Title: Sequence and epigenetic landscapes of active and silenced nucleolus organizers in *Arabidopsis*

Authors: Dalen Fultz[1,2,3,†], Anastasia McKinlay[1,2,3,†], Ramya Enganti[1,2,3], and Craig S. Pikaard[1,2,3,*]

Affiliations:
[1]Howard Hughes Medical Institute, Indiana University; Bloomington, IN, USA.
[2]Department of Biology, Indiana University; Bloomington, IN, USA
[3]Department of Molecular and Cellular Biochemistry, Indiana University; Bloomington, IN, USA
†These authors contributed equally to this work
* Corresponding author: cpikaard@iu.edu

Abstract:
Arabidopsis thaliana has two ribosomal RNA gene loci, nucleolus organizer regions *NOR2* and *NOR4*, whose complete sequences remain unknown. Using ultra-long DNA sequencing technology, we generated 5.5 and 3.9 Mbp sequence assemblies for *NOR2* and *NOR4* (in the reference strain, Col-0), revealing their distinctive gene subtype compositions. RNA sequencing, and gene identification within flow-sorted nucleoli, in both wild-type and *histone deacetylase 6* mutants, shows that most, but not all, *NOR4* genes are active whereas most, but not all, *NOR2* genes are epigenetically silenced. Long intervals of low CG and CHG methylation overlap regions of gene activity and gene subtype homogenization. Collectively, the data illuminate the genetic and epigenetic landscapes of the NORs and implicate transcription in rRNA gene concerted evolution.

One Sentence Summary:
Sequencing of Arabidopsis NORs reveals their organization, activity, and patterns of rRNA gene sequence homogenization.

Main Text:

In eukaryotes, ribosomal RNA (rRNA) genes transcribed by RNA Polymerase I (Pol I) are repeated in long tandem arrays, typically spanning millions of basepairs. These chromosomal loci are known as nucleolus organizer regions (NORs) because the nucleolus, the most prominent substructure of the nucleus, forms as rRNA genes are transcribed and processed into 18S, 25-28S and 5.8S catalytic rRNAs that assemble with ~80 ribosomal proteins and 5S rRNA to form ribosomes, the protein-synthesizing machines of cells (*1-4*).

rRNA gene repeats are extremely similar in sequence complexity yet can be differentially expressed, giving rise to the phenomenon known as nucleolar dominance (*5*). This epigenetic phenomenon was initially discovered in genetic hybrids in which NORs of one progenitor were preferentially active. However, nucleolar dominance is not unique to hybrids. Instead, it occurs in all species studied to date, regulating the number of rRNA genes that are active (*6*). In *Arabidopsis thaliana* Col-0, the inactive subset of rRNA genes is isilenced during early post-embryonic development (*7*) in a process involving DNA cytosine methylation and histone deacetylation (*8, 9*). Genetic segregation analyses of a small set of differentially expressed rRNA gene subtypes indicated that silenced and active rRNA genes may not be intermingled, with silenced subtypes mapping to the NOR on chromosome 2, *NOR2* and active subtypes mapping to the NOR on chromosome 4, *NOR4* (*10*) (*11*). However, in a Col-0 line in which a large portion of *NOR4* was replaced by sequences of *NOR2*, the normally silenced *NOR2* subtypes escaped silencing (*12*). Collectively, these studies suggested that rRNA gene activity is not dictated by individual gene sequences but, instead, somehow depends on NOR affiliation, highlighting a need to determine the sequences and functional organizations of the two NORs.

Variable length elements define at least 74 rRNA gene subtypes

At *NOR2* and *NOR4*, rRNA genes are repeated head-to-tail and oriented such that transcription occurs towards the centromere (Fig. 1A) (*13, 14*). Each transcription unit is separated by an intergenic spacer (IGS) that includes the gene promoter, spacer promoters (SPs), and variable numbers of 21nt Sal repeats, named for having *Sal* I restriction endonuclease sites (*15*). Each gene transcription unit has the potential to generate a 45S primary transcript, or pre-rRNA, whose sequence elements include a 5' External Transcribed Spacer (5' ETS), the 18S, 5.8S and 25S catalytic rRNA sequences, two Internal Transcribed Sequences, ITS1 and ITS2, and a 3' ETS. Gene promoter, rRNA coding sequences and ITS sequences are nearly identical among the hundreds of rRNA genes except that 270 or 334 bp deletions occur in the 18S rRNA sequences of a small number of genes (Fig. 1B) (*14*). Substantial variation occurs between genes within the IGS and ETS regions, including the number of Sal repeats, spacer promoters, ~310 bp repeats of the 5' ETS, and repeats within the 3' ETS region, making each of these structural gene elements variable in length (Fig. 1B). Some of

these variable-length elements (VLEs) have been described previously (*14, 16*), but we identified many more upon Pacific Bioscience SMRT sequencing of ~10 kbp rRNA gene units generated using the endonuclease I-*Ppo* I (Fig. 1A), which cuts once per rRNA gene repeat, thereby doubling the number of known IGS VLEs and revealing the VarC, Var D and VAR6 VLEs (fig. S1 and Fig. 1B). The 59 VLEs summarized in Fig. 1B (including reference sequences at the positions of the two deletions) occur in 74 permutations, thus defining 74 distinct rRNA gene subtypes based solely on these variable elements (fig. S2).

Sequencing, assembly, and composition analyses of NOR2 and NOR4

The high degree of sequence similarity among rRNA genes, especially within their transcribed regions, precludes their discrimination using conventional short-read DNA sequencing technologies, making NORs notoriously difficult loci to assemble. However, we hypothesized that rare combinations of specific VLEs within multiple adjacent genes might serve as landmarks that would enable NOR assembly. To this end, we performed Oxford Nanopore Technology (ONT) sequencing of *A. thaliana* genomic DNA, obtaining reads that averaged ~40-50 kbp but included reads as long as several hundred kbp (table S1). Whole genome assembly using the ONT sequence data recapitulated the current TAIR10 genome assembly (fig. S3), confirming the integrity and comprehensive coverage of the read set. However, the NORs failed to be assembled using the conventional genome assembly tools, as is also the case for the TAIR10 reference genome assembly.

