Supplemental Data

Cell, Volume *133*

Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*

Ryan Lister, Ronan C. O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker

Supplemental Experimental Procedures

Plant growth

All plants were grown in potting soil (Metro Mix 250; Grace-Sierra, Boca Raton, FL) at 23˚C under a 16-hour light/8-hour dark cycle. Immature (unopened) flower buds were removed and immediately frozen in liquid nitrogen.

MethylC-seq library generation

Genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA), and 5 µg of was fragmented by sonication to 50-500 bp with a Bioruptor (Diagenode Sparta, NJ), followed by end repair and ligation of methylated adapters provided by Illumina (Illumina, San Diego, CA) as per manufacturer's instructions for gDNA library construction. 100-200 ng of adapter-ligated gDNA of 120-170 bp was isolated by agarose gel electrophoresis, and subjected to two successive treatments of sodium bisulfite conversion using the EpiTect Bisulfite kit (Qiagen, Valencia, CA), using the subsequent FFPE purification step, as outlined in the manufacturer's instructions. The reaction was then purified once more using the PCR purification kit (Qiagen, Valencia, CA). Five ng of bisulfite-converted, adapter-ligated DNA molecules were enriched by 18 cycles of PCR with the following reaction composition: 2.5 U of uracil-insensitive *PfuTurboC*$_x$ Hotstart DNA polymerase (Stratagene), 5 µl 10X *PfuTurbo* reaction buffer, 25 µM dNTPs, 1 µl Primer 1.1, 1 µl Primer 2.1 (50 µl final). The thermocyling was as follows: 95˚C 2 min, 98˚C 30 sec, then 18 cycles of 98˚C 10 sec, 65˚C 30 sec and 72˚C 30 sec, completed with one 72˚C 5 min step. The enriched library was purified with the PCR purification kit (Qiagen, Valencia, CA)and quantity and quality examined by spectrophotometry, gel electrophoresis, and limited sequencing of cloned library molecules. A schematic of this procedure is presented in Figure S17.

Following isolation of adapter-ligated gDNA, three alternative bisulfite conversion methods were used to determine the optimal approach for whole-genome bisulfite sequencing. Firstly, the methylSEQr bisulfite conversion kit (Applied Biosystems, Foster City, CA) was used as per manufacturer's instructions. Secondly, the CpGenome Universal DNA modification kit (Upstate, Temecula, CA) was used as described by Meissner *et. al*. (2005), with the following modifications: alkali denaturation was performed for 20 min at 55 ˚C, the total reaction volume was 810 µl due to addition of 0.22 g urea, the mixture was incubated for 24 h at 55 ˚C. Thirdly, the bisulfite conversion protocol described by Clark *et.al.* (2006) was performed. Following bisulfite conversion, the libraries were enriched by 18 cycles of PCR and sequenced as described above. The sodium bisulfite non-conversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the chloroplast genome.

smRNA-seq library generation

Total RNA was isolated using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA). Immediately following RNA precipitation, the flow through from the anion-exchange chromatography column was further precipitated in another 2.5 volumes of 100% ethanol (smRNA fraction). The smRNA fraction was further purified by a phenol-chloroform extraction and an additional ethanol precipitation. Small RNAs were resolved by electrophoresis of 2.5 μg of the smRNA fraction and 7.5 μg of total RNA on 15% polyacrylamide gels containing 7 M urea in TBE buffer (45 mM Tris-borate, pH 8.0, and 1.0 mM EDTA). A gel slice containing RNAs of 15 to 35 nucleotides (based on the 10 base pair ladder size standard (Invitrogen, Carlsbad, CA)) was excised and eluted in 0.3 M NaCl rotating at room temperature for 4 hours. The eluted RNAs were precipitated using ethanol and resuspended in diethyl pyrocarbonate–treated deionized water. Gel-purified smRNA molecules were ligated sequentially to 5' and 3' RNA oligonucleotide adapters using T4 RNA ligase (10 units/μL) (Promega, Madison, WI). The 5' RNA adapter (5' - GUUCAGAGUUCUACAGUCCGACGAUC - 3') possessed 5' and 3' hydroxyl groups. The 3' RNA adapter (5'-pUCGUAUGCCGUCUUCUGCUUGidT-3') possessed a 5' mono-phosphate and a 3' inverted deoxythymidine (idT). The smRNAs were first ligated to the 5' RNA adapter. The ligation products were gel eluted and ligated to the 3' RNA adapter as described above. The final ligation products were then used as templates in a reverse transcription (RT) reaction using the RT-primer (5' - CAAGCAGAAGACGGCATACGA - 3') and Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA). This was followed by a limited (15 cycle) PCR amplification step using the PCR reverse (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3') and forward (5'-CAAGCAGAAGACGGCATACGA-3') primers and Phusion hot-start high fidelity DNA polymerase (New England Biolabs, Cambridge, MA). All oligonucleotides were provided by Illumina (San Diego, CA). The amplification products were separated by electrophoresis on a 6% polyacrylamide gel in TBE buffer, eluted in 0.3 M NaCl rotating at room temperature for 4 hours, precipitated using ethanol, and resuspended in nuclease-free water. A schematic of this procedure is presented in Figure S15.

mRNA-seq library generation

Total RNA was isolated using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA) and treated with DNaseI (Qiagen) for 30 min at room temperature. following ethanol precipitation the 18S and 28S rRNA molecules were depleted from 20 µg of total RNA in three sequential Ribominus (Invitrogen, Carlsbad, CA) reactions as per manufacturer's instructions, using 6 plant-specific biotinylated LNA oligonucleotide rRNA probes supplied by (Invitrogen, Carlsbad, CA). The 5' cap was removed from the rRNA-depleted RNA by treatment with 10 U/µl Tobacco Acid Pyrophosphatase for 1.5 h at 37˚C. This and all subsequent enzymatic reactions involving RNA used contained 2.5-4 U/µl RNaseOut ribonuclease inhibitor (Invitrogen, Carlsbad, CA). The RNA was purified by phenol:chloroform extraction and ethanol precipitation. This and all subsequent ethanol precipitations contained 20-40 µg/ml nuclease-free glycogen (Ambion, Austin, TX). De-capped RNA was fragmented by metal hydrolysis in 1X fragmentation buffer (Affymetrix, Santa Clara, CA) for 35 min at 94 ˚C then cooled on ice for 2 min and ethanol precipitated. The fragmented RNA was dephosphorylated using 10 U/µl Calf intestinal phosphatase (New England Biolabs, Cambridge, MA) for 1 h at 37˚C, then 10 µl Gel loading Buffer II (Ambion, Austin, TX) added, heated at 65˚C for 5 min, cooled on ice and then separated on a 10% polyacrylamide gel containing 7 M urea in TBE buffer (45 mM Tris-borate, pH 8.0, and 1.0 mM EDTA) by electrophoresis at 150 V for 2 h at 4˚C. The gel was stained in SYBR Gold (Invitrogen, Carlsbad, CA), and a gel slice containing RNAs of 35 to 50 nucleotides was excised, crushed, and the RNA eluted in 0.3 M NaCl rotating at room temperature for 4 hours. The eluted RNAs were ethanol precipitated and resuspended in nuclease-free water, after which the RNA fragments were heated to 65˚C for 5 min, cooled on ice for 5 min and then ligated to the Illumina 3' RNA oligonucleotide adapter (see smRNA library construction above) using 10 U/µl T4 RNA ligase (Promega, Madison, WI) in 10% DMSO, incubated at 20˚C for 6 h then 4 ˚C for 4 h. Nucleic acids in the ligation reaction were separated by electrophoresis and a gel slice containing 3' adapter-ligated RNA molecules from 50 to 80 nucleotides was excised and the RNA eluted and precipitated as described above. The gel-purified RNA was resuspended in nuclease-free water then phosphorylated in a reaction containing 1 U/µl T4 polynucleotide kinase (New England Biolabs, Cambridge, MA) and 1 mM ATP (Illumina, San Diego, CA) for 1 h at 37 ˚C. After purification by phenol:chloroform extraction and ethanol precipitation the 5' phosphorylated RNA fragments were ligated to the Illumina 5' RNA oligonucleotide adapter (see smRNA library construction above) under the same conditions used for the 3' adapter ligation. Nucleic acids in the ligation reaction were separated by electrophoresis and a gel slice containing 5' and 3' adapter-ligated RNA molecules from 80 to 125 nucleotides was excised and the RNA eluted as described above. The size-selected ligation products were then used as templates in a reverse transcription (RT) reaction, followed by a limited (20 cycle) PCR amplification step (see smRNA library construction above). The amplification products were separated by electrophoresis on a 6% polyacrylamide gel in TBE buffer and the 80 to 125 bp band excised. This cDNA was eluted in 1 X gel elution buffer (Illumina, San Diego, CA) rotating at room temperature for 3 hours, ethanol precipitated and resuspended in 15 µl elution buffer (Qiagen, Valencia, CA). A schematic of this procedure is presented in Figure S16.

High-throughput sequencing

MethylC-seq, smRNA-seq and mRNA-seq libraries were sequenced using the Illumina Genetic Analyzer (GA) as per manufacturer's instructions, except sequencing of methylC-seq libraries was performed for 49-56 cycles to yield longer sequences that are more amenable to unambiguous mapping to the Arabidopsis genome sequence.

