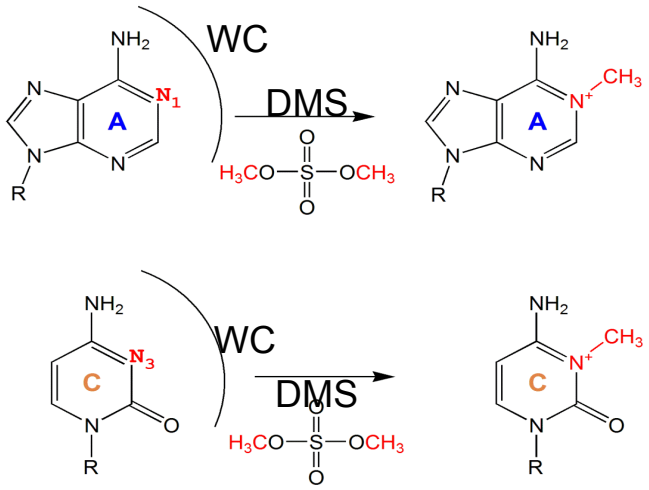
From experimental probing data to RNA secondary structure prediction

Matthew Norris

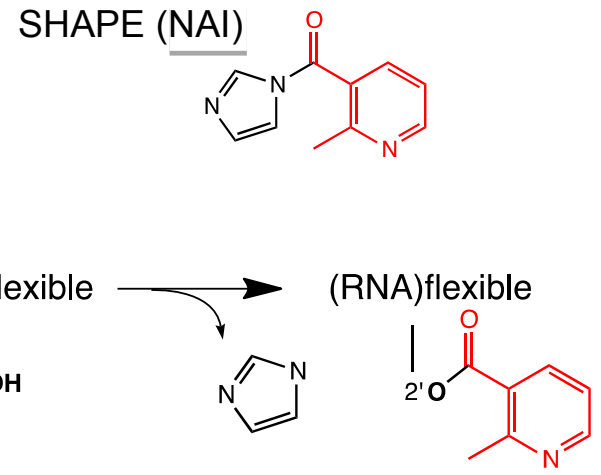
Structure probing using chemical modification

DMS



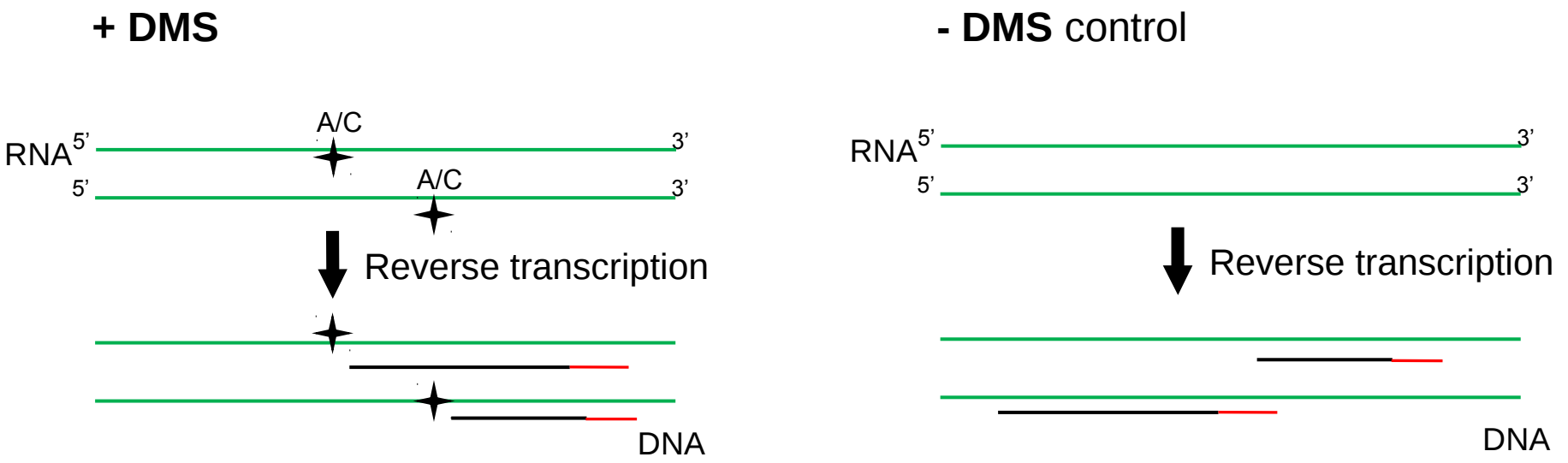
Methylates single-stranded RNA

SHAPE

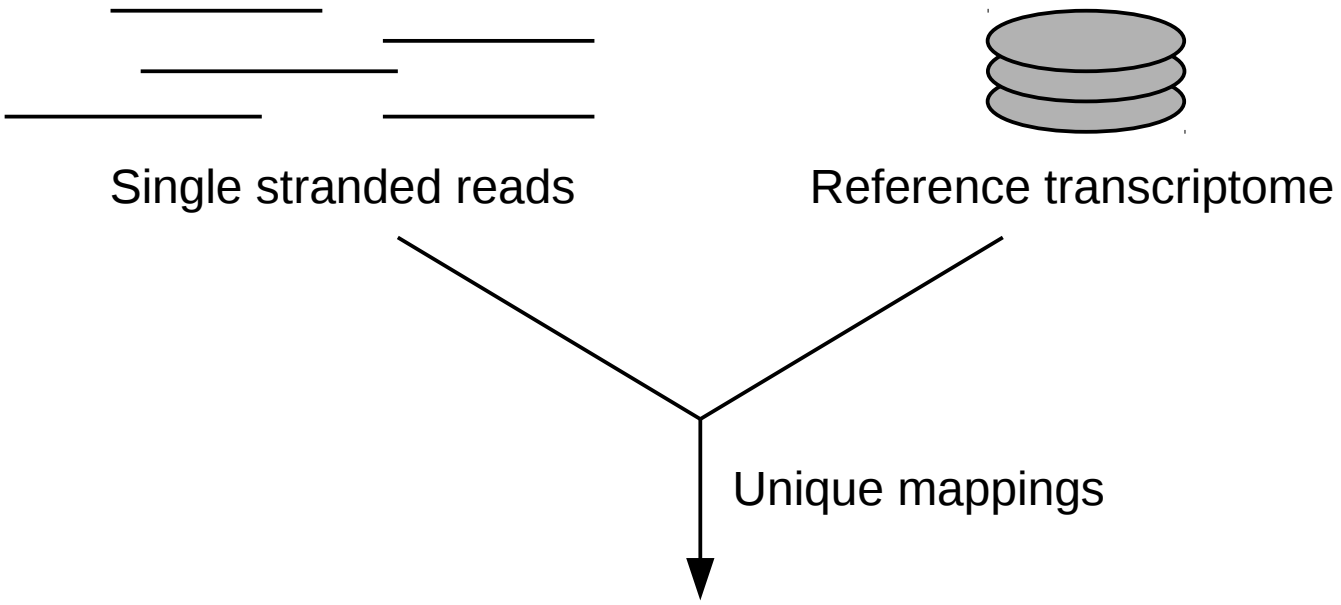


Acylates single-stranded RNA

Reverse transcription of DMS modified RNA



Mapping reads to obtain positional read counts



Normalisation - generating normalised read counts

1) Generate scaled read counts for **plus** and **minus.**

Log count / average log count

$$P(i) = \frac{\ln[P_r(i)]}{(\sum_{i=0}^{l} \ln[P_r(i)])/l} \qquad M(i) = \frac{\ln[M_r(i)]}{(\sum_{i=0}^{l} \ln[M_r(i)])/l}$$

