Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response
Ariadna Picart-Picolo, Stefan Grob, Nathalie Picault, Michal Franek, Christel Llauro, Thierry Halter, Tom Maier, Edouard Jobet, Julie Descombin, Panpan Zhang, et al.

To cite this version:
Ariadna Picart-Picolo, Stefan Grob, Nathalie Picault, Michal Franek, Christel Llauro, et al.. Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response. Genome Research, Cold Spring Harbor Laboratory Press, In press, 10.1101/gr.261586.120. hal-02962580

HAL Id: hal-02962580
https://hal-univ-perp.archives-ouvertes.fr/hal-02962580
Submitted on 9 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response

Ariadna Picart-Picolo, 1,2 Stefan Grob, 3 Nathalie Picault, 1,2 Michal Franek, 4 Christel Llauro, 1,2 Thierry Halter, 5 Tom R. Maier, 6 Edouard Jobet, 1,2 Julie Descombin, 1,2 Panpan Zhang, 2,7 Vijayapalani Paramasivan, 6 Thomas J. Baum, 6 Lionel Navarro, 5 Martina Dvořáková, 4 Marie Mirouze, 2,7 and Frédéric Pontvianne 1,2

1 CNRS, 2 UPVD, LGDP UMR5096, Université de Perpignan, 66860 Perpignan, France; 3 Institute of Plant and Microbial Biology, University of Zurich, CH-8008 Zurich, Switzerland; 4 Mendel Centre for Plant Genomics and Proteomics, CEITEC, Masaryk University, 625 00 Brno, Czech Republic; 5 ENS, IBENS, CNRS/INSERM, PSL Research University, 75005 Paris, France; 6 Department of Plant Pathology and Microbiology, Iowa State University, Ames, Iowa 50011, USA; 7 IRD, UMR232 DIADE, 34394 Montpellier, France

Rapid plant genome evolution is crucial to adapt to environmental changes. Chromosomal rearrangements and gene copy number variation (CNV) are two important tools for genome evolution and sources for the creation of new genes. However, their emergence takes many generations. In this study, we show that in Arabidopsis thaliana, a significant loss of ribosomal RNA (rRNA) genes with a past history of a mutation for the chromatin assembly factor 1 (CAF1) complex causes rapid changes in the genome structure. Using long-read sequencing and microscopic approaches, we have identified up to 15 independent large tandem duplications in direct orientation (TDDOs) ranging from 60 kb to 1.44 Mb. Our data suggest that these TDDOs appeared within a few generations, leading to the duplication of hundreds of genes. By subsequently focusing on a line only containing 20% of rRNA gene copies (20rDNA line), we investigated the impact of TDDOs on 3D genome organization, gene expression, and cytosine methylation. We found that duplicated genes often accumulate more transcripts. Among them, several are involved in plant–pathogen response, which could explain why the 20rDNA line is hyper-resistant to both bacterial and nematode infections. Finally, we show that the TDDOs create gene fusions and/or truncations and discuss their potential implications for the evolution of plant genomes.

[Supplemental material is available for this article.]
named 20rDNA L6) has a wild-type phenotype and retained a low amount of rRNA genes for five generations (referred as F5) (Pavlištová et al. 2016). In this study, we took advantage of this plant material to test the impact of a low amount of rRNA genes on plant genome stability during several generations. We found unexpected consequences on the genome structure and stability, as well as on its 3D genome organization. We also show the short-term consequences of gene copy number variation (CNV) on their expression and potentially their role in plant phenotypic traits such as pathogen responses.

Results

The 20rDNA L6 line accumulates features of genomic instability

The 20rDNA L6 line contains only 20% of rRNA genes compared to the wild-type Col-0 and was obtained as a wild-type segregant from a cross between fas1-4 and fas2-4 mutant lines as described (Fig. 1A; Pavlištová et al. 2016). To confirm and precisely map the potential consequence of genomic instability in 20rDNA L6F6 (F6 for the sixth generation after F1), we performed long-read resequencing using nanopore technology. We obtained 6.4 Gb of total sequences with a midsize of 6 kb. We then analyzed the sequencing coverage against the TAIR10 A. thaliana Col-0 reference genome to identify the highly covered regions (Fig. 1B). We have detected seven large duplications, corresponding to tandem duplications in direct orientation (TDDO), named TDDO1 to TDDO7. The largest region, TDDO4, represents 1.44 Mb, spanning the heterochromatic knob on the short arm of Chromosome 4 (hk4s), a large heterochromatic region outside the pericentromeres, and a euchromatic region distal to the knob. Other TDDOs range in size from 60 to 370 kb long and are present on Chromosomes 1, 2, 4, and 5 (Fig. 1B). The absence/presence of TDDO4, the largest duplication, was also confirmed by DNA-fluorescence in situ hybridization (FISH) (Fig. 1C). We used two probes generated from BAC clones: one recognizing a portion of TDDO4 (hk4s–T5H22) and one recognizing an unduplicated genomic region located between TDDO4 and the NOR4 (F5I10). Different cell types were analyzed from vegetative as well as reproductive tissues: in both, more signals corresponding to TDDO4 were detected in the 20rDNA L6F6 nuclei compared to wild-type Col-0 cells (Supplemental Fig. S1). Analyses of pachytene chromosomes clearly showed that the additional signal actually belonged to the same chromosome, which confirms the duplication hypothesis (Fig. 1C).