Processing Illumina GA sequences

Sequence information was extracted from the image files with the Illumina Firecrest and Bustard applications and mapped to the Arabidopsis (Col-0) reference genome sequence (TAIR 7) with the Illumina ELAND algorithm. ELAND aligns 32 bases or shorter reads, allowing up to two mismatches to the reference sequence. For reads longer than 32 bases, only the first 32 bases will be used for alignment, while the remaining sequence will be appended regardless of similarity to the reference sequence. A Perl script was used to truncate the appended sequence at the point where the next four bases contain two or more errors relative to the reference sequence. For reads that aligned to multiple positions in the reference genome at 32 bases we utilized a new version (1.080214) of the cross_match algorithm (P. Green personal communication) to map these non-unique reads to a reference sequence that was repeat-masked for 50 bp perfect repeat sequence.

Mapping methylC-seq sequences

When mapping reads generated from bisulfite converted genomic DNA, converted cytosines will score as a mismatch and will adversely affect the ELAND alignment ability. Therefore reads were mapped against computationally bisulfite converted and non-converted genome sequences. As bisulfite conversion of cytosine to thymidine results in non-complementarity of the two strands of a DNA duplex, reads were mapped against two converted genome sequences, one with cytosine changed to thymidine to represent a converted Watson strand, and a second with guanine changed to adenosine to represent the converted Crick strand. Reads that aligned to multiple positions in the three genomes were aligned to an unconverted genome using cross_match (see above).

Mapping smRNA-seq reads

Prior to alignment of the smRNA reads, a custom Perl script was used to identify the first seven bases of the 3’ adapter sequence, and the read was truncated up to the junction with the adaptor sequence. Each of the reads was then mapped to the genome with BLAST using a word size of 10 and expectation value of 10. Only perfect matches were accepted, as these shorter reads will have a higher tendency to falsely map than longer reads. No further analysis was performed on reads that do not contain the adapter sequence, as their size class could not be determined precisely.

Mapping mRNA-seq reads

In order to avoid omitting unannotated transcripts, 36 nucleotide transcriptome reads were aligned to the Arabidopsis reference genome sequence (TAIR 7) with the ELAND algorithm.

Post-sequencing processing of methylC-seq reads

To reduce clonal bias, short reads sequences that mapped to the same start position were collapsed into a single consensus read. Where a base call within the consensus was contentious, the base to be retained was randomly selected. A detailed statistical analysis of the clonal read bias is presented in the Supplementary Materials.

To identify the presence of a methylated cytosine, a significance threshold was determined at each base position using the binomial distribution, read depth and pre-computed error rate based on combined bisulfite conversion failure rate and sequencing error. Methylcytosine calls that fell below the minimum required threshold of percent methylation at a site were rejected. This approach ensured that no more than 5% of methylcytosine calls were false positives.

AnnoJ: A web 2.0 browser for visualization of wide range of genome data

We have developed an open-source web-based application called Anno-J for visualization of genomic data. Anno-J represents a significant step forward from existing web-based genome browsers, having been built using modern Web 2.0 technologies (REST, AJAX and DHTML) specifically to handle large amounts of data from next-generation sequencing projects. It is a distributed application, leveraging the ExtJS framework (http://www.extjs.com) and will run without manual installation in W3C compliant web-browsers. Visual presentation of data may be readily modified using CSS and track data may be sourced directly from any remote provider accessible via the internet.

The primary advantage of Anno-J over existing web-based genome browsers is simplicity of interaction for all parties. The program has been designed to cleanly separate the roles of user, engineer, website administrator, database administrator and graphic designer, and to lower barriers to entry for each. Language agnosticism ensures that back-end developers may use any server-side configuration and are not required to install specific server side software. Data structure is also agnostic, ensuring that database administrators do not have to morph data to suit the needs of the program. CSS usage permits designers to control the look and feel of tracks without having to master idiosyncratic presentation logic. Engineers can create new track plug-ins using defined interfaces without having to master database administration, graphic design or the core of the program. Finally, website administrators may quickly create instances of Anno-J by assembling an index page that points to remote program components, without having to understand how remote components were designed.

Supplemental Legends

Figure S1. Cytosine coverage for each genotype for methylC-seq. The average percentage of cytosines in each strand of the nuclear genome covered by at least 2 non-clonal, unambiguously aligned reads for each genotype and cytosine context.

Figure S2. Density of DNA methylation in wild type nuclear chromosomes 2 through 5. The density of methylcytosines of each context throughout each chromosome in 50 kb segments is presented.

Figure S3. Number of methylcytosines in DNA methyltransferase and DNA demethylase mutant plants at bases of equivalent sequencing read depth. Comparison of the number of methylcytosines identified in each genotype for each context, where the methylation status of a reference C position was only interrogated if the read depth for all four genotypes was between 6-10.

Figure S4. Ratio of methylcytosine density in each mutant versus wild type in nuclear chromosomes 2 trough 5. The ratio of the number of methylcytosines in each mutant versus Col-0 over 200 kb was calculated, where read depth was 6-10 in both mutant and Col-0. The

horizontal line represents Col-0, while the plotted line represents percentage methylation in the mutant versus Col-0.

Figure S5. Example of the increase in genic CHG methylation in met1. Gene body CHG hypermethylation in *met1*. Tracks are shown for gene annotation and DNA methylation sites, for which the color reflects the methylation context, as indicated. Abbreviations: mC, methylcytosine.

Figure S6. Hypermethylation in *rdd*.

A) - D) Regions of hypermethylation identified in *rdd*, indicated by arrows.

E) The positions of 1 kb regions in chromosome 1 that contain greater than 2 fold more DNA methylation in *rdd* relative to Col-0, represented as vertical bars. Tracks are shown for gene annotation and DNA methylation sites. The color of the DNA methylation bars represents the sequence context, as indicated. Abbreviations: mC, methylcytosine.

Figure S7. Relative abundance of smRNA sequences of each length and overlap with methylcytosines.

A) Percentages indicate the fraction of sequenced smRNAs of each size class relative to the total number of smRNAs sequenced for each genotype.

B) Number of genomic locations matched by unique smRNAs and the number of methylcytosines within each location.

C) Number of genomic locations matched by all smRNAs and the number of methylcytosines within each location.

Figure S8. DNA methylation associated with trans-acting small RNA generating loci. The smRNAs aligning to the tasiRNA generating loci are coincident with DNA methylation that is dependent on MET1 and/or DRM1, DRM2, CMT3. Sites of DNA methylation are indicated and the color reflects the methylation context, as indicated. Tracks are shown for gene annotation, DNA methylation sites and smRNAs. The color of the DNA methylation bars represents the sequence context, as indicated. Abbreviations: mC, methylcytosine.

Figure S9. Nucleotide distribution flanking and throughout sequences to which smRNAs align. Nucleotide fequency and distribution flanking and within uniquely aligning A) 21, B) 22, and C) 23-mer smRNAs. Abbreviations: mC, methylcytosine on the sense strand relative to the smRNA sequence; mC*, methylcytosine on the antisense strand.

Figure S10. Select examples of transposons/pseudogenes that display dramatic accumulation of new 21-mer smRNAs in *met1*.

Tracks are shown for gene annotation, DNA methylation sites and smRNA. smRNAs are colored internally by their uniqueness (red = maps to a single location, greyscale = maps to multiple locations), a surrounding box indicates the size class (orange = 21mer, black = 24mer), and the shading represents the copy number (darker = more copies, lighter = fewer copies). Abbreviations: mC, methylcytosine.

Figure S11. Unanotated transcripts dicovered by mRNA-seq.

Strand-specific shotgun sequencing of the Arabidopsis transcriptome revealed previously unannotated transcripts, as exemplified in panels A and B. Tracks are shown for gene annotation and mRNA-seq.

Figure S12. Mutator-like transposon DNA sequences were aligned with a progressive alignment algorithm {Feng, 1987 #209} with a gap open cost of 10 and gap extension cost of 1. The phylogenetic tree was constructed using a neighbor-joining algorithm. Transposons that displayed higher transcript abundance in met1 are highlighted according to the changes measured in the abundance of smRNAs and DNA methylation in each context, as indicated by the code prefix of each gene identifier, where U = up, D = down, E = equivalent. Code: position 1 = mRNA abundance, position 2 = smRNA abundance, position 3 = CG methylation abundance, position 4 = CHG methylation abundance, position 5 = CHH methylation abundance.

Figure S13. Transposon hypermethylation in *rdd*.

Examples of transposons that were observed to have higher densities of DNA methylation in the DNA demethylase mutant, rdd. Tracks are shown for gene annotation and DNA methylation sites, for which the color reflects the methylation context, as indicated. Abbreviations: mC, methylcytosine.

Figure S14. Integrated maps of the epigenome and its interaction with the transcriptome. The superimposition of the cytosine methylome, transcriptome and smRNAome clearly illustrates the diverse epigenetic and transcriptional landscapes encountered throughout the nuclear chromosomes.

A) Chromosome 1 euchromatic region, wild type.