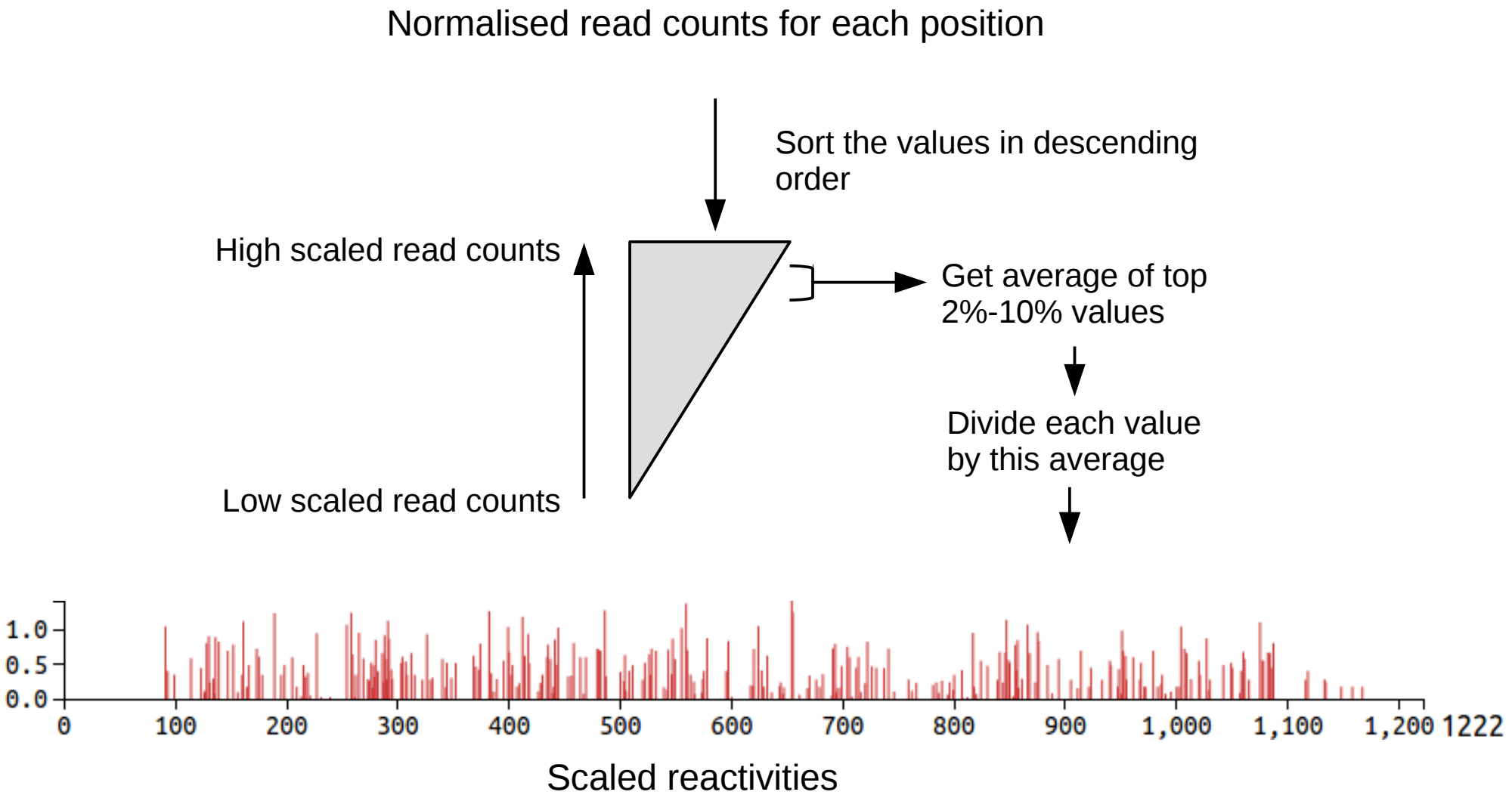
2) Subtract **minus** scaled read counts from **plus**

Produces normalised values for each position

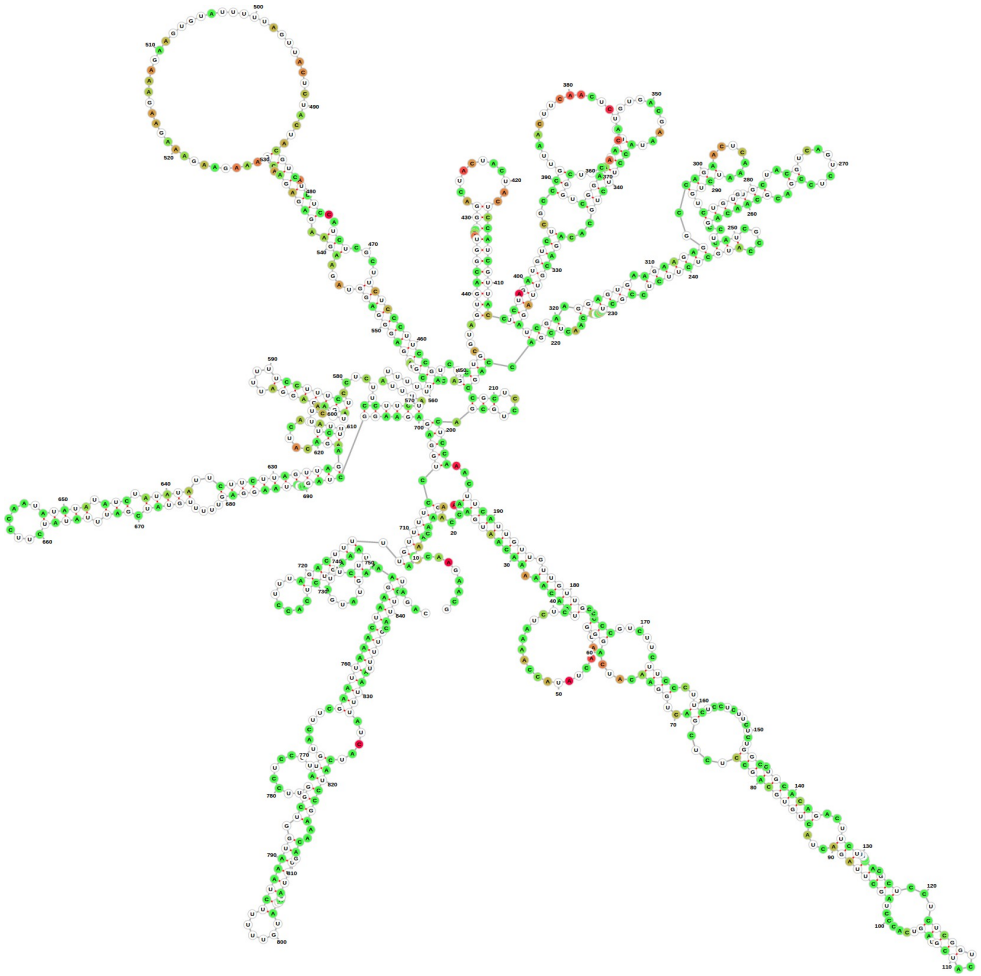
$$\theta(i) = \max((P(i) - M(i)), 0)$$

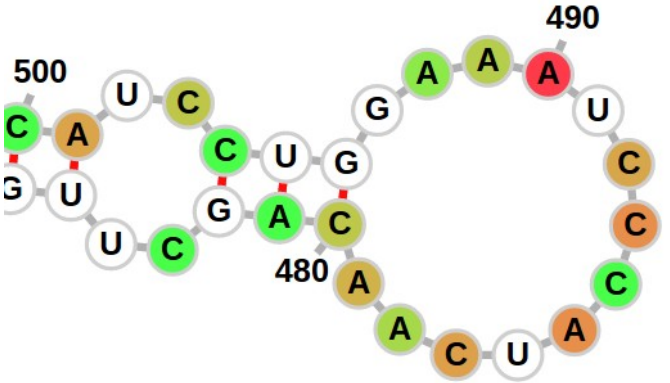
l	Sequence length
$P_r(i)$	Raw read count from **plus** lane at position *i*
$M_r(i)$	Raw read count from **minus** lane at position *i*
$P(i)$	Scaled **plus** read count at position *i*
$M(i)$	Scaled **minus** read count at position *i*
$\theta(i)$	Scaled read count with **minus** subtracted from **plus**