To assemble the NOR sequences, we exploited VLE patterns and other distinctive structural features to align rRNA gene reads, using two complementary approaches. In one approach, ONT reads containing rRNA gene sequences were compared to a single rRNA gene reference sequence, namely the sequence of gene subtype #4 (whose sequence is provided in file S1 and whose VLE composition is shown in fig. S1) to generate two-dimensional similarity matrices (dot-plots). Individual dot-plots displaying visually distinct and rare patterns were then manually assembled, like puzzle pieces, yielding longer contigs, as illustrated in Fig. 2A for three overlapping ONT reads from the centromere-proximal end of *NOR2*.

In parallel with the dot-plot puzzle-fitting approach, we developed a computational NOR assembly pipeline (Fig. 2B). The pipeline first identified the order of VLEs within a set of rRNA gene-containing ONT reads longer than 85kbp.This read set collectively represented ~100X coverage of the total rRNA gene pool (see methods). We then identified sets of VLEs predicted to be unique based on their frequency within the read set. These VLE patterns (file S2) were then used as landmarks to group overlapping ONT reads and ultimately connect them to other groups sharing different VLE landmarks, allowing contigs to be created (Fig. 2C and files S3-S7). Following multiple sequence alignment of the overlapping reads, consensus sequences were generated using custom scripts (link to Github tagged_longread_assembler) (Fig. 2C),

yielding full-length sequence assemblies of 5.5 Mbp and 3.9 Mbp for *NOR2* and *NOR4* respectively (Fig. 2D). These NOR sequences have been deposited at Genbank (BankIt2506918 and BankIt2708174; the data will become publicly available upon publication). Analysis of 18S and 25S rRNA sequences, which are essentially invariant among the hundreds of rRNA genes (*11*), indicates that the *NOR2* and *NOR4* sequence assemblies have a basepair accuracy of ~99.85% (file S8).

The *NOR2* and *NOR4* assemblies were quality checked in three ways. First, we compared the frequency of VLE patterns in raw ONT reads versus the NOR assemblies, yielding a correlation value of $R^2 = 0.985$ (file S9.) This indicates that the assemblies capture the patterns present in the raw data. Second, overlapping ONT reads of computationally assembled contigs were visually examined using the dot-plot puzzle-fitting approach. Third, we used custom-designed CRISPR guide RNAs complementary to different combinations of 3' ETS VLEs to cut genomic DNA, then separated the resulting fragments by contour-clamped homogeneous electric field (CHEF) electrophoresis and performed Southern blotting, revealing large fragments whose sizes are consistent with those predicted by the NOR assemblies (fig S4).

NOR2 has 518 rRNA gene repeats, based on the count of 25S rRNA sequences, whereas *NOR4* has 376 genes. Fig. 2D shows their organizations with respect to 3'ETS, IGS and 5'ETS VLEs, or internal deletions, each displayed as a vertical track. Another track depicts the number of times any single rRNA gene subtype is repeated within a sliding window spanning 9 genes (~90 kbp), thereby yielding a quantitative measure of local gene subtype homogeneity. Subtype content and distribution patterns are unique to each NOR, as illustrated in similarity matrices comparing VLE features within and between the NORs (Fig. 2E).

Although some VLEs are shared between the NORs (fig. S5), none of the rRNA gene subtypes resulting from unique combinations of VLEs are common to both NORs. *NOR2* is comprised of 47 rRNA gene subtypes (Figs 2F, 2G, 3A, and 3B) but a generalization is that *NOR2* consists of VAR1.1-class genes (subtypes 1-26 in Fig. 3A) interspersed with VAR3 genes of both the 3.1 subclass (subtypes 47-53) and 3.2 subclass (subtypes 54-70). Two regions of high local subtype homogeneity are present in *NOR2* and collectively account for ~25% of the NOR (Figs 2D and 2E). Homogenous region A is enriched for gene subtype 6 (Fig. 3A), the most abundant *NOR2* subtype, comprising 19% of the NOR (Fig. 2F). Region B is enriched for subtype 10, the second most abundant *NOR2* subtype (14% of the NOR, Figs. 2F and 3A).

NOR4 is comprised of 27 rRNA gene subtypes (Fig. 2G). Three subtypes account for 88% of the NOR (Fig. 2G) and are present within two large regions with high local gene subtype homogeneity, regions C and D. Telomere-proximal region C is composed of VAR3.2-class genes, primarily subtype 65 genes that account for 32% of *NOR4* (Figs 2G and 3B), interspersed with VAR4-class genes, primarily subtype 71 genes which comprise 6% of the NOR (Figs 2G and 3B). Centromere-proximal region D

is composed primarily of VAR2.1 class genes, specifically subtype 32, which comprises 50% of *NOR4* (Figs 2G and 3B).

At two positions within *NOR4*, marked by '#' symbols in Fig. 2D (and subsequent figures), the contigs on either side end in long stretches of subtype 32 repeats and cannot be joined unambiguously at specific positions. Comparison of the relative abundance of subtype 32 genes in raw rRNA sequencing reads versus the *NOR4* assembly suggests that an estimated ~31 copies of subtype 32 may be unaccounted for in the assembly (file S9). If so, the additional genes would be present at one or both locations indicated by the # symbols, potentially making *NOR4* slightly longer than indicated, but not shorter.

The telomere-proximal and centromere-proximal ends of *NOR2* and *NOR4* have distinct structural features (Figs 3C-F). At the telomeric end of *NOR4*, the genes are complete with the exception of the third gene, which has an internal deletion of several kbp (Fig. 3C). This is one of only four structural anomalies within the entirety of *NOR4*, the others being two genes with 25S deletions (Fig. 2D, 'Deletions' track) and one gene, near the centromere-proximal end, with a GP2 (Gypsy family) transposon insertion (Fig. 3E). By contrast, *NOR2* has numerous genes with structural anomalies. At the telomere-proximal end, *NOR2* has two rRNA genes in the opposite orientation due to an inversion of ~20 kbp (Fig. 3D). In the final 250 kbp at the centromere-proximal end, more than one-third of the genes have internal deletions and a complete GP2 transposon is present within the third gene from the end (Fig. 3F). In total, *NOR2* has ~20 genes with internal deletions that would likely render them non-functional (see Fig. 2D 'Deletions' track).

