Differences in firing efficiency, chromatin and transcription underlie the developmental plasticity of the *Arabidopsis* DNA replication origins

Sequeira-Mendes et al.

Supplemental Methods

Contents

Purification of short nascent strands (SNS)

Total genomic DNA and SNS preparations were obtained under RNase-free conditions, by an optimization of the protocol described (Sequeira-Mendes et al. 2009). Nuclei isolation from 4 or 10 days post-sowing (dps) *Arabidopsis* seedlings was performed prior to genomic DNA extraction as described (Chodavarapu et al. 2010), in order to minimize genomic DNA contamination with cytosolic polyphenols and other secondary metabolites. Twelve grams of whole seedlings were collected, frozen and ground in liquid nitrogen in the presence of 10% PVPP (Sigma). The ground material was resuspended in 10 ml per gram of Honda Buffer Modified for 30 min in a rotary shaker at 4 °C (HBM; 2% (p/v) PVP10 (Sigma), 25 mM Tris-HCl, pH 7.6, 440 mM sucrose (Merck), 10 mM magnesium chloride, 0.1% Triton X-100, 10 mM β-mercaptoethanol). To better release the nuclei, the resuspended material was processed in a dounce homogenizer twice with a loose and a tight pestles and filtered through a double miracloth mesh into corex tubes. The nuclei were centrifuged 10 min at 3000x*g* and 4 °C. The supernatant was discarded and nuclear pellet was resuspended in 5 ml per gram of Nuclei Isolation Buffer (NIB; 2% (p/v) PVP10 (Sigma), 20 mM Tris-HCl, pH 7.6, 250 mM sucrose (Merck), 5 mM magnesium chloride, 5 mM potassium chloride, 0.1% Triton X-100, 10 mM β-mercaptoethanol). The sample was loaded onto a 15/50% gradient of Percoll in NIB and centrifuged 20 min at 500xg and 4 °C with slow brake. The green upper layer was discarded and the same volume was replaced with NIB. Nuclei were centrifuged 5 min at 1100xg and 4 °C, washed twice with 10 ml of NIB and 4 °C, and resuspended in 20 ml of lysis buffer per 12 grams of starting material (0.5% (p/v) PVP10, 50 mM Tris-HCl, pH 8.0, 10 mM EDTA pH 8.0, 1% SDS, 10 mM β-mercaptoethanol) by agitation 15 min at 4 °C. To digest the proteins, 100 µg/ml proteinase K was added and incubated overnight at 37 °C with mild rotation. Total DNA was extracted twice, first using phenol, pH 8.0, then with phenol:chloroform:IAA and the aqueous phase containing genomic DNA was collected into polyallomer tubes (Beckman). DNA was precipitated by adding 1.5M sodium chloride and 2 volumes of absolute ethanol, incubated 1h at -80 °C and pelleted by centrifugation for 45 min at 52,000x*g* at 4 °C using an AH-627 rotor (Sorvall). DNA was washed twice with 70% ethanol, centrifuged 20 min at 52,000x*g* and room temperature in an AH-627 rotor (Sorvall), air dried and resuspended in 1 ml of TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) containing 160 U of RNase OUT (Invitrogen). DNA was incubated at 4 °C overnight without pippeting or vortexing.

Purified DNA was denatured by heating 10 min at 100 °C and size-fractionated in a seven-step neutral sucrose gradient (5-20% sucrose in TEN buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA and 100 mM sodium chloride), by centrifugation at 102,000xg in a SW-40Ti Beckman rotor for 20 h at 20 °C (Gomez and Antequera 2008). Fractions (1 ml) were collected from the top and the DNA was ethanol-precipitated. An aliquot of each fraction was

analyzed in a 1% alkaline agarose gel (50 mM sodium hydroxyde, 1 mM EDTA) to monitor size fractionation. Normally, fractions 3 (~100-600 nt), 4 (~300-800 nt) and 5+6+7 (~500-3000 nt) were processed further by treating with 0.67 U/µl of polynucleotide kinase (PNK, Fermentas) to phosphorylate 5'-hydroxyl ends in the presence of 1.34 mM dATP for 30 min at 37 °C. After PNK inactivation, phosphorylated DNA was extracted, precipitated and resuspended in water. SNS were distinguished from randomly broken DNA molecules based on the presence of 4-6 nt-long RNA primers at their 5'-ends, which made them resistant to λ-exonuclease treatment (Gerbi and Bielinsky 1997; Costas et al. 2011; Cayrou et al. 2015; Comoglio et al. 2015). The λ-exonuclease digestion was carried out with 5 U/µl of enzyme (Thermo Fisher Scientific) following the manufacture's instructions at 37 °C overnight. The efficiency of the digestion was monitored by adding 40 ng of phosphorylated linearized plasmid to an aliquot of each reaction tube. DNA from each λ exonuclease-treated fraction was extracted, precipitated and resuspended in TE. The phosphorylation and λ-exonuclease treatments were repeated at least twice. RNA was digested with 0.05 µg/ml RNase A (Roche) and 0.16 U/µl RNase I (Thermo Fisher Scientific) for 30 min at 37 °C. RNases were digested with 100 µg/ml proteinase K and DNA was extracted, precipitated and resuspended in Milli-Q water. The ssDNA of purified SNS was converted into dsDNA: first, SNS together with 2 pmol random hexamer primers (Roche) were denatured 5 min at 100 °C, then a slow annealing was achieved by cooling down the samples from 80 °C to room temperature; second, the dsDNA was synthesized by using 0.17 U/µl of Klenow fragment for 1h at 37 °C; third, the fragments were ligated with 2 U/µl of Taq DNA ligase (New England BioLabs) for 45 min at 45 °C; finally, dsDNA was extracted, precipitated, resuspended in Milli-Q water and quantified before proceeding to the library preparation. The same method of dsDNA conversion was applied to sheared and denatured genomic DNA to be used as sequencing control.