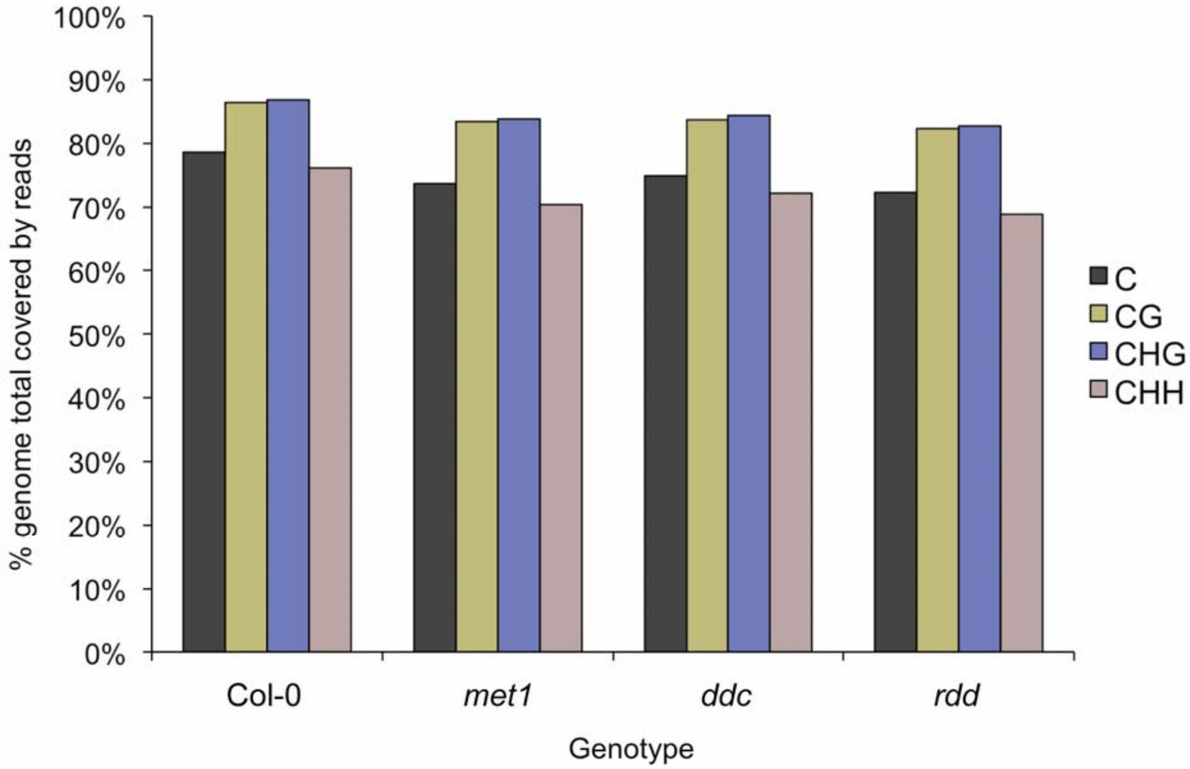
B) Chromosome 1 pericentromeric region, wild type. Tracks are shown for gene annotation, DNA methylation sites, smRNA-seq and mRNA-seq. Abbreviations: mC, methylcytosine.

Figure S15. Experimental procedure for generating smRNA-seq libraries.

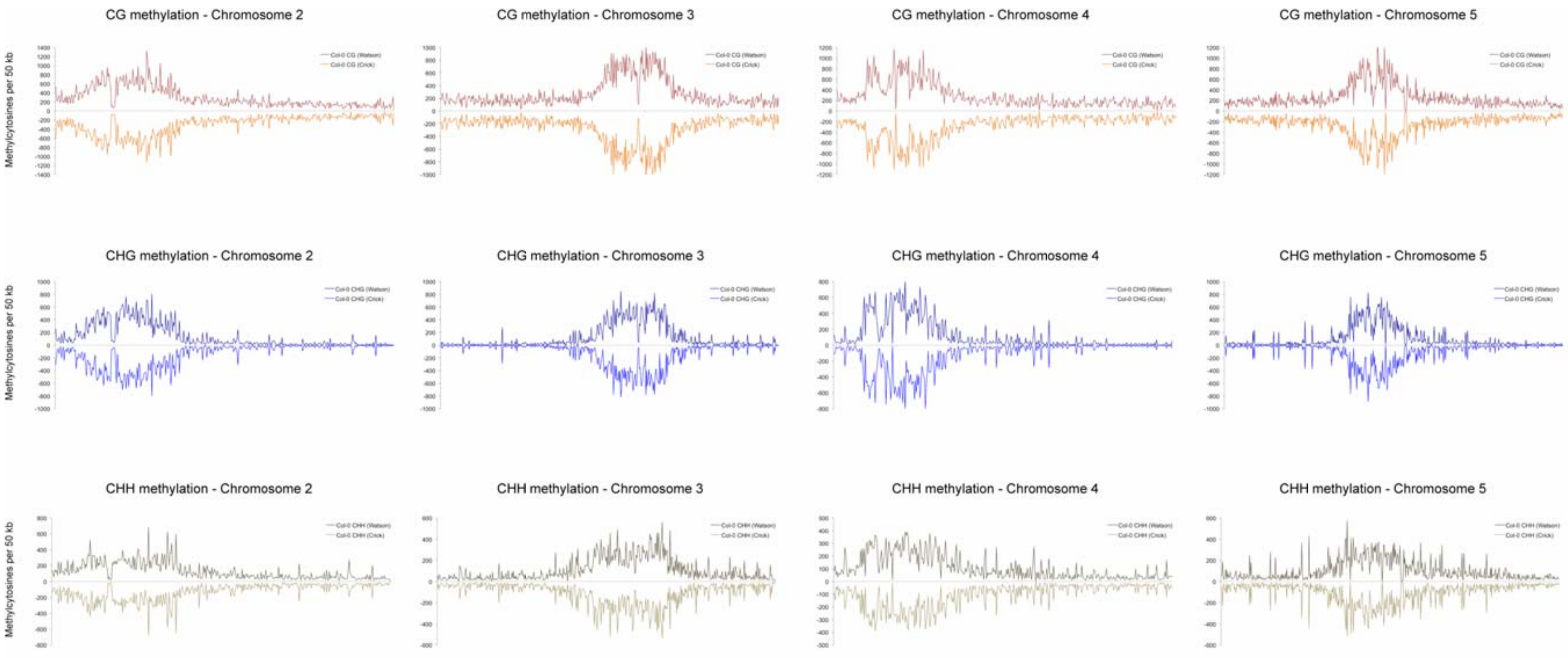
Figure S16. Experimental procedure for generating mRNA-seq libraries.

Figure S17. Experimental procedure for generating methylC-seq libraries.