Locations of active and silenced rRNA gene subtypes within NOR2 and NOR4

To determine the locations of transcriptionally active rRNA genes, we identified informative single nucleotide polymorphisms (SNPs) using Illumina whole genome sequencing, conducted at a depth of coverage (~163X) sufficient to detect low frequency SNPs (fig. S6A and file S10), and determined the positions of these SNPs in our NOR assemblies (fig. S6B). We then performed Illumina sequencing of reverse-transcribed leaf RNA and examined the expression of SNPs that mapped to specific positions within the NORs (Figs. 4A,B). Thirty-five SNPs examined were *NOR2*-specific, collectively representing genes distributed throughout the NOR. Only one was detected among expressed RNAs, C1886T, present within a group of genes near the centromere-proximal end of *NOR2*. By contrast, thirteen *NOR4*-specific SNPs were detected in RNA reads, mostly corresponding to genes of homogenous regions C and D. *NOR4*-specific SNPs that were not detected above threshold levels mostly corresponded to low abundance gene subtypes.

rRNA genes present within nucleoli are distributed throughout NOR4

The abundant nucleolar protein, Fibrillarin, when expressed as a fusion protein with YFP, enables the fluorescence-activated flow-sorting of whole nuclei or of nucleoli that persist upon disruption of nuclei by sonication (*9*). Using a PCR primer pair that flanks the 3' ETS variable region, amplification products of genes bearing the VAR1, 2, 3 or 4 VLEs (*10*) are all detected upon amplification of DNA isolated from whole nuclei, as expected (Fig. 4C lanes 3 and 5). However, VAR1-containing genes are depleted in isolated nucleoli (Fig. 4C, lane 4), reflecting their selective silencing (Fig 4C,lane 1). However, if silencing is prevented in an *hda6* mutant, VAR1 genes are expressed at high levels (Fig. 4C, lane 2) and are now present in purified nucleoli (Fig. 4C, lane 6). These experiments show that nucleolar localization can be used as a proxy for rRNA gene transcriptional activity, allowing active genes to be identified based on VLEs not present within the transcribed region and whose activity status thus cannot be determined using RNA-seq.

To identify the specific rRNA genes present within nucleoli, we performed ONT sequencing of DNA isolated from flow-sorted nucleoli of both wild-type plants and *hda6* mutant plants, as well as corresponding flow-sorted nuclei, which serve as controls. Reads averaged ~3 kbp in length (table S2), sufficient for VLE identification. The abundance of 3'ETS VLEs reads from the DNA of wild-type nuclei matched their relative abundance in the NOR assembly (fig S7), indicating a lack of bias in subtype detection by sequencing. Next, we calculated the level of nucleolar enrichment for NOR-specific VLEs, finding that *NOR4* rRNA genes are enriched in nucleoli, as demonstrated for nine *NOR4*-specific VLEs (Fig. 4D), two of which (VAR2.1 and 314_SP_1005) represent homogenous regions C and D (from Fig. 2D and E). For *NOR4* VLEs*,* nucleolus association was mostly unchanged between wildtype and *hda6*, except that VLEs at the centromere-proximal end showed decreased nucleolar association in *hda6*. By contrast, dramatic differences in wild-type versus *hda6* were apparent for *NOR2* –specific VLEs (Fig. 4E). In wild-type plants, all but two (marked with *) of 18 VLEs examined, which are distributed throughout *NOR2,* are depleted in nucleoli of wild-type plants but become enriched within the nucleoli of *hda6* plants, confirming that *NOR2* rRNA genes are subjected to HDA6-dependent silencing. However, several VLEs located near the centromere-proximal end of *NOR2* fail to become enriched in the nucleolus in *hda6* plants. Many genes in this region are missing promoters (see Fig. 3F), likely precluding them from participating in nucleolus formation.

Cytosine methylation landscapes of NOR2 and NOR4

We examined CG, CHG and CHH methylation throughout *NOR2* and *NOR4* using methylation calling tools (*17*) for ONT sequencing, calculating the percent methylation within 500 bp sliding windows (Figs 5A,B). Most of *NOR2* has very high CG methylation, moderate CHG methylation and low, but detectable CHH methylation (Figs 5A). The only exception is homogenous region B, which has low levels of methylation in all three contexts (Fig. 5A and fig. S8). At *NOR4,* a large central region has low

methylation levels, in all contexts, but is flanked, on both sides, by ~1 Mbp intervals, extending to the ends of the NOR, that have high methylation levels (Fig. 5D). Zooming in to achieve higher resolution shows that within regions of high CG methylation, at both NORs, the high levels of methylation occur across entire gene units except for a region upstream of the gene promoter, which contains low methylation in most but not all genes (fig. S8). Conversely, in low methylation regions, genes are uniformly hypomethylated except for the 3'ETS, where a peak of methylation occurs in most gene repeats (fig. S8). The highly methylated regions of *NOR2* and *NOR4* correlate with regions of low or undetectable gene expression (Figs. 5A-B, RNA track). Conversely, the hypomethylated region of *NOR4* overlaps the region in which actively transcribed gene subtypes are located (Fig. 5B, RNA track).

Discussion

Using an unconventional sequence assembly strategy, our determination of the *NOR2* and *NOR4* sequences, from their telomere-capped ends to their centromere-proximal ends, fills the last major gaps in the Arabidopsis genome. The NOR sequences add 9.4 Mbp of new sequence information, equivalent to ~6.5 % of the genome.

