRNA.



Supplementary Figure 5: The frequency of AWUAAA and UGUA motifs found within 100nt upstream of both proximal (pPAS) and distal (dPAS) poly(A) sites for shortened or lengthened mRNAs are shown. Analyses showed similar trends when considering mRNAs with **A)** multiple PASs and >20% APA; **B)** only two PASs and >20% APA; **C)** only two PASs and >50% APA; and **D)** only two PASs and >80% APA.



Supplementary Figure 6:

Motif enrichment anlaysis using MEME suite illustrates the position of enriched motif relative to detected poly(A) sites. This reveals AWUAAA sites ~20-40 nts upstream of detected poly(A) sites found in Drosophila melanogaster S2 cells either with or without CF25d knockdown. Although UGUA are significantly enriched, there is no positional preference. Traces are shown for all found poly(A) sites (**All**), and for proximal (**pPAS**) and distal (**dPAS**) sites found in mRNAs that were either lengthened (**Long**) or shortened (**Short**) upon CF25d knock-down.

Supplementary Table 1: Mapping statistics for Poly(A)-ClickSeq of total cellular RNA from either wild-type or CFIm25 KD HeLa cells.

	Control		CFIm25 KD	
Total Raw Reads	**97532979**		**94717471**	
Rep 1	*35950672*		*36447774*	
Rep 2	*29912811*		*26317115*	
Rep 3	*31669496*		*31952582*	
Number Processed Reads	**45565507**	**46.72%**	**44291920**	**46.76%**
Rep 1	*16871383*	46.93%	*17082420*	46.87%
Rep 2	*14097898*	47.13%	*11819393*	44.91%
Rep 3	*14596226*	46.09%	*15390107*	48.17%
Reads Mapped to Human Genome	**44049869**	**96.67%**	**42398266**	**95.72%**
Rep 1	*16326587*	96.77%	*16372478*	95.84%
Rep 2	*13597357*	96.45%	*11293094*	95.55%
Rep 3	*14125925*	96.78%	*14732694*	95.73%
Unmapped Reads	**1515638**	**3.33%**	**1893654**	**4.28%**
Rep 1	*544796*	3.23%	*709942*	4.16%
Rep 2	*500541*	3.55%	*526299*	4.45%
Rep 3	*470301*	3.22%	*657413*	4.27%
Detected Poly(A) Sites				
Rep 1	*37434*		*47811*	
Rep 2	*29544*		*50650*	
Rep 3	*24256*		*21346*	
Unique Poly(A) Sites found in one or more replicates	**56937**		**76176**	
Unique Poly(A) Sites found in two or more replicates	**24937**		**33008**	
Unique Poly(A) Sites found in all three replicates	**12501**		**13580**	

Supplementary Table 2: Locations of detected Poly(A)site and comparison to Poly(A)DB in HiSeq dataset. Reads count are shown and Poly(A)sites are shown in *italics*.

Present in two+ replicates	31752990 *24937*		30993891 *33008*	
UCSC Genes	29534658	93.01%	28804998	92.94%
	22066	88.49%	*29434*	89.17%
Of which:				
Exons	28470681	89.66%	27600003	89.05%
	21002	84.22%	*27605*	83.63%
3' prime exon	28182639	88.76%	27314472	88.13%
	20660	82.85%	*26964*	81.69%
Within 500 nts down-stream of UCSC annotation	1733592	5.46%	1655001	5.34%
	1879	7.53%	*2188*	6.63%
Remaining	484740	1.53%	533892	1.72%
	992	3.98%	*1386*	4.20%
Poly(A)DB	23207205	73.09%	22466619	72.49%
	14457	57.97%	*26359*	79.86%
Poly(A)DB+/-10nts	27658533	87.11%	26305071	84.87%
	20856	83.63%	*26172*	79.29%

Supplementary Table 3: Locations of detected Poly(A)site and comparison to Poly(A)DB in MiSeq dataset. Reads shown and Poly(A)sites shown in *italics*.

Present in three replicates	1279116 *10691*	63.47%	1361517 *11154*	60.52%
UCSC Genes	1200186	93.83%	1279629	93.99%
	9877	92.39%	*10370*	92.97%
Of which:				
Exons	1154973	90.29%	1233759	90.62%
	9637	90.14%	*10110*	90.64%
3' prime exon	1146066	89.60%	1223145	89.84%
	9491	88.78%	*9939*	89.11%
Within 500 nts down-stream of UCSC annotation	64806	5.07%	68211	5.01%
	619	5.79%	*601*	5.39%
Remaining	14124	1.10%	13677	1.00%
	195	1.82%	*183*	1.64%
Poly(A)DB	960018	75.05%	1015152	74.56%
	7047	65.92%	*7209*	64.63%
Poly(A)DB+/-10nts	1119714	87.54%	1188249	87.27%
	9660	90.36%	*9997*	89.63%

Supplementary Table 4: Locations of detected Poly(A)site and comparison to Poly(A)DB in MiSeq dataset in drosophila. Reads shown and Poly(A)sites shown in *italics*.

