

**DamIT: A novel method for simultaneous detection of transient
TF-gene binding and regulation in *Arabidopsis***

Austen Jack

Dr. Gloria Coruzzi

May 6, 2024

A thesis in fulfillment of the Master's in Biology Degree

Summary

It is necessary to measure transcription factor (TF)-DNA interactions to more accurately characterize gene regulatory networks (GRNs). TFs which operate in a “hit and run” fashion may interact with targets too quickly to be detected with standard DNA-binding assays such as ChIP-seq, which is biased towards stable TF-DNA interactions. Thus, current GRNs may be missing important transient TF-gene interactions. To address this, the DamIT method has been developed to better measure both stable and transient TF-DNA interactions at TF-regulated genes in the same cells. DamIT allows for the simultaneous capture of DNA and RNA from the same sample. This decreases experimental workload and eliminates batch effect issues when comparing TF-DNA binding and gene regulation datasets. Three candidate “hit and run” TFs were tested with DamIT to validate direct interaction with regulated genes: *NLP7*, *GBF2*, and *bZIP1*. DamIT successfully identified hundreds of directly regulated target genes that are also TF-bound in the same cells. DamIT detected DNA-binding at TF-regulated genes that ChIP-seq failed to capture. This verifies the efficacy of using DamIT to study TFs which likely operate under the “hit and run” model.

Introduction

Transcription factors (TFs) bind DNA to regulate gene transcription. After binding a specific DNA sequence, TFs recruit other proteins to either induce or repress gene transcription. Identifying where TFs bind and how they regulate gene expression is integral to our understanding of gene regulatory networks (GRNs). GRNs link TF-DNA binding to gene expression responses that underlie cell development and plasticity at a genome-wide scale.

One interesting insight from GRN research is that a majority of genes bound by a given TF are not regulated (O’Malley et al., 2016). Genes bound but not regulated by a TF may be

poised for regulation in specific conditions, or may lack important co-regulators. Conversely, some TFs have known regulated target genes that do not appear to be bound by that TF. These regulated (but not bound) genes may instead be bound by partner TFs, or may be transient target genes. Transient target genes are hypothesized to fall into a “hit and run” model of gene regulation, where rapid (~seconds) TF-DNA binding (the “hit”) recruits transcriptional machinery to change gene expression. The TF quickly releases from the DNA (the “run”) and the gene expression change is maintained after the TF is no longer bound to DNA (Doidy et al., 2016). This transient TF-gene interaction cannot be detected with conventional TF-DNA binding assays such as ChIP-seq. Thus, current GRNs may be biased towards stably bound target genes, and may be missing important transient targets. Validation of this “hit and run” model of gene regulation would lead to more accurate and comprehensive GRNs.

ChIP-seq is one of the most commonly used methods to identify specific loci where TFs bind. In ChIP-seq, cells are treated with formaldehyde to fix TFs to the genome where they bind. Cells are then sonicated to shear chromatin, and TF-bound DNA is immunoprecipitated (IP) with TF-specific antibodies. Sequencing the precipitated DNA and mapping reads to the genome reveals loci where the TF binds relative to an input (no TF IP) control. Transient targets cannot be detected with ChIP-seq because of the time it takes to fix TFs to the genome. Novel experimental approaches are needed to better validate transient TF gene interactions.

We previously developed the TARGET system (Transient Assay Reporting Genome-wide Effects of Transcription factors) to identify the direct target genes of a TF (Bargmann et al., 2013). In TARGET, glucocorticoid receptor-(GR)-TF-fusion proteins are expressed in plant protoplasts. Cells are treated with cycloheximide (CHX) to prevent translation of any secondary TFs that might disrupt the transcriptome. Dexamethasone (DEX) treatment induces GR:TF

nuclear import. Transfected cells are collected with FACS, and RNA-seq is performed to analyze changes in mRNA expression caused by the TF relative to an empty vector (EV) control.

4-thio-uracil (4tU) can be added to the sample, which gets incorporated into transcribed mRNA. Precipitation of 4tU captures actively transcribed genes in response to GR:TF nuclear import (Para et al., 2014).

We previously combined 4tU-TARGET with time-series ChIP-seq to validate the “hit and run” model for the TF bZIP1, an important *Arabidopsis* nitrogen-response regulator. ChIP-seq with anti-GR antibody revealed genes bound by *bZIP1* in response to GR:bZIP1 nuclear import in short time intervals (time from DEX treatment). In separate samples, pull down of 4tU-containing transcripts revealed the active transcription of GR:bZIP1 target genes that were not bound by GR:bZIP1 with ChIP-seq. This revealed that *bZIP1* can regulate target genes without stably binding to them (Para et al 2014). However, the transient TF-DNA binding of *bZIP1* remains unvalidated. Newer genomic methods are needed to better characterize transient TF-DNA binding.

DNA-Adenine Methyltransferase-Sequencing (DamID-seq) was developed to detect transient TF-DNA binding (Orian et al., 2009). DamID involves tethering a TF of interest to a bacterial DNA adenine methyltransferase (DAM). The DAM domain methylates adenine bases within GATC motifs near the sites of DAM:TF DNA binding. The adenine methylation mark is permanent, resulting in a stable footprint of TF-DNA binding that can be isolated with enzyme digestion and sequenced. DamID provides a more comprehensive map of TF-DNA binding events, many of which are transient interactions that are missed by ChIP-seq (Alvarez et al., 2020).

We recently incorporated DamID technology into the TARGET system to allow for high-throughput detection of transient TF-DNA binding in plant cells (Alvarez et al., 2023). We modified the TARGET experiment to express DAM:GR:TF fusion proteins, which when combined with DNA extraction and DamID-seq, reveals genome-wide TF-DNA binding in response to DEX treatment. We used this DamID approach to validate the “hit and run” model of gene regulation for an important plant nitrogen GRN hub TF, *NLP7* (Alvarez et al., 2020). *NLP7* was successfully shown to transiently bind with loci which it regulated in TARGET but did not stably bind in ChIP-seq.

A major limitation to this experiment is that the transient TF-DNA binding (DamID) data is generated from a separate experiment than the TF-gene regulation assay (TARGET). Confounding issues of batch effect (protoplast cell population variation, presence of important partner TFs, time of day signaling, etc) may lead to discrepancies between binding and regulation datasets. There is also a large workload required to collect binding and regulation datasets for the same TF. DamID has been modified to study other aspects of gene regulation such as TF-TF interactions and chromatin accessibility (Aughey et al., 2019). For example, Markodimitraki et al. recently developed scDam&T-seq, which captures chromatin protein-DNA contacts and gene expression from the same cells using a unique barcoding approach in combination with single cell sequencing (Markodimitraki et al., 2020). scDam&T-seq uses stable cell lines expressing DAM:chromatin protein fusions and allows for robust comparisons between the chromatin and gene expression landscapes of different cell types. This approach is difficult to scale up as it requires transgenic cell lines. Furthermore, the long-term DAM activity can be toxic in cells and affect cell phenotypes. We sought to improve upon this method to allow for

higher-throughput simultaneous detection of TF-gene binding and regulation in response to TF overexpression.

Here we present DamIT (DamID-seq Incorporating Transcriptomics), which allows for simultaneous capture of direct TF-gene regulation and genome-wide TF-DNA binding from the same cell population. Briefly, DAM:GR:TF fusion proteins are transfected into protoplasts that are treated with CHX and DEX. FACS sorting isolates transfected cells, and a kit is used to simultaneously isolate RNA and DNA without the need for splitting cells. RNA is sequenced to identify direct target gene regulation relative to a DAM:GR:EV control. DNA is processed through the standard DamID-seq pipeline to identify genome-wide transient TF-DNA binding enrichment relative to the DAM:GR:EV control. When combined, genome-wide TF-DNA binding and target gene regulation can be studied from the same cells. This greatly reduces experimental workload that typically comes from separate ChIP-seq and RNA-seq experiments. DamIT offers a rapid and scalable approach to studying transient TF gene interactions in any species amenable to cell transfection. DamIT is also a more accessible approach to the study of “hit and run” transcription as it does not rely on single-cell sequencing platforms. We developed DamIT and applied it to three *Arabidopsis* TFs to show that the “hit and run” model is a prevalent mechanism of transient gene regulation.

Methods

Aim 1: TF Selection and Cloning

TF Selection and Sequence Optimization

Selecting TFs to validate as transient with DamIT was based on the criterion that they show <50% bound/regulated targets when TARGET data from Brooks et al. was combined with ChIP-seq data from Song et al. (Brooks et al., 2021 and Song et al., 2016). The following TFs

met this requirement: *GBF2*, *FBH3*, *DREB2A*, *MYB3*, *HSFA6A*, and *HAT22*. The coding sequences of each of these genes were retrieved from TAIR.org for optimization. The sequence optimization script from Alvarez et al. was used on the coding sequences and showed how many GATC sites were replaced with synonymous codons as well as whether to clone with BsaI or BsmBI enzymes (Alvarez et al. 2023). The indicated cloning enzyme's digestion site was added to one end of the optimized sequence and the reverse complement of the digestion site was added to the other. The final gene fragment sequence was sent for synthesis from Synbio. Gene fragments were PCR amplified in order to have sufficient quantities for cloning.