A surprising finding is that none of the 74 gene subtypes defined by their unique combinations of VLEs is common to both NORs. Instead, each NOR has a unique set of rRNA gene subtypes. This allowed the unambiguous assignment of subtype-specific gene expression data, determined by RNA-seq or subtype association with flow-sorted nucleoli (*9*) to specific positions within *NOR2* or *NOR4*. Collectively, the data show that rRNA genes broadly distributed throughout *NOR2* are generally inactive whereas genes within broad intervals of *NOR4* are active. Inactive *NOR2*-specific gene subtypes are epigenetically silenced, as demonstrated for 14 different NOR2 VLE markers (see Fig. 4B) that become nucleolus-associated when silencing is prevented in an *hda6* mutant (Fig. 4). By contrast, *NOR4* subtype expression is mostly unchanged in wild-type versus *hda6* plants.

Prior studies that focused on the detection of RNAs bearing VAR1-type 3' ETS, sequences, which are present in genes that primarily map to NOR2 (*10*) showed that their developmentally regulated silencing requires HDA6, Histone H3 lysine 9 methyltransferases and the DNA methyltransferases MET1 and CMT2 (*7, 9, 18, 19*). These, and other studies, established that repressive histone modifications and DNA methylation act in partnership to silence rRNA genes whereas active rRNA genes display cytosine hypomethylation (*20*). In our current study, we find an approximately 2 Mbp hypomethylated region at the center of *NOR4* that overlaps the positions of rRNA gene subtypes that are transcribed and associated with the nucleolus. Our interpretation is that this hypomethylated central region is the most highly transcribed region of *NOR4*, at least in the leaf and inflorescence tissues that were the source of the DNA and RNA sequenced in our study. Gene subtypes near the telomeric and centromere-proximal ends of *NOR4* are inactive or expressed at low levels and these regions of the NOR are

hypermethylated, like most of *NOR2*. Interestingly, a small region of *NOR2*, within homogenous region B, is hypomethylated, possibly marking a small interval of gene activity that may be the source of trace levels of VAR1-containing transcripts detected in wild-type plants (see Fig. 4, lane 1).

An intriguing finding is that the regions of high gene activity and/or low cytosine methylation within both *NOR2* and *NOR4* are also regions with high degrees of gene subtype homogenization. This suggests that transcription may be a driving force for local rRNA gene homogenization in the concerted evolution of rRNA genes (*21*) (*22*), which is thought to occur via gene conversion or recombination following slipped mispairing of NOR sequences (*23*) (*23*). Consistent with this hypothesis, mitotic recombination and gene conversion events within the single NOR of budding yeast are stimulated by active Pol I transcription (*24, 25*). The greater diversity of gene subtypes at *NOR2* versus *NOR4* (47 versus 27 subtypes), as well as the occurrence at *NOR2* of numerous genes with deletions or rearrangements, may thus be a consequence of reduced gene homogenization at *NOR2* as a consequence of selective silencing during vegetative development. Our results also show that exchange of sequences between the two NORs is rare, with common sequences limited to short intervals on the scale of the VLEs, but not complete rRNA genes. We thus deduce that the concerted evolution of the two NORs occurs through events involving short sequence intervals, either though gene conversion or double crossovers, followed by local homogenization of the sequences among adjacent genes. The correlation between hotspots of transcription and hotspots of gene subtype homogenization implicates transcription as a driver of the homogenization process.

References and Notes:

Acknowledgments: We thank the Indiana Center for Genomics and Bioinformatics for Illumina sequencing. We thank Dr. Feng Wang for writing a script for the design of CRISPR guide RNAs (sgRNAs). This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

Funding: Howard Hughes Medical Institute Investigator funds to CSP; Carlos O. Miller endowed professorship funds provided to CSP by Indiana University.

Author contributions:

Conceptualization: DF, AM, CSP
Data curation: DF, AM
Methodology: DF, AM, RE, CSP
Investigation: DF, AM, RE
Visualization: DF, AM, RE
Funding acquisition: CSP

Project administration: DF, AM, RE, CSP
Software: DF
Supervision: CSP
Writing – original draft: DF, AM, CSP
Writing – review & editing: DF, AM, RE, CSP

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: Sequencing data have been deposited at GenBank under BankIt IDs 2506918 and 2708174 and will become publicly available upon publication. NOR assembly tools have been submitted to GitHub and will become publicly available upon publication.