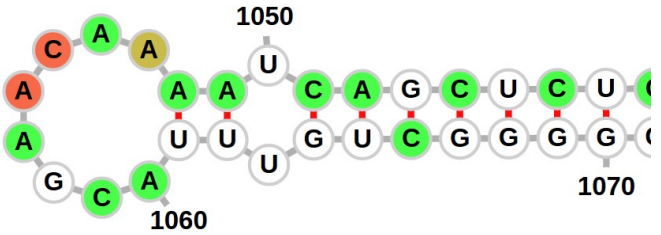
Normalisation - 2-8% scaling

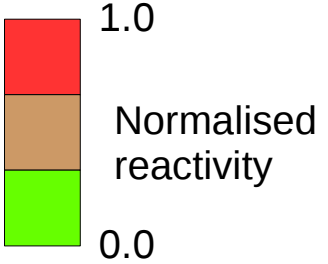


Structure prediction using probing data and *RNAstructure* method



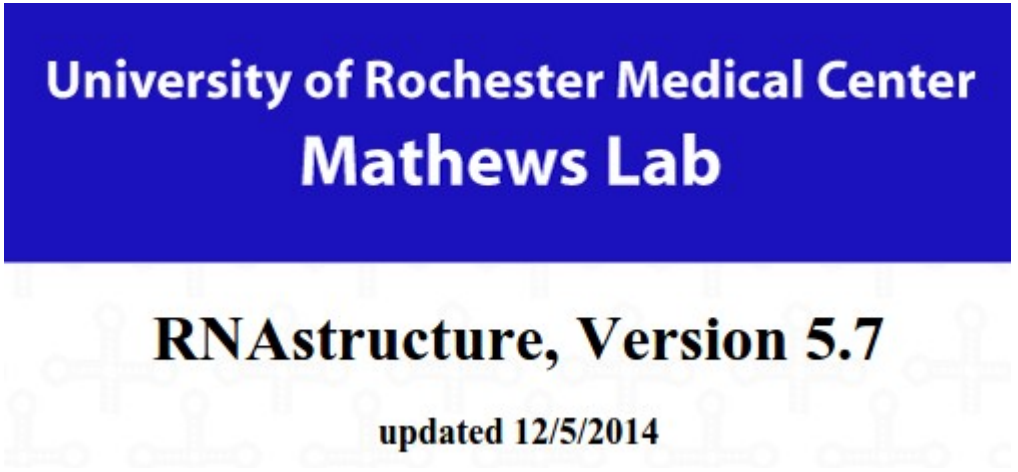




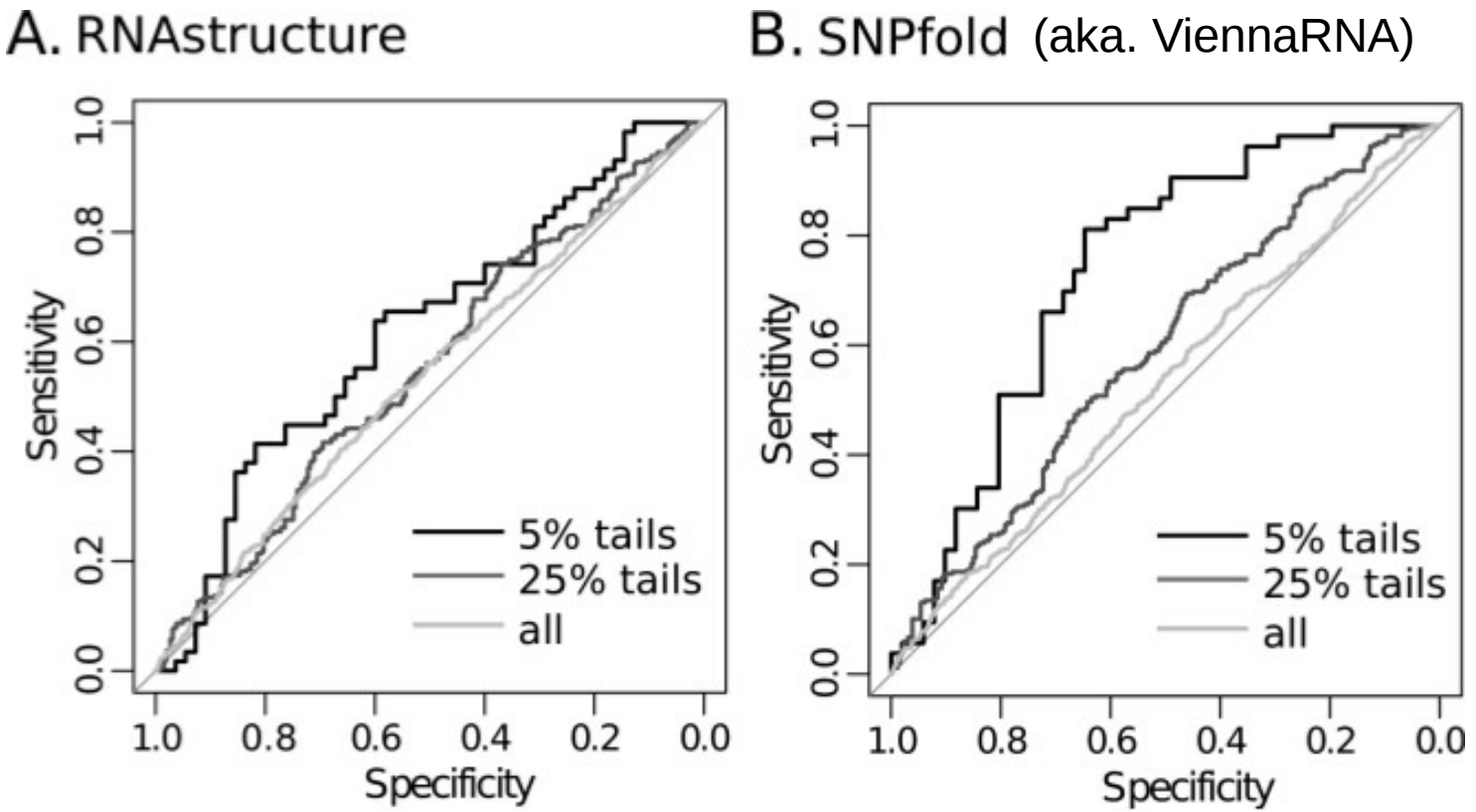


Structure prediction methods

- RNAstructure package
  - Can accept quantitative probing data as “soft” constraints.
  - Soft constraints: continuous values between 0 and 1, used as pseudo free energies in the folding model.
  - Used to generate structure predictions in FoldAtlas.
- ViennaRNA package
  - Can accept probing data, but only as “hard” constraints - “must be paired” or “must not be paired”.
  - Must use a threshold to determine single stranded or base paired categories.



Comparing RNAstructure with ViennaRNA for predicting structure-disrupting single nucleotide polymorphisms (SNPs)



Corley *et al*., 2015. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Res. 43, 1859–1868.

Using probing constraints in *RNAstructure* to improve predictive performance

Table 2. Prediction accuracies for nonribosomal RNAs

		No constraints		SHAPE	
RNA	Nucleotides	Sensitivity	PPV	Sensitivity	PPV
Yeast $tRNA^{Asp}$	75	95.2	95.2	100.0	100.0
HCV IRES domain II	95	56.5	59.1	95.7	100.0
P546 domain, group I intron	155	42.9	44.4	96.4	98.2

Deigan *et al*., 2008. Accurate SHAPE directed RNA structure determination. PNAS 106, 97-102.