Cloning Optimized TF Gene Fragments into pDamBOB Plasmid

Each synthesized gene fragment was cloned into the pDamBOB plasmid. The plasmid backbone was digested with BsaI and Quick CIP following directions provided in the enzyme kits from NEB. Each synthesized fragment was digested with the assigned enzyme from optimization following standard methods from NEB. The gene fragments and plasmid backbones were ligated together using T4 DNA ligase from NEB. The final ligation reactions were run on 2% agarose gels, bands equal to the expected size of the final plasmids were cut out of the gels and purified using the Monarch DNA Gel Extraction kit from NEB. Nanodrop was used to measure DNA concentration and confirm if the DNA had been successfully purified.

Transforming Pir1 cells with Cloned pDamBOB Plasmids

OneShot Pir1 cells were thawed on ice for 30 minutes, then 100 ng of plasmid DNA for each gene fragment was incubated over ice with its own aliquot of 25 μ L of thawed cells for 30 minutes. The cells were heat-shocked at 42 degrees celsius for 30 seconds. The cells were put back on ice for 5 minutes before being added to 450 μ L of LB stock media. The LB + cells mixture was placed in a 37 degree celsius incubator for 1 hour outgrowth. At the end of the

outgrowth period, 3 aliquots of 100 μ L of the LB + cells combination were plated onto 3 LB + carbenicillin plates. The plates were placed in a 37 degree celsius incubator overnight. If there was growth the following day, the plasmid was successfully cloned, otherwise the previous steps would be repeated.

Quality Control of Cloned Plasmids

From each plasmid sample, six colonies were picked and put into 5 mL of LB + carbenicillin broth. These were grown in a 37 degree incubator overnight. The following day, plasmid DNA was extracted from these samples using the Zymogen MiniPrep kit. Nanodrop was performed on the final elution from the MiniPrep kits.

Each plasmid was digested with an enzyme found on the plasmid backbone only once, and not in the gene fragment at all. This is to linearize the plasmid and ensure only one DNA band be seen on the gel. The digestions were run on 2% agarose gels, and any samples that showed a single strong band at the expected plasmid size were sent for sequencing. Two plasmids with quality gels from each gene fragment sample were sent to Plasmidsaurus for whole-plasmid sequencing. The sequencing results for each plasmid showed if cloning was successful and if the cloned fragment was in-frame with the Dam methyltransferase and glucocorticoid receptor domains. All TF gene fragments were successfully cloned into the pDamBOB plasmid.

Aim 2: DamIT Assay

DamIT Assay Preparation: Plasmid Bulking and Protoplasting

Transformed Pir1 cells were preserved in glycerol stocks in order to grow for plasmid bulking. Each transformed cell population was grown in 5 mL of LB + carbenicillin in a 37 degree celsius incubator for 5 hours. The 5 mL culture was transferred to 300 mL of LB +

carbenicillin and grown in a 37 degree incubator overnight. The following day, plasmid DNA was extracted from the cell cultures following the Zymogen MaxiPrep vacuum protocol. Final eluted plasmid DNA concentration was measured using Nanodrop. Plasmid volumes equal to 120 ng of DNA were collected.

Col-0 *Arabidopsis thaliana* seeds were sterilized with 10% bleach and grown on 2% agar plates following the protocol outlined by Alvarez et al. The roots were removed from the plates after 10 days of growth. Protoplasts were prepared from the roots using 1% cellulase solution. Protoplasts were stained with FDA and counted on a hemocytometer under a fluorescent microscope. Enough protoplasts were produced for 3,000,000 cells for each treatment replicate.

DamIT Procedure and Optimization

3,000,000 *Arabidopsis* protoplasts cells were aliquoted for each treatment replicate. Each aliquot was transfected with 120 ng of plasmid DNA mediated by PEG. The transfected protoplasts were transferred to a 24-well cell culture plate and wrapped in foil to prevent degradation of mCherry fluorescence and incubated on a shake table overnight at room temperature for production of the desired DAM:GR:TF fusion proteins. Each replicate was treated with 5 μ L of 1 mM CHX after the incubation and then treated with 1.25 μ L 1 mM DEX two hours prior to FACS sorting. FACS sorting was performed into 350 μ L Qiagen AllPrep Buffer RLT+ which lyses cells and stabilizes nucleic acids for processing and analysis.

DNA and RNA from the FACS sorted cells were separated with the Qiagen AllPrep micro kit following the cell suspension protocol verbatim. DamID was performed on the DNA from each replicate as described in Alvarez et al., except for the addition of the previously described FACS sorting and T7 exonuclease degradation of background DNA for all replicates (Alvarez et al., 2023). DNA libraries were prepared with the NEBNext Ultra II DNA library prep

kit following the protocol as described. RNA libraries were prepared with the NEB Ultra II RNA Library Prep kit following the protocol as described. Both the DNA and RNA libraries were analyzed on TapeStation for the proper library fragment sizes. DNA and RNA libraries were sequenced with Illumina. Single-end sequencing was used for DNA, and paired-end sequencing was used for RNA. Library fragment length was set to 150 bp for DNA library sequencing and 100 bp for RNA library sequencing.

Aim 3: Computational Analysis

DNA-seq Computational Analysis

DNA library fastq files were first trimmed with CutAdapt with the default parameters as well as the following: -q 30 --times 4 --trim-n -m 25. The trimmed files were processed in Fastqc with default parameters to ensure quality of trimmed sequences. Trimmed files were aligned to the TAIR10 Arabidopsis genome (Araport11 annotation) using HiSat2 with the default parameters, then sorted into BAM files in SamTools. DeepTools bamcompare was used to produce BigWig files with default parameters as well as the following: --binSize 1 --minMappingQuality 30 --normalizeUsing RPKM -e 300. These BigWig files were used to visualize peaks in JBrowse. Sorted BAM files were quality filtered in SamTools with SamTools view and the following parameter: -q 30. For each sample, the two replicates of filtered BAM files were merged with SamTools merge, which would be analyzed further. All filtered, merged, and intersect BAM files were processed in MACS2 to call peaks. MACS2 predictd was used to determine the --extSize parameter for each sample which will be used with MACS callpeak. MACS2 callpeak was used for each sample bam file relative to DAM-GR-EV with the following parameters: -f BAM -g 1.2e8 --nomodel --extsize (predictd value) -q e-8. BedTools intersect was used to find common peaks between the two filtered BAM files for each replicate. The first three

columns of resulting narrowPeak files were processed in BedTools with the Araport11_protein_coding.201606.bed *Arabidopsis* gene annotation. BedTools window was run for each bed file with the following parameters: -sw -r 1000 -l 0 to intersect peaks with genes. The final gene lists were analyzed in R for peak visualization and intersection with RNA-seq data.

RNA-seq Computational Analysis

RNA library fastq files were trimmed with CutAdapt with the default parameters as well as the following: -q 30 --trim-n -m 25. Quality of trimmed RNA library reads was analyzed with Fastqc at default parameters. The trimmed reads were aligned to the TAIR10 *Arabidopsis* genome with STAR paired-end alignment on default parameters. Subread FeatureCounts with the paired-end parameter was used to count how many reads were aligned to each gene in the *Arabidopsis* genome. Final gene counts files were analyzed in R with DESeq2 for differentially expressed genes (DEGs, $p < 0.05$). The clusterProfiler package was used to gene ontology, and EnrichedHeatmap was used to plot DNA binding over the TF-regulated genes.

Results

TF's in *Arabidopsis* which likely operate in a “hit and run” fashion were determined with high-quality TARGET and ChIP-seq data (Brooks et al. 2021 and Song et al. 2016). The combination of these datasets shows the ratio of regulated genes which are also stably bound in ChIP-seq. TFs with a low ratio of bound/regulated targets were candidates for operating as “hit and run” TFs. The goal of this research is to validate candidate “hit and run” TFs as operating under this transient model of gene regulation using DamIT. Given the low ratio of bound/regulated targets in the candidate “hit and run” TFs, it is hypothesized that they will be

validated as true “hit and run” TFs if DamIT is able to detect higher levels of interaction between TFs and genes they regulate compared to ChIP-seq.

Aim 1: Production of Viable Plasmids for DamID Assays

Determination of Likely “Hit and Run” TFs and Codon Optimization

We considered a TF with a bound/regulated ratio at or below 50% as a possible candidate for operating under the “hit and run” TF model in *Arabidopsis*. The following TFs were selected for DamID: *GBF2*, *NLP7*, *bZIP1*, and a Dam-GR-EV control. *GBF2* was found to have exactly 50% bound/regulated targets in *Arabidopsis*, meaning it may operate as a “hit and run” TF for half of its targets, and stably bind the other half (Fig. 1a). Verification of this model has not been tested for *GBF2*. *bZIP1* is a known “hit and run” TF based on our previous studies integrating time-series ChIP-seq and TARGET data (Para et al, 2014). DamID has not yet been completed on *bZIP1*, thus we are including *bZIP1* to further validate its “hit and run” activity. We recently validated the transient “hit and run” model of action for *NLP7* using DamID, thus we are including *NLP7* as a positive control (Alvarez et al. 2020).

This collection of TF’s gives a full range of known/unknown transient TFs: *NLP7* is a fully known transient “hit and run” TF, *bZIP1* is known as transient but has not been validated with DamID, and *GBF2* has not yet been tested.