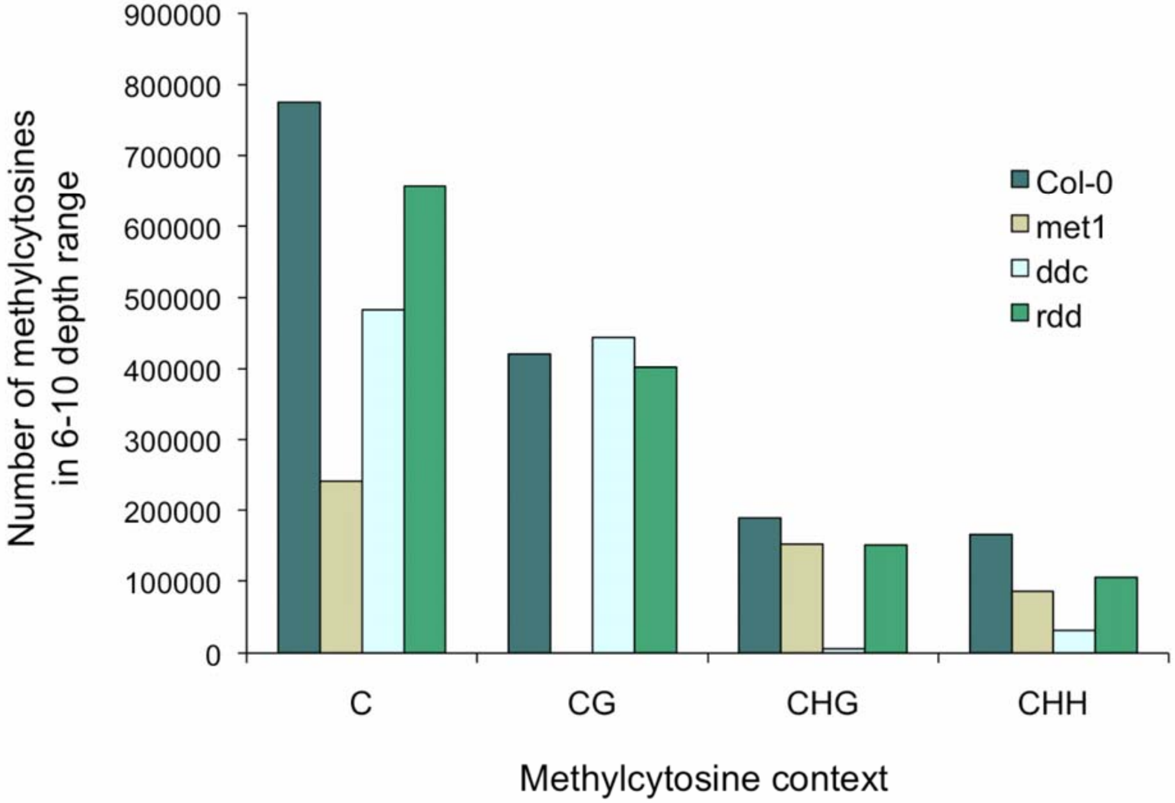
Supplementary Figures



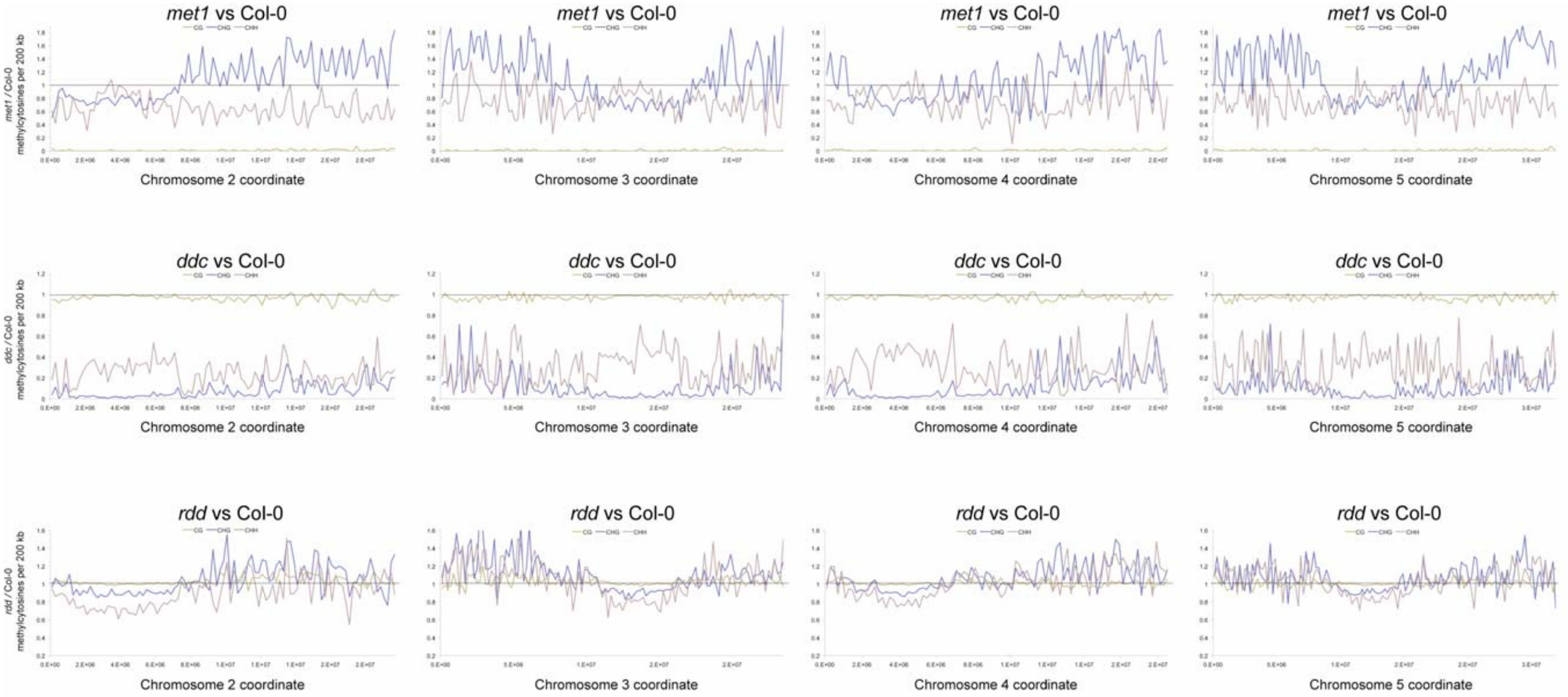
Lister et. al. Supplementary figure 1.



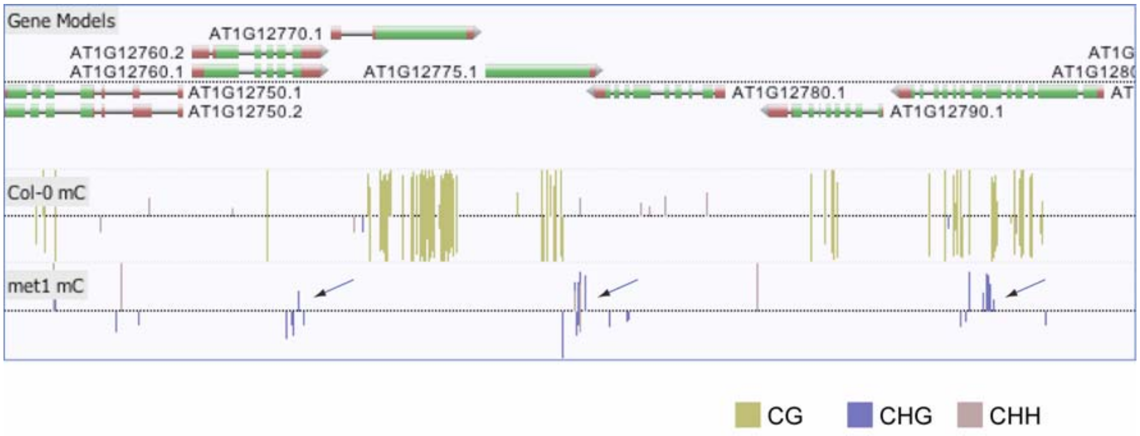
Lister et. al. Supplementary figure 2.



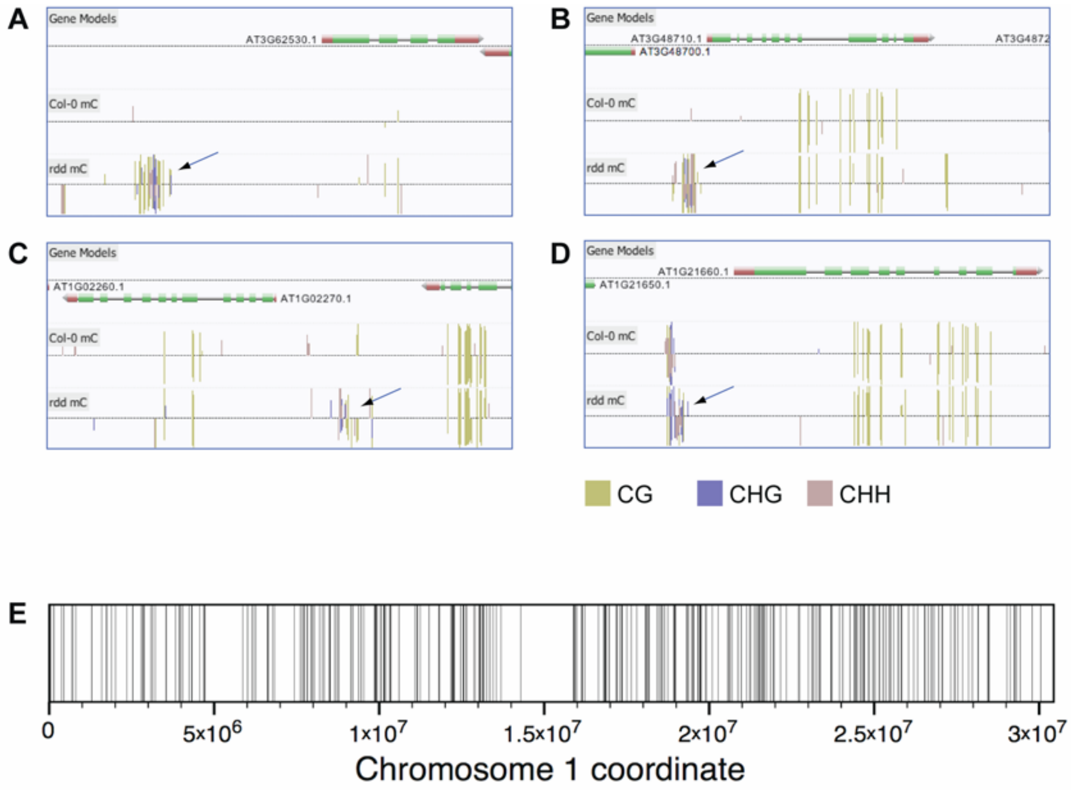
Lister et. al. Supplementary figure 3.