Using probing constraints to improve predictive performance

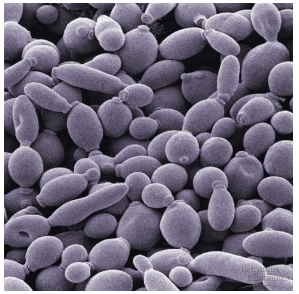
Row		PPV/Sensitivity
1	*in silico* vs. phylogenetic structure	0.27/0.31
2	*in vivo* vs. phylogenetic structure	0.41/0.45
3	*in vivo* vs. phylogenetic structure, omitting false negatives	0.50/0.52
4	ideal A/C constraint vs. phylogenetic structure	0.63/0.63
5	ideal A/C/U/G constraint vs. phylogenetic structure	0.68/0.65
6	*in vivo* vs. *in silico*	0.48/0.46

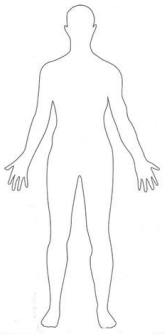
18S rRNA

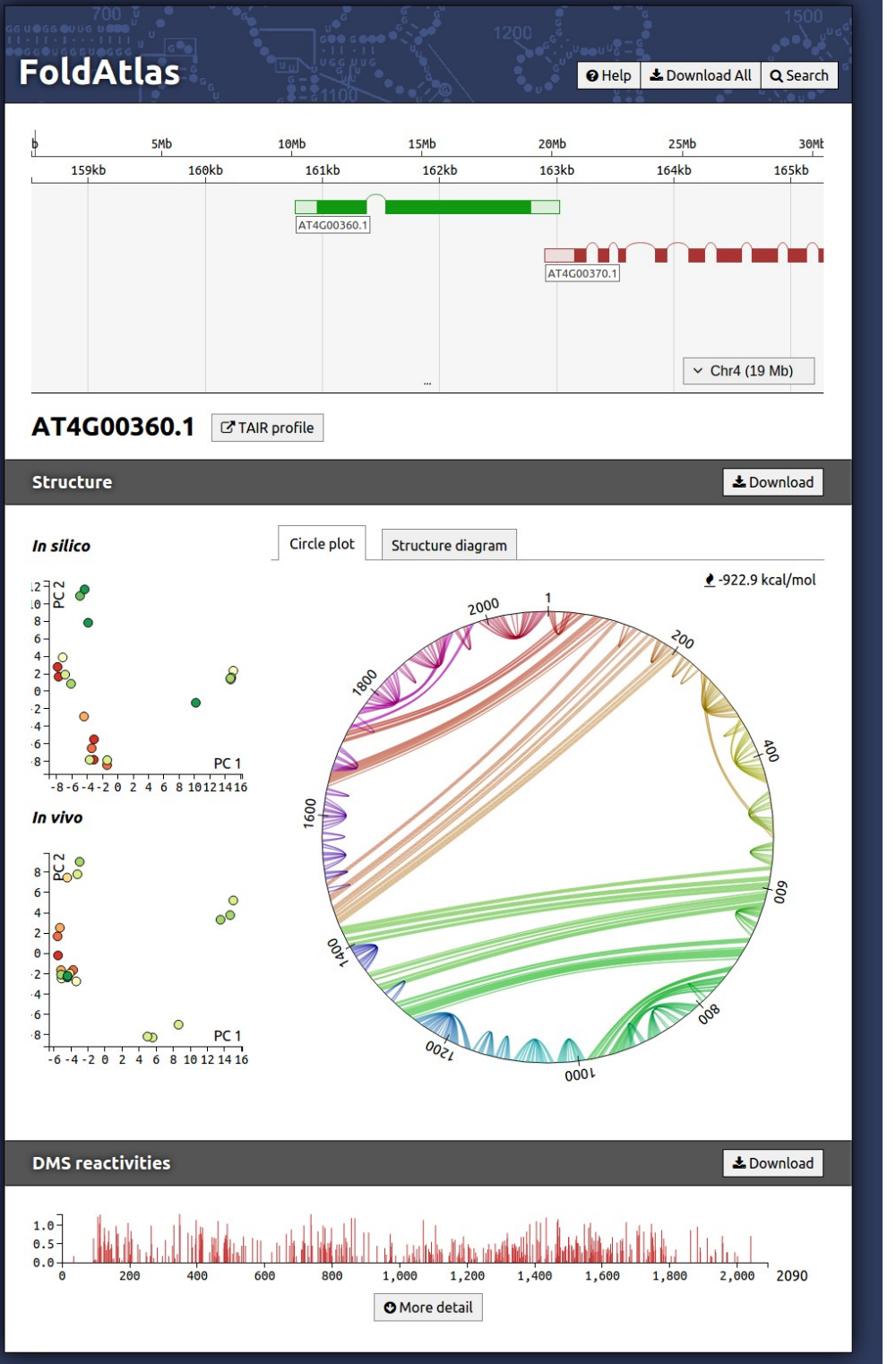
Ding *et al*., 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505, 696–700.

FoldAtlas - Aims

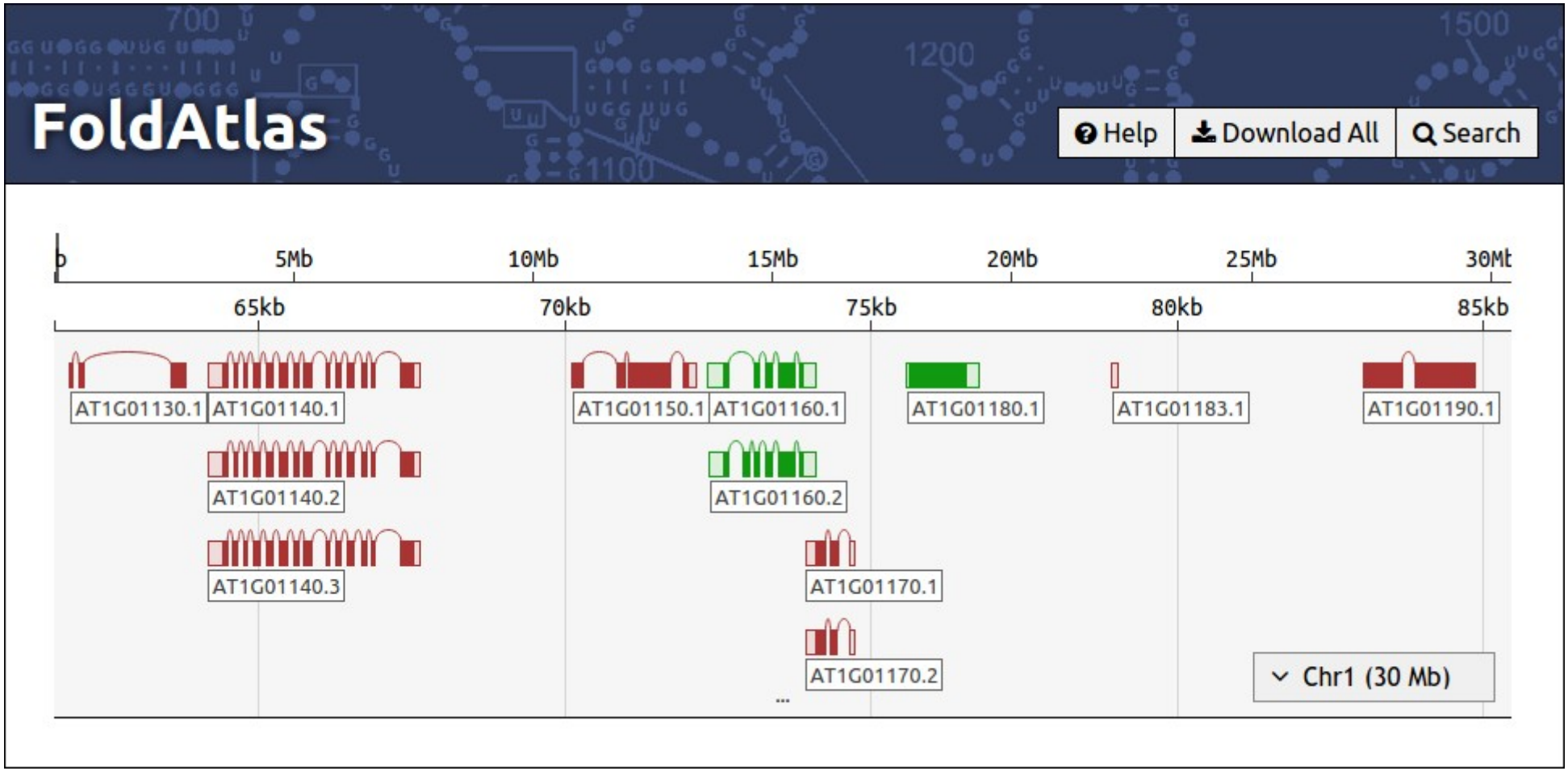
- Provide a repository for RNA structure probing data from DMS and SHAPE experiments.
- Store RNA structure predictions that have been made using probing data.
- Store data for a range of organisms and conditions.
  - Current: *A. thaliana* data
  - Future: *C. elegans*, *S. cerevisiae, …*
- Provide a pipeline for calculating normalised reactivities.
- Provide easy access to all of the data.