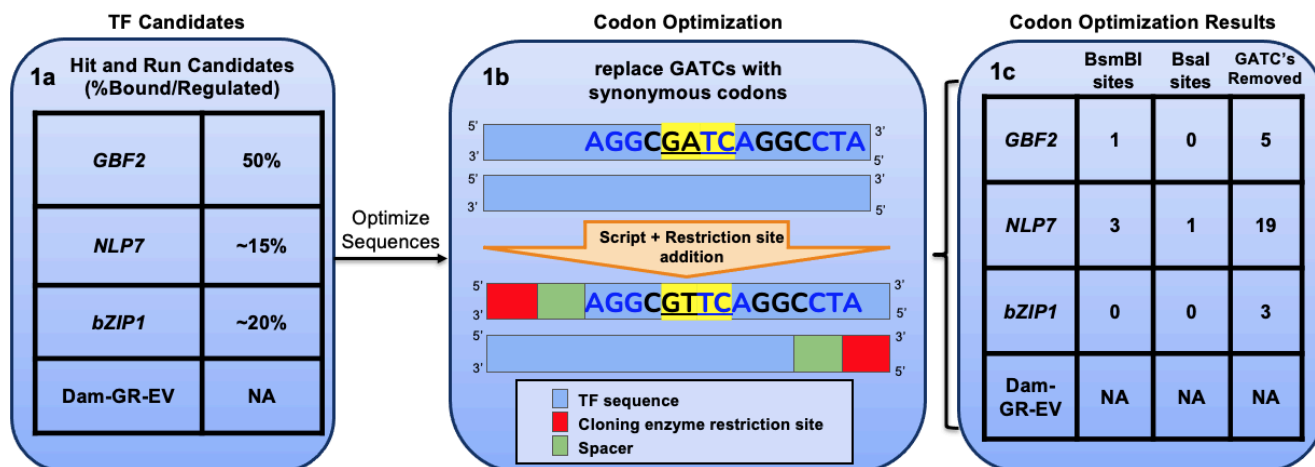


Figure 1. Synthesis and cloning of “hit and run” TFs into the DamID vector. 1a) Ratio of bound/regulated targets in candidate “hit and run” TFs. Candidates were selected if they had a regulated/bound ratio <50%. 1b) Schematic of how selected TFs are sequence-optimized for DamIT experiments. GATC’s are replaced with synonymous codons, and cloning enzyme sequences are added to the ends. 1c) Output statistics of codon optimization script for each selected TF.

Because DamID relies on GATC methylation, the sequences of these “hit and run” candidate TFs are optimized to replace any GATC sites with synonymous codons so that no plasmid DNA methylation contaminates the final data. Codon optimization also indicates which enzymes to use for golden-gate cloning into the pDamBOB vector. Each TF sequence is run through a script from Alvarez et al. that is able to achieve this and provide the necessary statistics upon completion (Alvarez et al., 2023) (Fig. 1b, 1c). The script reports how many GATC’s are replaced and how many BsaI or BsmBI sites are in the sequence. If a TF contains only one of either the BsaI or BsmBI sites but not the other, the enzyme that does not have any digestion sites in the fragment will be added to the end of its sequence for cloning. If the TF does not contain any BsaI digestion sites, BsaI digestion sites are added to flank the codon-optimized TF sequence to facilitate golden gate cloning. If one or more BsaI digestion sites are found within the TF sequence, BsmBI sites are added to the TF sequence instead. If both BsaI and BsmBI sites are found within the TF sequence, the BsaI site will be codon optimized and thus removed, and BsaI digestion sites will be added to the ends of the codon optimized sequence.

NLP7 and *bZIP1* had already been run through this optimization. In *NLP7*, 19 GATC sites were replaced with synonymous codons, and there were 3 BsmBI sites and 1 BsaI site (Fig. 1c). The BsaI site was replaced with a synonymous codon. In *bZIP1*, there were 3 GATC sites replaced with synonymous codons, and 0 BsmBI and BsaI sites (Fig. 1c). The *GBF2* sequence had not been run through this script, so it was optimized to continue with cloning. *GBF2* had 5 GATC sites replaced with synonymous codons, and it contained 1 BsmBI site and 0 BsaI sites (Fig. 1c). A BsaI site was added to the beginning of the optimized sequence, and the reverse complement of a BsaI site was added to the end of the optimized sequence for cloning. The optimized fragment sequence was synthesized by Synbio to continue with the cloning process.

Cloning and Pir1 Transformation

The TF gene fragments were cloned into the pDamBOB vector in-frame with a dam-methyltransferase (DAM) domain and a glucocorticoid receptor (GR) domain. The pDamBOB vector contains the R6K origin of replication. R6K contains no GATC sites and is only compatible with Pir1 *E. coli* cells, which is what cloned plasmids were transformed into. This plasmid also confers carbenicillin resistance to Pir1 cells by expressing the *Bla* gene. pDamBOB expresses an mCherry gene which causes root protoplasts to fluoresce in FACS if they are successfully transfected. Lastly, the vector also contains many digestion sites to linearize it for running on screening gels during cloning quality control. *NLP7*, *bZIP1*, and the Dam-GR-EV control were all already produced and ready for DamID; *GBF2* needed to be cloned into the vector.

The synthesized TF fragments were digested with either BsaI or BsmBI depending on the results of optimization. The pDamBOB plasmid was digested with BsaI and Quick CIP to prevent unwanted re-ligation of the plasmid. These separate digestions were ligated together with

T4 DNA ligase to produce final cloned plasmids. Once added to the vector, the ligation reaction was transformed into Pir1 *E. coli* cells. Any cells successfully transformed gained resistance to carbenicillin from the *Bla* fragment on the plasmid. The transformed cells were grown on LB plates that contained carbenicillin. Any colonies that grew were known to contain plasmids, but they needed to be screened in order to ensure that the correct plasmid had been transformed into the cells.

To prepare the screening gel, 6 colonies were selected and mini-prepped. Each of the bulked plasmids were digested with a unique cutting restriction enzyme (*Mfe*I) to linearize the plasmid. Linearized plasmids are easier to run and compare on an agarose gel. Linearized plasmids were analyzed on gels to look for single bands at the expected DAM-GR-GBF2 plasmid size (6,000 BP) (Fig. 2b). Any screened plasmids that had multiple bands or had a single strong band far outside the expected length were considered improperly cloned.

Quality Control of DamIT Plasmids

For *GBF2*, Two MiniPrep samples were selected and sent for sequencing with Plasmidsaurus to verify which is the intended plasmid (Fig. 2b). Plasmidsaurus performs whole-plasmid long-read sequencing, and returns a report of what fragments are on a sequenced plasmid. The reports have high accuracy and are the primary method of quality control for TF gene fragment cloning. *GBF2* was successfully cloned in-frame with the DAM and GR fragments that will produce the fusion protein used in later experiments. The other plasmid that had been sent for sequencing contained *bZIP1* because the original cloning vector contains *bZIP1* as a filler sequence before digestion with *Bsa*I. This particular plasmid had either not been digested, or the *bZIP1* fragment had re-ligated. The successfully cloned *GBF2* plasmid was used to prepare a glycerol stock of transformed Pir1 cells for storage and subsequent bulking of higher

quantities of plasmids for the DamIT experiments. This cloning pipeline was also completed on 5 other candidate “hit and run” TFs (*FBH3*, *MYB3*, *DREB2A*, *HSFA6A*, *HAT22*) so they may be tested later with DamIT.

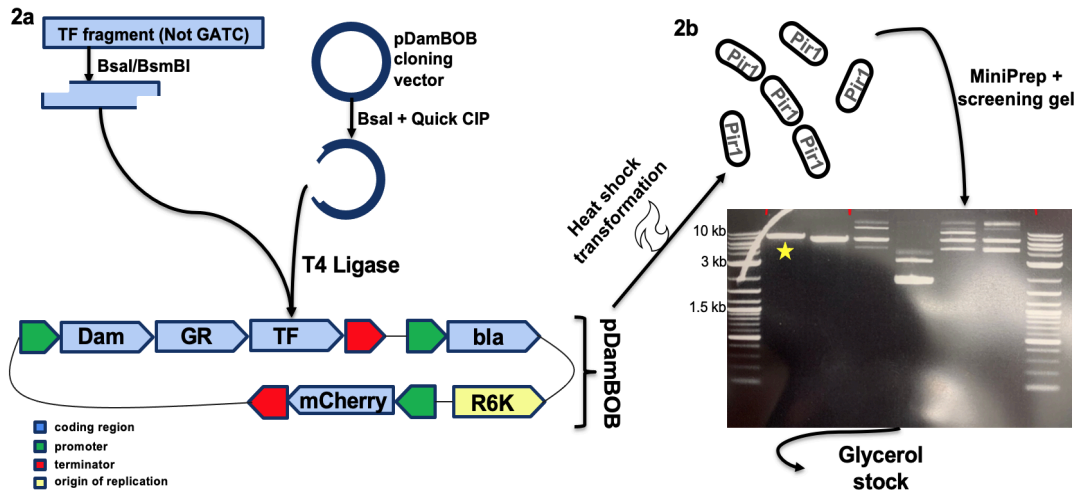


Figure 2. Mapping cloned pDamBOB vector and performing Pir1 transformation. 2a) Synthesized codon-optimized TF fragment is digested with BsaI or BsmBI, and pDamBOB backbone is digested with BsaI and QuickCIP. T4 ligase is used to clone the fragment into the plasmid in-frame with the Dam and GR domains. A Bla sequence confers carbenicillin resistance to transformed Pir1 cells. R6K is the origin of replication which is compatible with Pir1 cells. mCherry is the fluorescent protein transfected *Arabidopsis* cells express for FACS sorting. 2b) Cloned plasmids are transformed into Pir1 cells with heat shock. Transformed cells are screened on a gel to determine which to send for sequencing and prepare glycerol stocks of. A star indicates the *GBF2* sample which was verified as successfully cloned.

