The handle http://hdl.handle.net/1887/38624 holds various files of this Leiden University dissertation.

**Author:** Tol, Niels van  
**Title:** Phenotypic engineering of photosynthesis related traits in Arabidopsis thaliana using genome interrogation  
**Issue Date:** 2016-03-24
Chapter 2

Artificial transcription factor-mediated regulation of gene expression

Niels van Tol and Bert J. van der Zaal

Abstract

The transcriptional regulation of endogenous genes with artificial transcription factors (TFs) can offer new tools for plant biotechnology. Three systems are available for mediating site-specific DNA recognition of artificial TFs: those based on zinc fingers, TALEs, and on the CRISPR/Cas9 technology. Artificial TFs require an effector domain that controls the frequency of transcription initiation at endogenous target genes. These effector domains can be transcriptional activators or repressors, but can also have enzymatic activities involved in chromatin remodeling or epigenetic regulation. Artificial TFs are able to regulate gene expression in trans, thus allowing them to evoke dominant mutant phenotypes. Large scale changes in transcriptional activity are induced when the DNA binding domain is deliberately designed to have lower binding specificity. This technique, known as genome interrogation, is a powerful tool for generating novel mutant phenotypes. Genome interrogation has clear mechanistic and practical advantages over activation tagging, which is the technique most closely resembling it. Most notably, genome interrogation can lead to the discovery of mutant phenotypes that are unlikely to be found when using more conventional single gene-based approaches.
1. Introduction

The phenotype of any given organism results from a complex interplay between its genome and the mechanisms that led to the expression of its genes. This interplay is characterized by intricate feedback loops that generate the essential robustness of the phenotype. The feedback loops must also allow for flexibility when endogenous or exogenous stimuli demand for specific phenotypic adaptations. The metaphor of Waddington’s epigenetic landscape [1], a model describing the different developmental paths that an embryonic cell can take towards differentiation, is still very much relevant to modern developmental genetics. The stability of gene expression patterns controlled by established epigenetic cues enables cells to withstand most of the random biotic and abiotic noise. However, when a key determinant is able to induce a crucial epigenetic change, cells and organisms might be forced into a different state or developmental program. This epigenetic view of the regulation of gene expression complements the view where genetic variation is the source of phenotypic variation; genetic variation is futile when not expressed. The phenotype of a cell can be regarded as being the product of the epigenetic landscape, genome wide transcription patterns and variation at the sequence level at any given stage of development. Fundamental research on these processes has allowed us to gather knowledge on which genes or sets of genes are involved in phenotypes of interest. In this review, we address several means of placing phenotypes under artificial control by employing artificial transcription factors (TFs) as tools for regulating the expression of endogenous genes in plants.

1.1. Regulation of gene expression

The short sequence upstream of the transcription start site that in eukaryotic genes contains the binding sites for general transcription factors and RNA polymerase II [2] is often referred to as the “minimal promoter” of a gene. More gene-specific regulatory sequences can be found in the DNA sequence upstream of this minimal promoter. It has become common practice in the field of plant molecular biology to designate a rather arbitrary DNA fragment of one to a few kilobase (kb) pairs long and located upstream of the translational start site as the “promoter” of a gene. Plant molecular biologists are usually aware of the fact that many more regulatory sequences exist at greater distances at both the 5’ and the 3’ ends of a gene as well as within its coding sequence that contribute to the precise level of gene expression. Short statements regarding “promoter activity” usually refer to the contribution of at most a few kb of upstream DNA sequence on to the regulation of transcription levels. Within the context of artificial TF-mediated regulation of gene expression, it would be better to employ the term “gene control region” rather than “promoter”. This control region is usually defined as the portion of a eukaryotic gene containing the core promoter as well as any other regulatory sequences that control or influence transcription of that gene. Within the control
region, the eukaryotic core promoter is defined as the region that can be bound by the general transcription factors required for RNA Polymerase II-dependent transcription initiation at the transcription start site, thus equaling the “minimal promoter” mentioned above. Apart from the core promoter, the control region contains enhancer and silencer sequences [3]. These regulatory sequences are potential docking sites for more specific transcription factors that can affect the number of transcription starts at the core promoter per unit of time. The regulatory sequences can be present in cis of the start site, within a distance of a few kb from the core promoter, or be located at much larger genomic distances where the term “in cis” gradually becomes practically irrelevant. In the latter cases, these regulatory elements are absent from the relatively short PCR-generated DNA sequences taken for the “promoter” in more pragmatic approaches. When discussing the effects of artificial TFs, it is much more appropriate to acknowledge all interactions that are formed within the larger gene control region.