The occurrence of duplication events is a sign of genomic instability. Thus, the chromosomal rearrangements observed in 20rDNA L6F6 could be the consequence of double-stranded breaks (DSBs). To test this hypothesis, we compared the amount of spontaneous DSBs between 20rDNA L6F6 and wild-type Col-0 cells by performing immunostaining of serine 139-phosphorylated H2Ax histone variant (P-γ-H2Ax) on the short arm of Chromosome 4 (hk4s), a large heterochromatic region outside the pericentromeres, and a euchromatic region distal to the knob. Other TDDOs range in size from 60 to 370 kb long and are present on Chromosomes 1, 2, 4, and 5 (Fig. 1B). The absence/presence of TDDO4, the largest duplication, was also confirmed by DNA-fluorescence in situ hybridization (FISH) (Fig. 1C). We used two probes generated from BAC clones: one recognizing a portion of TDDO4 (hk4s–T5H22) and one recognizing an unduplicated genomic region located between TDDO4 and the NOR4 (F5I10). Different cell types were analyzed from vegetative as well as reproductive tissues: in both, more signals corresponding to TDDO4 were detected in the 20rDNA L6F6 nuclei compared to wild-type Col-0 cells (Supplemental Fig. S1). Analyses of pachytene chromosomes clearly showed that the additional signal actually belonged to the same chromosome, which confirms the duplication hypothesis (Fig. 1C).

The occurrence of duplication events is a sign of genomic instability. Thus, the chromosomal rearrangements observed in 20rDNA L6F6 could be the consequence of double-stranded breaks (DSBs). To test this hypothesis, we compared the amount of spontaneous DSBs between 20rDNA L6F6 and wild-type Col-0 cells by performing immunostaining of serine 139-phosphorylated H2Ax histone variant (P-γ-H2Ax), which is a marker of DSB (Charbonnel et al. 2010). P-γ-H2Ax foci were detected at a higher rate in 20rDNA L6F6 nuclei compared to wild-type Col-0 cells (Supplemental Fig. S2A). Accumulation of DSB foci can potentially be associated with a DNA repair defect. This hypothesis is supported by an increased susceptibility of the...
20rDNA L6F6 line to a treatment with the genotoxin bleomycin (Supplemental Fig. S2B).

Appearance of the duplication events in the 20rDNA line

To show a potential link between low rDNA copies and TDDO appearance, it is crucial to know when these duplication events occurred. Like 20rDNA L6F6 line, 20rDNA L9F6 is an independent inbred line deriving from the cross between fas1-4 and fas2-4 mutants that both also display low amounts of rDNA copies (Supplemental Fig. S3; Mozgová et al. 2010; Pavlišová et al. 2016). We then performed long-read resequencing using nanopore technology and identified TDDO in the 20rDNA L9F6 line, as well as in the offspring of the parental lines fas1-4 and fas2-4 used to generate the initial cross (Fig. 2A).

In fas1-4, none of the seven TDDOs identified in 20rDNA L6F6 were detected, but we found six new TDDOs ranging from 57 to 175 kb long (Fig. 2A; Supplemental Fig. S4A). In fas2-4, only TDDO4 is present, as well as two additional duplications named TDDO8 (286 kb) and TDDO9 (106 kb) (Fig. 2A; Supplemental Fig. S4B). We also identified a deletion of 9.75 kb named DEL1 on Chromosome 4 (Fig. 2A; Supplemental Fig. S5). Analyses of 20rDNA L9F6 revealed that TDDOs 1, 5, and 7 are shared between L9F6 and L6F6, which suggests their presence in the F1 (Fig. 2A; Supplemental Fig. S4C). We could not find TDDO4, although this duplication is in one of the parents. Further analyses by quantitative PCR and DNA-FISH revealed that TDDO4 has been segregated out between the L9F2 and L9F4 (Supplemental Fig. S4D). This hypothesis is supported at least for TDDO3 by qPCR analyses (Fig. 2B). However, owing to a lack of long-read sequences obtained for L9F6, we are not able to determine the existence of novel TDDO.