Aim 2: DamIT Assay Description and Experimental Design

TARGET/DamIT Assay

The overall purpose of this assay is to transfect *Arabidopsis* root protoplasts with DAM:GR:TF plasmids and analyze the resulting DNA and RNA which come from the same cells. The DNA is methylated on the adenines of any GATC sequence near loci the TF transiently interacts with. RNA is sequenced to determine which genes are upregulated or downregulated in response to TF nuclear import.

The first two steps of the experiment are to prepare *Arabidopsis* root protoplasts and bulk the DAM:GR:TF plasmids from glycerol stocks prepared in Aim 1. With root protoplasts and

plasmid DNA prepared, the TARGET portion of DamIT can be performed. This begins with PEG-mediated transfection of the protoplasts, followed by treatment with cycloheximide (CHX) and dexamethasone (DEX). CHX inhibits translation within the protoplasts in order to prevent translation of any downstream secondary TFs that may affect transcription of RNA not directly influenced by the TF being tested. DEX treatment induces nuclear transport of the 35S-overexpressed DAM:GR:TF fusion protein via the GR domain. Once in the nucleus, the DAM:GR:TF interacts with DNA and the DAM domain methylates the adenines of any GATC sequences near the sites of TF-DNA binding regardless of how stable the interactions are. Any successfully transfected protoplasts also expresses the mCherry fluorescent protein which is used for FACS after 2 hours of DEX treatment. FACS is an addition we have made in this experiment compared to the previous DamID method used in Alvarez et al. This excludes any cells that are not successfully transfected, removing their background gDNA and mRNA profiles from our final dataset. The Qiagen All-Prep kit (Cat ID 80284) is used to separate the DNA and RNA collected from FACS so that they can be processed for DamID and RNA-seq respectively. Conceptually, FACS sorts cells into a buffer that stabilizes both DNA and RNA from the cells. Then all sorted cells are spun through a collection column where only DNA is collected and RNA is preserved in the flow-through for isolation. From that point DNA and RNA can be processed separately.

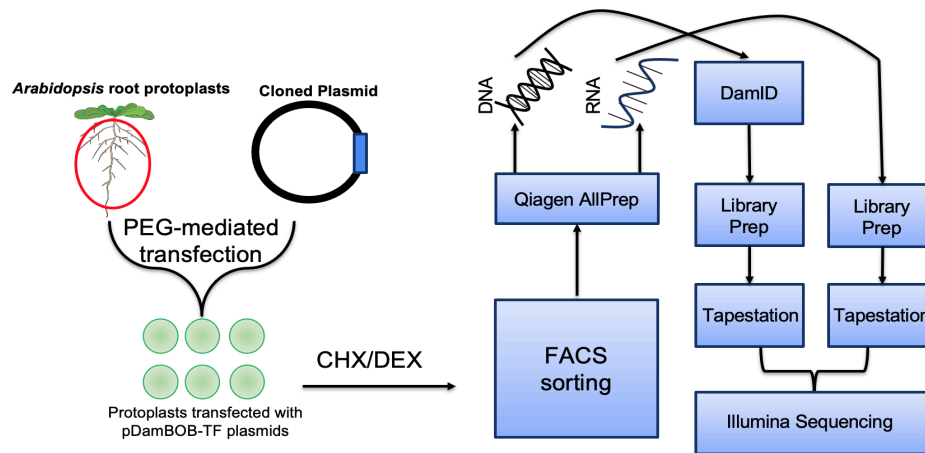
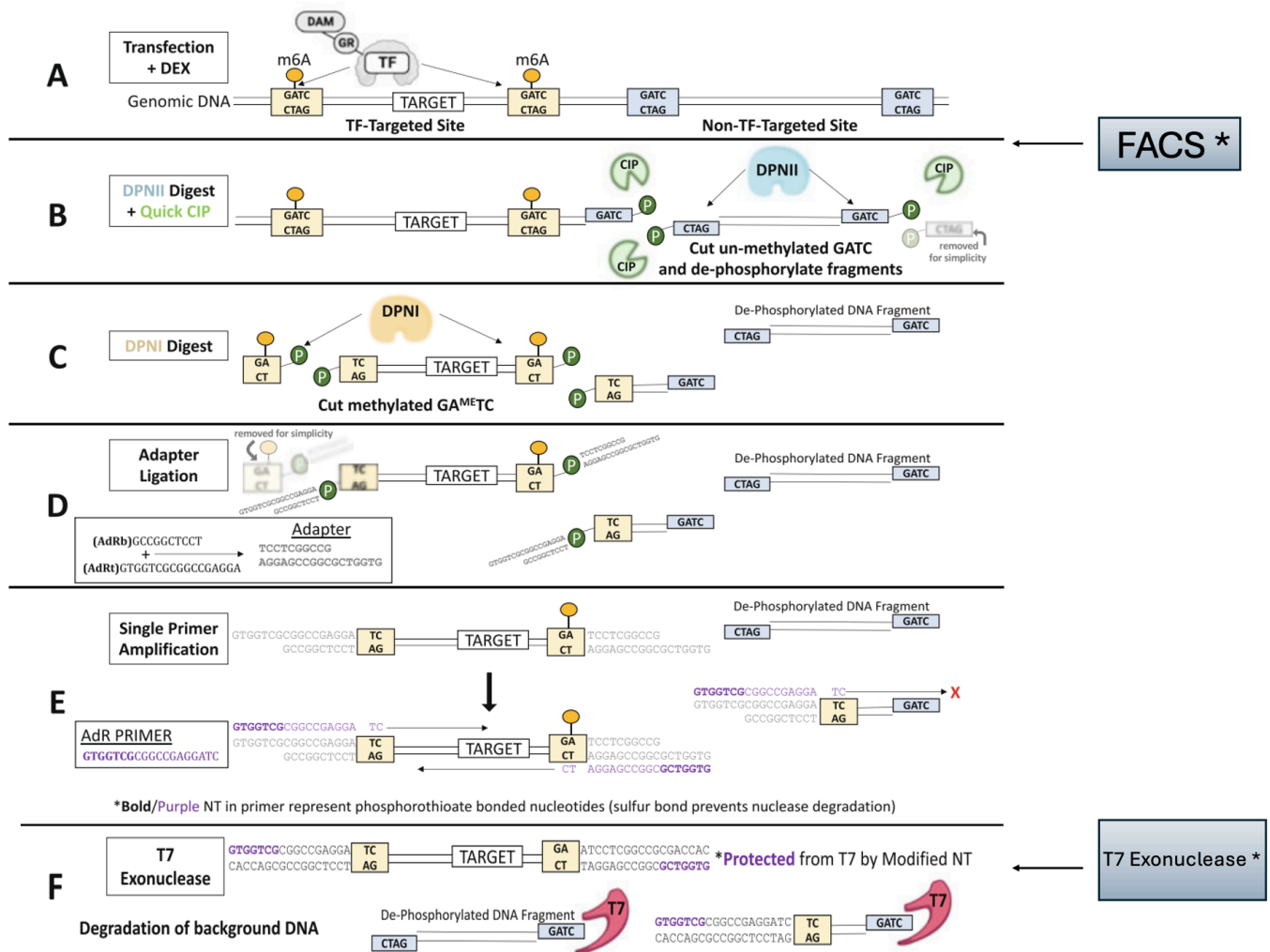


Figure 3. Sample processing for the DamIT protocol. *Arabidopsis* root cell protoplasts are prepared and transfected with cloned DAM:GR:TF plasmids. Transfected protoplasts are treated with CHX and DEX before cell sorting with FACS. DNA and RNA from sorted cells are separated using the Qiagen AllPrep kit. DNA is processed through DamID to collect and amplify methylated DNA. Final DNA and RNA libraries are prepared and analyzed in Tapestation for quality control before sequencing with Illumina.

DamID Procedure and Optimization

Upon collection from the Qiagen All-Prep kit, RNA is prepped for Illumina sequencing with the NEBNext Ultra II RNA Prep kit. DNA is processed through the standard DamID pipeline to isolate methylated DNA associated with TF and DNA binding. Methylated DNA is first digested with DpnII and dephosphorylated with Quick CIP. DpnII digests all unmethylated GATC sequences, and Quick CIP dephosphorylates digested fragments to prevent adaptor ligation in later steps. All methylated GATC sequences are digested with DpnI, which must be completed second because DpnI is known to sometimes digest unmethylated GATC sequences as well. If any unmethylated GATC sites are digested by DpnI, they will be included in the final library and affect the analysis by causing false positives. Adaptors are ligated to the phosphate groups of DpnI-digested sites. Once adaptors are ligated, AdR primers are used to PCR amplify the adaptor-ligated DNA fragments. The nucleotides in these primers are phosphorothioate-bonded and cannot be degraded by T7 exonuclease, which is used following PCR amplification to degrade all background DNA. A final DNA library is produced with the

NEBNext Library Prep Kit for Illumina Sequencing. All libraries are analyzed with TapeStation to ensure that the majority of fragments are within the desired range of about 200-500 bp in length and of high enough quantity for sequencing.