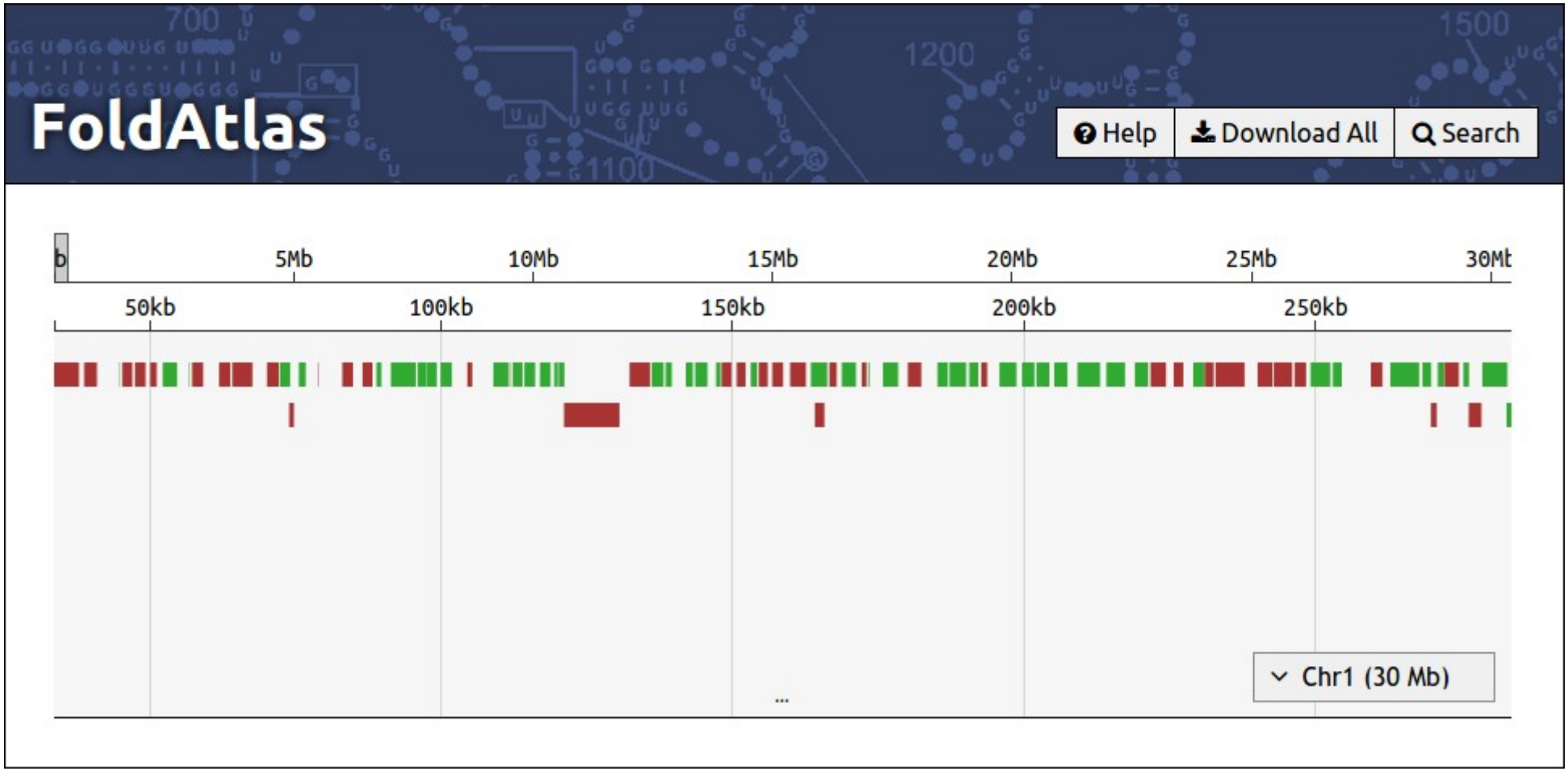




d3nome genome browser



d3nome genome browser



Transcript ID search

e.g. “AT1G01470.1”

