José M. Franco-Zorrilla[1] and Salomé Prat

Department of Plant Molecular Genetics, Centro Nacional de Biotecnología-CSIC, Darwin 3, 28049-Madrid, Spain.

[1]Corresponding author (e-mail: jmfranco@cnb.csic.es)

Running Head: Determining the DNA binding specificity of potato regulators

DAP-seq identification of transcription factor binding sites in potato

SUMMARY

Plant growth and adaptation to environmental fluctuations involves a tight control of cellular processes which, to a great extent, are mediated by changes at the transcriptional level. This regulation is exerted by transcription factors (TFs), a group of regulatory proteins that control gene expression by directly binding to the gene promoter regions via their cognate TF-binding sites (TFBS). The nature of TFBS defines the pattern of expression of the various plant loci, the precise combinatorial assembly of these elements being key in conferring plant's adaptation ability and in domestication. As such, TFs are main potential targets for biotechnological interventions, prompting in the last decade notable protein-DNA interaction efforts towards definition of their TFBS. Distinct methods based on *in vivo* or *in vitro* approaches defined the TFBS for many TFs, mainly in *Arabidopsis*, but comprehensive information on the transcriptional networks for many regulators is still lacking, especially in crops. In this chapter, detailed protocols for DAP-seq studies to unbiased identification of TFBS in potato are provided. This methodology relies on the affinity purification of genomic DNA-protein complexes *in vitro,* and high-throughput sequencing of the eluted DNA fragments. DAP-seq outperforms other *in vitro* DNA-motif definition strategies, such as Protein Binding Microarrays and SELEX-seq, since the protein of interest is directly bound to the genomic DNA extracted from plants yielding all the potential sites bound by the TF in the genome. Actually, data generated from DAP-seq experiments are highly similar to those out of ChIP-seq methods, but are generated much faster. We also provide a standard procedure to the analysis of the DAP-seq data, addressed to non-experienced users, that involves two consecutive steps: (i) processing of raw data (trimming, filtering and read alignment); and (ii) peak calling and identification on enriched motifs. This method allows identification of the binding profiles of dozens of TFs in crops, in a timely manner.

KEY WORDS

Transcription factor (TF), DAP-seq, DNA-binding sites, transcriptional regulation.

1. INTRODUCTION

Plants are exposed to an ever-changing environment and therefore adapt their growth and development to these environmental changes for survival. These adaptive responses are tightly controlled, and this regulation is mostly exerted by changes at the gene expression level. Transcriptional regulation relies on the interaction of sequence-specific DNA binding proteins, referred to as transcription factors (TFs), to short DNA motifs (6–12 bp) in the gene regulatory regions. Thus, unbiased analysis of protein-DNA interactions towards identification of the TF DNA cognate motifs (referred to as TF binding sites, TFBS), is crucial to the understanding of the transcriptional events underlying different cellular processes, no matter whether these are triggered by external stimuli or are the result of an intrinsic developmental program.

Conserved TFBS are particularly important in the study of agriculturally relevant adaptive traits, since variation in these elements has been associated with phenotypic diversity, and many loci important in crop domestication involve changes in their transcriptional activity [1][2]. Furthermore, many quantitative trait loci QTLs found to contribute to crop yield, food quality or adaptation to novel environmental conditions, correspond to TFs [3], evidencing the relevance of diversification of both *cis*- (i.e., TFBS) and *trans*- (TF) regulators during crop evolution.

During the last decade, thousands of transcriptomic data have been generated, mostly in *Arabidopsis*, providing an exhaustive view of the plant transcriptome dynamics. Moreover, the number of transcriptional studies in potato and other crops has increased exponentially during last years, thanks to the ease of RNA-seq experiments. A better description of the TF DNA-binding motifs is then needed to define more precisely their transcriptional networks in plants, and to better understand the biology of complex traits for breeding.

Chromatin immunoprecipitation coupled to deep sequencing (ChIP-seq), is considered as the gold-standard technique to identify TFBS and the TF target genes *in vivo*. Actually, this methodology is becoming more accessible, and the number of ChIP-seq data in model species or crop plants is ever increasing. Apart from a number of technical drawbacks associated with ChIP-seq as indicated elsewhere [3], this methodology is highly time consuming, particularly in crops, given the requirement of genetic backgrounds expressing translational fusions of the TFs to adequate tags, preferably under their native promoters and in a loss-of-function background.

A valid alternative to ChIP-seq for direct target gene identification is the analyses of DNA-binding specificities *in vitro*. Regardless the SELEX methodology was initially described 30 years ago [4], it has only been during this decade that several high throughput *in vitro* strategies were developed, hence providing a broad vision of the intrinsic DNA binding properties of distinct plant TF families [3]. One of these more recent strategies has been the use of protein binding microarrays (PBMs), that enabled the identification of the core motifs recognized by hundreds of plant TFs [5–8]. PBMs offer several advantages that made them the method of choice during the last years, such as quickness, its relative low cost and universality. This last property made possible the identification of the binding elements for several TFs from crop plants, including potato [8–16].

A new sequencing-based methodology introduced in 2016, known as DAP-seq, opened the possibility of determining more directly the TFBS on plant genomic DNA [17, 18]. This technique is based on the affinity purification of DNA-protein complexes *in vitro*, and sequencing of the eluted DNA. One important aspect of this methodology is that the DNA to be incubated is not a synthetic pool of oligonucleotides, but the plant genomic DNA previously fragmented into relatively small pieces by sonication. Importantly, there is no restriction with the origin of DNA, as it may come from any species, including crops with very large genomes [17, 19]. That means that the eluted fraction corresponds to plant DNA fragments bound by the TF and, therefore, the identification of TFBS is relatively straightforward, independently of being derived from large genome crops. Data obtained after DAP-seq is very similar to that from ChIP-seq studies –i.e., enriched peaks corresponding to TF-bound DNA— with the difference that there is no an initial crosslinking step (and thus the risk on false positive interactions is reduced), and the experiments are performed *in vitro*, with a considerable saving in time.

Here we present a detailed protocol for DAP-seq studies in *Solanum tuberosum*. Taking advantage of our expertise on PBMs [6, 7], we adapted the use of *E. coli* soluble protein extracts as the source of recombinant protein for DAP-seq, therefore shortening and simplifying the hands-on process in relation to the initial protocol [17, 18]. We also describe in detail a pipeline for the analysis of the results, especially prepared for researchers with no computational skills.