Alvarez et al. 2023

Figure 4. Schematic of the DamID protocol from Alvarez et al. 2023 and changes implemented. 4A) After transfection the DAM domain of the DAM:GR:TF fusion protein methylates adenines of GATC sequences. FACS is incorporated after this step, which is not included in the original protocol by Alvarez et al. 4B) DpnII digest with Quick CIP cleaves and dephosphorylates all unmethylated GATC sites. 4C) DpnI digest cleaves all methylated GATC sites. 4D) Ligation of adaptors to methylated (phosphorylated) fragments. 4E) PCR amplification with phosphorothioate-bonded primers. 4F) T7 exonuclease degradation of background DNA, which is optional in the Alvarez et al. protocol but used for all samples in this experiment.

Experimental Design with the DamIT Assay

There were four samples run through the DamIT assay with three treatment replicates each; *NLP7*, *bZIP1*, *GBF2*, and a Dam-GR-EV control. Each replicate comprised >3,000,000 protoplasts and was transfected with plasmid DNA. FACS sorting showed low cell counts in two out of three replicates for each sample. The two lowest cell-count replicates for each sample were pooled together after FACS, resulting in two total replicates per sample for the remainder of the experiment. Upon completion of DamIT with the sorted cells, libraries were prepared for both the DNA and RNA. These libraries were analyzed with Tapestation, and had the proper fragment sizes in sufficient quantities for Illumina sequencing.

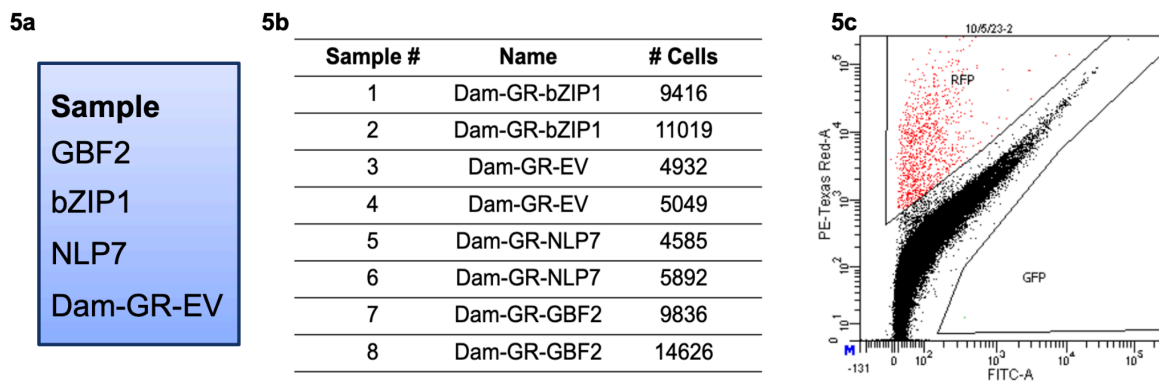


Figure 5. Experimental design and FACS data. 5a) *GBF2*, *bZIP1*, and *NLP7* were selected for DamIT experimentation with a Dam-GR-EV negative control. Each sample had two biological replicates and was processed through the DamIT protocol with methylated DNA and RNA transcriptional profiles as the final products. 5b) A table of cell counts of sorted cells from each sample replicate. 5c) An image of the gating from a FACS sample. There was a strong RFP (mCherry) signal, showing success of transfection.

Aim 3: Computational Analysis of DamIT Data

Computational pipelines as outlined in Figure 6 were used to analyze DamIT DNA-binding data and TF-DEG data. These pipelines were performed for each of the *NLP7*, *GBF2*, and *bZIP1* samples. The goal of intersecting the DamIT TF DNA-binding data with DamIT TF-DEG data is to show which loci regulated by a TF are also bound by the TF in the same cells, validating such loci as being directly influenced by the TF.

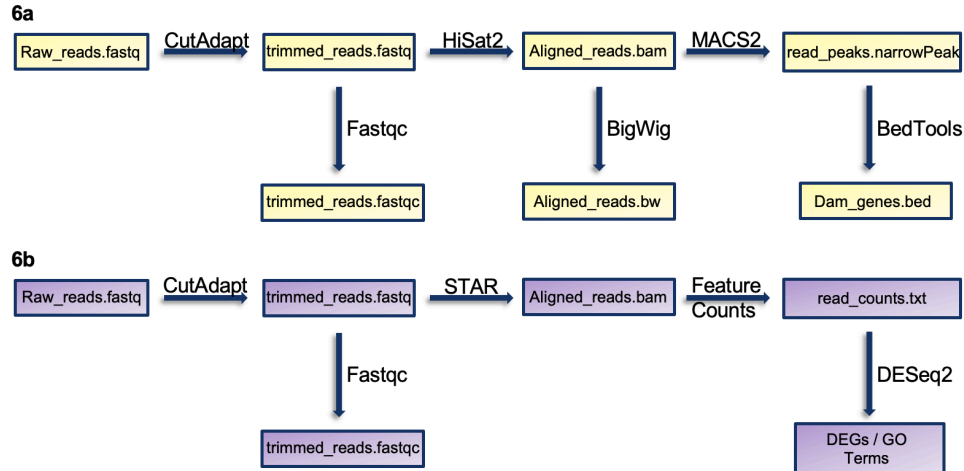


Figure 6. Computational analysis pipelines for both DNA and RNA libraries. 6a) DNA computational analysis pipeline. Raw reads were trimmed with CutAdapt, then analyzed for quality with FastQC. Trimmed reads were aligned to the *Arabidopsis* genome using HiSat2. Aligned reads were analyzed in BigWig for visualization, and peaks were called in MACS2. BedTools associated final peaks with *Arabidopsis* genes. 6b) RNA computational analysis pipeline. Raw reads were trimmed with CutAdapt, then analyzed for quality with FastQC. Trimmed reads were aligned to the *Arabidopsis* genome with STAR. Aligned reads were assigned to genes with FeatureCounts, and then analyzed for DEGs in DESeq2.

Differential Gene Expression in GBF2, bZIP1, and NLP7 with DamIT

Each of *GBF2*, *bZIP1*, and *NLP7* showed TF overexpression relative to EV and other experiment TFs samples (Fig. 7a, 7d, 7g). All samples had significant DEGs both induced and repressed relative to the DAM-GR-EV controls. *GBF2* regulated a total of 1,122 genes, where 490 were induced and 662 were repressed (Fig. 7b). *bZIP1* regulated a total of 399 genes, where 241 were induced and 158 were repressed (Fig. 7e). For *NLP7*, 458 genes were regulated, where 337 were induced and 121 were repressed (Fig. 7h). Expression levels of DEGs specific to each TF were compared across all samples to validate that a TF was responsible for the highest rate of regulation for its DEGs (Fig. 7b, 7e, 7h). The three most significantly enriched gene ontology terms from the induced and repressed gene lists were collected for each sample, and were consistent with processes each TF is known to influence (Fig. 7c, 7f, 7i) (Verma et al., 2022; Para et al., 2014; Alvarez et al., 2020).