In summary, our analyses identified 15 TDDOs that appeared independently, either in the parental line or in the two independent inbred lines resulting from the fas1-4 and fas2-4 cross (Fig. 2A; Supplemental Figs. S4A,D, S7). This hypothesis is supported by the absence of TDDO4 in all generation 1 (G1) mutants fas2-4 and fas2-5 analyzed by PCR (Supplemental Fig. S4E,F), suggesting the also very recent appearance of TDDO4 in the parental fas2-4 (G4) line.

Impact of low rDNA and TDDO on 3D genome organization

The nucleolus plays an important role in the spatial organization of the chromosomes (Bersaglieri and Santoro 2019; Pontvianne and Liu 2020; Santos et al. 2020). Nucleolus-associated chromatin domains (NADs), essentially composed of repressed chromatin domains, localize at the nucleolar periphery (Németh et al. 2010; van Koningsbruggen et al. 2010; Pontvianne et al. 2016b). Because rRNA gene nuclear distribution has a critical impact in NADs identity both in plant and animal cells (Quinodoz et al. 2018; Picart-Picolo et al. 2019, 2020), we analyzed NADs composition and 3D organization in 20rDNA L6F6. The fifth generation of the 20rDNA L6 (20rDNA L6F5) was transformed with a transgene ectopically expressing the FIBRILLARIN 2 nucleolar protein fused to the yellow fluorescent protein (FIB2:YFP). Using the FIB2:YFP nucleolar marker, we isolated nuclei and nucleoli from the transformants and identified NADs nuclear and nucleolar DNA sequences as previously described (Pontvianne...
et al. 2016a; Carpentier et al. 2018). Preceding studies have clearly shown that in wild-type Col-0 leaf cells, NOR4-derived rRNA genes are expressed and associate with the nucleolus. Conversely, NOR2 is excluded from the nucleolus, and NOR2-derived rRNA genes are silent (Pontvianne et al. 2013; Chandrasekhar et al. 2016). As a result, NADs are essentially distributed in the entire short arm of Chromosome 4 (kr4s), which juxtaposes the active NOR4 and associates with the nucleolus (Pontvianne et al. 2016b). Compared to the wild-type, NADs in 20rDNA L6F6 are enriched from genomic regions located on both Chromosomes 2 and 4 short arms (Fig. 3A; Supplemental Fig. S8). Among the 434 genes that gained nucleolar association, 144 belong to Chromosome 2 (33%) (Supplemental Table S1). In contrast, only 19 genes on Chromosome 4 gain nucleolar association. These results are consistent with the rDNA transcriptional state, as all leftover NOR2 and NOR4-derived rRNA genes are actively transcribed and associate with the nucleolus (Pavlíštová et al. 2016). We also detected an enrichment of centromeric sequences associating with nucleoli in L6F6 compared to wild type. This type of reorganization was previously shown to associate with changes in NOR subnuclear organization (Pontvianne et al. 2016b; Pontvianne and Grob 2020). As in wild type, subtelomeric regions remain associated with the nucleolus in the 20rDNA L6F6 line (Fig. 3A; Supplemental Fig. S8). In summary, NAD identification in 20rDNA L6F6 revealed that 5.6 Mb of chromatin domains mainly enriched in silent epigenetic marks changed their subnuclear distribution, which suggests a substantial reorganization of the nuclear genome.

Figure 3. 3D genome organization in L6F6. (A) Chromosome plots displaying the relative enrichment of a given genomic segment with the nucleolus. The y-axis displays the fold change nucleolus enrichment between wild-type Col-0 and the 20rDNA L6F6. Each dot represents a 100-kb window. Nucleolus-enriched genomic regions above the threshold are red, and depleted regions are violet. (B) Coverage-normalized t-test difference matrix (50-kb bins). The color of each pixel of the matrix is defined by the result of a t-test using the triplicate contact frequencies from wild-type and 20rDNA coverage-normalized Hi-C samples. The two magnified areas correspond to the two regions displaying the highest level of contact frequency changes. (C,D) Non-normalized Hi-C snapshot showing the contact frequencies on the short arm of Chromosome 4 in wild-type Col-0 (C) versus the 20rDNA L6F6 (D). TEs and genes are annotated to illustrate the occurrence of euchromatin and heterochromatin, respectively. (E) Ratio between Hi-C contact frequencies from wild-type and 20rDNA L6F6. Negative ratios correspond to more contacts in the wild type, whereas positive ratios correspond to more contacts in the 20rDNA L6F6. (F) Eigenvector of the wild-type Co-0 Hi-C data set and annotation of the TDDO4 affecting the knob hk4s. Note the central duplication breakpoint exactly coincides with a change between LSD and a CSD.