The conserved Mediator complex is also required for successful initiation of RNA Polymerase II-dependent transcription at core promoters in eukaryotes. The Mediator complex functions as a highly complex co-activator of transcription, interacting with the protein domains of RNA polymerase II holoenzyme and general transcription factors. Mediator also interacts with the more specific transcription factors binding to sequences outside of the core promoter. Without the stimulatory contribution of the latter proteins, RNA polymerase II is unable to initiate gene transcription [4,5]. The Mediator complex can thus be thought of as a platform for integrating or relaying signals that can stimulate the initiation of transcription in the regulation of gene expression [4]. However, once the factors conducive for transcription are present and the expression of genes has been switched on in a stable manner, one could imagine that further information and activity is needed to subsequently decrease transcriptional activity or even switch off the expressed genes when this would be required, such as during developmental processes. Accumulating evidence connects the Mediator complex with epigenetic regulation, recruiting factors and enzymes that lead to the deposition of epigenetic molecular markers associated with gene silencing [6,7].

1.2. Chimeric transcription factors
Transcription factors contain a DNA binding domain and a domain that is able to affect transcriptional regulation. Such “effector” regulatory domains increase or decrease the number of transcriptional starts of a gene when bound to DNA at an appropriate position in the gene control region. The effector domain can be envisaged as directly interacting with one or more of the general transcription factors and/or RNA polymerase subunits at the transcription start site or indirectly by recruiting proteins that make these essential contacts.
The use of these effector domains has been reported in connection with natural transcription factors. Plant transcription factors equipped with signature DNA binding domains were fused to a small C-terminal peptide domain that inhibits gene expression [8,9]. This strategy is aimed at turning natural transcriptional regulators into dominant repressors of gene expression that specifically bind to the gene control region of their natural target genes. Changes in the phenotype are readily observed due to loss-of-function mutations resulting from the reduced expression of the genes that are under control of the transcription factors being experimentally manipulated. This strategy is termed Chimeric REpressor gene Silencing Technology (CRES-T) [10]. A system involving fusions with activating effector domains instead of repressing domains could also be envisaged, where an enhancing transcription factor would then affect transcription at its natural target loci in a positive manner.

In the CRES-T technology, as well as in its possible derivatives, DNA binding properties of natural TFs form the basis for the mode of action of these chimeric proteins. The artificial TFs discussed below allow for recognition of any target site of choice to affect the transcriptional activity of genes of interest at the control regions of their normal genomic position. However it is necessary to address relevant target sites within the control region to specifically regulate the expression of endogenous genes of interest. A technique that employs naturally occurring DNA binding domains is hardly an option. Even if a binding site for a known transcription factor would be present, such sites are usually of low complexity and occur at many positions within the genome. This could possibly affect the transcriptional regulation of a host of genes that are normally under control of this particular transcription factor. Custom made site-specific DNA binding domains are required to address unique sites within the genome. The molecular details of systems that allow for site-specific protein-DNA recognition have become understood during the last 15 years to such an extent that it has become possible to design and produce sequence specific DNA binding domains.

2. Artificial DNA binding domains

2.1. Zinc finger domains

From the late 1990s onwards, the DNA recognition code for Cys2His2 (C2H2)-type zinc finger domains has largely been elucidated. Each zinc finger domain (approximately 30 amino acids long) interacts with a triplet of consecutive bases on one strand of the DNA through one amino acid residue just before its alpha helix, and two amino acids within its alpha helix [11]. A fourth contact is made with a base on the opposite strand [11]. Changes in the amino acid composition of the alpha helix change the DNA binding specificity [11].
For site-specific recognition, complex DNA binding domains are required. With each ZF module recognizing a triplet of base pairs, it can easily be calculated that one will need fusions of at least five or six ZF domains to define a cognate contiguous 15 or 18 base pair sequence that is unique within a complex genome, encompassing a billion base pairs or more. The design of complex polydactyl zinc fingers and the consequences for their association or dissociation constants ($K_a$ and $K_d$, respectively) have been studied [12-15]. The dissociation constants of the contiguous target sequences of six-fingered polydactyl zinc fingers are mostly in the low nanomolar range [13,16,17]. For the shorter three-fingered domains, the dissociation constants were in the range of 10 - 80 nM [17,18], but methods for determining their $K_a/K_d$ values have never been standardized. Thus, the biochemical affinity of a polydactyl zinc finger domain for its target sequence increases with extra zinc finger-DNA triplet interactions.