Removal of PCR duplicate reads of NGS data

PCR duplicate reads were removed using the in-house script specified below.

```
#!/usr/bin/perl -w
use strict;

my $input_file=shift;
my $output;
my $last_plus_strand_chromosome;
my $last_neg_strand_chromosome;
my $last_plus_strand_position;
my $last_neg_strand_position;
my $line;
my @data_point;
my $bitwise_flag;
my $chromosome;
```

```
my $position;
my $last_chromosome;
my $last_position;

if ($input_file=~/(.+\.bam)/){
   $output=$1."_nosibs_sam";
};

open INPUT, "samtools view -h $input_file |" or die "Cannot open $input_file\n";
open OUTPUT, ">$output" or die "Cannot open $output\n";
#open OUTPUT, "| samtools -S -b > $output" or die "Cannot open $output\n";

while ($line=<INPUT>){
   if ($line!~/^@/){
      @data_point=split /\t/, $line;
      $bitwise_flag=$data_point[1];
      $chromosome=$data_point[2];
      $position=$data_point[3];
      if ($bitwise_flag & 16){
         if (($last_neg_strand_position!=$position) ||
                  ($last_neg_strand_chromosome ne $chromosome)){
            print OUTPUT "$line";
            $last_neg_strand_position=$position;
            $last_neg_strand_chromosome=$chromosome;
         };
      }else{
         if(($last_plus_strand_position!=$position) ||
               ($last_plus_strand_chromosome ne $chromosome)){
            print OUTPUT "$line";
            $last_plus_strand_position=$position;
            $last_plus_strand_chromosome=$chromosome;
         };
      };
   }else{
      print OUTPUT "$line";
   }
};
```

Peak-calling

For each sample and each fraction, we call ORIs with our own peak calling algorithm ZPeaks (U. Bastolla, R. Peiro, J. Sequeira-Mendes, Z. Vergara, C. Gutierrez, in preparation) that can be accessed at https://github.com/ugobas/Zpeaks. ZPeaks (i) provides a well defined, genome-wide profile of Nascent Strand Score (NSS), instrumental for weighting candidate ORIs and generic genomic locations, and (ii) localizes an ORI at the local maximum of the NSS over the ORI box called, needed for centering the metaplots. We tested ZPeaks by visual inspection of the overlap between experiment and control reads and candidate ORIs as well as by the statistical analysis of the ORIs properties. Furthermore, our procedure was robust with respect to false positive ORIs because (i) it requires that each ORI is detected in several independent experiments and (ii) it weights each ORI with its NSS, so that spurious ORIs have low NSS and contribute little to the average properties.

Thus, ZPeaks computes optimally smoothed profiles of the reads of the experiment and