Supplementary Materials

Materials and Methods
Figs. S1 to S8
Tables S1 to S2
Data S1 to S10

Literature Cited

1. D. Hernandez-Verdun, P. Roussel, M. Thiry, V. Sirri, D. L. Lafontaine, The nucleolus: structure/function relationship in RNA metabolism. *Wiley interdisciplinary reviews. RNA* **1**, 415-431 (2010).
2. T. Pederson, The nucleolus. *Cold Spring Harb Perspect Biol* **3**, DOI 10.1101/cshperspect.a000638 (2011).
3. J. Saez-Vasquez, M. Delseny, Ribosome Biogenesis in Plants: From Functional 45S Ribosomal DNA Organization to Ribosome Assembly Factors. *Plant Cell* **31**, 1945-1967 (2019).
4. T. W. Turowski, D. Tollervey, Cotranscriptional events in eukaryotic ribosome synthesis. *Wiley interdisciplinary reviews. RNA* **6**, 129-139 (2015).
5. B. McStay, Nucleolar dominance: a model for rRNA gene silencing. *Genes Dev* **20**, 1207-1214 (2006).
6. S. Tucker, A. Vitins, C. S. Pikaard, Nucleolar dominance and ribosomal RNA gene silencing. *Curr Opin Cell Biol* **22**, 351-356 (2010).
7. K. W. Earley *et al.*, Mechanisms of HDA6-mediated rRNA gene silencing: suppression of intergenic Pol II transcription and differential effects on maintenance versus siRNA-directed cytosine methylation. *Genes Dev* **24**, 1119-1132 (2010).
8. K. Earley *et al.*, Erasure of histone acetylation by Arabidopsis HDA6 mediates large-scale gene silencing in nucleolar dominance. *Genes Dev* **20**, 1283-1293 (2006).
9. F. Pontvianne *et al.*, Subnuclear partitioning of rRNA genes between the nucleolus and nucleoplasm reflects alternative epiallelic states. *Genes and Development* **27**, 1545-1550 (2013).
10. C. Chandrasekhara, G. Mohannath, T. Blevins, F. Pontvianne, C. S. Pikaard, Chromosome-specific NOR inactivation explains selective rRNA gene silencing and dosage control in Arabidopsis. *Genes Dev* **30**, 177-190 (2016).
11. F. A. Rabanal *et al.*, Epistatic and allelic interactions control expression of ribosomal RNA gene clusters in Arabidopsis thaliana. *Genome Biol* **18**, DOI 10.1186/s13059-13017-11209-z (2017).
12. G. Mohannath, F. Pontvianne, C. S. Pikaard, Selective nucleolus organizer inactivation in Arabidopsis is a chromosome position-effect phenomenon. *Proc Natl Acad Sci U S A* **113**, 13426-13431 (2016).
13. G. P. Copenhaver, C. S. Pikaard, RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of Arabidopsis thaliana adjoin the telomeres on chromosomes 2 and 4. *Plant J* **9**, 259-272 (1996).
14. J. Sims, G. Sestini, C. Elgert, A. von Haeseler, P. Schlogelhofer, Sequencing of the Arabidopsis NOR2 reveals its distinct organization and tissue-specific rRNA ribosomal variants. *Nat Commun* **12**, DOI 10.1038/s41467-41020-20728-41466 (2021).
15. P. Gruendler, I. Unfried, K. Pascher, D. Schweizer, rDNA intergenic region from Arabidopsis thaliana. Structural analysis, intraspecific variation and functional implications. *J Mol Biol* **221**, 1209-1222 (1991).
16. K. Havlova *et al.*, Variation of 45S rDNA intergenic spacers in Arabidopsis thaliana. *Plant Mol Biol*, (2016).
17. P. Ni *et al.*, Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. *Nat Commun* **12**, 5976 (2021).
18. F. Pontvianne *et al.*, Histone methyltransferases regulating rRNA gene dose and dosage control in Arabidopsis. *Genes Dev* **26**, 945-957 (2012).

19. G. Mohannath *et al.*, DNA hypermethylation and condensed chromatin correlate with chromosome-specific rRNA gene silencing in Arabidopsis. *BioRxiv*, (2023).
20. R. J. Lawrence *et al.*, A concerted DNA methylation/histone methylation switch regulates rRNA gene dosage control and nucleolar dominance. *Mol Cell* **13**, 599-609 (2004).
21. G. Dover *et al.*, in *Genome evolution,* G. A. Dover, R. B. Flavell, Eds. (Academic Press, London, 1982), pp. 343-372.
22. E. A. Zimmer, S. L. Martin, S. M. Beverley, Y. W. Kan, A. C. Wilson, Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci U S A* **77**, 2158-2162 (1980).
23. D. Liao, Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet* **64**, 24-30 (1999).
24. S. E. Stewart, G. S. Roeder, o. B. Y. U. N. H. C. (Department, Transcription by RNA polymerase I stimulates mitotic recombination in Saccharomyces cerevisiae. *Molecular And Cellular Biology; Mol Cell Biol* **9**, 3464-3472 (1989).
25. M. K. Voelkel, G. S. Roeder, o. B. Y. U. N. H. C. (Department, Gene conversion tracts stimulated by HOT1-promoted transcription are long and continuous. *Genetics; Genetics* **126**, 851-867 (1990).

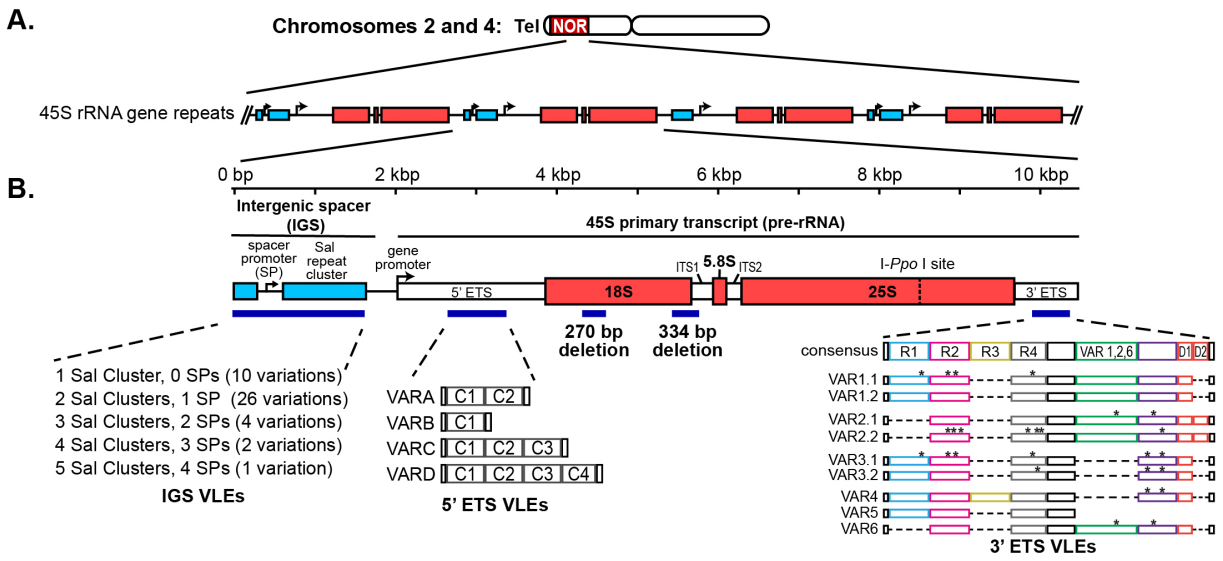

Figure 1. Organization of the *Arabidopsis thaliana* NORs.
(**A**) The NORs abut the telomeres on the short arms of chromosomes 2 and 4 and consist of tandemly repeated copies of 45S rRNA genes. (**B**) Diagram of a single rRNA gene repeat, showing the locations of Variable Length Elements (VLEs). IGS VLEs are named according to the length and arrangement of their Sal clusters, for example 294-SP-1272 contains Sal clusters of 294 bp and 1272 bp separated by a single spacer promoter. Asterisks in the 3'ETS VLEs denote single nucleotide variations from the consensus.