Zinc finger-DNA interactions are not fully specific [19]. Zinc fingers cannot be expected to bind to their cognate DNA triplets with equal affinity. For several of the 64 possible DNA triplets, specific zinc finger domains are unlikely to become available as there seem to be insufficient structural means to establish the required protein-DNA contacts [20,21]. Moreover, the published triplet specificities of zinc fingers have been based on selection criteria where the particular zinc finger of interest was embedded within a larger structural context, such as being the middle one within a three finger framework aimed at binding a 9 bp target sequence. Altogether, established zinc finger-DNA recognition lexicons should be regarded as reflecting a preferred interaction rather than an exclusive one.

When aiming for highly specific, high affinity DNA binding, a positive bias towards selection of ZFs recognizing 5’-GNN-3’ triplets (with N being any of the four DNA bases) [22] might be advisable, since the GNN-based ZF code provides for robust interactions with nearly all of its constituents [19]. A protocol such as Oligomerized Pool ENgineering (OPEN) provides for many details as how to generate arrays of zinc fingers [23]. Although the triplet recognition code might suggest otherwise, zinc finger-DNA interactions involve an extra contact with a fourth base in the opposite strand of DNA, as mentioned above. This phenomenon, known as “target site overlap”, makes certain consecutive zinc finger combinations unfavorable for high affinity interactions. Despite all available data and having taken precautions regarding target site overlap, the precise affinity and specificity of polydactyl zinc finger domains for their intended target sites remains hard to predict [24]. The cognate 9 bp target sequence of three finger domains is likely to be more of a consensus sequence rather than an actual target sequence that is partly shared by all 9 bp sequences that can interact with the three finger module.

For the more complex six finger domains, it is tempting to assume that in vivo interactions only involve 18 bp contacts. However, when just looking at contiguous polydactyl zinc finger
domains and their target sites, each six finger domain provides for 2 five finger interactions, 3 four finger interactions, 4 three finger interactions, and 5 two finger interactions with 15, 12, 9, and 6 bp of DNA, respectively. The more complex interactions, involving three finger domains and more, could very well result in \textit{in vivo} off-target interactions when expressed in target cells, as could combinations of two finger-two finger or two finger-three finger interactions within a single six finger domain. It seems unlikely that one or more mismatches with unrecognized DNA triplets will completely abolish DNA binding to subsets of an 18 bp target site, unless such mismatches will severely distort the binding at neighboring triplets as well. In any case, whether non-specific interactions occur \textit{in vivo} depends on the affinity for the shorter and/or interrupted target sites as well as on the concentration of the polydactyl zinc finger proteins within a cell. Therefore, after assembly of a zinc finger-based DNA binding domain it remains to be seen whether it is able to bind at its intended genomic target site within the organism of interest.

Chromatin structure at or around the intended binding site of polydactyl zinc fingers is very likely to have a role in their interaction with DNA. The DNaseI sensitivity of chromatin embedded target sites positively correlates with the possibility to affect gene regulation by means of zinc finger artificial TFs [25,26], but correlations are not guarantees. The position of the target site within the gene control region is also crucial for any \textit{in trans} regulation of target gene expression. Choosing target sites close to the core promoter, about 50-150 bp upstream of the transcription start site of the gene of interest, offers the best chances for artificial modulation of transcription. Still, the gene control region usually contains one or more binding sites for natural transcriptional regulators that are able to contribute to or repress the assembly of the general transcriptional machinery and the transcription initiation complex. Zinc finger artificial TFs targeting this region might therefore compete with or even replace the normal regulatory factors. The most practical application of the zinc finger technology is construction of site-specific zinc finger nucleases (ZFNs) [27].

2.2. TALEs

Some 10 years after the elucidation of the greater part of the DNA recognition code of ZFs, the DNA recognition code of transcription activator-like effectors (TALEs) became available [28,29]. TALEs are produced by the bacterial pathogen \textit{Xanthomonas} and are injected into plant cells, where they bind to the regulatory regions of specific plant genes, activating their transcription [30]. The core DNA binding domains of TALEs consist of repeats of modules of 34 amino acids that each bind to 1 bp of DNA. Further insights and refinements of the TALE technology have led to the establishment of highly efficient protocols for the construction of designed TALE domains (dTALEs) [31,32]. As goes for ZF domains, TALE domains have been predominantly used for the construction of site-specific nucleases (TALENs). The “Golden
Gate” assembly allows for the relatively easy construction of dTALE domains with 15-31 repeats, thus in principle recognizing 15-31 bp targets [31]. The TALE technology presently seems to offer better options for making DNA binding domains with sufficient specificity, especially to investigators who do not have access to company-owned polydactyl zinc finger data.