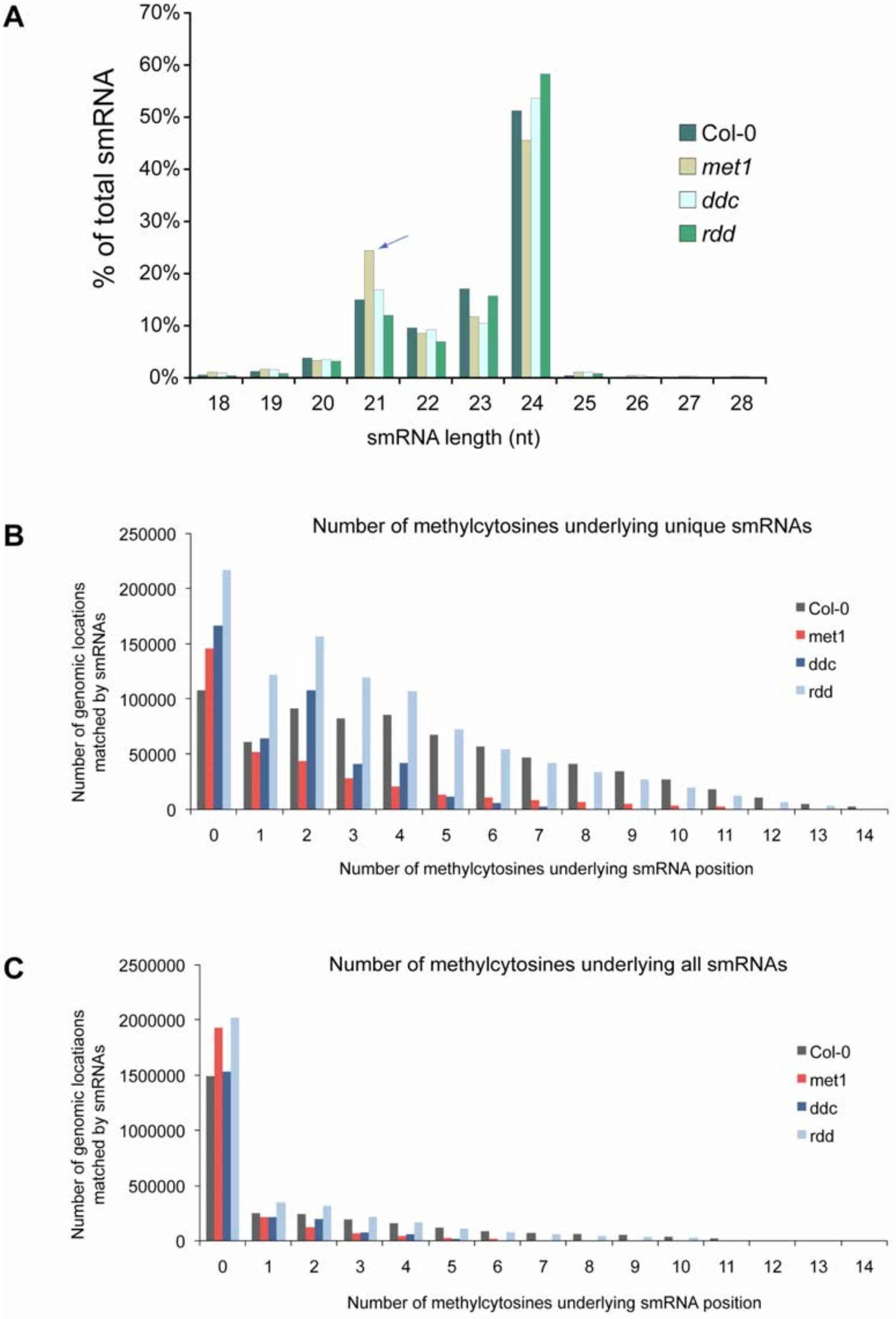
Lister et. al. Supplementary figure 4.



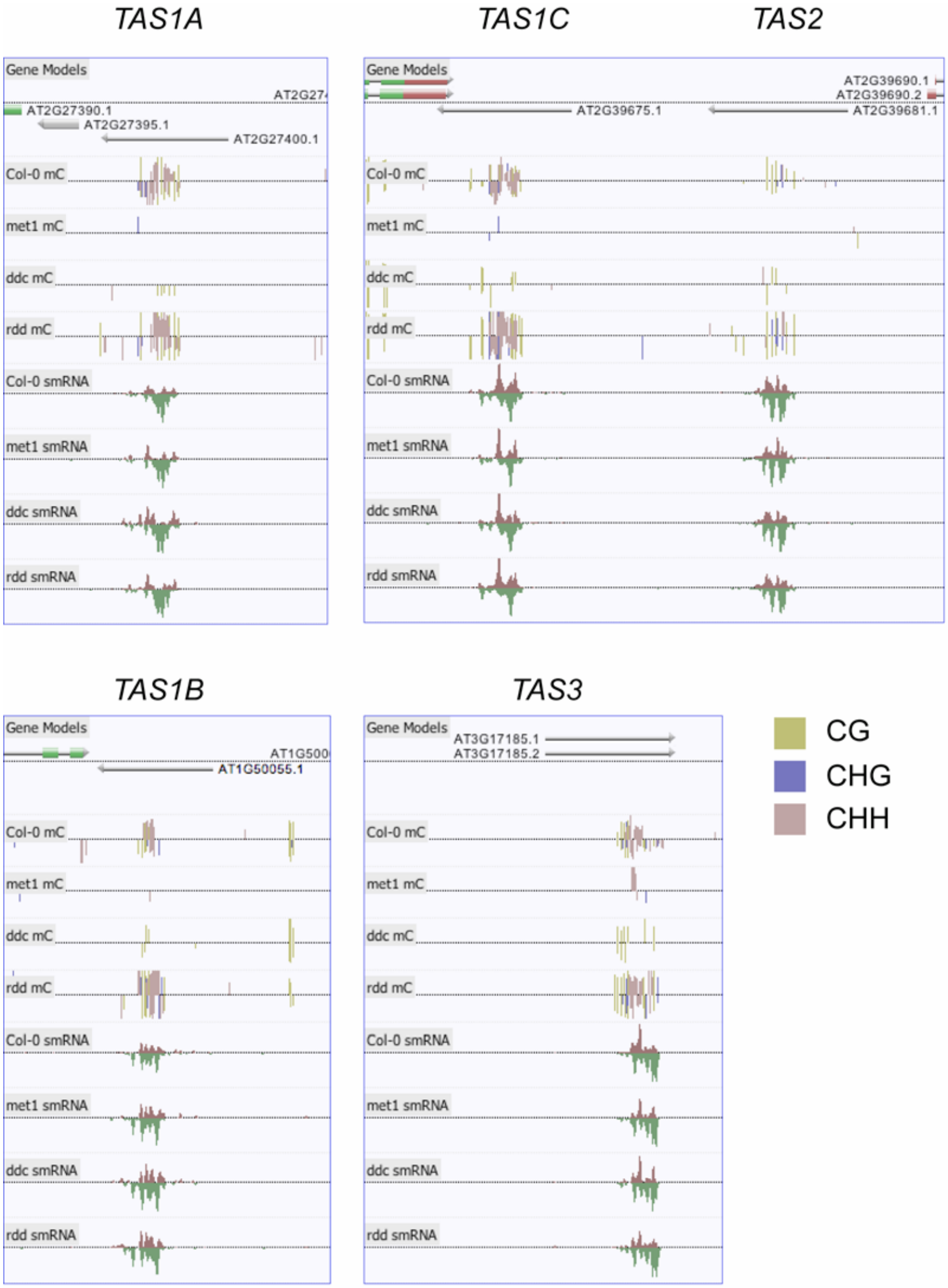
Lister et. al. Supplementary figure 5.



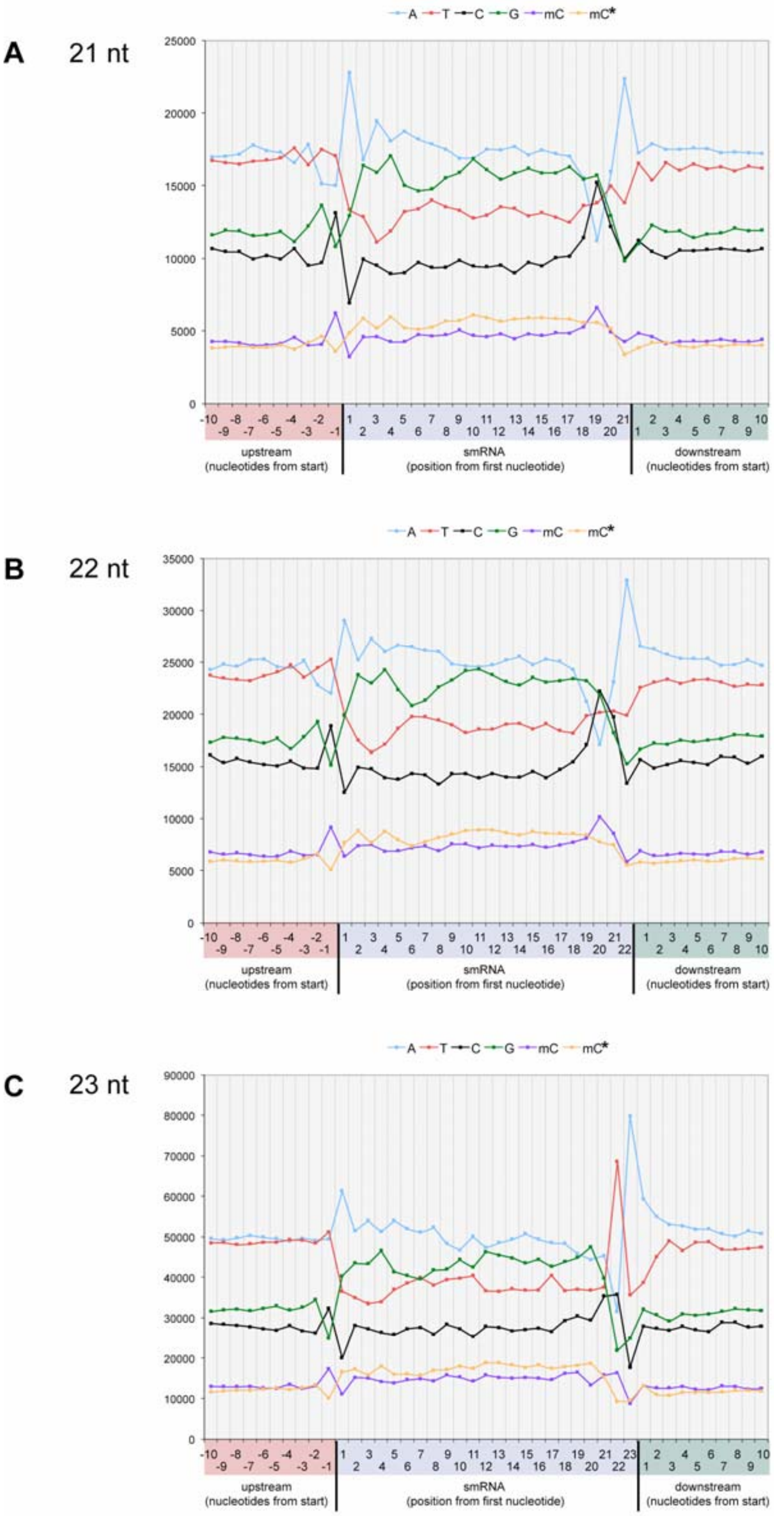
Lister et. al. Supplementary figure 6.