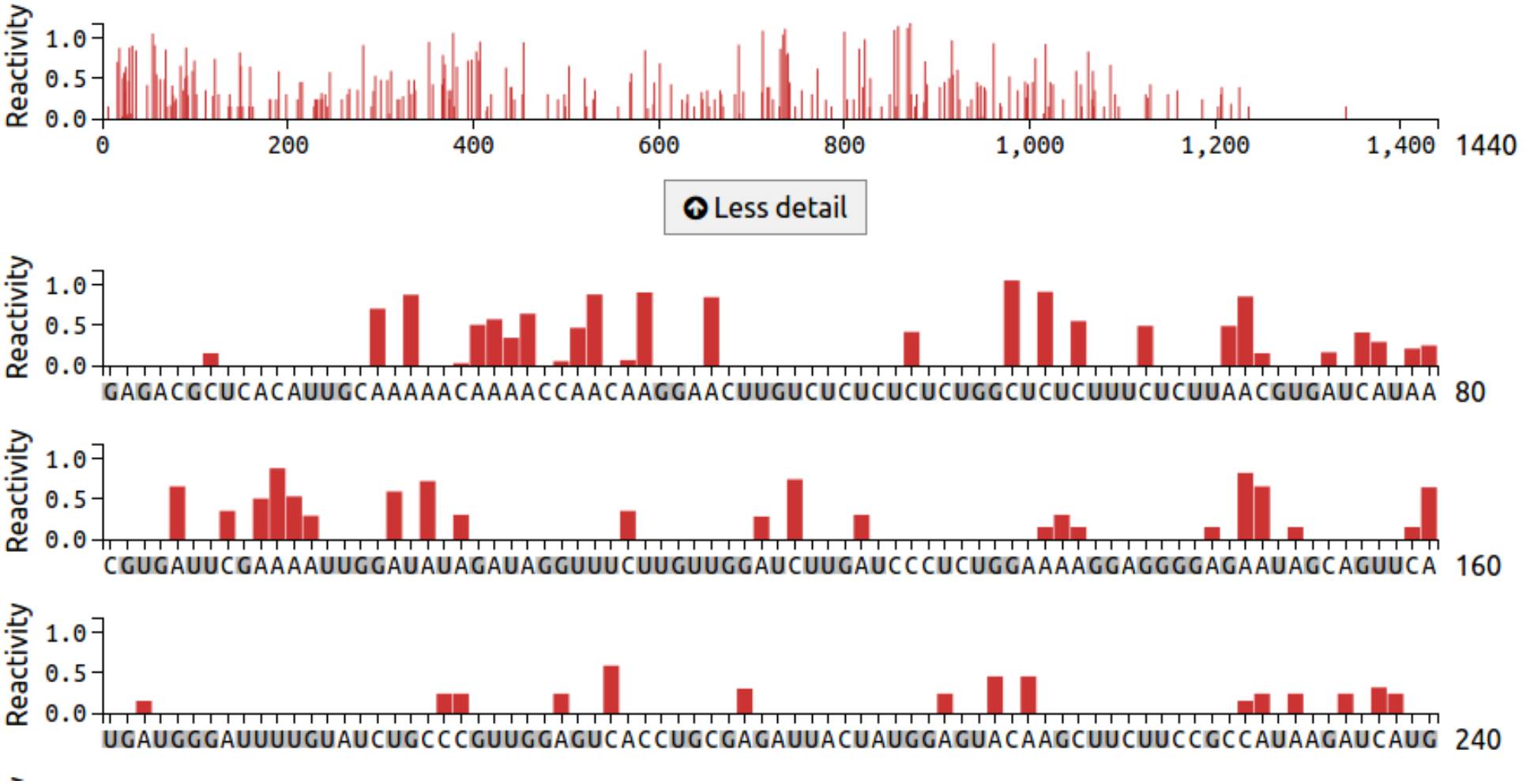
Coverage search

Transcript ID | Coverage

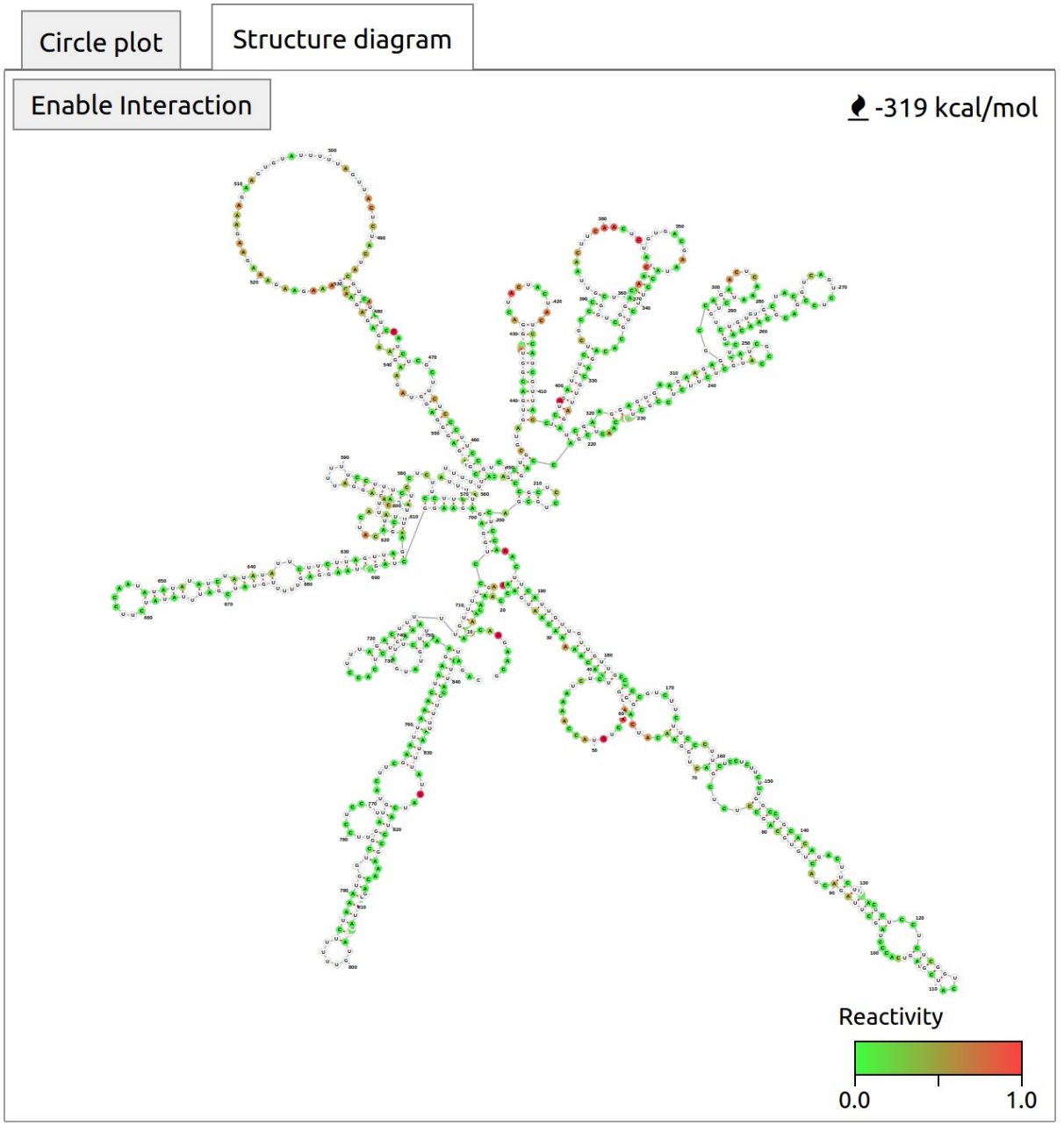
« ‹ Page 1 of 26 › »

Transcript ID	Gene length	Coverage / base	*In vivo* structure?
AT2G01021.1	102	734.127	yes
AT1G01620.1	1638	260.515	yes
AT1G01620.2	1638	244.142	yes
AT4G00360.1	2256	98.6177	yes
AT1G01120.1	1899	62.5034	yes
AT5G01530.1	1509	44.7905	yes
AT1G01100.2	1125	37.0903	yes
AT1G01100.3	1125	36.1926	yes
AT1G01100.4	1125	33.9951	yes
AT1G01100.1	1125	30.3118	yes
AT4G00720.1	3674	18.8726	yes
AT3G01500.1	3296	17.4623	yes
AT3G01500.2	3296	17.1501	yes
AT2G01150.1	949	16.6151	yes
AT4G00430.2	2178	16.2997	yes
AT5G01800.1	1941	16.2347	yes
AT3G01500.3	3296	16.0783	yes
AT4G00100.1	1229	15.5143	yes
AT3G01472.1	2008	15.3028	yes
AT3G01470.1	2008	15.3028	yes
AT4G00430.1	2178	14.3303	yes
AT1G01090.1	1802	13.0037	yes
AT4G00810.2	1248	12.5273	yes
AT5G01750.2	1562	11.8509	yes
AT1G01470.1	803	11.641	yes