To get a global view of the chromatin 3D organization, we analyzed all chromatin–chromatin interactions using genome-wide chromosome-conformation capture (Hi-C). We generated triplicate Hi-C samples from both wild-type and 20rDNA L6F6 14-d-old seedlings (Fig. 3B; Supplemental Fig. S9). To assess differences between two given sets of Hi-C samples statistically, we took advantage of our triplicate Hi-C data sets and performed student t-tests on each contact frequency (pixel of the Hi-C matrix) and determined whether contact frequencies significantly changed between the wild-type and 20rDNA L6F6 (Fig. 3C,E). Contact frequencies assayed by Hi-C can be used to detect chromosomal rearrangements (Himmelbach et al. 2018). In our case, a duplication would lead to a twofold increase in coverage of the affected region, thus doubling of interaction frequencies at this region. We indeed found several regions displaying a significant (P<0.01) increase of contact frequencies at several chromosomal locations, all corresponding to the previously described TDDO1 to TDDO7. Analyzing the genome-wide coverage using unpaired raw Hi-C sequencing reads confirmed the presence of significant increase in coverage of the affected regions (Supplemental Fig. S10). We subsequently normalized our Hi-C matrices for the assayed coverage. However, coverage-normalized Hi-C data showed that short-range contact frequencies within the duplicated regions are significantly depleted. Whether this depletion of contact frequencies is biologically significant or represents an artifact of the normalization procedure is extremely difficult to determine.

To further examine potential differences in 3D folding principles between wild-type Col-0 and 20rDNA L6F6, we performed a
principal component analysis (PCA) to retrieve the eigenvector, which is characteristic of 3D folding patterns of a Hi-C data set (Grob et al. 2014; Lieberman-Aiden et al. 2009). Sign changes in the eigenvector delineate basic 3D folding domains, known as loose structural domains (LSDs) and closed structural domains (CSDs), which are analogous to animal A and B compartments (Lieberman-Aiden et al. 2009). We could not identify significant changes in the eigenvectors between wild-type Col-0 and the 20rDNA L6F6. Moreover, outside the duplicated regions, no changes in genomic bin contact frequencies could be observed. We therefore focused on the duplicated regions and analyzed duplication breakage points with the eigenvector obtained by the PCA analysis of the wild-type Col-0 Hi-C data (Supplemental Fig. S11). We observed that in a majority of the TDDOs in L6F6, at least one of the breakage points coincides with sign changes (CSDs to LSDs) or directional changes (valleys and peaks within a structural domain) in the eigenvector, with the exception of TDDO3. Hence, the changes in 3D conformation may have facilitated the occurrence of the TDDOs. This was most prominent for TDDO4, where the more central breakpoint exactly colocalizes with the change between the CSD and the LSD, which defines the ancient inversion breakpoint that gave rise to the knob (Fig. 3F; Zapata et al. 2016). This suggests the existence of continuously fragile chromosomal regions, the borders between structural domains being diagnostic for these regions.

**Duplication events create chimeric genes**

Most of the time, TDDOs keep genes intact and do not lead to gene loss. However, truncated genes can be generated at the breakpoint junction, while keeping intact genes on the edges of duplication (Newman et al. 2015). Besides, when breakpoints are located in two different genes in the same orientation, gene fusion can take place if the open reading frame (ORF) is preserved. In 20rDNA L6F6, we systematically analyzed the TDDO breakpoint junctions (Supplemental Figs. S7, S12). Of the seven cases of TDDO identified in this line, three potentially created fused or truncated proteins (Fig. 4A,F). On Chromosome 1, TDDO1 fused the first exon of AT1G53325 that encodes the N-terminal domain of the MEDiator 13-like with four of the five exons of AT1G54770 that encodes the FC2 pre-rRNA processing factor. On Chromosome 2, although genes are in the opposite orientation, TDDO3 creates a shorter ORF of the AT2G38460 gene that potentially produces a truncated FERROPORNTIN 1 protein. Finally, on Chromosome 4, TDDO4 fused the AT4G05475 gene to a transposable element (TE) (AT4G02960), leading to the potential expression of three new ORFs, including one that encodes a protein with two leucin rich repeats (LRR) (Fig. 4A,F). We then systematically analyzed the presence of these chimeric genes in the genome of the parental fas mutant lines and in 20rDNA L6 and L9, respectively (Fig. 4G). The TDDO1-derived chimeric gene can be detected in both L6 and L9, which confirm the appearance of TDDO1 after the cross between fas 1–4 and fas 2–4 (Figs. 2, 4G). The chimeric gene generated from TDDO3 was specifically detected in L6, whereas the chimeric gene generated by TDDO4 was detected in fas 2–4, L6, and L9 plants, confirming the results obtained earlier (Fig. 2) but also suggesting that some generations of 20rDNA L9 inbreeds plants may still segregate TDDO4.