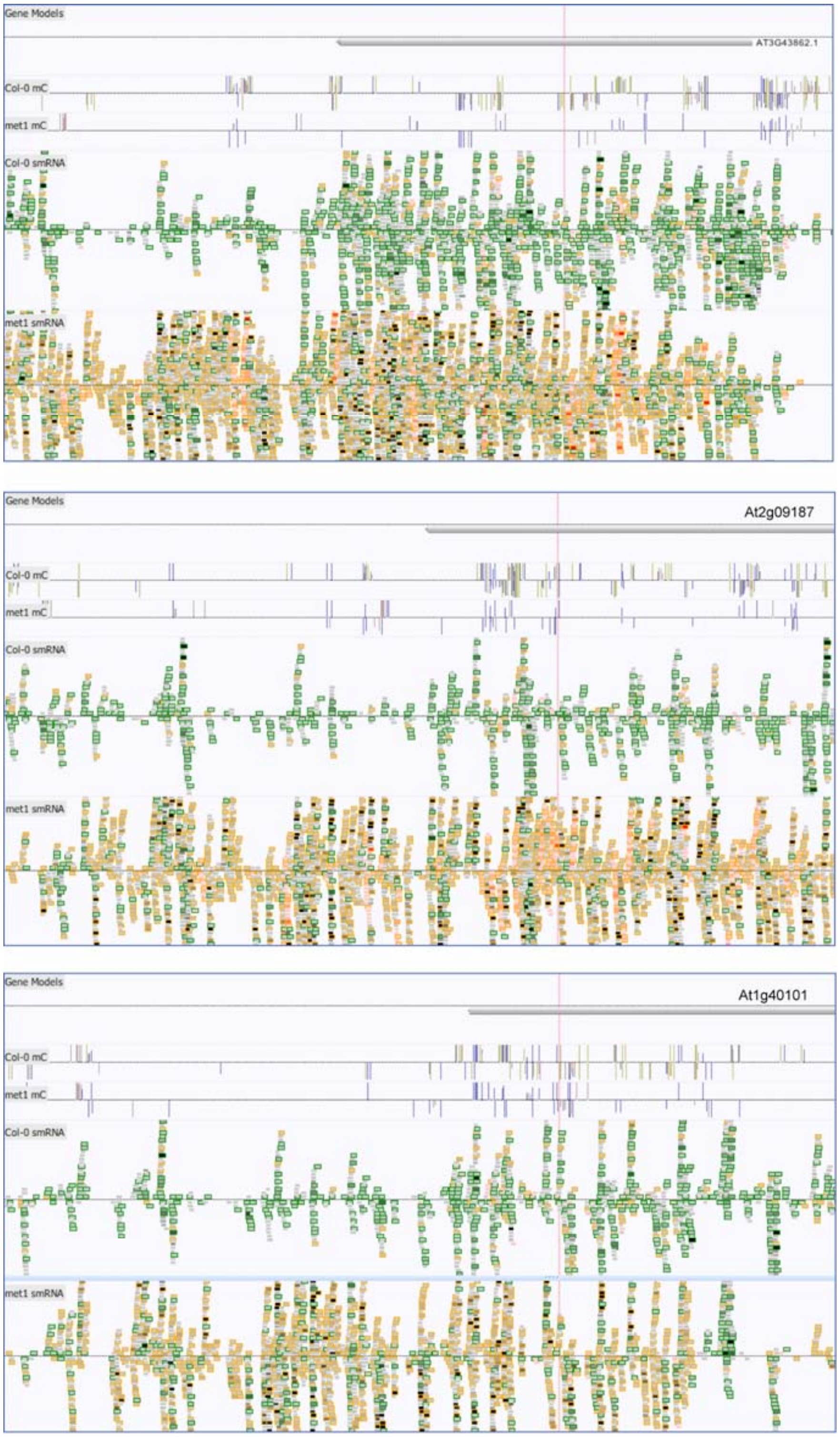
Lister et. al. Supplementary figure 7.



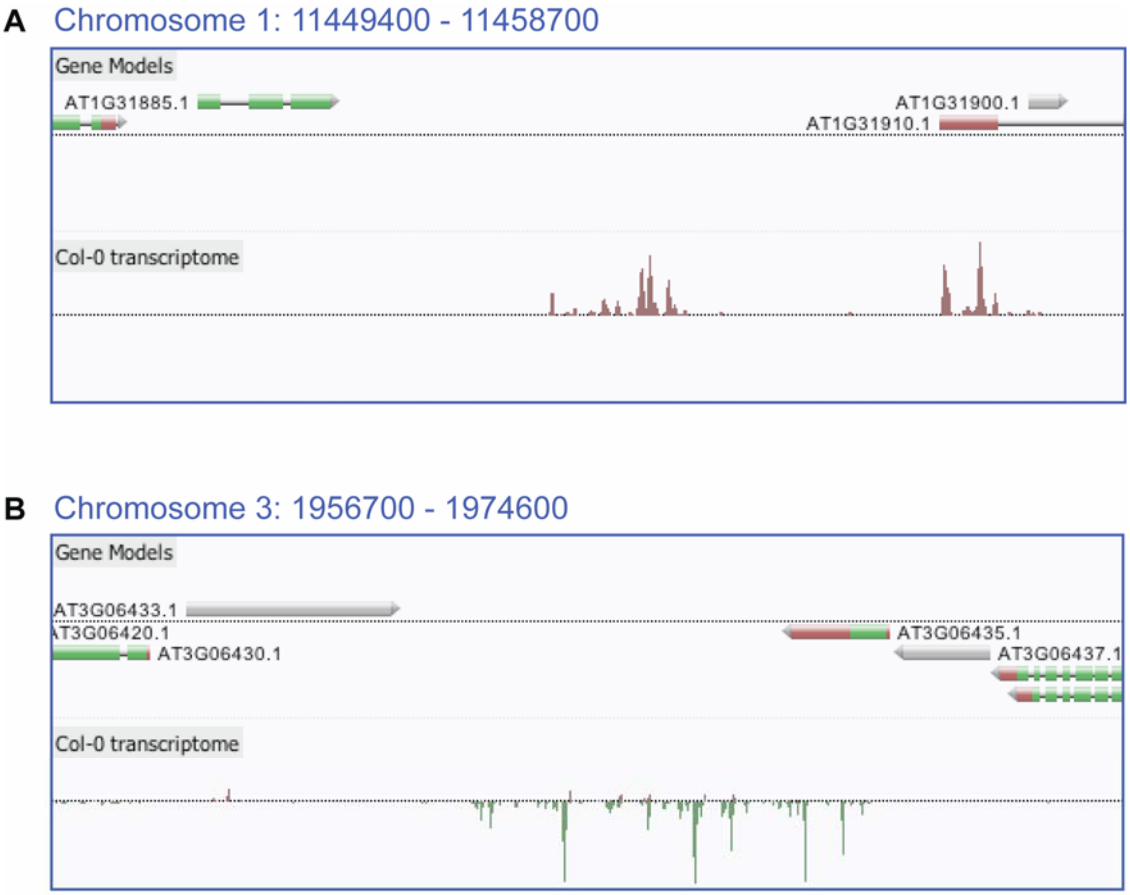
Lister et. al. Supplementary figure 8.