2. MATERIALS

2.1. Reagents and solutions

1. Elution Buffer (EB): 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8.
2. 3 M NaOAc, pH 5.2.
3. Ethanol.
4. 100 mM dNTPs PCR Grade Set.
5. NP-40.
6. 10X Phosphate Buffered Saline (PBS): 1.37 M NaCl, 27 mM KCl, 100 mM Na_2HPO_4, 18 mM KH_2PO_4.
7. 0.1 M Phenylmethanesulfonyl fluoride (PMSF) in ethanol
8. LB medium supplemented with 0.2% glucose.
9. 10x Annealing Buffer: 100 mM Tris, pH 8.0, 500 mM NaCl, 10 mM EDTA.

2.2. Enzymes and kits

1. End-It DNA End-Repair Kit (Epicentre).
2. Klenow (3′–>5′ exo-).
3. T4 DNA Ligase.
4. Phusion High-Fidelity DNA Polymerase.
5. Amylose Magnetic Beads 25 mg, 10 mg/ml.
6. Qubit™ 1X dsDNA HS Assay Kit (ThermoFisher).
7. Agencourt AMPure XP beads (Beckman Coulter).
8. Agilent High Sensitivity DNA chip (Agilent Technologies)

2.3. Oligonucleotides (*see* Note 1)

1. Adaptor TruSeq strand A: 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3′ (asterisk indicates phosphorothioate bond).

2. Adaptor TruSeq strand B: 5′-P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′ (where 'P' indicates a 5′ phosphate group).
3. Prepared 30 μM dilution of Annealed Y-adaptor: dissolve oligonucleotides Adaptor Strand A and Strand B to 100 μM with sterile water. Mix 75 μl of each oligonucleotide, 25 μl of 10x Annealing and 75 μl of sterile water. Incubate the mixture in a thermal block for 5 min at 95 ºC, switch the heat off and allow to slowly cool to room temperature (should take 60-90 min) (*see* **Note 2**).
4. TruSeq Universal Primer: 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCT ACACGACGCTCTTCCGATCT-3′.
5. TruSeq Index Primer: 5′-CAAGCAGAAGACGGCATACGAGAT-NNNNNN-GTGACTGGAGTTCAGACGTGTGCTCTTCCG-3′ (where 'NNNNNN' represents the 6mer index sequence for multiplexing) (*see* **Note 3**)

2.4. Equipment

1. M220 Focused Ultrasonicator (Covaris) (*see* **Note 4**).
2. MicroTUBE AFA Snap-Cap 6x16mm (Covaris) and M220 Holder XTU Insert microTUBE 130 µl (Covaris).
3. Qubit Fluorometer (ThermoFisher).
4. Table microcentrifuge.
5. 1.5 ml and 2.0 ml microcentrifuge tubes.
6. 50 ml conical-bottom plastic tubes.
7. Thin wall 0.2 ml and 0.5 ml tubes.
8. Neodinium magnet or magnetic rack suitable for 1.5 ml tubes.
9. Tube rotator mixer.
10. Orbital platform shaker.
11. Incubator shaker.

12. Probe sonicator for cell disruption.

13. Agilent 2100 Bioanalyzer (Agilent Technologies)

3. METHODS

3.1. Preparation of a genomic DNA library

DAP-seq requires high quality genomic DNA (gDNA) with no visible shearing. We have found that the conventional cetyl trimethylammonium bromide (CTAB) and modified Dellaporta methods [20, 21] yield high amounts of pure gDNA suitable for downstream modifications. In order to increase yields and avoid co-extraction of phenolics and polysaccharides, it is recommended using young tissues as the source of DNA. Whatever the method of DNA extraction is used, the RNA should be removed by incubation with RNase A for at least 30 min at 37 ºC and further extraction with phenol:chloroform:isoamyl alcohol (25:24:1). The method for quantification is also important: whereas UV absorbance at 260 nm (such as in Nanodrop systems) is very simple and quite precise for quantification of DNA, it is also susceptible to contaminants in solution, and the excess of salts. Whenever possible, Qubit fluorometric quantitation using the specific assay for dsDNA is recommended.

1. Start procedure with 5 μg of purified gDNA in Elution Buffer (EB) in a total volume of 125 μl (*see* **Note 5**).
2. Transfer gDNA in solution to a Covaris microTUBE and shear DNA using the 200bp target peak protocol. In a M220 Covaris sonicator, this protocol is: 50W Peak Incident Power; 20% Duty Factor; 200 Cycles per Burst; 150 sec.
3. Cleanup sample with AMPure XP Beads at a 1:2 ratio (125 μl DNA and 260 μl beads) (*see* **Note 6**).
4. Mix thoroughly DNA and beads by pipetting 5-10 times. Incubate 10 min at room temperature.
5. Place tube in the magnetic rack, let stand for at least 2 min until the solution is clear and carefully remove supernatant.

6. Remove tube from the magnet and wash the beads with 500 μl 80% ethanol freshly prepared. Pipet up and down 5 times for resuspension of the pellet.

7. Place tube in the magnet, let stand for 1 min and remove supernatant.

8. Repeat wash with 80% ethanol once more. After the second wash, let the sample to completely dry at room temperature for 5 min.

9. Resuspend the beads in 34 μl EB and mix well by pipetting. Incubate for 10 min.

10. Place tube in the magnet and transfer supernatant containing the DNA to a new tube.

11. End repair of the 34 μl of DNA sample using the End-It DNA End-Repair Kit. Add to DNA 5 μl 10x End-It buffer, 5 μl 10 mM dNTP mix, 5 μl 10 mM ATP and 1 μl End-It enzyme mix.

12. Incubate for 45 min at room temperature.

13. Cleanup the sample with 80 μl (1.6x) AMPure XP Beads and repeat steps 4-8 from this section (*see* **Note 6**).

14. Resuspend the DNA in 32 μl EB and proceed with A-tailing reaction.

15. Add to the 32 μl of DNA the following reagents: 5 μl 10× NEBuffer2, 10 μl 1 mM dATP and 3 μl Klenow Fragment (3'–>5' exo-; 5 U/μl).

16. Incubate at 37 °C for 30 min.

17. Cleanup the sample with 90 μl (1.8x) AMPure XP Beads and repeat steps 4-8 (*see* **Note 6**).

18. Resuspend the DNA in 30 μl EB and proceed with adaptor ligation.

19. Add to the 30 μl of DNA the following reagents: 5 μl T4 DNA Ligase 10× Buffer, 10 μl 30 μM annealed Y adaptor and 5 μl T4 DNA Ligase (3 U/μl).

20. Incubate at room temperature for 3 hours.

21. Cleanup the sample with 50 μl (1.0x) AMPure XP Beads and repeat steps 4-8 (*see* **Note 6**).

22. Resuspend the DNA in 30 µl EB and proceed with quantitation in a Qubit Fluorometer using Qubit 1X dsDNA HS Assay Kit. Final yield of the gDNA library should be in the 1.5-2.5 µg range.