We finally investigated whether these chimeric genes were transcribed. A first analysis of our RNA-seq data revealed that these genes were all able to accumulate transcripts. Using RT-qPCR, we confirmed the expression of the TDDO1- and TDDO3-derived chimeric genes, as well as the ability of the TDDO1-derived chimeric gene to be properly spliced (Supplemental Fig. S13). However, although reads could be detected in the RNA-seq data, we did not detect any signals for the TDDO4-derived chimeric gene by RT-

**Figure 4.** TDDOs provoke chimeric genes formation. (A–C) Schematic representation of TDDO1, TDDO3, and TDDO4 that provoked TE and/or gene fusion in the 20rDNA L6F6. Genes or TEs present in the breaking junction and their orientation are shown. (D–F) Open reading frames (ORFs) potentially generated at the breaking points. TDDO1 provokes the fusion of the first exon of AT1G53325 that encodes for an ATPase motif of MEDiator 13 and the last four exons of AT1G54770 that contain an RNA processing domain (D). The chimeric gene created between AT2G38460 and AT1G54770 potentially encodes for a truncated FERROPORNTIN protein (E). The breaking points at TDDO4 fuse the S’ sequence of a TE (AT4G02960) with the second and last exon of the gene AT4G05475, which sequence encodes two Leucine Rich Repeats (LRR) (F). (G,H) PCR was performed with primers flanking the breaking junctions of TDDO1, TDDO3, and TDDO4 in the wild-type Col-0, the two mutants fas 1–4 (G5) and fas 2–4 (G5), and in 20rDNA lines L6 (generations F6 and F8) and L9 (generation F7). All PCR products were confirmed by Sanger sequencing. Genomic DNA (G) and cDNA (H) were used as templates. Amplicons from the locus encoding the elongation factor EF1ALPHA was used as a loading control.
PCR (Fig. 4H). In conclusion, our data show that TDDOs can promote the expression of chimeric genes.

Characterization and impact of duplication events on gene expression

All TDDOs gained in 20rDNA L6F6 correspond to a gain of 2.31 Mb per haploid genome and induce CNVs of 626 genes and 851 transposable elements (TEs) (Supplemental Table S1). Changes in the 3D genome organization and CNVs can have an impact on chromatin marks and gene expression. We therefore analyzed the global gene expression pattern by poly(A)+ RNA-seq and the methylome by whole-genome bisulfite sequencing (WGBS) in wild-type Col-0 versus 20rDNA L6F6. We analyzed four replicates per samples by RNA-seq and identified differentially accumulating transcripts: 321 up-regulated genes and 14 up-regulated TEs, as well as 37 down-regulated genes but no down-regulated TEs in 20rDNA L6F6 compared to the wild-type Col-0 (with an adjusted P-value < 0.01 and log2(fold change) > 1.5 or < 1.5) (Fig. 5A; Supplemental Table S1). We confirmed these results by quantitative RT-PCR (RT-qPCR) on nine randomly chosen genes and TEs (Supplemental Fig. S14). We did not find any correlation between differentially expressed genes and genes located in the newly arisen NADs of 20rDNA L6F6 (Supplemental Fig. S15A,C).

However, we found that duplicated genes and TEs were significantly more expressed (Supplemental Fig. S15D,E). Of the up-regulated TEs, 57% (8) are also duplicated. If we consider the 321 up-regulated genes with a log2 fold change enrichment of 1.5, we found that 22% of these genes (71) belonged to duplicated genes, but the TDDOs only represent 2% of the genome. Conversely, no genes present in TDDO are down-regulated. Higher expression can only be observed from initially expressed genes in wild-type plants. Only 286 duplicated genes are actually expressed, and 160 of them are at least twice more expressed in 20rDNA L6F6 than in wild-type Col-0 (Fig. 5B). Depending on their genomic location, TDDOs perform differently. For instance, most of the TDDO3-derived genes produced at least twice as many transcripts in 20rDNA L6F6 (71 up-regulated genes of the 80 expressed genes), whereas genes present in TDDO4, enriched in genomic regions with heterochromatic features, were less up-regulated (61 fold change >2 genes of the 142 expressed genes) (Fig. 5B). Finally, box-plot analyses of all genes versus the duplicated genes indeed revealed their overall ability to overaccumulate more transcripts in 20rDNA L6F6 (Fig. 5C). Thus, our data strongly suggest that gene duplication often leads to an increased expression, often higher than the twofold change expected in the hypothesis of additive expression.

