FoldAtlas: a repository for genome-wide RNA structure probing data

Matthew Norris[1*], Jitender Cheema[1], Chun Kit Kwok[2], Matthew Hartley[1], Richard J Morris[1], Sharon Aviran[3], Yiliang Ding[1*]
[1]John Innes Centre, Norwich Research Park, Norwich, UK
[2]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW (UK)
[3]Department of Biomedical Engineering and Genome Center, UC Davis, Davis, USA
[*]to whom correspondence should be addressed

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY RESULTS

Library construction and mapping

The *FoldAtlas* dataset was generated using a previously established *Structure-seq* method (Ding *et al*., 2014; Ding *et al*., 2015). Five-day-old *Arabidopsis thaliana* etiolated seedlings were treated with DMS at 22ºC, followed by dithiothreitol quench. Cellular RNA was purified and subjected to two-round of poly(A)-selection to enrich the proportion of mRNA. As a consequence of enrichment, some types of RNA molecule are under-represented in our data, with examples including enhancer and RNA polymerase III transcripts.

After reverse transcription, first-strand complementary DNAs were ligated at their 3' ends to a DNA linker and PCR was performed. Different barcode indices were used for the (-)DMS and (+)DMS libraries, which were subjected to sequencing using the Illumina HiSeq 2500 platform. Data were generated for three independent biological replicates. Therefore the new Structure-seq dataset in this work provides a higher depth of transcriptome-wide mRNA structure information than earlier experiments.

Mapping was performed using *bowtie* (Langmead *et al*., 2009) to obtain counts of chemically modified positions, as described in the earlier work (Ding *et al*., 2014). When counting reverse transcriptase (RT) stop points at a position, the value was incremented by 1 when the read uniquely mapped to a genomic position. If a read mapped to more than one genomic position, the RT stop count at the position was incremented by 1 / *n*, where *n* was the number of genomic positions mapped.

After mapping, unique, non-unique and unmapped reads were counted, which indicated that a high proportion of reads mapped to the reference sequences (Table S1). We then counted the numbers of reads that were assigned to each type of RNA molecule, indicating a high proportion of mRNA in the sample (Table S2). We observed the expected enrichment of A and C nucleotides in the (+)DMS library (Table S3).

After obtaining the RT stop counts for each biological replicate, correlation analyses between the replicates were carried out (Fig. S1). High correlation was observed between biological replicates R1, R2, and R3, hence these were summed to form (+)DMS and (-)DMS data sets. We also calculated the average read count per A or C nucleotide for each RNA molecule. The average coverage distributions are visualised across the RNA molecules in Fig. S2. Using RT stop counts for (-)DMS and (+)DMS sets, normalised reactivities were calculated as described in the earlier work (Ding *et al*., 2013). The normalised reactivities were then used for secondary structure prediction.

RNA secondary structure prediction.

We predicted RNA secondary structures for each RNA molecule at a temperature of 22 ºC using the *Fold* program from version 5.7 of the *RNAstructure* package (Reuter and Mathews, 2010). Folding was carried out using the default slope and intercept parameters of 1.8 and -0.6 kcal/mol respectively, generating up to 20 structures, including a minimum free energy (MFE) structure and a set of suboptimal structures. Unconstrained *in silico* predictions were made for all RNA molecules. We also generated *in vivo* predictions, using the normalised reactivities as soft constraints. The *in vivo* predictions were made using 11138 RNA molecules, or 40.29% of the *Arabidopsis thaliana* transcriptome (Fig. S2).

	Total	**Uniquely mapped**		**Non-uniquely mapped**		**Unmapped**	
Lane	**Reads**	**Reads**	**%**	**Reads**	**Reads**	**Reads**	**%**
(-)DMS Biological Replicate 1	3.60 x 10^{7}	2.89 x 10^{7}	80.37	1.49 x 10^{6}	3.60 x 10^{7}	2.89 x 10^{7}	15.50
(+)DMS Biological Replicate 1	3.93 x 10^{7}	3.25 x 10^{7}	82.74	1.61 x 10^{6}	3.93 x 10^{7}	3.25 x 10^{7}	13.16
(-)DMS Biological Replicate 2	2.60 x 10^{7}	2.11 x 10^{7}	81.09	1.10 x 10^{6}	2.60 x 10^{7}	2.11 x 10^{7}	14.68
(+)DMS Biological Replicate 2	2.80 x 10^{7}	2.27 x 10^{7}	81.17	1.29 x 10^{6}	2.80 x 10^{7}	2.27 x 10^{7}	14.22
(-)DMS Biological Replicate 3	3.35 x 10^{7}	2.69 x 10^{7}	80.35	1.50 x 10^{6}	3.35 x 10^{7}	2.69 x 10^{7}	15.17
(+)DMS Biological Replicate 3	2.84 x 10^{7}	2.35 x 10^{7}	83.00	1.28 x 10^{6}	2.84 x 10^{7}	2.35 x 10^{7}	12.49