3.2. Preparation of input DNA sample

This step is necessary after every new gDNA library is obtained, and it should be sequenced together with the DAP samples. Sequencing results will be used as negative (non-enriched) control for the DAP libraries. However, there is no need to repeat this step for every protein, since a single sequenced input sample will serve for all proteins analyzed with the same gDNA library.

1. Assemble a PCR reaction by mixing 50ng of gDNA library (see step 22 in section 3.1), 10µl of 5× Phusion HF Buffer, 2.5µl of 10 mM dNTPs, 1µl of 25 µM TruSeq Universal Primer, 1µl of 25 µM Primer TruSeq Index Primer (*see* **Note 8**), 1µl of 2 U/µl Phusion DNA Polymerase and sterile water up to 50µl.

2. Set the tube in a thermocycler and run the following program: 95 °C for 2 min, 98 °C for 30 sec, 11 cycles of 98 °C for 15 sec, 60 °C for 30 sec and 72 °C for 2 min, followed by final extension at 72 °C for 10 min.
3. Cleanup the sample with 50 µl (1.0x) AMPure XP Beads and repeat steps 4-8 from Section 3.1 (*see* **Note 6**).

4. Resuspend the DNA in 31 µl of EB and use 1 µl of sample to measure the DNA concentration using the Qubit 1X dsDNA HS Assay Kit.

3.3. Recombinant protein extracts (*see* Note 9)

1. Induce expression of the recombinant transcription factor of interest fused to maltose binding protein (MBP) at the optimized conditions in a 30 ml *E. coli* culture (*see* **Note 10**).

2. Save 1 ml aliquots of induced and non-induced cultures. Pellet cells by centrifugation at top speed for 1 min and discard supernatants. Store aliquots at –20 ºC until use (*see* **Note 11**).
3. Harvest cells by centrifugation at 4,000 x *g* for 15 min at 4 ºC in 50 ml conical-bottom tubes and discard the supernatant.
4. Snap freeze the pellet with liquid nitrogen or store at –80 ºC if is not going to be used immediately (*see* **Note 12**).
5. Resuspend by vortexing the frozen cellular pellet in 1 ml 1x PBS supplemented with 1mM PMSF. Transfer the lysate to a 2 ml centrifuge tube and keep on ice.
6. Disrupt cells with a probe sonicator three times for 30 sec each, keeping the tube on ice. To avoid over-heating of the sample, keep the tube on ice for 1 min between different rounds of sonication.
7. Centrifuge at 15,000 x g for 15 min at 4º C. Transfer the supernatant to a new centrifuge tube. Repeat this step once again.
8. Keep the protein extract on ice and use the sample immediately for step 4 in Section 3.4.

3.4. Affinity purification of protein-bound DNA

The following procedure is an adaptation of the "Small-scale DAP-seq protocol using *E. coli*-expressed recombinant protein" by [17, 18], to be used with *E. coli* extracts expressing MBP-fusion proteins.

1. Pipet 25 µl of Amylose Magnetic Beads into a 1.5-ml microcentrifuge tube per protein extract.

2. Place the tube on a magnet for 1 min and remove the supernatant.

3. Wash the beads three times with 500 µl of 1x PBS. After the last wash, remove as much supernatant as possible.

4. Remove the tube from the magnet and add 400 μl of the protein extract (see step 8 in Section 3.3). Be sure to completely resuspend the Amylose Magnetic Beads by gentle pipetting.

5. Rotate for 1 h at 4ºC with the tube rotator mixer, to bind the protein to the beads.

6. Spin down solution by centrifugation at low speed (5 sec at 1,000 g), place on the magnet and remove the supernatant.

7. Wash the beads with 500 μl 1x PBS+NP-40 (0.005%) and place the tube back on the magnet to remove supernatant.

8. Repeat last step for four washes in total.

9. Wash the beads twice with 500 μl 1x PBS without detergent. Remove as much supernatant as possible after the last wash.

10. Add 40 μl of 1x PBS and gently resuspend beads.

11. Add 500 ng of the potato gDNA library (see step 22 in section 3.1,) diluted in 40 μl 1x PBS.

12. Rotate the microcentrifuge tube horizontally to keep the beads in suspension and to allow protein-DNA binding. Incubate for 1 h.

13. Place the tube on the magnet and wash the beads with 200 μl 1x PBS+NP-40 (0.005%).

14. Repeat last step for four washes in total.

15. Perform two additional washes with 200 μl 1x PBS without detergent. During the final wash, transfer beads in solution to a new 1.5-ml microcentrifuge tube.

16. Place the tube on the magnet and aspirate as much as supernatant as possible.

17. Add 25 μl EB and thoroughly resuspend the beads by vortexing.

18. Incubate in a thermal block at 98 ºC for 10 min.

19. Quick-spin for 5 s at 3,000 g and place the tube on the magnet.

20. Recover the supernatant that contains the protein-bound DNA.

21. Proceed with PCR enrichment by assembling a PCR reaction as follows: 25 μl of enriched DNA, 10 μl of 5× Phusion HF Buffer, 2.5 μl of 10 mM dNTPs, 1 μl of 25 μM Primer A, 1 μl of 25 μM Primer B, 1 μl of 2U/μl Phusion DNA Polymerase and 8.5 μl of sterile water.
22. Place the tube in a thermocycler and run the following program: 95 °C for 2 min, 98 °C for 30 sec, 20 cycles of 98 °C for 15 sec, 60 °C for 30 sec, 72 °C for 2 min, followed by extension at 72 °C for 10 min.
23. Cleanup the sample with 50 μl (1.0x) AMPure XP Beads and repeat steps 4-8 from Section 3.1 (*see* **Note 6**).
24. Resuspend the DNA in 31 µl of EB and use 1 µl of sample to measure the DNA concentration using the Qubit 1X dsDNA HS Assay Kit. Typical DAP-seq libraries may yield 10-30 ng/µl.
25. Store the DAP-seq library at –20 ºC until use.

3.5. Sequencing of pooled DAP-seq libraries

One advantage of the DAP-seq method is that it does not require a very high coverage to detect the enriched peaks. However, it is very difficult to establish *a priori* the number of sequencing reads required to identify all protein-bound genomic fragments, as this will depend on several factors, such as the affinity of the protein for binding to DNA and the number of targeted elements in the genome, plus some technical issues (i.e. concentration of the protein available in the soluble extracts), and on the size and complexity of the genome among others. The *S. tuberosum* genome has an overall length of approximately 840 Mb and most cultivars and landraces are tetraploid [22, 23], which makes the potato genome at least 12-fold larger to that of *Arabidopsis.* In consequence, the number of sequencing reads required to identify enriched peaks shall be at least 10-fold higher. Aside all these factors, the number of significantly enriched peaks differ with the protein itself and we have observed that 8 million pairs (PE 75 bp or 150 bp) are enough to detect the majority of the theoretical binding sites (Fig. 1).