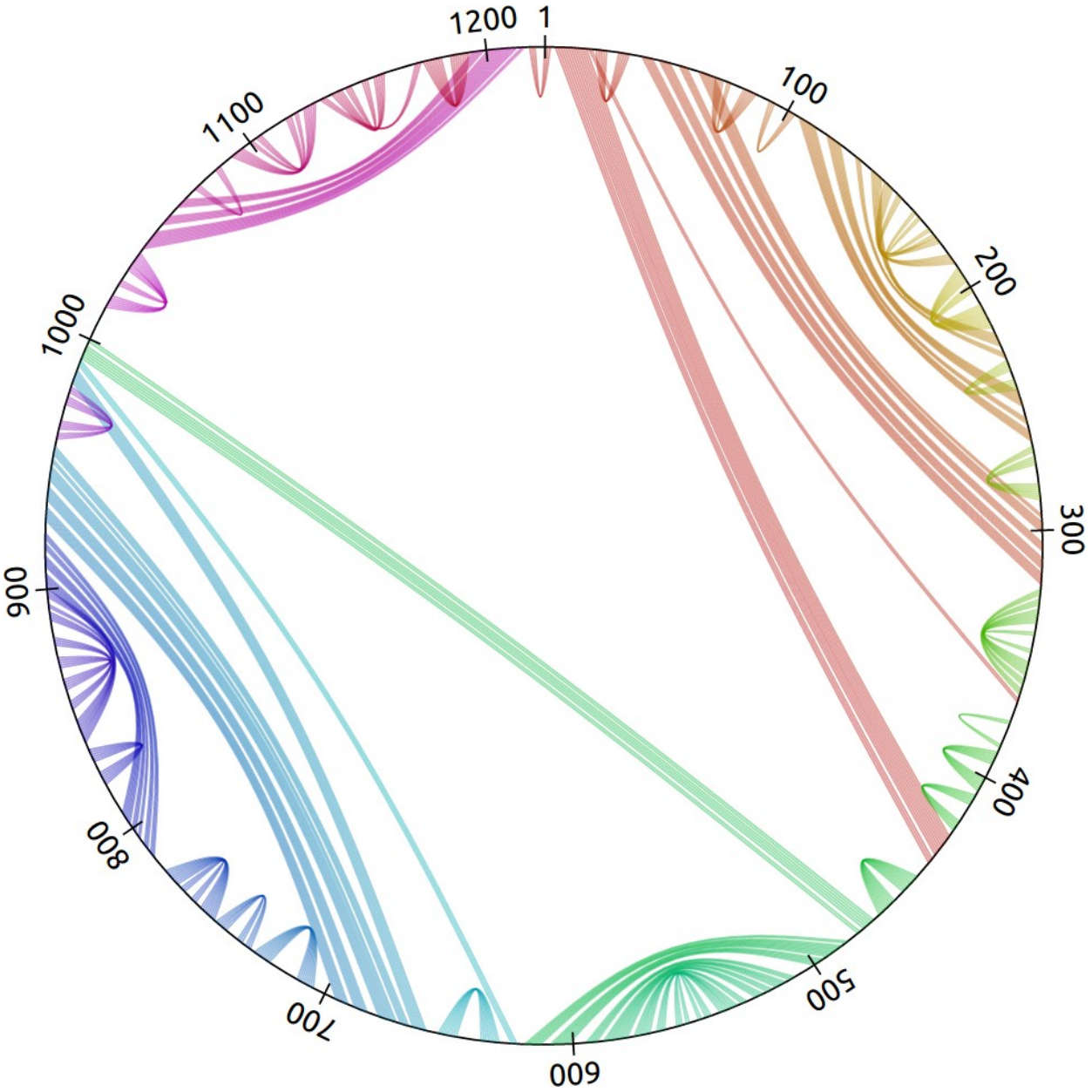
Visualising normalised DMS reactivities



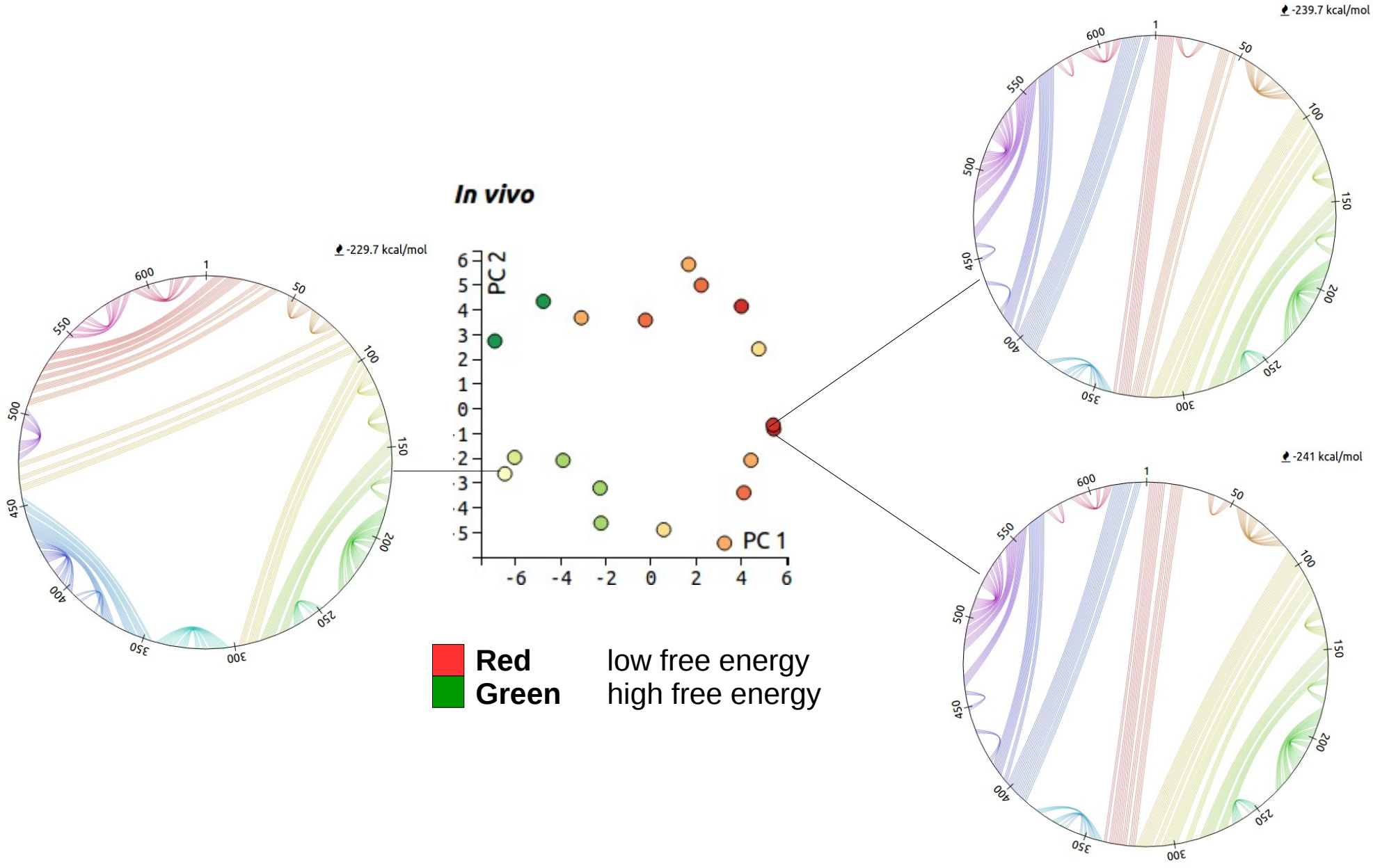
Structure visualisation using an interactive structure diagram (*forna* package)



Structure visualisation using a circle plot



Structure similarity visualisation using principal components analysis (PCA)



Downloading normalised DMS reactivities

1) The nucleotide position. Numbers start from 1.
2) The base type letter.
3) The reactivity. 0 indicates no reactivity whilst 1 or more indicates the strongest reactivity. NA indicates a G or U nucleotide, where DMS reactivities are not relevant.

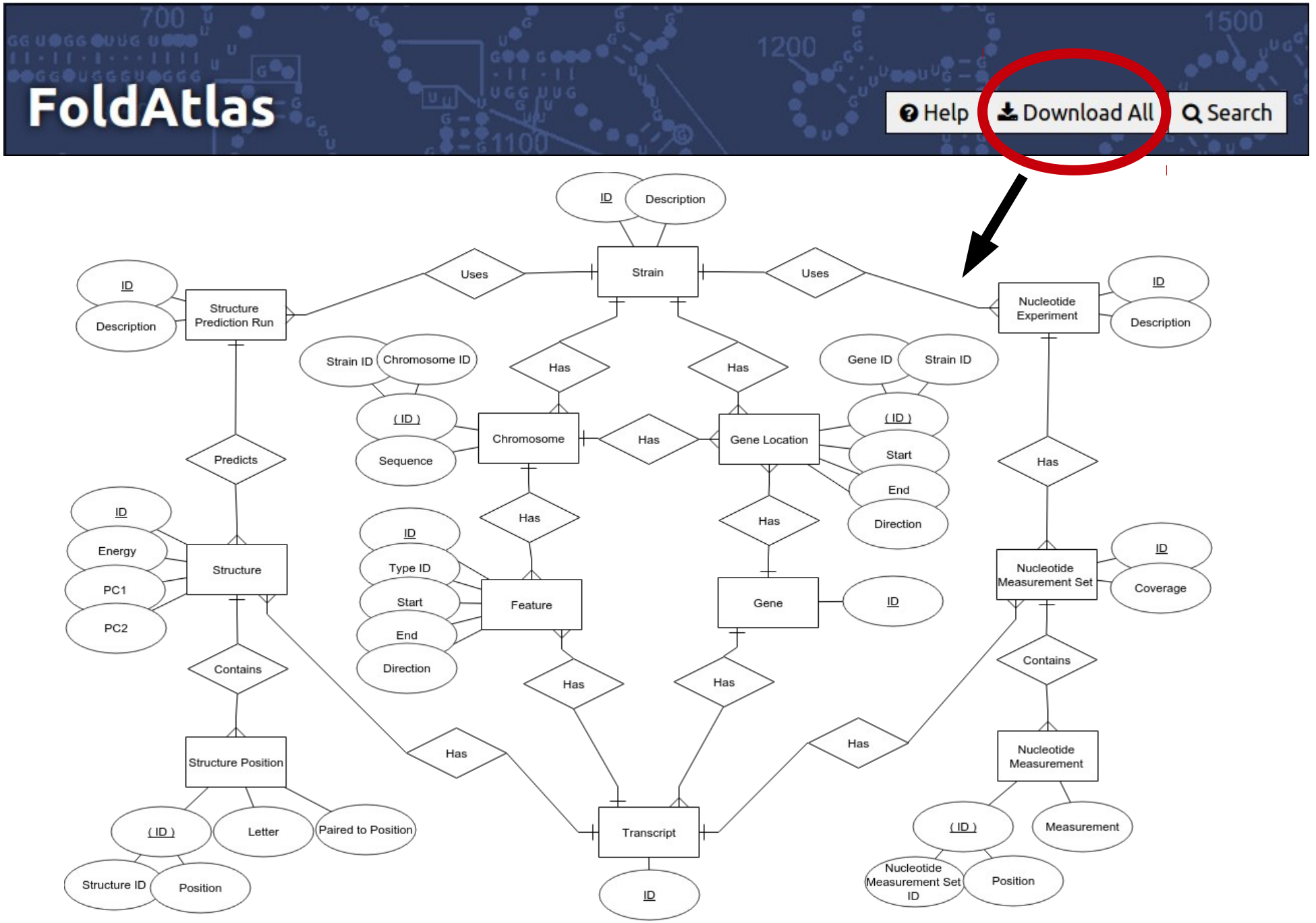
1	2	3
75	U	NA
76	G	NA
77	U	NA
78	A	0.492188
79	G	NA
80	A	0.767697
81	A	0.940973
82	A	0.45815
83	A	0.0
84	G	NA
85	A	0.823105