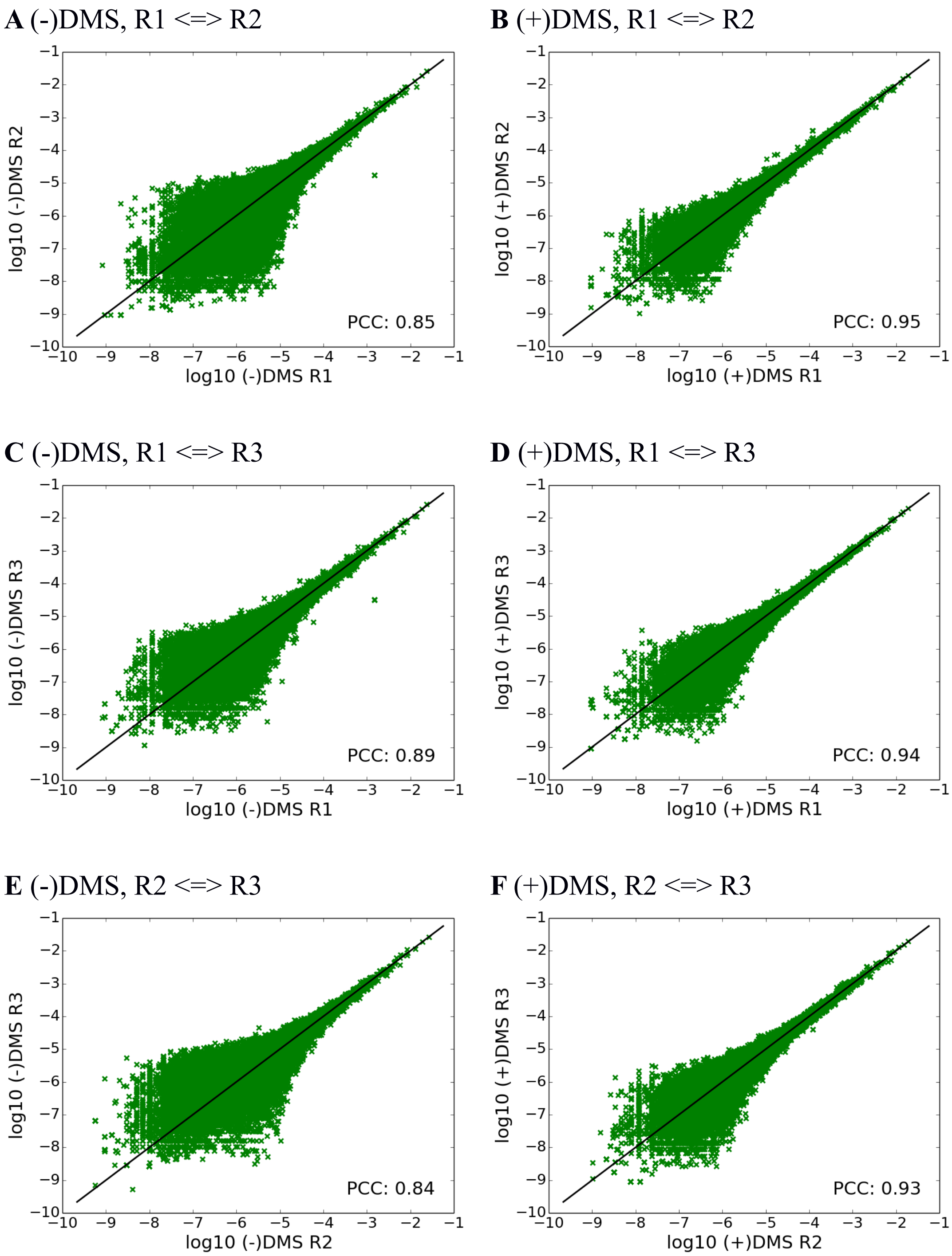
Supplementary Table S1: Counts and percentages for uniquely mapped, non-uniquely mapped, and unmapped reads, across each biological replicate.