The steps below are designed to obtain a pooled DNA library corresponding to several samples ready to be sequenced by an in house facility or specialized company and, therefore, the protocols involving qPCR quantification of the pool and processing with the Illumina platform are omitted.

1. Pool the DAP-seq libraries by mixing identical concentrations of each sample. A range of 30-50 ng per sample should be enough for sequencing (*see* **Note 13** and **Note 14**).
2. Load 1 μl in an Agilent High Sensitivity DNA chip, or similar. The DNA pool should peak at 320-350 bp (average shearing at 200 bp + 120 bp adapter) (*see* **Note 15**).
3. Measure DNA concentration using Qubit 1X dsDNA HS Assay Kit. Pooled DAP-seq libraries are ready for precise qPCR quantification prior Illumina sequencing.

3.6. Analysis of results

The analysis of DAP-seq library sequencing does not require specific software and consists of two consecutive steps: (i) processing of raw data (trimming, filtering and read alignment) and (ii) peak calling and identification of enriched motifs. Whereas the first step is common to all the sequencing strategies, the second step involves the identification of protein-bound DNA, and the software required is the same as in other strategies, such as ChIP-seq. In this protocol we offer some cues for processing the raw data using Galaxy public servers, whenever possible. Galaxy is an open, web-based platform to perform sequencing analysis in an accessible way to researchers non-experienced in computing [24].

3.6.1. Processing of raw data

1. Download the latest versions of the genome (.fa file) and annotations (.gff file) of *S. tuberosum* (*see* **Note 16**).
2. Create a personal account in the Main Galaxy Server (https://usegalaxy.org/).

3. Upload reference files and demultiplexed FASTQ files from input and DAP-seq libraries. Once uploaded, the names of the files will appear in the History (right) panel in green.

4. Select the tool “Trim Galore!” from the Tools panel on the left. This tool automatically detects and trims the adapter sequence from reads and filters out low quality reads.

5. Select in the Tool interface (middle panel) the type of library (typically paired-end) and the name(s) of the FASTQ file(s) to be trimmed. Run with default parameters by clicking the Execute button (*see* **Note 17**).

6. Select the tool “Bowtie2” from the Tools panel (*see* **Note 18**).

   a. Select in the type of library (typically paired-end) and the name(s) of the trimmed FASTQ files.

   b. Select the *S. tuberosum* reference genome (fasta file) from the history.

   c. Click “Yes” to save mapping statistics to the history.

   d. Run with default parameters by clicking the Execute button.

7. Perform the steps 4-6 with input and experimental datasets before calling the peaks.

8. Generate BigWig files for visualization of the genomes (*see* **Note 19**).

   a. Look for the tool “bamCoverage” in the Tools menu on the left.

   b. Select the alignment file to convert and the bin size (the shorter, the slower will be the conversion and the larger the output file). Default values are convenient.

   c. Select the normalization method. If selected “normalize coverage to 1x”, the effective genome size should be 7.5E+08.

   d. Execute will generate a compressed BigWig file from the alignment.

3.6.2. Peak calling and identification of enriched motifs (*see* Note 20).

1. Select the tool “MACS2 callpeak” from the Tools panel (*see* **Note 21**).

   a. Select the experimental DAP-seq and control BAM alignment files from the history.

b. Select the format of the files (typically, paired-end BAM)

c. Define the Effective genome size as 7.5E+08 (*see* **Note 22**).

d. Run with default parameters by clicking the Execute button.

e. The output of MACS2 will be a tabular file with the genome coordinates of significant peaks and summits, corresponding fold-enrichment, p-values and FDR-adjusted p-values.

2. Download the GEM package [25] from http://groups.csail.mit.edu/cgs/gem/, uncompress and create a folder with the package (*see* **Note 23** and **Note 24**).

a. Split the *S. tuberosum* genome fasta file into 13 files, one per chromosome (12 chromosomes in the potato genome + unanchored chr0 genes, and store in a folder named “StubGenome” by saving the different sequences as separated files.

b. Generate a text file including the length of the chromosomes and name it “Stub-genome_sizes.txt” (*see* **Note 25**).

c. Download from Galaxy the BAM files from the enriched and input samples and store them into the GEM folder

d. Open the command console in your system, go to the GEM folder and paste the following script (*see* **Note 26**):

```
java -jar gem.jar --d Read_Distribution_default.txt --g Stub-genome_sizes.txt --genome
./StubGenome/ --s 750000000 --expt ./DAPseq-exp.bam --ctrl ./Input.bam --out ResultsExp1 --f SAM -
-outNP --range 150 --smooth 0 --mrc 1 --fold 2 --q 1.30 --k_min 6 --k_max 20 --k_seqs 600 --
k_neg_dinu_shuffle --pp_nmotifs 1
```

e. The output folder (“ResultsExp1” in this example) will contain all the results for prediction of binding events and sequences. Consider the file GEM_events.narrowPeakas the final list of binding events (*see* **Note 27**).

Provided is a description of the primary steps for handling the sequencing files and for identifying enriched peaks in DAP-seq experiments, as these use to be the most difficult for researchers non-experienced in bioinformatics. Once the enriched peaks are obtained, there are several tools for annotation of the genes nearby the peaks, such as R package ChIPseeker [26] or the tool annotatePeaks in Homer [27]. The *de novo* discovery of sequence motifs enriched in the peaks is also of interest. In this respect, MEME Suite ((http://meme-suite.org/); [28]) offers several possibilities for motif discovery and their similarity to other known motifs available in databases.

3.7. Expected results

Initial description of DAP-seq in *Arabidopsis* estimated the experiments as successful when they produced substantially enriched peaks and when >5% of the reads fell in peaks [18]. The experiments carried out in potato using the protocol detailed here, yielded between 15 and 35% reads in peaks, and the number of significant peaks varied between a few thousands to near 100,000, in a range similar as observed in other species, including Arabidopsis. This indicates that the size of the genome is not inconvenient to detect TFBS.

We observed a high disparity between different proteins, even from the same species, suggesting that the higher or lower number of enriched peaks depends on the intrinsic DNA-binding properties of the proteins and/or some technical issues, mostly their solubility in PBS-based extracts. Even so, the protocol offers the possibility of identifying in a simple way the comprehensive TFBS repertoire for a given regulator in *Solanum tuberosum*.