To analyze the impact of CNVs at the DNA methylation level, we performed triplicate WGBS in wild-type Col-0 versus 20rDNA L6F6 lines. At the genome-scale, we observed a modest increase in CG, CHG, and CHH methylation in 20rDNA L6F6 at genes (Fig. 5E,F; Supplemental Fig. S16). However, methylation at TEs was affected in both CHG and CHH contexts, but not in the CG context (Fig. 5G,H; Supplemental Fig. S17). This observation is also true if we only analyze duplicated or up-regulated genes, with the exception of gene body methylation that is unaffected for up-regulated genes (Fig. 5D; Supplemental Fig. S16). Finally, differentially methylated regions (DMRs) identified in 20rDNA L6F6 compared to wild-type Col-0 did not show a potential overlap between up-regulated genes and hypomethylated regions.

Duplication events are linked to higher pathogen resistance

In the pool of up-regulated genes in 20rDNA L6F6, genes implicated in biotic and stress responses are particularly enriched (Fig. 6A).
We performed RT-qPCR experiments and confirmed the overexpression of key genes involved in the plant–pathogen response (Fig. 6B). Among these genes are PATHOGENESIS-RELATED GENE 1 (PR1) and PATHOGENESIS-RELATED GENE 5 (PR5), whose higher expression levels are usually correlated with increased resistance against bacteria and nematodes (Wubben et al. 2008). Some of these genes were found in TDDO3 and TDDO4. Their higher expression rate could therefore be a consequence of the duplication events (Supplemental Fig. S17). Among them, ASYMMETRIC LEAVES 1 (AS1), present in TDDO3, has an evolutionarily conserved role in plant–pathogen interactions (Yang et al. 2008). AS1 indeed acts as a positive regulator of extracellular defenses against bacterial pathogens in a salicylic acid-independent manner (Nurnberg et al. 2007). In addition, genes encoding four cysteine-rich receptor-like kinases (CRKs), located in TDDO4, also overaccumulate transcripts in the L6F6 (Supplemental Fig. S18). Among these genes is CRK36, whose overexpression is sufficient to enhance pattern-triggered immune response and bacterial pathogen resistance (Yeh et al. 2015).

_A. thaliana_ is susceptible to various pathogens, from prokaryotes to multicellular organisms. To test their resistance capabilities, we first infected both the wild-type Col-0 and 20rDNA L6F6 with the sugar beet cyst nematode _Heterodera schachtii_ (Fig. 6C). We observed that only half the number of females was able to develop on 20rDNA L6F6 plants in comparison with wild-type Col-0 plants. However, we did not observe a change in the syncytium feeding site size, that is, the plant feeding structure induced by these nematodes (Fig. 6C). Secondly, we tested the ability of 20rDNA L6F6 to be infected by the virulent bacteria _Pseudomonas syringae_ strain DC3000. Three days after inoculation, bacterial growth was significantly lower in 20rDNA L6F6 (Fig. 6D) than in wild-type Col-0. Single-nucleotide polymorphisms (SNPs) could also explain changes in plant–pathogen responses, but our analyses revealed that among the 196 SNPs found in genes L6F6 compared to wild-type Col-0, none correspond to genes implicated in biotic stress response (Supplemental Fig. S19). Considering that _fas2-4_ mutant is hyper-resistant to _P. syringae_ (Mozgová et al. 2015), and we identified TDDO4 in some lines of this mutant, one hypothesis is that the overexpression of pathogen response genes present in TDDO4 rather than _FAS2_ gene mutation is directly implicated in the resistance against _P. syringae_. However, we cannot exclude that _fas2-4_ and L6F6 pathogen resistance is mediated independently of TDDO4, which could also explain why very little overlap can be observed among the up-regulated genes in both lines (Supplemental Fig. S20).

In conclusion, we showed that higher accumulation of transcripts from genes implicated in the plant–pathogen response correlate with the plant’s ability to resist against at least two types of distinct pathogens.

**Discussion**

Genomic structural variations shape animal and plant genomes (Krasileva 2019). Within a period of several millions of years, numerous rearrangements have occurred to shape the _Arabidopsis thaliana_ genome, including duplications, translocations, inversions, and deletions (Blanc et al. 2000; Henry et al. 2006). Recently, genome analysis of seven accessions of _A. thaliana_ revealed that they contain, on average, 15 Mb of rearranged sequences, generating CNVs for thousands of genes (Jiao and Schneeberger 2020). In this case, deletions, gain, or loss of copies are considered as important sources of CNVs and have potentially occurred in tens of thousands of years of evolution (Fuligoni and Hancock 2018). CNVs occurring in the context of tandem duplication events represent between 3 and 4 Mb of genomic sequences in each of the seven accesses sequenced (Jiao and Schneeberger 2020). In our case, only a few generations were necessary to gain up to several megabases of genomic sequences by tandem duplications.