the control, obtains from them a normalized smoothed profile, calls peaks when the profile is above an user-specified threshold, and sets the ORI location at the maximum of the normalized profile. More in detail, the algorithm works as follows, once the sequencing reads have been aligned to the reference *Arabidopsis* TAIR10 genome: (1) The wig files (normalized read counts) are input to ZPeaks and the number of reads is rescaled so that its mean number over each chromosome is the same both for the experiment *e* and the control *c*. If the control is not available, a constant profile is used. To increase the reliability of bins where the control is low, values of the control below the mean are interpolated between the current value and the mean: if $c_i < \langle c \rangle$, then we use $c'_i = (c_i + \langle C \rangle)/2$, where i indicates the genomic location and $\langle C \rangle$ is the mean value of C over the chromosome where i is located; (2) The profiles of the rescaled experiment and control are smoothed as $c_i' = \Sigma_k \, c_i \, w_{ik}$ where the weights w_{ik} are given by $w_{ik} = exp(-d_{ik}/d_0)/\Sigma_l \, exp(-d_{il}/d_0)$, d_{ik} is the distance between the center of bin i and bin k, d_0 is a parameter that is optimized as described below. A cut-off on distance is used to accelerate the computation, whose value is optimized alongside d0; (3) From the smoothed experiment and control, the difference score $d_i = e_i' - c_i'$ is constructed and it is transformed into the Z score $z_i = (d_i - \langle d \rangle)/s_d$, where <d> is the mean value of d over the chromosome of i and s_d is the standard deviation; (4) For the chosen threshold T, the program counts the number of bins with $z_i > T$, $N(T)$. The smoothing parameter d_0 that yields the largest $N(T)$ for the chosen threshold is chosen as the optimal parameter. The rationale is that, if the profiles are smoothed too much, then the experiment and the control will tend to become equal to their mean values and d_i will tend to be zero, thus decreasing N(T), whereas if the profiles are smoothed too little the standard deviations will be large, also decreasing N(T). We can always determine numerically an optimal parameter d_0 for which $N(T)$ is maximum, which justifies our procedure. We defined the nascent strand score (NSS) profile of the experiment *e* as $NSS_{ei} = z_i$; (5) We then joined together consecutive bins with $NSS_{ei} > T$ separated by less than 200 nucleotides, obtaining boxes that represent candidate origins; (6) Finally, the putative DNA replication origin is set at the bin where z_i is maximum within the box, and the limits of the box are reduced in such a way that the ORI is at the center and the new box is contained into the original one. It must kept in mind that the NSS showed a continuous distribution without any sign of saturation, perhaps suggesting that some bias introduced by the amplification step prior to sequencing may have some contribution.

One may expect that the threshold parameter T may be objectively determined by clustering all genomic bins in two clusters through some clustering algorithm such as *k*-means, Expectation Maximization (that assumes that the scores z_i are distributed according to a Gaussian distribution) or Hidden Markov Models (that also exploits the positional order of the bins along the chromosome). We followed such strategies, but the thresholds that we obtained were low, a sizable fraction of the genome satisfied $z_i > T$, and visual inspection

showed that most candidate ORIs were not reliable. Thus, we had no better choice than selecting an arbitrary threshold T and determining *bona fide* ORIs by combining different experiments, as explained below.

ZPeaks code

The peak calling algorithm ZPeaks can be accessed at https://github.com/ugobas/Zpeaks. It is also provided here as Supplemental Code.

Combining peaks into consensus boxes (potential ORIs)

Our strategy consisted of determining a robust set of ORIs detected in at least two independent experiments and two fractions for each experiment and weighting each candidate ORI with the NSS value of each experiment in such a way that the results are little dependent of false positives with low score.

We analyzed two developmental stages (4 and 10 day-old seedlings) and 3 experiments for each stage (exp1, exp2, exp3), obtaining six different samples. For each of them, either two (F3 and F4) or three (F3, F4, F5+6+7) consecutive fractions of the sucrose gradients for size selection of nascent strands were sequenced. Fractions were equivalent between 4 and 10 day-old seedlings, matching F3, F4 and F5+6+7. We called candidate ORIs with a tolerant threshold ($z>1.8$) and for each sample we selected candidate regions, or boxes, that were identified in at least two fractions of the same gradient. Boxes with size smaller than 200 bp were eliminated, and boxes closer than 200 bp were joined. In this way we obtained six datasets of high quality ORIs, which numbers were: 842 (4d_exp1), 1938 (4d_exp2), 3008 (4d_exp3), 3298 (10d_exp1), 1686 (10d_exp2), 3107 (10d_exp3).

To increase the reliability of candidate ORIs, we selected only those boxes that had been found in at least two out of six independent samples, obtaining a total of 2374 highly reliable candidate ORIs. We matched the boxes with non-vanishing overlap and if an ORI had multiple overlaps, we selected the largest overlap. The center of the combined box was computed as the weighted average of the location with maximum score present in the associated boxes, weighting more the boxes with high NSS and small size. When we matched different fractions, the fraction F5+6+7, which contains larger nascent strands, was used to confirm boxes but not to locate their center, in order to obtain better resolution. The limits of the combined box were set in such a way that all of the bins are above the threshold in all fractions.

Supplemental References

Cayrou C, Ballester B, Peiffer I, Fenouil R, Coulombe P, Andrau JC, van Helden J, Mechali M. 2015. The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* **25**: 1873-1885.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ et al. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature* **466**: 388-392.

Comoglio F, Schlumpf T, Schmid V, Rohs R, Beisel C, Paro R. 2015. High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep* **11**: 821-834.

Costas C, de la Paz Sanchez M, Stroud H, Yu Y, Oliveros JC, Feng S, Benguria A, Lopez-Vidriero I, Zhang X, Solano R et al. 2011. Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nature Struc Mol Biol* **18**: 395-400.

Gerbi SA, Bielinsky AK. 1997. Replication initiation point mapping. *Methods* **13**: 271-280.

Gomez M, Antequera F. 2008. Overreplication of short DNA regions during S phase in human cells. *Genes Dev***22**: 375-385.

Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, Gomez M. 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet* **5**: e1000446.