3.8. Biological potential of generated data

DAP-seq offers many advantages over other high-throughput methods for the identification of TFBS. For example, PBMs yield a consensus recognition sequence inferred from binding to a collection of synthetic oligonucleotides covering all the 10mer or 11mer possibilities [5, 7], but direct extrapolation to plant genomes to identify TFBS is not straightforward. Similarly, SELEX-

Seq also provides a binding motif derived from a synthetic degenerate oligonucleotide, longer in size than in PBMs, but with identical drawbacks. Thus, DAP-seq covers the advantages of *in vitro* methods (feasibility, quickness, parallelization) and of ChIP-seq (direct identification of TFBS in plant genomes).

These features make DAP-seq the method of choice to define TF target genes and consensus binding sequences in crops. Fig. 2 illustrates an example of the possibilities for definition of TFBS. This case corresponds to the binding profile of a potato MYC2-related factor, a master regulator of gene expression in response to jasmonates and wounding [29, 30]. In this example, bound peaks were located in the promoter regions of the proteinase inhibitor coding genes, normally induced in response to wounding and supposed to be targeted by MYC2 or related TFs. Interestingly, the peak summits match to DNA motifs recognized by MYCs, i.e. G-box (5'-CACGTG-3'), PBE (5'-CATGTG-3') and T/G-box (5'-AACGTG-3'), reflecting the binding specificities of the protein. Thus, by using this method we were able to achieve similar results as those obtained from ChIP-seq experiments, but they were generated in shorter time.

Rapid identification of TFBS by DAP-seq open new possibilities in the study of agriculturally relevant traits, as it was proposed that domestication and improvement of crops largely occurred on *cis*-regulatory regions for candidate genes [2, 31]. For example, two QTLs that affect fruit size and shape in tomato, *locule number* (*lc*) and *fascinated* (*fas*), correspond to two independent *cis*-regulatory mutations that synergistically control the size of floral meristems [32, 33]. In the case of *lc*, the phenotype is caused by two single nucleotide polymorphisms (SNPs) affecting expression of the *SlWUSCHEL* homeodomain TF (Fig. 3A). The phenotypic effect of those changes is an increase of the locules number and fruit size in varieties carrying the SNPs [33]. Interestingly, these SNPs disrupt a CArG-box recognized by the MADS-box family of TFs, and hinders repression of *SlWUSCHEL* by TOMATO AGAMOUS1 (TAG1), leading to higher number of locules [34].

Examples of sequence variation in TFBS responsible for agriculturally relevant traits were not yet described in potato. Systematic DAP-seq cultivar analyses with candidate transcription factors would facilitate the discovery of new regulatory genes to be used in breeding. In example, enzymatic browning of potato tubers is caused by polyphenol oxidase (PPO) activity [35], and downregulation of *PPO* genes reduces browning [36]. Given that browning occurs directly after wounding, some *PPO* genes may be targeted by MYC2-related TFs. In fact, we observed that MYC2 strongly binds the promoter of *StuPPO1*, at a position displaying two T/G-boxes in tandem, pointing to a direct regulation by this TF on wounding (Fig. 3B and C). Notably, while these DNA elements are conserved in all sequenced potato cultivars, they vary in sequence in other Solanaceae species (Fig. 3C). This would suggest that *PPO1* has a lower contribution to enzymatic browning in these species, an aspect that may deserve further investigation.

Identification of these T/G motifs would not have been possible without DAP-seq studies, illustrating the importance of a comprehensive characterization of TFBS for a candidate transcription factor by using this methodology, in order to understand its function in modulating a particular trait. This strategy, in combination with directed genome editing to create allelic variation in *cis*-regulatory elements [37], may considerably speed the selection of novel desired traits in crop breeding.

4. NOTES

1. Oligonucleotides should be high-performance liquid chromatography (HPLC)-purified to avoid incompletely synthesized primers.

2. Annealed Y-adaptor should be prepared before starting the genomic DNA library protocol. Store the annealed Y adapter at -20 ºC in 50 µl aliquots.

3. Index sequences (6 bases) can be found in the "Illumina Adapter Sequences" document under the epigraph "TruSeq Single Indexes". For reference, sequences corresponding to index 1 to 12 are:

Index 1: ATCACG
Index 2: CGATGT
Index 3: TTAGGC
Index 4: TGACCA
Index 5: ACAGTG
Index 6: GCCAAT
Index 7: CAGATC
Index 8: ACTTGA
Index 9: GATCAG
Index 10: TAGCTT
Index 11: GGCTAC
Index 12: CTTGTA

4. This equipment can be replaced by any other AFA (Adaptive Focused Acoustics) Covaris Ultrasonicator.

5. When gDNA concentration is lower than 5 µg in 125 µl, fragmentation can be carried out in two steps and later pool the samples.

6. Cleanup steps with AMPure XP Beads can be replaced by NaOAc/Ethanol precipitations.

7. Annealed Y-adaptor should be prepared before starting the genomic DNA library protocol. Store the annealed Y adapter at -20 ºC in 50 µl aliquots.

8. Primer B contains an index sequence. Be sure not to use the same index in other DAP/input samples sequenced in the same lane.

9. N-terminal fusions to Maltose Binding Protein (MBP) are particularly suitable when *E. coli* lysates are used for protein binding, since MBP improves the solubility and promotes the proper folding of its fusion partners [38]. Other tags can otherwise be used (GST, 6xHis, etc.).

10. It is recommended to stick to standard methodologies for protein expression in *E. coli*. In the case of MBP-fused proteins, follow pMAL Protein Fusion and Purification System instructions (New England Biolabs). In our experience, induction of protein expression with 1 mM IPTG in LB medium supplemented with 0.2 % glucose for 4-6 h at 28º C works well for most proteins.

11. Expression and solubility of recombinant proteins in the final culture must be checked by 8-12% SDS-PAGE before performing the DNA-binding assays. Use frozen aliquots and follow steps 5 to 7 but resuspend pellets in 0.2 ml 1xPBS instead. Centrifuge only once at step 7 and keep pellet and supernatant separated. Resuspend pellet with 0.2 ml 1xPBS and add to all the samples the corresponding volume of SDS-PAGE sample buffer and load to gels. The recombinant protein should be detected after Coomassie staining of soluble protein extracts.

12. We have used bacterial pellets frozen for several weeks at -80 ºC with no appreciable impact on binding activity.

13. DAP-seq (or input) libraries to be pooled must be prepared with different Primer B index sequences.

14. The number of samples that can be mixed in a single pool will depend on the number of samples to be analyzed, the desired final output per sample, and the available sequencing platform. For reference, 20 samples from *S. tuberosum* can be pooled and sequenced on a single Illumina HiSeq 4000 lane, giving approximately 15 million read-pairs per sample.

15. Make sure that there is no appreciable adapter dimer (sharp peak at 120 bp) or that its total amount is less than 0.5% of molecules. In case of detecting adapter dimer, perform an extra purification with AMPure XP Beads.