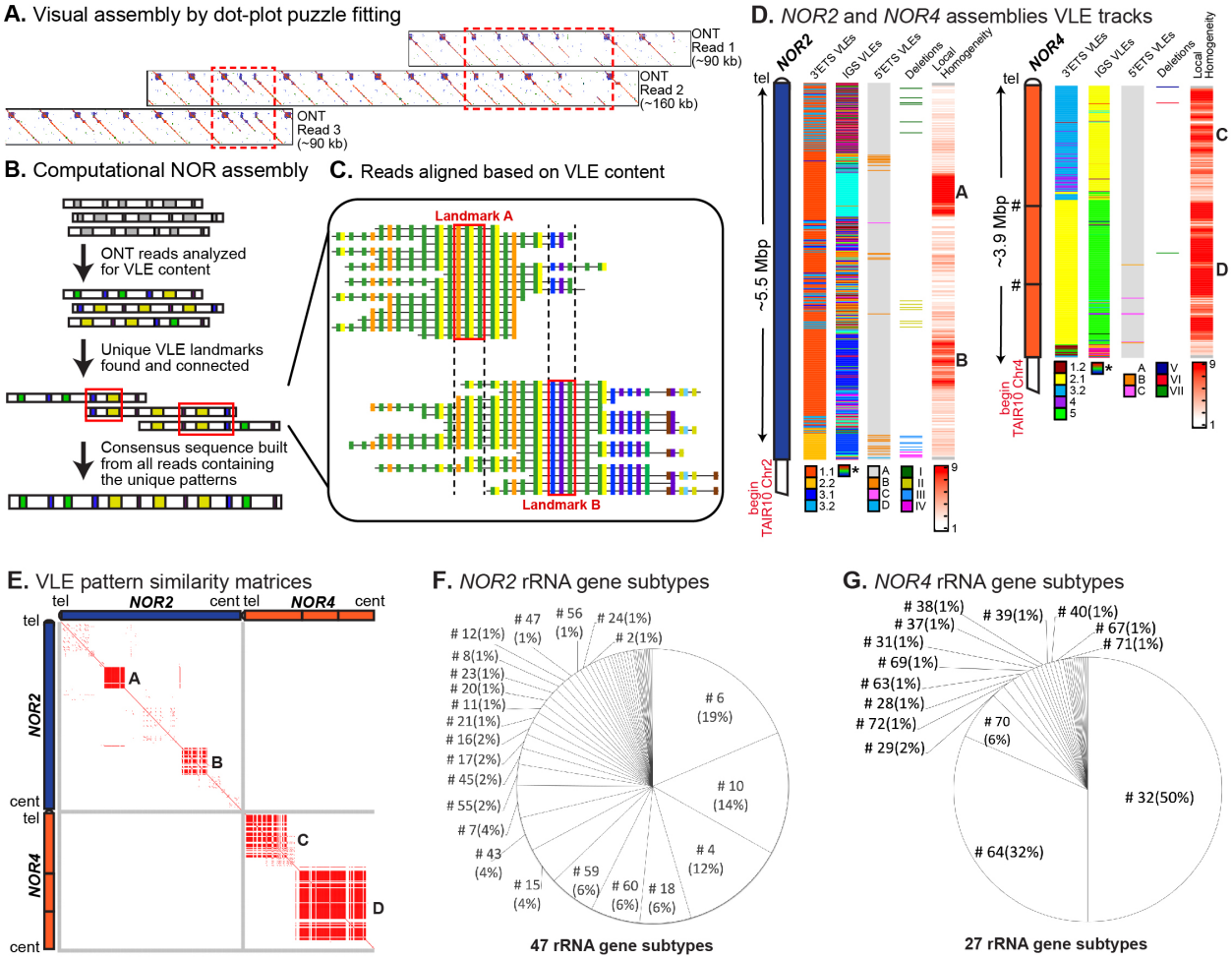

Figure 2. Assembly of NOR2 and NOR4.
(**A**) Example of assembly using overlapping dot-plots. (**B**) Schematic of computational assembly pipeline. (**C**) Diagram of aligned reads that include adjacent VLE landmarks. (**D**). Depiction of the NOR assemblies with tracks for distinct VLEs, deletions, and local homogeneity (defined as number of identical gene subtypes within a nine gene sliding window). Legends below each track indicate colors of each VLE, except for the IGS VLEs which have too many colors (43) to visually distinguish each. (**E**) Similarity matrices comparing VLE content within and between NORs. Each red point represents a perfect match of VLEs within a three gene window. (**F-G**). Pie charts representing the relative abundance of gene subtypes within *NOR2* or *NOR4*. See fig. S2 for subtype details.