Downloading structure predictions

1	2	3	4	5	6	7	8	9
...								
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	U	1294	130
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	C	1295	0
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	C	1296	126
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	A	1297	125
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	C	1298	124
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	C	1299	0
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	A	1300	121
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	A	1301	120
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	C	1302	119
5273	In vivo	AT1G01010.1	-850.4	9.10729	-1.76294	A	1303	0
...								

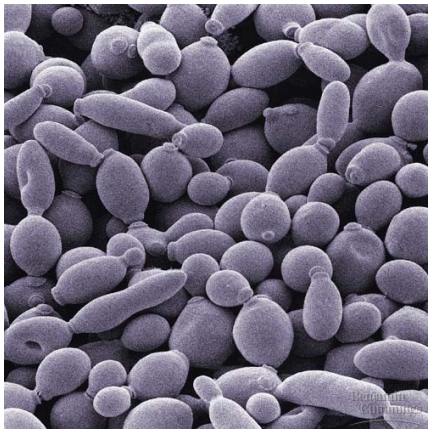
1) A unique identifier for the structure, used internally by FoldAtlas
2) Description of the prediction method; can be in silico or in vitro.
3) The TAIR transcript indentifier.
4) Free energy estimate, in kcal/mol, generated by RNAstructure.
5-6) Principle components 1 and 2.
7) Nucleotide letter.
8) Position of this nucleotide.
9) The position that this nucleotide pairs to. 0 indicates no base pairing.

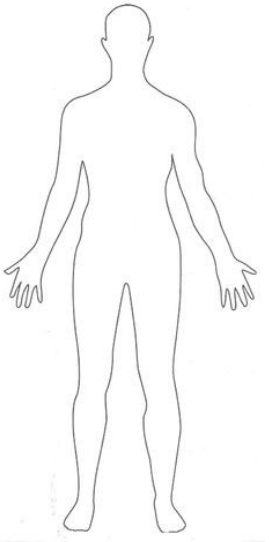
Downloading the entire FoldAtlas MySQL database



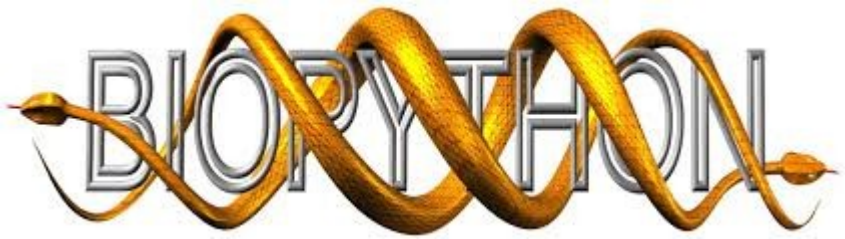
Future work

- Add SHAPE data.
- Add more normalisation methods.
- Add other data visualisations.
- Add more species.
- Add other deep sequencing data (e.g. ribosome profiling data).

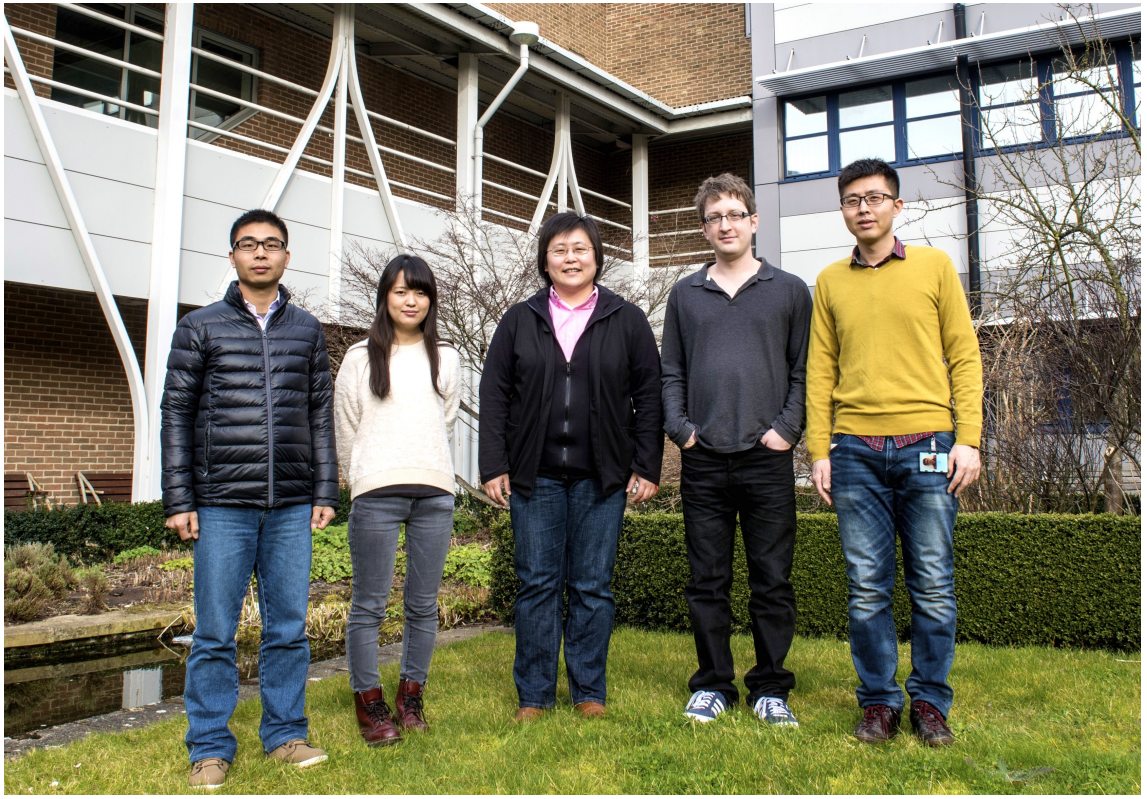




Technology used



Acknowledgements



- Yiliang Ding
- Xiaofei Yang
- Qi Liu
- Matthew Hartley

BBSRC
20 Years of Pioneering
Great British Bioscience

John Innes Centre
Unlocking Nature's Diversity