16. Reference sequence files can be obtained from several sources: Spud DB (https://solanaceae.plantbiology.msu.edu); Solgenomics (https://solgenomics.net/);

Ensembl Plants (http://plants.ensembl.org/Solanum_tuberosum/Info/Index); Phytozome (https://phytozome.jgi.doe.gov/). Files from different sources do not always correspond to identical versions. It is very important use reference files from the same source for different samples to get uniform results. Same for .fa and .gff files.

17. This tool allows customizing several parameters, such as the adapter sequence, the quality threshold or the minimum length of the read. Running with default parameters fits with most of the experiments.

18. Bowtie 2 [39] is the most commonly used aligner software, since it is very fast and memory efficient for short read mapping to long reference sequences.

19. BAM files (output of Bowtie2) can be directly loaded into any genome browser application (such as Integrative Genome Viewer, IGV: http://software.broadinstitute.org/software/igv/; or Integrated Genome Brower, IGB: https://bioviz.org/. However, these files are very large, which may interfere with visualization of the peaks if we want to open several files at once, and not enough memory is available in our computer. To avoid this, the BAM files can be converted into BedGraph or BigWig files that summarize the number of aligned reads per genomic interval (usually in 50 bp bins). Both formats contain the same data but BigWig is a compressed version, and therefore, more amenable.

20. Several peak calling algorithms have been described, but in this protocol we are using either MACS2, available in Galaxy, or GEM that should be downloaded and run locally.

21. MACS2 (Model-based Analysis of ChIP-Seq) [40] is a peak calling software widely used in ChIP-seq experiments and, therefore, can also be used in DAP-seq to identify enriched peaks. It is able to detect wide peaks, many of them corresponding to consecutive or nearby binding elements. MACS2 allows the analysis of the experimental DAP sample without a comparison to the control input sample. However, this is not recommended if we want to avoid a positional bias due to non-uniform shearing of DNA.

22. Effective genome size is the portion of the genome that is mappable, thus discarding stretches of Ns and repetitive regions. The value here indicated (750 Mb) corresponds to 90% of the genome.

23. GEM is a different peak calling algorithm suited for ChIP-seq and, specially, for ChIP-exo data. GEM links binding peak calling and motif discovery to vey accurately predict the binding events. This method tends to predict sharper peaks and, therefore, can discriminate between closely spaced binding events. GEM was the method of choice in the initial description of the DAP-seq methodology [17, 18].

24. GEM is not available in Galaxy and needs to be run locally. There are no specific computing requirements for running GEM and the only prerequisite is Java that comes with all the major operating drivers on each system (Windows, Mac, Linux).

25. There are several tools or scripts downloadable from the web to obtain the size of the chromosomes. For Galaxy users the tool "Compute sequence length" can be used.

26. Check the GEM documentation in http://groups.csail.mit.edu/cgs/gem/ for a detailed description of the parameters.

27. Refer to GEM documentation for a detailed description of the output.

ACKNOWLEDGEMENTS

Research in J.M.F.-Z.'s lab was funded by the Spanish Ministry for Science and Innovation, grant BIO2017-86651-P (MINECO/FEDER). Potato research in Prat's lab was supported by grants BIO2015-3019-EXP grant from the Spanish Ministry of Economy, Industry and Competitiveness and the PCIN-2017-032 CONCERT-JAPAN project financed by the Ministry of Science, Innovation and Universities and by the European Union’s Horizon 2020 research and innovation program under grant agreement No. 862858.

REFERENCES

1. Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14:840–852
2. Swinnen G, Goossens A, Pauwels L (2016) Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. Trends Plant Sci. 21:506–515
3. Franco-Zorrilla JM, Solano R (2017) Identification of plant transcription factor target sequences. Biochim Biophys Acta - Gene Regul Mech 1860:21–30
4. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249:505–510
5. Franco-Zorrilla JM, López-Vidriero I, Carrasco JLJL, et al (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci 111:2367–2372
6. Franco-Zorrilla JM, Solano R (2014) High-throughput analysis of protein-DNA binding affinity. In: Methods in Molecular Biology. Humana Press Inc., pp 697–709
7. Godoy M, Franco-Zorrilla JM, Pérez-Pérez J, et al (2011) Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. Plant J 66:700–711
8. Weirauch MT, Yang A, Albu M, et al (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158:1431–1443
9. Vélez-Bermúdez I-C, Salazar-Henao JE, Fornalé S, et al (2015) A MYB/ZML Complex Regulates Wound-Induced Lignin Genes in Maize. Plant Cell 27:3245–3259
10. Raines T, Blakley IC, Tsai Y-C, et al (2016) Characterization of the cytokinin-responsive transcriptome in rice. BMC Plant Biol 16:260
11. Hichri I, Muhovski Y, Žižková E, et al (2014) The Solanum lycopersicum zinc finger2 cysteine-2/histidine-2 repressor-like transcription factor regulates development and tolerance to salinity in tomato and arabidopsis. Plant Physiol 164:1967–1990

12. Hichri I, Muhovski Y, Žižková E, et al (2017) The Solanum lycopersicum WRKY3 transcription factor SLWRKY3 is involved in salt stress tolerance in tomato. Front Plant Sci 8:1343

13. Molina-Hidalgo FJ, Medina-Puche L, Cañete-Gómez C, et al (2017) The fruit-specific transcription factor FaDOF2 regulates the production of eugenol in ripe fruit receptacles. J Exp Bot 68:4529–4543

14. Medina-Puche L, Molina-Hidalgo FJ, Boersma M, et al (2015) An R2R3-MYB transcription factor regulates eugenol production in ripe strawberry fruit receptacles. Plant Physiol 168:598–614

15. Abelenda JA, Cruz-Oró E, Franco-Zorrilla JM, Prat S (2016) Potato StCONSTANS-like1 Suppresses Storage Organ Formation by Directly Activating the FT-like StSP5G Repressor. Curr Biol 26:872–881

16. Nicolas M, Rodríguez-Buey ML, Franco-Zorrilla JM, Cubas P (2015) A Recently Evolved Alternative Splice Site in the BRANCHED1a Gene Controls Potato Plant Architecture. Curr Biol 25:1799–1809

17. O’Malley RC, Huang SC, Song L, et al (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 166:1598

18. Bartlett A, O’Malley RC, Huang SC, et al (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc 12:1659–1672

19. Galli M, Khakhar A, Lu Z, et al (2018) The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat Commun 9:1–14

20. Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: Version II. Plant Mol Biol Report 1:19–21

21. Liu W, Zhou Y, Liao H, et al (2011) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Am J Bot 98:202–222

22. Xu X, Pan S, Cheng S, et al (2011) Genome sequence and analysis of the tuber crop potato.

Nature 475:189–195

23. Hardigan MA, Laimbeer FPE, Newton L, et al (2017) Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc Natl Acad Sci U S A 114:E9999–E10008