Quantitative assessments of dTALE affinities for DNA are rare, in contrast to polydactyl zinc fingers. The ability to predict the in vivo performance of dTALEs is not as straightforward as might be expected. In an extensive analysis, the recognition of DNA by dTALEs was shown to still be surprisingly complex [33]. dTALEs constructed with 19 repeats had apparent $K_r$’s for their cognate 19 bp target sites ranging from (sub)nanomolar to low micromolar values, depending on the choice of alternative TALE repeats for the recognition of particular nucleotides. Furthermore, the N-terminal repeats contributed more to the DNA binding affinity than the C-terminal ones. An issue that might raise concern for some applications is that a DNA target sequence only providing for an interaction with 10 out of 19 bases could still interact with low nanomolar $K_r$. Such a phenomenon is likely to result in many off-target interactions within a complex genome [33].

2.3. CRISPR/Cas9
The most recently discovered DNA binding domains were found in the CRISPR (Clustered regularly Interspaced Short Palindromic Repeats)/Cas9 system, which is a defense system employed by a range of bacterial species aimed at the degradation of viral DNA [34]. In this system, specific guide RNAs direct the Cas9 endonuclease protein to their target DNA sequence, leading to subsequent cleavage of that sequence. Manipulation of this naturally occurring system demonstrated that the guide RNAs base pair with complementary DNA sequences at their 5’ end, and interact with Cas9 through their 3’ end [35]. The length of the homology-searching 5’ RNA sequence is usually about 20 bases, but shorter sequences have recently been reported to have less off-target effects [36]. This might be due to the fact that relatively long RNA sequence can allow for more than one high affinity RNA-DNA interaction, while shorter ones can not. The CRISPR/Cas9 system has made an extremely rapid entry into biotechnology, predominantly for making site-specific double strand breaks and thereby targeted mutations within a genome, analogous to zinc finger nuclease and TALEN technology. For such genome engineering purposes, plasmids with genes encoding the guide RNA as well as Cas9 have to be introduced into target cells. The templates for guide RNAs can easily be edited by inserting oligonucleotide sequences complementary to the cognate target DNA region. Thus, the use of the CRISPR/Cas9 system for the induction of site-specific double strand breaks has gained tremendous attention during the last two years [37-39].
Derivatives of the Cas9 protein lacking nuclease activity (dCas9) can also be made amendable for generating artificial TFs. Induction of gene expression was achieved via dCas9 fusions to the powerful transcriptional activator VP64 [40,41]. Specific repression was observed by targeting just a dCas9 protein to potentially regulatory target sites [42]. The elegant and methodologically simple RNA-based targeting strategy of the CRISPR/Cas9 system might make CRISPR/Cas9-based ATFs very attractive as tools for the regulation of endogenous gene expression by offering further prospects for simultaneously controlling the regulation of sets of genes. This would require just the introduction of relatively simple guide RNA expression cassettes and one expression cassette encoding the dCas9 protein. It can be expected that further modifications of the dCas9 protein can provide better scaffolds for translational fusions to different types of effector domains.

3. Effector domains

We will not discuss all possible details about the types of effector domains that can be part of artificial TFs. Generally speaking, these domains can either have stimulatory or inhibitory effects on transcriptional regulation. Within the conceptual framework of gene regulation as mentioned above, both types of effects can be regarded as a consequence of the contribution of effector domain-mediated signals to the assembly of a transcription initiation complex at the gene control region. Effector domains can be chosen for their enzymatic activities, for instance the activities of proteins involved in establishing or removing epigenetic markings or in chromatin remodeling. Changes in epigenetic marks and/or chromatin structure can result in the recruitment of different kinds of transcriptional modulators to the gene control region [43,44]. Effector domains should be able to recruit accessory factors to a particular gene of interest if not active by themselves. Effectors such as (multimers of) the Herpes simplex virus-derived VP16 domain are thought to open chromatin structure by recruiting histone acetylating enzymes and enzymes with SWI/SNF chromatin remodeling activity [45,46]. The class of EAR (-like) repression domains offers the opportunity for gene silencing in plants. The classical LxLxL type EAR domains [47], as well as the more recently identified domain with an R/KLFGV motif [48], act by recruiting co-repressors such as TOPLESS (TPL) or the four TPL-related proteins. Expression of these proteins in Arabidopsis is largely constitutive, but enhanced during floral transition and flowering [49,50].