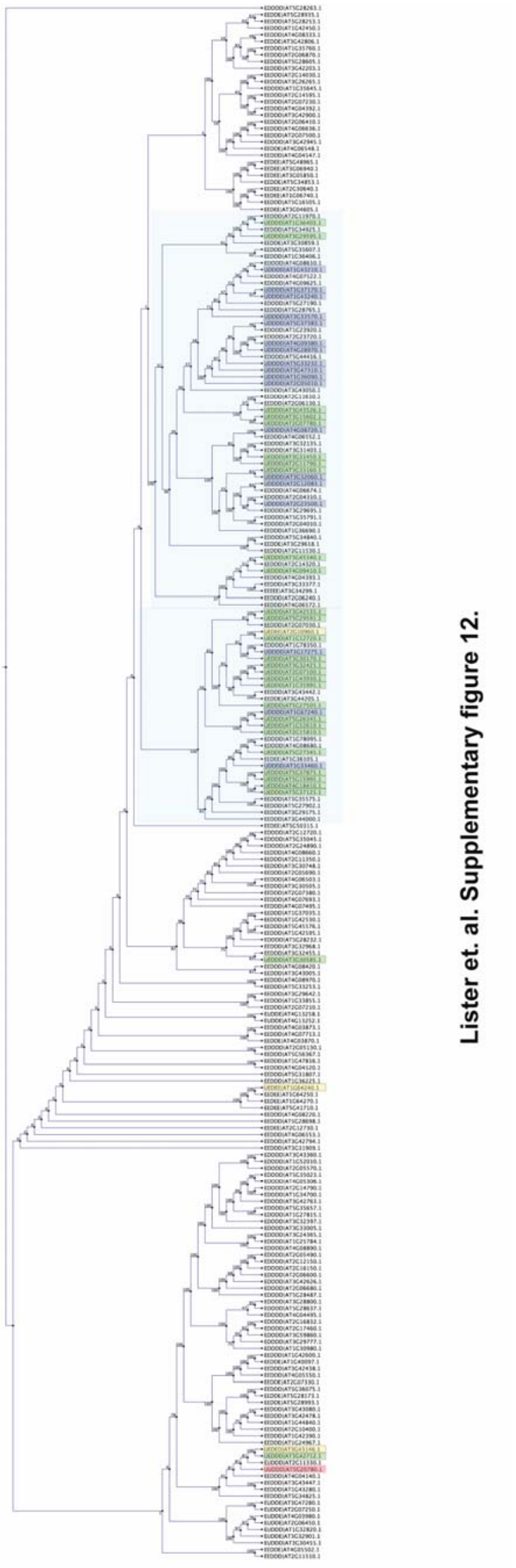
Lister et al. Supplementary figure 9.



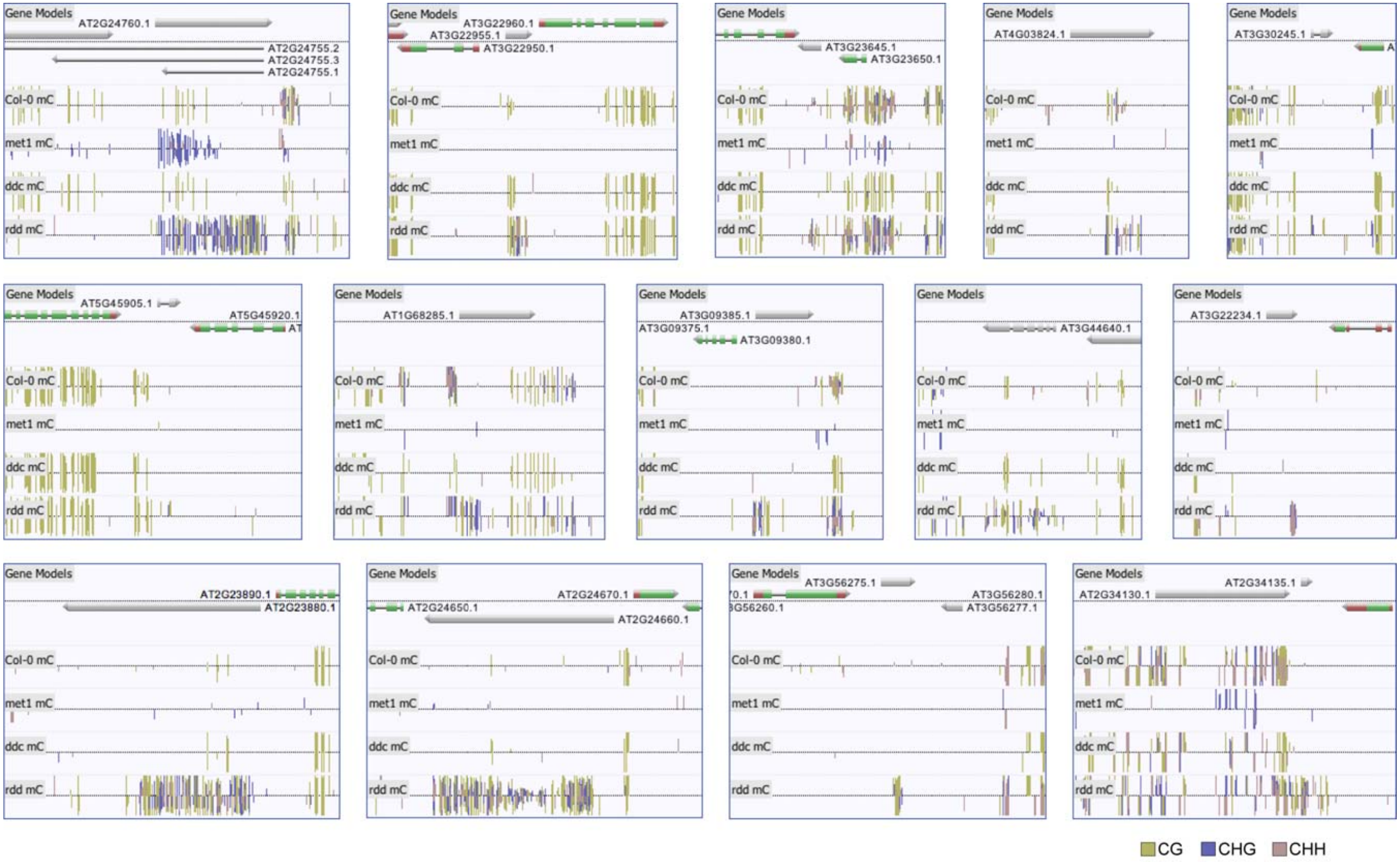
Lister et. al. Supplementary figure 10.



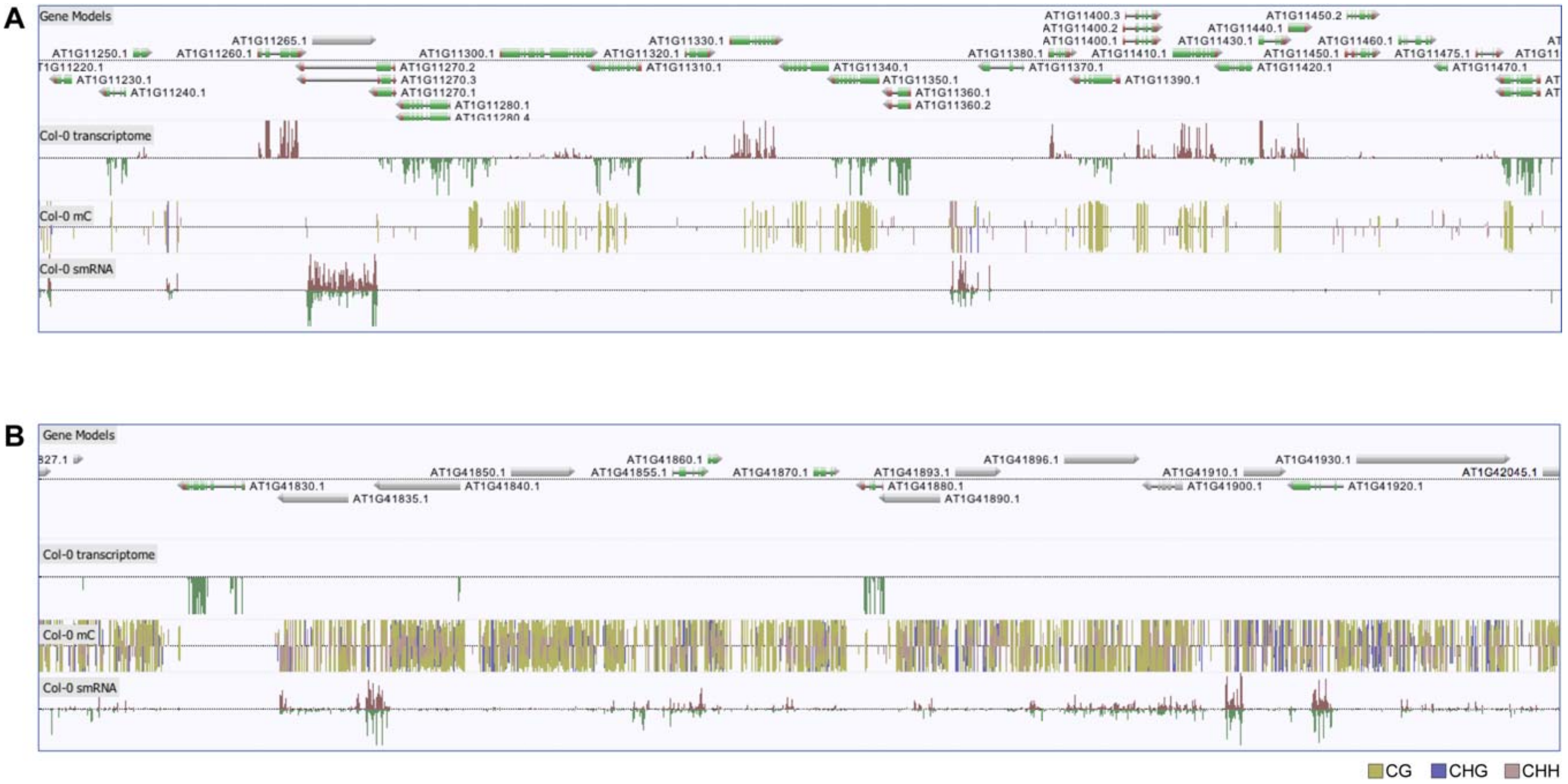
Lister et. al. Supplementary figure 11.