24. Goecks J, Nekrutenko A, Taylor J, et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:

25. Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Comput Biol 8:e1002638

26. Yu G, Wang LG, He QY (2015) ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31:2382–2383

27. Heinz S, Benner C, Spann N, et al (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 38:576–589

28. Bailey TL, Boden M, Buske FA, et al (2009) MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res 37:

29. Boter M, Ruíz-Rivero O, Abdeen A, et al (2004) Conserved MYC transcription factors play a key role in jasmonate signaling both in tomato and Arabidopsis. Genes Dev 18:1577–1591

30. Lorenzo O, Chico JM, Sanchez-Serrano JJ, et al (2004) Jasmonate-insensitive1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in Arabidopsis. Plant Cell 16:1938–1950

31. Hufford MB, Xu X, Van Heerwaarden J, et al (2012) Comparative population genomics of maize domestication and improvement. Nat Genet 44:808–811

32. Cong B, Barrero LS, Tanksley SD (2008) Regulatory change in YABBY-like transcription

factor led to evolution of extreme fruit size during tomato domestication. Nat Genet 40:800–804

33. Muños S, Ranc N, Botton E, et al (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL. Plant Physiol 156:2244–2254

34. van der Knaap E, Chakrabarti M, Chu YH, et al (2014) What lies beyond the eye: the molecular mechanisms regulating tomato fruit weight and shape. Front Plant Sci 5:227

35. Thygesen PW, Dry IB, Robinson SP (1995) Polyphenol oxidase in potato: A multigene family that exhibits differential expression patterns. Plant Physiol 109:525–531

36. Chi M, Bhagwat B, Lane WD, et al (2014) Reduced polyphenol oxidase gene expression and enzymatic browning in potato (Solanum tuberosum L.) with artificial microRNAs. BMC Plant Biol 14:62

37. Rodríguez-Leal D, Lemmon ZH, Man J, et al (2017) Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell 171:470-480.e8

38. Kapust RB, Waugh DS (1999) Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused . Protein Sci 8:1668–1674

39. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359

40. Feng J, Liu T, Qin B, et al (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728–1740

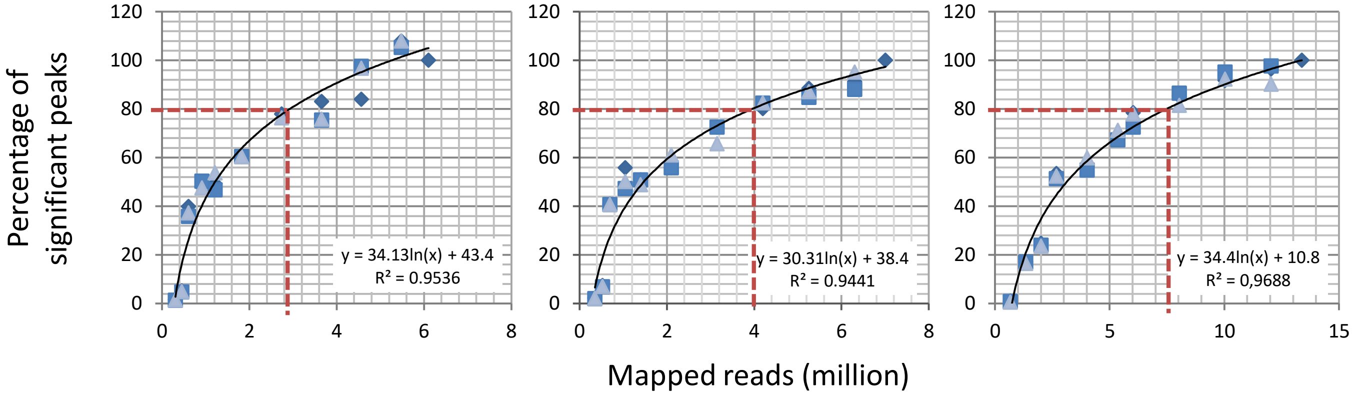
FIGURE LEGENDS

Fig. 1. Effect of depth coverage in the identification of TF-binding sites in *Solanum tuberosum*. Plots represent the proportion of significant peaks detected in three DAP-seq experiments with different proteins at decreasing sequencing depths. The number of significant peaks obtained after analysis of each DAP-seq experiment was considered as 100%. Decreasing sequencing depths were obtained by randomly subsampling the original FASTQ files, and computing for the identification of significant peaks. Three independent subsamples were obtained at each depth point. Thanks to the logarithmic distribution of the DAP-seq experiments, only 40-50% of the sequencing reads in each experiment would be enough for detection of 80% of the most significant peaks.

Fig. 2. Binding of a *Solanum tuberosum* MYC2-related TF to candidate genes. Peaks in blue correspond to enriched DAP-seq signal (i.e., normalized ratio [TF/input]) of a *Solanum tuberosum* MYC2-related TF at different genomic regions. Each panel represents a different proteinase inhibitor gene, presumably regulated by MYC TFs. All the four genes are bound by the TF at their promoter regions in DAP-seq assays. Tracks '+' and '-' represent the Watson and Crisck strands of the genome, respectively. Track 'motifs' shows the positions of the different MYC2-cognate elements, G-box (red), PBE (green) and T/G-box (yellow). Horizontal bar represents 1 kb.

Fig. 3. *Cis*-regulatory elements associated with agriculturally relevant traits. **A**, single nucleotide polymorphisms (SNPs) in *locule number* (*lc*) locus, located 1,080 bp downstream the 3' end of StWUSCHEL. The sequence at this region from several varieties differing in their number of locules is shown, as well as the genotypes for the SNPs (0, homozygous for 'low locules' SNPs; 1, homozygous for 'high locules'; and 2, heterozygous; each number represents the two variable SNPs). Y, C or T; R, G or A. Sequence data were downloaded from NCBI, according to results from [33]. Data on locule number were obtained by [33].

B, binding of a *Solanum tuberosum* MYC2-related TF to *StuPPO1 gene*. Peak summit corresponds to two T/G-boxes in tandem (yellow bars) located 95 bp upstream the start codon. **C,** alignment of MYC2-bound *StuPPO1* region encompassing the two T/G-boxes from several Solanaceae species. Stub, *Solanum tuberosum*; Slyc, *S. lycopersicum*; Smel, *S. melongena*; Spen, *S. pennellii*; Spin, *S. pinnatisectum*; Snum, *S. nummularium*; Cann, *Capsicum annum*; Paxi, *Petunia axillaris*.