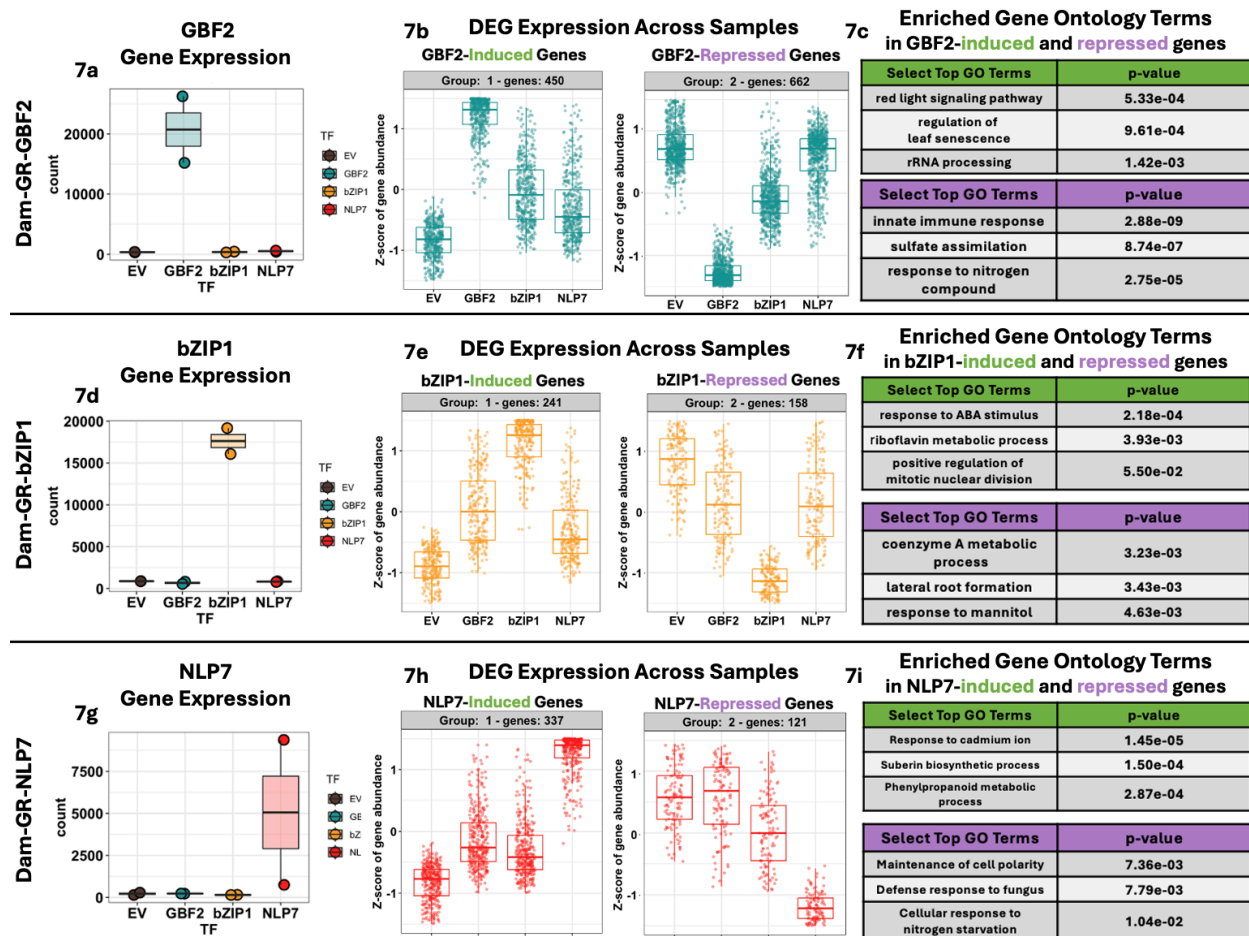


Figure 7. DamIT captures genes directly regulated by TFs. TF overexpression for 7a) *GBF2*, 7d) *bZIP1*, and 7g) *NLP7*. TF-specific DEG expression relative to other samples for 7b) *GBF2*, 7e) *bZIP1*, and 7h) *NLP7*. Enriched gene ontology terms of most significant induced and repressed genes for 7c) *GBF2*, 7f) *bZIP1*, and 7i) *NLP7*.

DNA Binding of *GBF2*, *bZIP1*, and *NLP7* with DamIT

DamIT-DNA binding peaks were called relative to the DAM-GR-EV control with a stringent q-value threshold of $1e-8$. Peaks collected from all samples were about 2,500 bp in length (Fig. 8b, 8e, 8h). Two *GBF2* replicates called 8,616 and 2881 peaks, and merging these two replicates together resulted in 10,054 peaks (Fig. 8a). One example of a significant *GBF2* peak called in one sample is visualized with JBrowse in Figure 8c. Two *bZIP1* replicates called 10,799 and 15,476 peaks; merging these two replicates together called 16,378 total peaks (Fig. 8d). One example of a significant *bZIP1* peak is visualized with JBrowse in Figure 8f over an F-Box gene. Two *NLP7* replicates resulted in 10,183 and 13,766 peaks; merging these two

replicates together called 16,339 peaks (Fig. 8g). One example of a highly reproducible *NLP7* peak over the promoter of a transmembrane protein is visualized with JBrowse in Figure 8i.

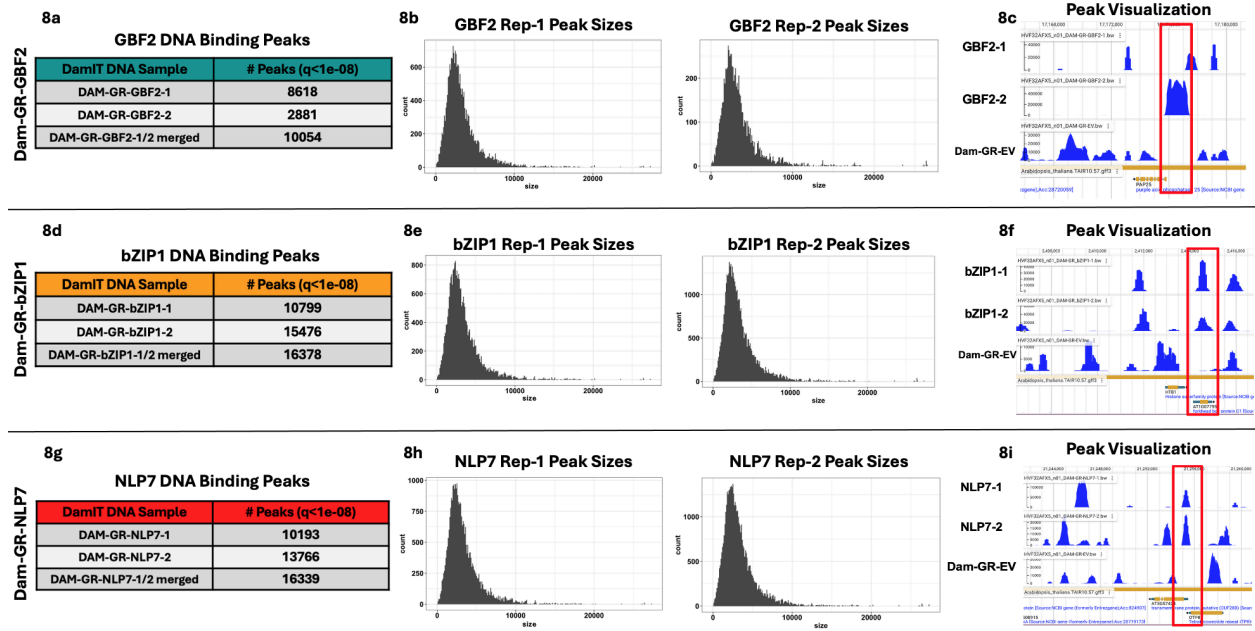


Figure 8. DamIT captures genome-wide TF-DNA interactions. Quantities of DNA-binding peaks for individual replicates and merged datasets from 8a) *GBF2*, 8d) *bZIP1*, and 8g) *NLP7*. Distributions of DNA-binding peak sizes for individual treatment replicates from 8b) *GBF2*, 8e) *bZIP1*, and 8h) *NLP7*. Select examples of visualized significant DNA-binding peaks for 8c) *GBF2*, 8f) *bZIP1*, and 8i) *NLP7*.

Intersection of DEGs and DNA Binding Data for *NLP7*, *GBF2*, and *bZIP1*

Combining DEG and DNA binding datasets for each TF shows how many genes regulated by a TF are also TF-bound in DamIT. Of 1,122 genes regulated by *GBF2*, 407 were bound by *GBF2* in at least one replicate, giving a binding/regulation rate of 36% (Fig 9a). Plotting *GBF2* DNA binding over *GBF2* DEGs relative to the gene TSS shows a similar binding profile compared to the EV control (Fig. 9b). Of 399 genes regulated by *bZIP1*, 224 were bound by *bZIP1* in at least one DamIT replicate, giving a binding/regulation rate of 56% (Fig. 9c). Plotting *bZIP1* DNA binding over DEG TSSs shows an average binding profile with clear binding enrichment at DEGs gene-bodies compared to the EV control (Fig. 9d). For the 461

NLP7 DEGs, 298 were bound by *NLP7* in at least one replicate, giving a binding rate of 65%

Compared to the DAM-GR-EV, there is enrichment of *NLP7* binding at NLP7 DEGs (Fig 9e, 9f).