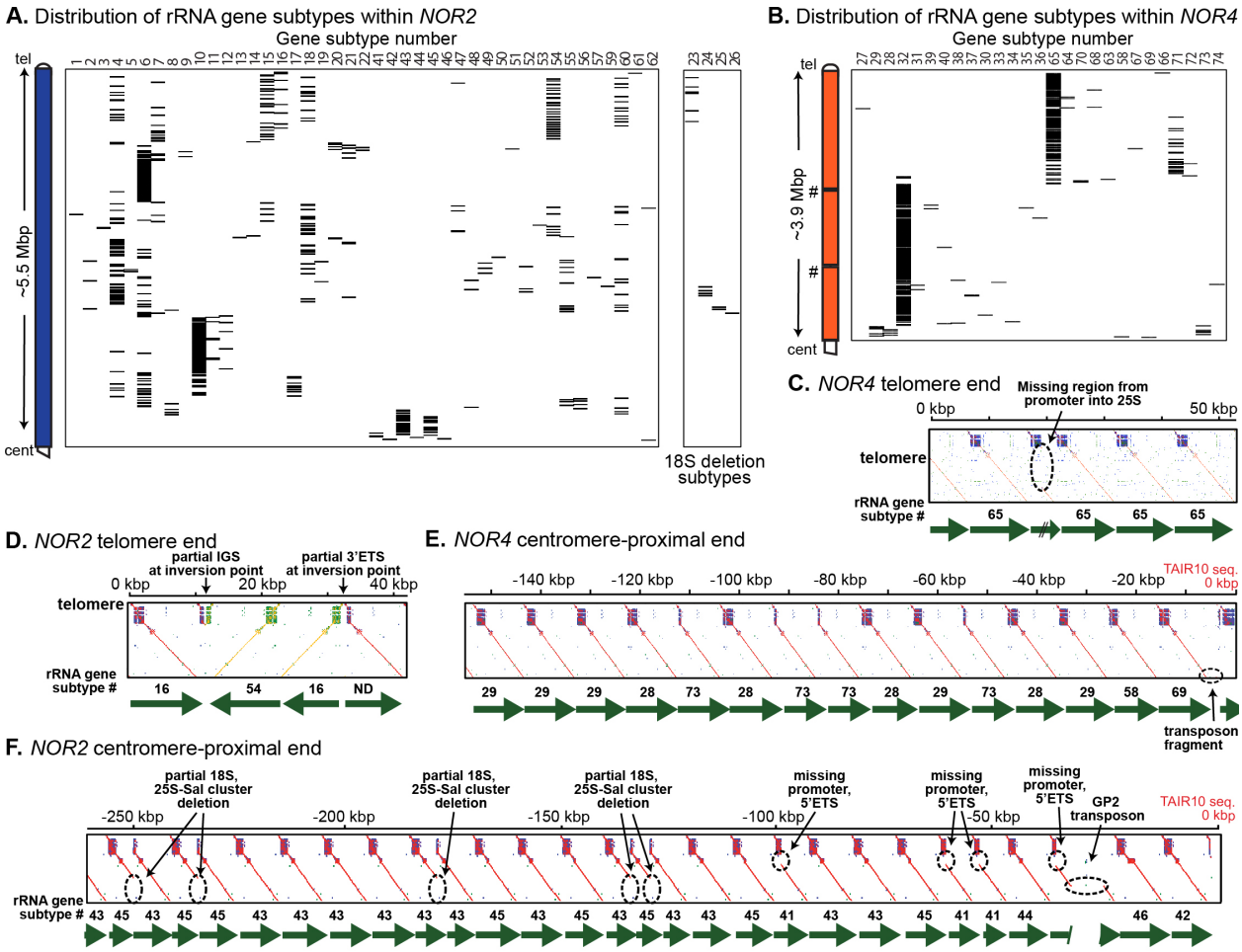

Figure 3. Structural features of NOR2 and NOR4.
(**A-B**) Positions of individuall gene subtypes throughout *NOR2* and *NOR4*. The right panel of (A) shows subtypes containing 18S deletions. (**C-F**) Similarity matrices for the telomere and centromere-proximal ends of *NOR2* and *NOR4*. Distinctive features are highlighted with dashed circles. Numbering is relative to the telomere-rRNA gene junction in panels (C) and (D) and relative to the start of the TAIR10 chromosome 2 and 4 assemblies for (E) and (F).