Percentage of significant peaks
120
100
80
60
40
20
0
0
2
4
6
8
y = 34.13ln(x) + 43.4
R² = 0.9536
y = 30.31ln(x) + 38.4
R² = 0.9441
0
5
10
15
y = 34.4ln(x) + 10.8
R² = 0,9688
Mapped reads (million)

Figure 1

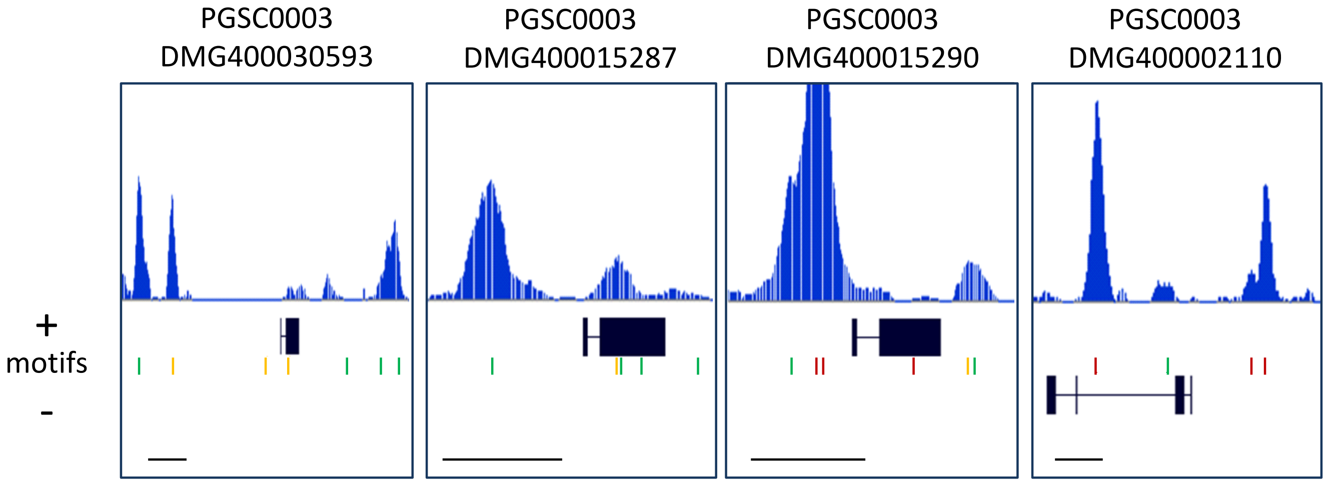

PGSC0003
DMG400030593
PGSC0003
DMG400015287
PGSC0003
DMG400015290
PGSC0003
DMG400002110
+
motifs
-

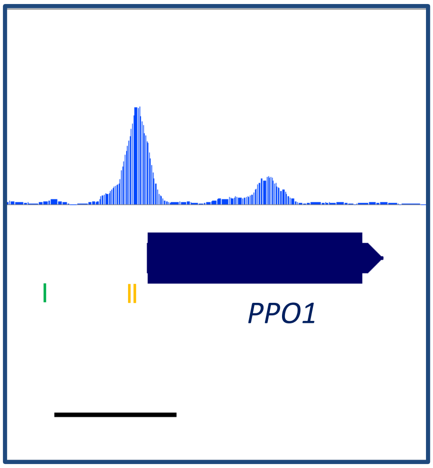
Figure 2

A

		genotype	locule number
Cervena Kapka	GGCATGATGTTTACTAATTGGACAATTCGTACT	0/0	2.0
Cerise Rouge	GGCATGATGTTTACTAATTGGACAATTCGTACT	0/0	2.0
Red Cherry Small	GGCATGATGTTTACTAATTGGACAATTCGTACT	0/0	2.1
Banjul 2	GGCATGATGCTTACTAATTGGACAATTCGTACT	1/0	2.3
Osu 4014-4	GGCATGATGYTTACTRATTGGACAATTCGTACT	2/2	3.1
LA1320	GGCATGATGYTTACTRATTGGACAATTCGTACT	2/2	3.3
Fenhong Tianrou	GGCATGATGYTTACTRATTGGACAATTCGTACT	2/2	4.3
Marmande Anzani	GGCATGATGCTTACTGATTGGACAATTCGTACT	1/1	10.4
LA1251-PI365901	GGCATGATGCTTACTGATTGGACAATTCGTACT	1/1	15.0

B

StuPPO1
(PGSC0003DMG400029575)



C

```
Stub  AGAGAGAGTGAGTAATTACTCCAAGATAAGATCTACAATT
Cann  CATCATAATTTTCATGATTTACAAGACAAGAT---GAACA
Paxi  GATGTCTTGCTTTTTTTATTGCATGATAATAC---AAACA
Smel  AAAGAGAGTA---ATTTAATACAAGATAAGATTTAGACT-
Snum  ------------TAATTACTCCAARATAAGATCTACAATT
Spen  GGAGAGAGTGAGTAAATACTCCAAGATAAGATCTACAATT
Slyc  GGAGAGAGTGAGTAAATACTCCAAGATAAGATCTACAATT
Spin  GGAGAGAGTGAGTAAATACTCCAAGATAAGATCTACAATT
                         * **  * ** *     *

Stub  ATCACCAACGTGTTACACGTTTTGTGCTACA-TATACCTT
Cann  ATCAGCCTTGTGTTTCACGTTTGTTTTTAGGTACCACAAT
Paxi  AGCCTCCACGTGCTTCACGTTTCTAC------TTACCATT
Smel  --CAGCAACATGTTTCACGTTTGTGT---CA-TATTACCT
Snum  ATCACCAATGTGTTACACGTTTTGTGCTACA-TATACCTT
Spen  ATCACCAACGTGTTACACATTTTGTGCTACA-TATACCTT
Slyc  ATCAGCAACGTGTTACACATTTTGTGCTACAATATACCTT
Spin  ATCAGCAACGTGTTACACATTTTGTGCTACAATATACCTT
        *  *    ** * *** ***                 *

Stub  CACCATTTTGTGTATAAATAAAGTTGC-AACTCTTCTAAC
Cann  ATAAGTTGTGTCCATAAATACTGATGAA-ACCATGCAGAG
Paxi  ACCCCTTTTGTGTATAAATAAAGTTGAA-ACCCTTCAAAC
Smel  TTATCAATTGTGTATAAATAAAGGTTACATCC-TTCAACC
Snum  CACCATTTTGTGTATAAATAAAGTTGC-AACTCTTCTAAC
Spen  CACCATTTTGTGTATAAATAAAGGTTGCATCTCTTCAAAC
Slyc  CACCATTTTGTGTATATATAAAGGTTGCATCTCTTCAAAC
Spin  CACCATTTTGTGTATATATAAAGGTTGCATCTCTTCAAAC
              ***  *** ***  * *     *  * *
```

Figure 3