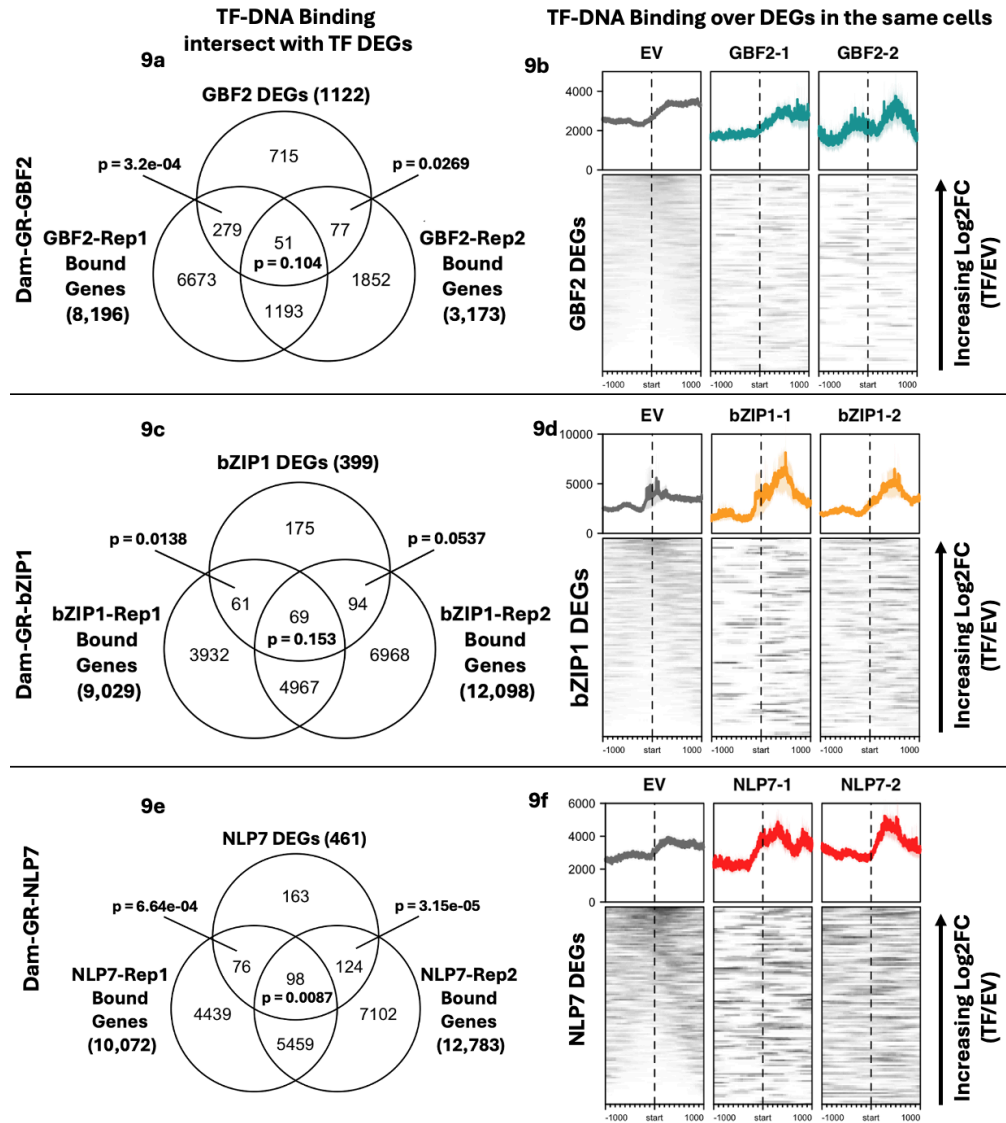


Figure 9. DamIT identifies hundreds of TF-regulated targets which are directly bound by the TF. Venn diagram of TF DEGs intersected with both replicates of TF bound genes for: 9a) GBF2, 9c) bZIP1, and 9e) NLP7. DNA binding profiles over DEG lists compared DAM-GR-EV (EV), centered at DEG TSS +/- 1kb for 9b) GBF2 9d) bZIP1 and 9f) NLP7. DEGs are sorted in descending order starting from the highest Log2 fold-change. Enrichment profiles represent average binding from RPKM normalized alignment data.

Intersection of DamIT Data with Published Data

We next compared our DamIT DNA Binding data to published datasets. Para et al. had not performed DamID on *bZIP1*, but did perform TARGET and ChIP-seq (Para et al., 2014). Of

1,308 genes regulated by *bZIP1* from Para et al, 845 were bound by *bZIP1* in this DamIT experiment, giving a binding rate of 65% (Fig. 10a). Comparing DNA-binding data of *bZIP1* from this experiment to the genes regulated in the Para et al. TARGET experiment shows an enrichment of binding to *bZIP1* DEGs which were not bound by *bZIP1* in ChIP-seq (Fig. 10b Class III). Alvarez et al. has already published TARGET data on *NLP7*, showing which genes it regulates. Of the 492 *NLP7* DEGs collected by Alvarez et al., 353 were bound by *NLP7* in this DamIT experiment, giving a binding rate of 72% (Fig. 10c).

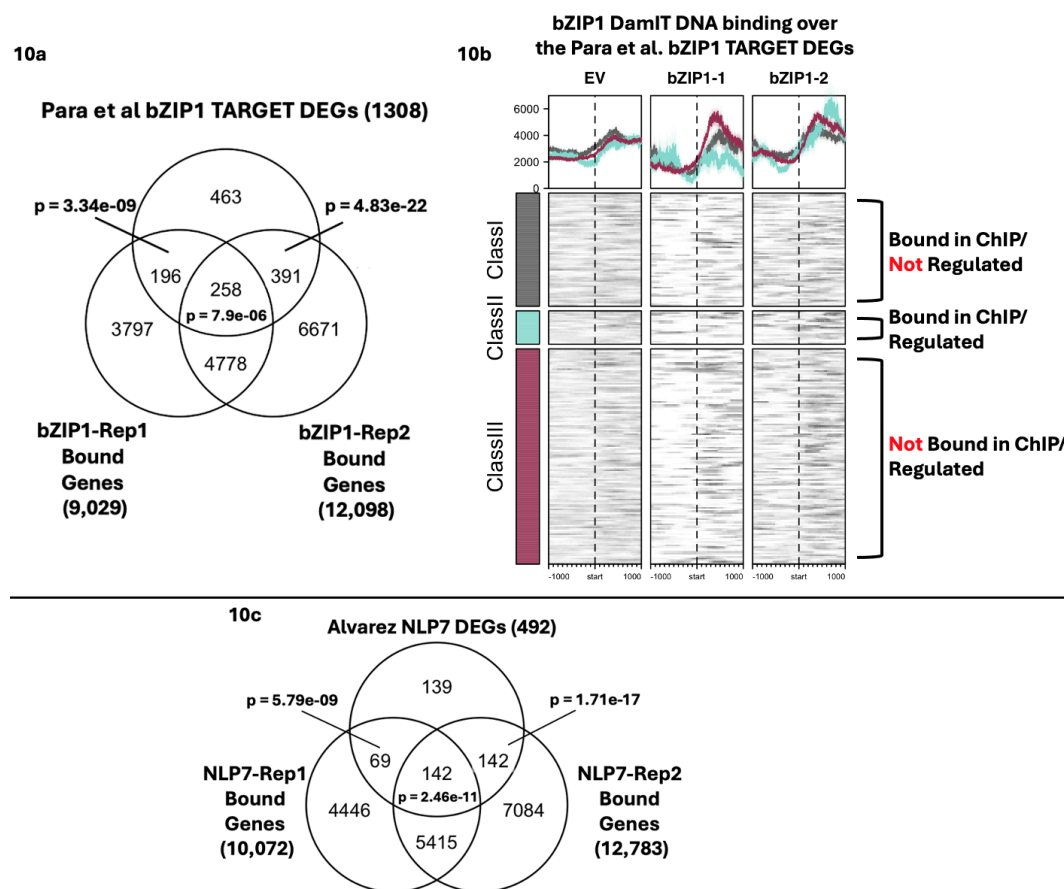


Figure 10. Intersection of data from this experiment to data from Alvarez et al. and Para et al. 10a) Venn diagram of DEGs from Para et al and genes bound by *bZIP1* from this DamIT experiment. 10b) Imposing DNA binding profiles from this experiment over DEGs collected by Para et al. Higher levels of DNA binding peaks are demonstrated by darker regions within DEG heatmaps relative to TSS. DEGs are organized by Class I (bound in ChIP-seq and not regulated in TARGET), Class II (bound in ChIP-seq and regulated in TARGET), and Class III (not bound in ChIP-seq and regulated in TARGET). Class III has the most regulated genes, and shows binding enrichment with data from this DamIT experiment. 10c) Venn diagram of DEGs from Alvarez et al. and genes bound by *NLP7* from this DamIT experiment.

Discussion

NLP7 DamIT Data is Reproducible and Offers Proof of DamIT Efficacy

The purpose of including *NLP7* in this experiment is to get new DamID data that can be compared to published *NLP7* data. Alvarez et al. gathered 492 *NLP7* DEGs with TARGET, and this experiment uncovered 461 *NLP7* DEGs with DamIT (Fig. 10c). Of the 492 DEGs from Alvarez et al., 72% were bound by *NLP7* with DamIT; of the 461 DEGs from this DamIT experiment, 65% were bound by *NLP7* (Fig. 10c, 9e). We uncovered a consistent binding rate compared to the Alvarez et al. experiment, both far surpassing the 15% regulated genes being bound by *NLP7* in ChIP-seq (Fig. 1a). The new *NLP7* DamIT DNA binding data serves as independent confirmation of the ability of DamID technology to capture transient TF-DNA binding.

It is important to note that DAM-GR-EV control binding signal is correlated with DEG log2 fold-change (Fig. 9f). Thus, as DEGs get more upregulated relative to the DAM-GR-EV control, more DAM-GR-EV DNA binding signal is detected centered at the gene TSS. DAM-GR fusion proteins are known to preferentially methylate open chromatin (Aughey et al., 2018). Therefore, with proper sequencing depth, DamIT may simultaneously provide TF-DNA binding, TF-Target gene regulation, and chromatin accessibility all from the same experiment. Further validation is needed to confirm this. While DAM-GR-EV DNA binding correlates with DEG Log2 fold-change, Dam-GR-NLP7 DNA binding is more dispersed among the *NLP7* DEGs, which is more indicative of TF-DNA binding.