The rapid occurrence of these rearrangements is particularly intriguing. The relative sensitivity to genotoxic stress and the detection of a higher rate of spontaneous DSB in our 20rDNA lines is certainly one source of their appearance (Fig. 1D; Supplemental Fig. S2), but the precise mechanisms remain to be determined. One possibility is the implication of nonallelic homologous recombination (NAHR), usually responsible for TDDO (Zhang et al. 2013; Krasileva 2019). This mechanism can generate segmental duplications or deletions. In the 20rDNA L6F6, we detected duplications but no deletions, probably because of their deleterious effects.

Two other particular aspects of the detected TDDOs are their large sizes and locations, ranging from 57 kb to 1.44 Mb (Fig. 2; Supplemental Fig. S7). The TDDO borders do not share any genetic elements or particular genes. However, our Hi-C data revealed that...
sign changes in the eigenvector seem to be overrepresented at breaking junctions, suggesting a potential link between the 3D genome folding and the occurrence of TDDOs. The systematic identification and characterization of additional TDDOs would be necessary to strengthen this hypothesis.

It is also intriguing that more than half of the 15 TDDOs are located on NOR-bearing chromosomes (Fig. 2; Supplemental Fig. S7). Because of their tandemly repeated nature, NORs are indeed subjected to an inherent instability. Therefore, the existence of a sensing system monitoring their abundance has been proposed (Nelson et al. 2019), potentially via unequal sister chromatid exchange (Tartof 1974a,b). The 20rDNA lines derive from the cross between fas1 and fas2 mutants, whose mutations provoked a gradual loss of rRNA genes copies (Mozgová et al. 2010). Importantly, L6 and L9 are the only siblings in which the number of rRNA genes remained stable at a low level, whereas all other lineages quickly acquired rRNA genes (Pavlíšťová et al. 2016). However, our data actually show that rRNA gene copies are increasing progressively throughout the inbreeding of 20rDNA L6F6 (Supplemental Fig. S6), suggesting that the CNVs are found not only at the level of the TDDOs, but also at the level of the NORs. It remains to elucidate whether a link between the rRNA gene gains and the appearance of TDDO exists and if the same mechanisms are involved. Nevertheless, a loss of rRNA gene copies also associates with genomic instability and hypersensitivity to DNA damage in cancer cells (Wang and Lemos 2017; Xu et al. 2017). Moreover, DNA damage sensitivity and rDNA replication defects also occur in budding yeast low rDNA copy strains (Ide et al. 2010).

Short-term consequences of gene duplications have been studied in animals, especially in cancer cells, where multiple de novo tandem duplication events induce gene CNVs (Quigley et al. 2018; Wee et al. 2018). The 20rDNA L6F6 line is an unprecedented opportunity to study the transcriptional behavior of newly duplicated genes. Globally, duplicated genes tend to be more expressed (Fig. 5C). Previous observations suggest that the expression of tandem genes recently duplicated is often greater than twofold (Loehlin and Carroll 2016). Although we cannot exclude that the detected transcripts come from only one of the duplicated genes, it is more likely that equivalent additive expression occurs for the duplicated genes. During evolution, duplicated gene expression can quickly lead to specialized expression patterns, often in a tissue-specific manner, although a significant number retain correlated transcriptional profiles (Blanc and Wolfe 2004; Guschanski et al. 2017). In our case, we were able to correlate this change in gene expression with the acquisition of increased resistance to different pathogens (Fig. 6). Analyzing gene expression in the future generation will allow us to evaluate if rapid transcriptional regulation occurs.

Plant genomes are rapidly evolving and their capacity to adapt to environmental changes is crucial. Like genome hybridization and TE mobilization, CNV is one important tool of genome evolution (Kondrashov 2012; Gabur et al. 2019; Quadra et al. 2019). Together with previous observation, our data show the importance of systematically detecting CNVs. CNVs can indeed associate with adaptive traits (Kondrashov 2012; Gabur et al. 2019; Alonge et al. 2020). In our case, we found a potential link between CNV and pathogen resistance (Fig. 6). CNVs were already shown to be implicated in nematode resistance in soybean (Cook et al. 2012), but also in potato cultivar genome heterogeneity (Pham et al. 2017). We showed that the CNVs in the 20rDNA lines occurred only in a few generations in controlled growing conditions. This last point is particularly interesting in the context of plant breeding. In addition, TDDOs have the potential to create chimeric genes (Fig. 4). TDDO events can promote cancer cell formation, via the activation of oncogenes (Quigley et al. 2018). In that case, breaking junctions can affect the expression of an oncogene by modifying its regulation by enhancers, for example. In our study, the chimeric genes created are expressed and properly spliced. Although we do not have evidence concerning their potential ability to be translated or if the resultant protein would be functional, it is tempting to speculate that TDDO-mediated chimeric genes can lead to gene novelty as previously described (Chen et al. 2013). Studying the consequences of TDDOs in future generations will certainly shed light on their potential impact on genome evolution and plant adaptation.