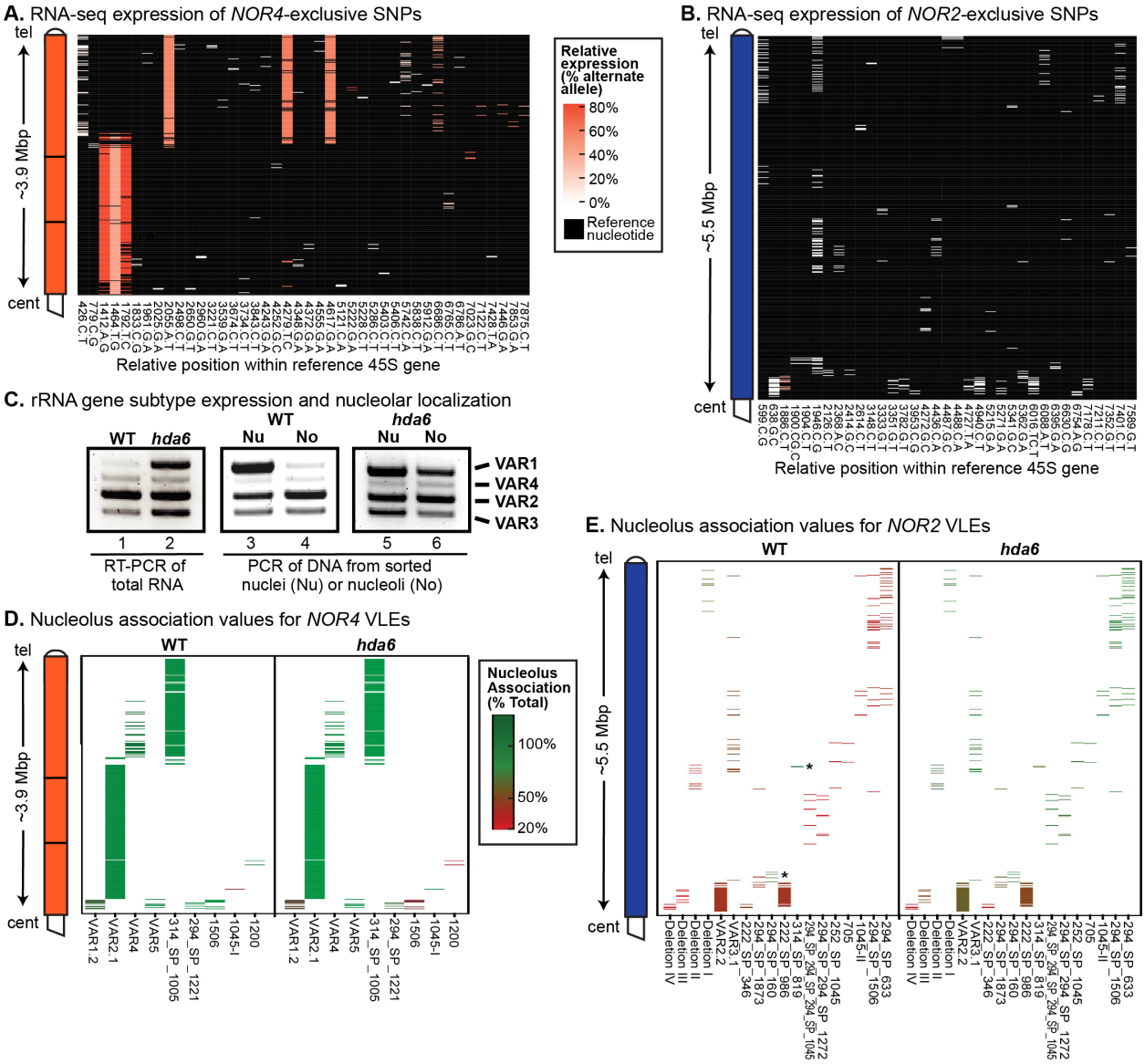

Figure 4. rRNA gene activity and nucleolar association maps for NOR2 and NOR4

(**A-B**) Heatmaps of the expression level of NOR4 or NOR2-specific SNPs at gene positions represented by lines on the y-axis. SNP positions are numbered according to their position in the ribosomal gene reference sequence. White to red colors indicate level of expression based on normalized RNA-seq read count. Black indicates reference nucleotide. (**C**) RT-PCR of the 3'ETS VAR region, comparing WT and hda6 mutants (left panel) and comparing nuclear versus nucleolar associated genes (middle and right panel). (**D-E**) Heatmaps depicting degree of nucleolus association for *NOR4* or *NOR2*-specific VLEs. Asterisks denote rare *NOR2* VLEs that are nucleolus associated. All experiments were performed using mature leaf tissue.

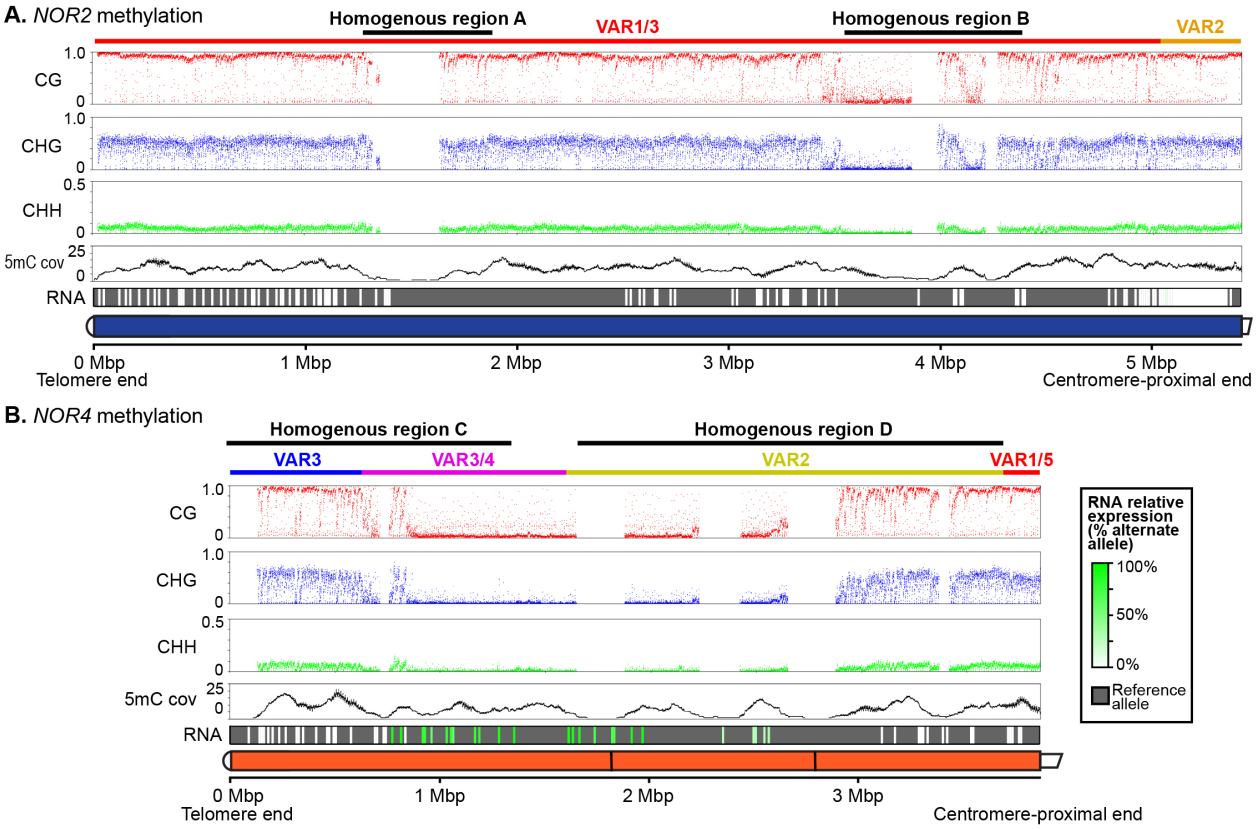

Figure 5. Methylation landscapes of NOR2 and NOR4.
(**A-B**) 5mC frequencies in the CG, CHG, and CHH contexts at *NOR2* and *NOR4*. Each colored point represents a sliding window of 500bp with a 250bp step. Regions where there are no data points in the methylation tracks were below our coverage threshold for 5mC calling, as seen in the 5mC cov track, which shows the average number of reads for sites within a window that were retained by the methylation calling pipeline. The RNA track shows the relative expression levels of local SNPs. Experiments were performed using inflorescence apex tissue.