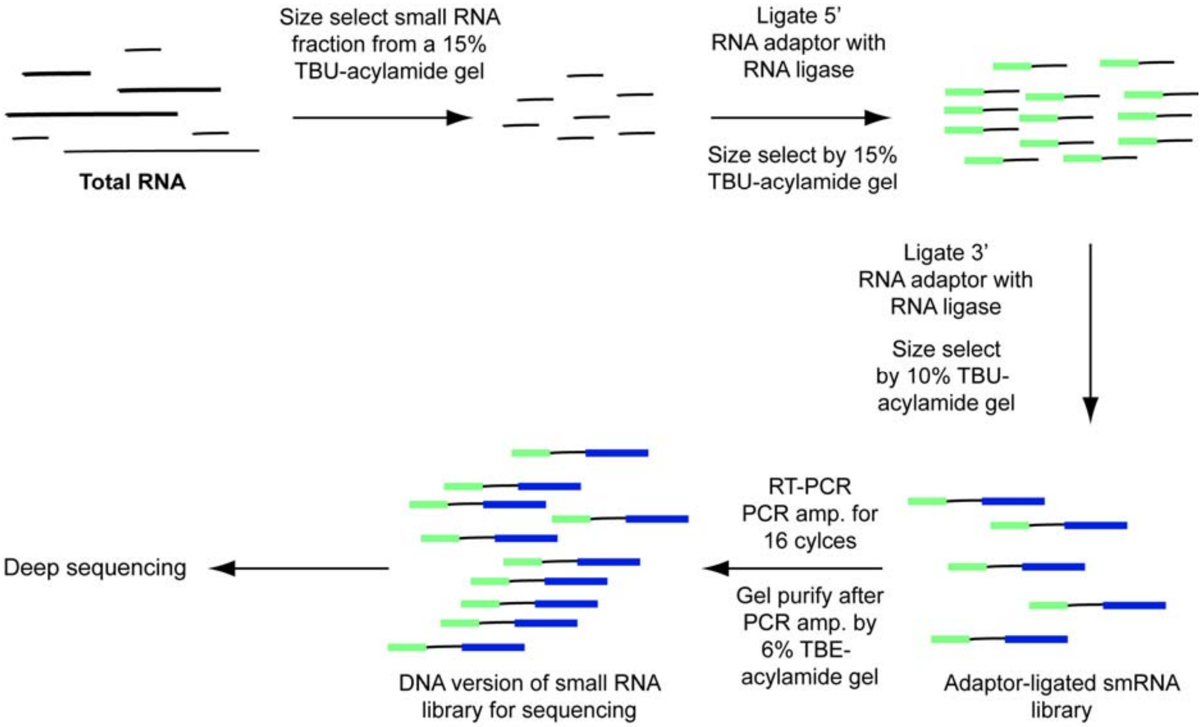
Lister et. al. Supplementary figure 12.



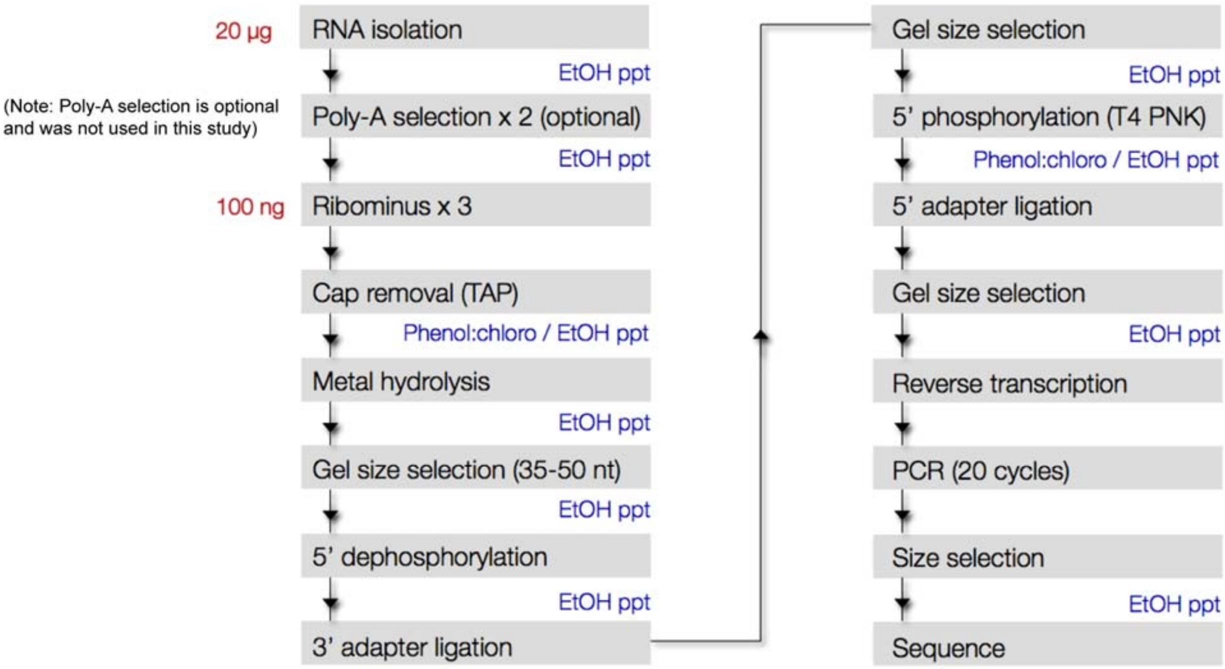
Lister et. al. Supplementary figure 13.



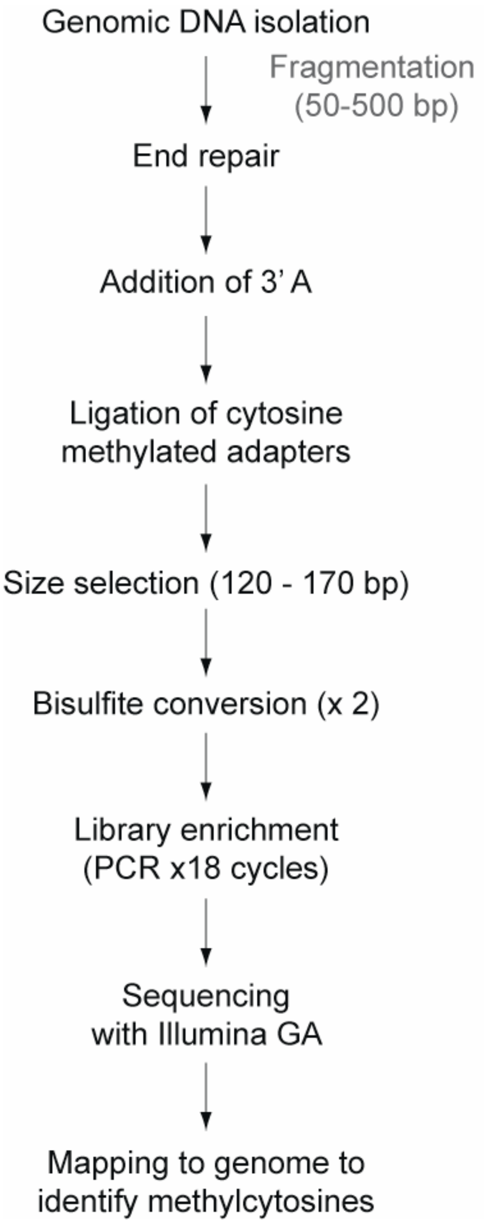
Lister et. al. Supplementary figure 14.


Total RNA
Size select small RNA fraction from a 15% TBU-acylamide gel
Ligate 5' RNA adaptor with RNA ligase
Size select by 15% TBU-acylamide gel
Ligate 3' RNA adaptor with RNA ligase
Size select by 10% TBU-acylamide gel
Adaptor-ligated smRNA library
RT-PCR PCR amp. for 16 cylces
Gel purify after PCR amp. by 6% TBE-acylamide gel
DNA version of small RNA library for sequencing
Deep sequencing

Lister et. al. Supplementary figure 15.



Lister et al. Supplementary figure 16.



Lister et al. Supplemental figure 17.