	(+)DMS library		**(-)DMS library**	
RNA type	**Read assignments**	**Read assignments (%)**	**Read assignments**	**Read assignments (%)**
mRNA	1.19 x 10^{8}	95.99	1.13 x 10^{8}	97.22
rRNA	4.18 x 10^{6}	3.37	2.69 x 10^{6}	2.31
ncRNA	6.05 x 10^{5}	0.49	4.64 x 10^{5}	0.40
snRNA	6.74 x 10^{4}	0.054	2.91 x 10^{4}	0.025
tRNA	1.33 x 10^{3}	0.0011	9.40 x 10^{2}	0.00081
miRNA	1.24 x 10^{4}	0.010	7.11 x 10^{3}	0.0061
snoRNA	1.17 x 10^{5}	0.094	4.18 x 10^{4}	0.036
Total	1.24 x 10^{8}	100	1.16 x 10^{8}	100

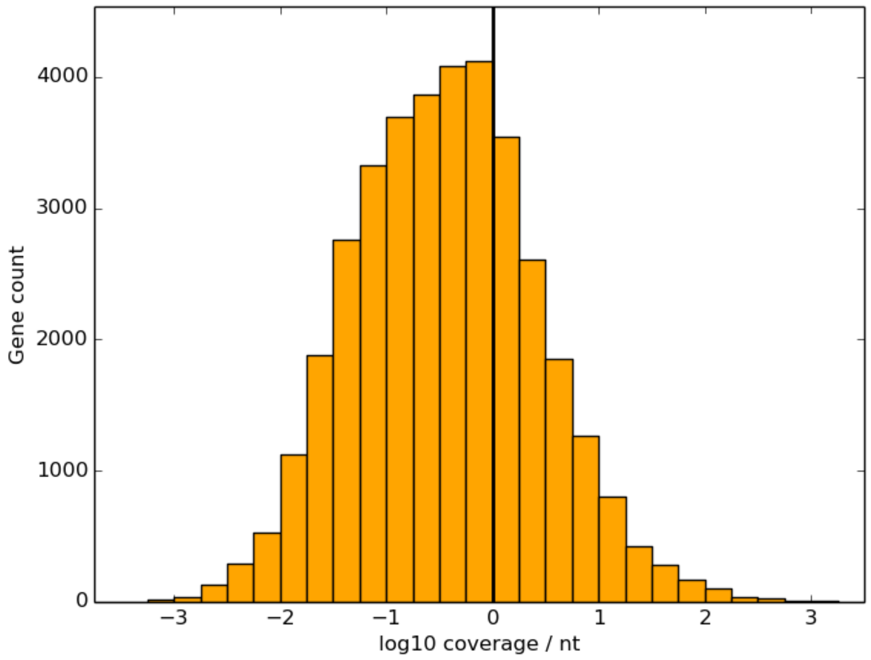
Supplementary Table S2: Raw reads and percentages for read assignments to 7 types of RNA. These values include assignments to multiple splice isoforms, hence the total number of assignments is higher than the total number of reads. Categories include RNA with coding sequence (mRNA), ribosomal RNA (rRNA), small nuclear RNA (snRNA), transfer RNA (tRNA), micro RNA (miRNA) and small nucleolar RNA (snoRNA). Non-coding RNA that is not rRNA, snRNA, tRNA, miRNA or snoRNA is indicated as ncRNA.

Nucleotide	**(+)DMS library**	**(-)DMS library**
A	43.85	21.11
C	24.38	29.58
G	14.24	18.56
U	17.54	30.75

Supplementary Table S3: Percentages of reads mapped to A, C, G and U nucleotides in the (+)DMS and (-)DMS libraries, showing enrichment of A and C nucleotides in the (+)DMS library.



Supplementary Fig. S1: Correlations between scaled read counts of biological replicates. Each point is a single transcript. A transcript's scaled read count is the read count assigned to the transcript across all bases, divided by the total number of read assignments in the entire biological replicate. These scaled read counts are placed on a log10 scale. The Pearson's correlation coefficient (PCC) of the log10 transformed scaled read counts is indicated in the bottom right corner. **(A)** Biological replicate R1 compared against R2, (-)DMS. **(B)** R1 vs. R2, (+)DMS. **(C)** R1 vs R3, (-)DMS. **(D)** R1 vs R3, (+)DMS. **(E)** R2 vs R3, (-)DMS. **(F)** R2 vs R3, (+)DMS.



Supplementary Fig. S2: Distributions of transcripts in terms of their average read count per base, in the (+)DMS set. Only A and C nucleotides are considered, since these react with DMS. Read counts are summed across biological replicates. A total of 37882 transcripts were analysed; transcripts with no coverage are not shown. Secondary structure predictions were made for 11,138 transcripts, where the average number of RT stop points per A or C base was greater than 1 in the (+)DMS set.

SUPPLEMENTARY REFERENCES

- Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10, R25.
- Ding,Y. *et al*. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature, **505**, 696–700.
- Ding,Y. *et al*. (2015) Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. Nat. Protocols, **10**, 1050–1066.
- Reuter,J.S. and Mathews,D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics, **11**, 129.