GBF2 May Not be able to Access Closed Chromatin On Its Own

GBF2 was included in this experiment as a novel “hit and run” candidate TF to test DamIT on. *GBF2* showed a 50% binding rate to regulated targets (Song et al.) and may or may

not operate as a “hit and run” TF (Fig. 1a). Of all binding profiles examined in this experiment, both *GBF2* replicates showed the most similar DNA binding signal to the EV control (Fig. 9b). While DAM-GR-EV binding signal again correlates to DEG Log2 fold-change, the DAM-GR-*GBF2* binding is much more dispersed throughout the genes. However, the DAM-GR-*GBF2* strength of binding (average binding signal) is not greater than that of the DAM-GR-EV. This indicates that *GBF2* binds the DNA of its TARGET genes but may be constrained by chromatin accessibility. It may be the case that *GBF2* is simply not a pioneer TF that can access closed chromatin the way that other TFs might.

DamIT Validates bZIP1 as a Hit and Run TF

bZIP1 was included in this experiment to further validate the transient interactions it has with its targets. Para et al. showed that *bZIP1* operates transiently because the majority of the genes it regulates (using 4tU TARGET) were not bound in ChIP-seq. The rate of *bZIP1* regulated genes being bound with ChIP-seq was 10% (Fig. 1a). DamIT detected a much higher rate of bound/regulated genes at 54% (Fig. 9c). The same DamIT-DNA binding data was intersected with the *bZIP1* DEGs collected from Para et al. and showed an interaction rate of 65% (Fig. 10a). In both cases DamIT is able to capture more DNA interactions with regulated genes compared to ChIP-seq. While we predicted that DamIT would capture a higher rate of bound/regulated genes from the same cells, the lower rate in DamIT may be caused by the low power to call DEGs with only two replicates per sample. Para et al. likely uncovered higher confidence DEGs with more replicates. Repeating this experiment with more treatment replicates may improve the rate of bound/regulated DEGs.

Validation of *bZIP1* as a “hit and run” TF is further demonstrated when measuring the DamIT *bZIP1* DNA binding signal over the Para et al DEGs (Fig 10b). Para et al. assigned

classes to *bZIP1* DEGs where class 3 represents *bZIP1* TARGET DEGs not bound by *bZIP1* in ChIP-seq. We detected robust enrichment of DAM-GR-*bZIP1* DNA binding among the class 3 DEGs relative to the DAM-GR-EV control (Fig 10b). This demonstrates that transient gene interactions are measurable with DamIT, and that *bZIP1* likely regulates genes through a “hit and run” process. Additionally, the pioneering activity of *bZIP1* is supported by the strong binding of DAM-GR-*bZIP1* to genes with relatively low chromatin accessibility in the DAM-GR-EV sample (Fig 9d). Together, DamIT on *bZIP1* revealed novel transient DNA binding of *bZIP1* to its regulated genes, and that *bZIP1* can bind genes with relatively low chromatin accessibility.

Conclusion

The data presented indicate that DamIT is a viable method for measuring TF-target gene interactions that has higher sensitivity than ChIP-seq. DamIT was able to detect 3- to 7-times more interactions between TFs and regulated genes compared to ChIP-seq. DamIT validated DNA binding at 65% of *NLP7* regulated genes and 54% of *bZIP1* regulated genes, whereas ChIP-seq only detected 20% and 10% respectively (Fig. 9c, 9e, 1a). DamIT only detected interaction between *GBF2* and 32% of the *GBF2*-regulated genes (compared to 50% from ChIP-seq), meaning *GBF2* may not display as much transient activity compared to *NLP7* or *bZIP1* (Fig. 9a). Repeating the experiment with more treatment replicates may improve this result.

Intersecting data from this experiment with data from Alvarez et al. and Para et al. demonstrate the consistency of this assay as well (Fig. 10a, 10c). A similar number of *NLP7* DEGs were identified with DamIT from this experiment compared to *NLP7* TARGET in Alvarez et al. 74% of the Alvarez et al. *NLP7* DEGs were bound in the *NLP7* DamIT dataset. Since *NLP7* was included as a positive control, our results indicate successful use of the DamIT protocol. The

Para et al. data also showed consistent results, since a significant portion (65%) of their *bZIP1* TARGET DEG dataset was bound by *bZIP1* in this DamIT experiment. DamIT was able to validate direct binding of these TFs to their targets at a higher rate than ChIP-seq. Thus DamIT is a novel method to measure TF-DNA binding, TF gene regulation, and possibly chromatin accessibility in the same experiment. DamIT will enable the construction of more accurate and comprehensive GRNs that incorporate transient TF-gene interactions, and is a novel method for validating the “hit and run” model of gene regulation.

References

- Alvarez JM, Schinke AL, Brooks MD, Pasquino A, Leonelli L, Varala K, Safi A, Krouk G, Krapp A, Coruzzi GM. Transient genome-wide interactions of the master transcription factor NLP7 initiate a rapid nitrogen-response cascade. *Nat Commun.* 2020 Mar 2;11(1):1157. doi: 10.1038/s41467-020-14979-6. PMID: 32123177; PMCID: PMC7052136.
- Alvarez, José M., et al. “Damid-Seq: A genome-wide DNA methylation method that captures both transient and stable TF-DNA interactions in plant cells.” *Methods in Molecular Biology*, 2023, pp. 87–107, https://doi.org/10.1007/978-1-0716-3354-0_7.
- Aughey GN, Cheetham SW, Southall TD. DamID as a versatile tool for understanding gene regulation. *Development.* 2019 Mar 15;146(6):dev173666. doi: 10.1242/dev.173666. PMID: 30877125; PMCID: PMC6451315.
- Aughey, Gabriel N, et al. “Catada reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo.” *eLife*, vol. 7, 26 Feb. 2018, <https://doi.org/10.7554/elife.32341>.
- Bargmann BO, Marshall-Colon A, Efroni I, Ruffel S, Birnbaum KD, Coruzzi GM, Krouk G. TARGET: a transient transformation system for genome-wide transcription factor target discovery. *Mol Plant.* 2013 May;6(3):978-80. doi: 10.1093/mp/sst010. Epub 2013 Jan 18. PMID: 23335732; PMCID: PMC3660954.
- Brooks, Matthew, et al. The Target System: Rapid Identification of Direct Targets of Transcription Factors by Gene Regulation in Plant Cells V2, MIMB. 9 May 2022, <https://doi.org/10.17504/protocols.io.rm7vz32yrgx1/v2>.
- Brooks MD, Juang CL, Katari MS, Alvarez JM, Pasquino A, Shih HJ, Huang J, Shanks C, Cirrone J, Coruzzi GM. ConnecTF: A platform to integrate transcription factor-gene interactions and validate regulatory networks. *Plant Physiol.* 2021 Feb 25;185(1):49-66. doi: 10.1093/plphys/kiaa012. PMID: 33631799; PMCID: PMC8133578.

- Doidy J, Li Y, Neymotin B, Edwards MB, Varala K, Gresham D, Coruzzi GM. "Hit-and-Run" transcription: de novo transcription initiated by a transient bZIP1 "hit" persists after the "run". *BMC Genomics*. 2016 Feb 3;17:92. doi: 10.1186/s12864-016-2410-2. PMID: 26843062; PMCID: PMC4738784.
- Markodimitraki CM, Rang FJ, Rooijers K, de Vries SS, Chialastri A, de Luca KL, Lochs SJA, Mooijman D, Dey SS, Kind J. Simultaneous quantification of protein-DNA interactions and transcriptomes in single cells with scDam&T-seq. *Nat Protoc*. 2020 Jun;15(6):1922-1953. doi: 10.1038/s41596-020-0314-8. Epub 2020 Apr 29. PMID: 32350457; PMCID: PMC7779467.
- O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*. 2016 May 19;165(5):1280-1292. doi: 10.1016/j.cell.2016.04.038. Erratum in: *Cell*. 2016 Sep 8;166(6):1598. PMID: 27203113; PMCID: PMC4907330.
- Orian A, Abed M, Kenyagin-Karsenti D, Boico O. DamID: a methylation-based chromatin profiling approach. *Methods Mol Biol*. 2009;567:155-69. doi: 10.1007/978-1-60327-414-2_11. PMID: 19588092.
- Para A, Li Y, Marshall-Colón A, Varala K, Francoeur NJ, Moran TM, Edwards MB, Hackley C, Bargmann BO, Birnbaum KD, McCombie WR, Krouk G, Coruzzi GM. Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in Arabidopsis. *Proc Natl Acad Sci U S A*. 2014 Jul 15;111(28):10371-6. doi: 10.1073/pnas.1404657111. Epub 2014 Jun 23. PMID: 24958886; PMCID: PMC4104873.
- Song L, Huang SC, Wise A, Castanon R, Nery JR, Chen H, Watanabe M, Thomas J, Bar-Joseph Z, Ecker JR. A transcription factor hierarchy defines an environmental stress response network. *Science*. 2016 Nov 4;354(6312):aag1550. doi: 10.1126/science.aag1550. PMID: 27811239; PMCID: PMC5217750.
- Verma S, Attuluri VPS, Robert HS. Transcriptional control of Arabidopsis seed development. *Planta*. 2022 Mar 23;255(4):90. doi: 10.1007/s00425-022-03870-x. PMID: 35318532; PMCID: PMC8940821.