Methods

Plant materials

Seeds corresponding to the fas1-4 (SAIL_662_D10) and fas2-4 (SALK_033228) were previously reported (Exner et al. 2006). All 20rDNA seeds that include fas1-4 and fas2-4 parental lines, as well as L6 and L9 lines used in this study correspond to stock previously reported (Pavlíšťová et al. 2016). For NADs identification, wild-type Col-0 expressing the FIB2:YFP fusion protein was described in Pontvianne et al. (2013). The 20rDNA L6F5 line was transformed by agroinfiltration to insert a transgene expressing FIB2:YFP fusion protein as described previously (Pontvianne et al. 2013).

Nanopore sequencing and data analyses

Genomic DNA preparation was performed as previously described (Debladis et al. 2017). After Qubit dosage (dsDNA High Sensitivity, Thermo Fisher Scientific), a second step of DNA purification was performed with the Genomic DNA Clean and Concentrator kit (Zymo Research) and precipitated. A last Qubit dosage was performed before library preparation using the 1D Genomic DNA by ligation kit SQK-LSK109 (Oxford Nanopore Technologies), following the manufacturer’s instructions. The R9.5 ONT flow-cell FLO-MIN106D (Oxford Nanopore Technologies) was used. We obtained 6.4 Gb of sequences for L6F6, 0.7 Gb for L9F6, 5.9 Gb for fas1-4, and 11.4 Gb for fas2-4.

ONT reads were mapped on the TAIR10 reference genome using minimap2 with -a -Q -map-ont options (Li 2018). The alignment files were converted into BED files using BEDTools, and the coverage per 100-kb window was calculated using coverageBED (Quinlan and Hall 2010). For each 100-kb window, the ratio r = 20%rDNA coverage/wild-type Col-0 coverage was calculated. The mean (m) and standard error (SE) were calculated across the entire genome. Differentially covered regions in the 20%rDNA line were defined as regions for which r ≥ m+2SE or r ≤ m−2SE.

Additional methods can be found in the Supplemental Material.

Data access

All raw and processed sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) under accession number PRJEB35832. Code used to produce Hi-C figures is available as Supplemental Code.
Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Réné Merret and Michèle Ludié for Illumina sequencing and laboratory members for fruitful discussions. We also thank the flow cytometry facility, the microscopic facility, and the sequencing facility of Perpignan University Via Domitia Bioenvironnement (Perpignan, France). This work and A.P.-P.’s PhD fellowship are supported by the Agence Nationale de la Recherche (ANR), JCJC NucleoReg (ANR-15-CE12-0013-01) to F.P. F.P. was supported by the French Laboratory of Excellence project TULIP (ANR-10-LABX-41 and ANR-11-IDEX-0002-02). Work conducted at Iowa State University was supported by Hatch Act and State of Iowa funds. M.S. is a member of the European Training Network “EpiDiverse” that receives funding from the European Union Horizon 2020 program under Marie Skłodowska-Curie grant agreement No. 764965. M.D. and M.F. were supported by the Czech Science Foundation project 19-11880Y; by Ministry of Education, Youth and Sports of the Czech Republic INTER-COST (LITC18048); and by European Regional Development Fund, Project “SINGING PLANT” (CZ.02.1.01/0.0/0.0/16_026/0008446). A.P.-P., S.G., N.P., M.F., M.D., and F.P. are part of the European Cooperation in Science and Technology COST ACTION CA16212 INDEPD.


References


Received January 23, 2020; accepted in revised form September 15, 2020.
Large tandem duplications affect gene expression, 3D organization, and plant–pathogen response

Ariadna Picart-Picolo, Stefan Grob, Nathalie Picault, et al.

*Genome Res.* published online October 8, 2020
Access the most recent version at doi:10.1101/gr.261586.120

### Supplemental Material

http://genome.cshlp.org/content/suppl/2020/10/05/gr.261586.120.DC1

### P<NPublished online October 8, 2020 in advance of the print journal.

### Open Access

Freely available online through the *Genome Research* Open Access option.

### Creative Commons License

This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

### Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to: http://genome.cshlp.org/subscriptions