Present in three replicates	1349733 *6910*	63.33%	1476633 *7473*	60.91%
UCSC Genes	1107774	82.07%	1220520	82.66%
	5467	79.12%	*5919*	79.21%
Of which:				
Exons	1060974	77.73%	1162635	78.29%
	5250	68.16%	*5646*	71.34%
3' prime exon	1043196	76.43%	1142496	76.93%
	5005	64.97%	*5362*	67.75%
Within 500 nts down-stream of UCSC annotation	221052	16.19%	232782	15.67%
	1297	16.84%	*1380*	17.44%
Remaining	20907	1.53%	23331	1.57%
	146	1.90%	*174*	2.20%
Poly(A)DB	1163220	85.22%	1273905	85.78%
	4681	60.77%	*4927*	62.26%
Poly(A)DB+/-10nts	1250631	91.62%	1363980	91.85%
	5538	71.89%	*5850*	73.92%

Supplementary Datafile 1:

Compilation of Scripts for processing raw Poly(A)-ClickSeq data. Last edited: 29th Nov 2016. Please contact alrouth@utmb.edu with questions/requests. All scripts have been successfully executed on Cygwin workstation and on Linux server using python version 2.7. Required software packages and the last confirmed working version are: HiSat2 v2.0.4 (1), samtools v1.2 (2), cutadapt v1.9.1 (3), fastx_toolkit v0.0.14 http://hannonlab.cshl.edu/fastx_toolkit/. Different packages/versions may require adjustments.

Scripts include:

1) Extract_nts.py:
    - Uses *samtools(2)* to extract nucleotide either before or after poly(A) sites provided in a BEDGraph in the format generated using the pAz-Seq scripts.
2) Extract_pA_Lens_Ad.py:
    - Required during read processing to measure and extract poly(A) length in individual reads and append this information on to the read name
3) MakeBEDGRAPH_pALenAr.py
    - Required to make the BEDGraph from a mapped SAM file.
4) Mask_ints.py
    - Required to remove mapped reads that are likely present due to non-specific/internal priming. Requires *samtools(2)*
5) Merge_Reps.py
    - Allows merging of multiple BEDGraph files
6) Remove_5prime_IDtag.py
    - Required during read processing to remove nucleotides derived from the 5' Click adaptor. Usually only six nucleotides. This can function as limited ID tag.

The following are examples of batch recipes that can be run locally on a stand-alone workstation. Adjustments must be made for (e.g.) SLURM queue submission on a server. Folder containing individual scripts must be in PATH, otherwise recipes must be adjusted to point to each script.

1) pAz_Prep.txt : processing raw reads
2) pAz_Map.txt : maps processed reads
3) pAz_BED.txt : generates BEDGraph files

Supplementary Datafile 2: BEDgraph files of HiSeq analysis of Wt HeLa and CFIm25 KD poly(A) sites, Human hg19. Individual BEDgraphs for each replicate (3x) for both wild-type and CF25Im KD cells are provided, as well as the merged datasets requiring a unique poly(A) site to be present in two or more replicates (as used in this manuscript). Additionally, BEDgraphs of the coverage of reads over the reference genome found in Poly(A)ClickSeq datasets are provided.

Supplementary Datafile 3: BEDgraph files of MiSeq analysis of Wt HeLa and CFIm25 KD poly(A) sites, Human hg19. Individual BEDgraphs for each replicate (3x) for both wild-type and CF25Im KD cells are provided, as well as the merged datasets requiring a unique poly(A) site to be present in two or more replicates (as used in this manuscript).

Supplementary Datafile 4: BEDgraph files of MiSeq analysis of Wt S2 and CFIm25 KD poly(A) sites, Drosophila Dm6. Individual BEDgraphs for each replicate (3x) for both wild-type and CF25Im KD cells are provided, as well as the merged datasets requiring a unique poly(A) site to be present in two or more replicates (as used in this manuscript). Additionally, BEDgraphs of the coverage of reads over the reference genome found in Poly(A)ClickSeq datasets are provided.

Supplementary References

1. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**, 357-360.
2. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
3. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10-12.