The choice of an effector domain largely depends on the investigator's imagination. The increasing knowledge about epigenetics and transcriptional regulation renders it easier to make an informed choice for a particular effector domain. If any further considerations need to be mentioned, they are rather generic and in fact rather obvious. When attempting
to rewrite epigenetic codes, the effector domain should preferably be active as a monomer, and possess the required enzymatic activity without being hampered by translational fusion with a DNA binding domain. When enzymatic partners need to be recruited, they and their substrates should be available in the cell types of interest. The enzymatic reactions often depend on preceding steps, such as di- and tri-methylation of lysine residue 36 in histone protein 3 of *Arabidopsis* requiring previous monomethylation [51]. Finally, the activity of all site-specific artificial TFs that harbor effector domains is only relevant at the intended target site. Interactions with partners at other cellular positions might lead to a phenomenon known as ‘squelching’, which is the sequestration of partners prohibiting the intended effect at the site of interest.

4. Applications of artificial transcription factors

The scope of present and future applications is in principle the same for any type of artificial TF, whether it is zinc finger-, TALE-, or CRISPR/Cas9-based. Some issues regarding the ease of construction and possible off-target effects have briefly been mentioned above and have recently been reviewed [27]. It seems to be too early to make the clear statement that one type of DNA binding domain, whether it is protein or RNA-based, is to be preferred over any of the others. Several studies have demonstrated the *in vivo* activity of zinc finger-based artificial TFs [52-54]. Designing new zinc finger-based domains for specific target site recognition currently has its problems, certainly for academic researchers with limited funds for outsourcing the synthesis of polydactyl zinc fingers to parties that have exclusive access to company-protected information. If more information will become publicly available and/or the prices of outsourcing drop, the zinc finger technology might very well regain the key position that it had prior to the rapid entry of dTALE domains into this field. The original guidelines for the design of dTALEs might have been an oversimplification [33], but TALE-based artificial TFs have also been the subject of several studies [55-57]. The current pitfalls and problems in dTALE design might be accounted for in the future. The possibilities are very promising for CRISPR/Cas9-based ATFs. At the time of writing this review however, only very few papers have demonstrated their *in vivo* activity [40,41]. A simplified overview of the artificial TFs based on the different DNA binding domain technologies and their mode of action is presented in Fig. 1.
Artificial transcription factor-mediated regulation of gene expression

4.1. Regulation of the expression of specific endogenous genes

Dominant regulatory control of the expression of selected endogenous genes became possible with the emergence of zinc finger artificial TFs. The type of DNA binding domain used for the design of artificial TFs is a practical choice rather than a fundamental one. Thus, it is not surprising that the ideas that were put forward for zinc finger artificial TFs more than 10 years ago [58] also hold for dTALE- or CRIPR/Cas9-based artificial TFs. Several studies with zinc finger- and dTALE-based artificial TFs in the plant field are mentioned in a recent review about advanced genetic tools for plant biotechnology [59]. They all demonstrate the possibility of changing the expression pattern of a plant gene within its native context. Of course, as described above, there are various reasons why other attempts to do so could have failed. Lack of success might be due to the failure of generating a site-specific DNA binding
domain or to an inability of the respective DNA binding domain to access its intended target site. In a yeast model system, only the more complex five and six fingered artificial TFs were able to transactivate a glucose repressed reporter gene [15]. The position of the target site within the gene control region is very important. Nevertheless, with the choice of tools presently available, there is a reasonably high probability that it will be possible to generate an artificial TF influencing the expression of a gene of interest.

4.2. Genome interrogation

Instead of aiming at gene-specific regulation via complex DNA binding domains intended to interact with high affinity at a single cognate target site, artificial TFs with more simple DNA binding domains also allow for interesting applications. In a seminal paper, novel mutant phenotypes could be evoked in yeast cells as well as in mammalian cells when such cells expressed members of a library of about 100,000 zinc finger artificial transcription factors with different three finger and four finger domains [60]. The gist of this method was that single gene constructs encoding a particular three finger or four finger artificial TF were introduced and expressed in transformed cells, as well in their clonal offspring. Three and four finger domains, might allow for multiple genome-wide interactions due to their relatively short 9 bp or 12 bp cognate target sequences. For instance, each randomly chosen 9 bp sequence will on average occur about once every 130,000 bp of double stranded DNA, and a 12 bp sequence about once per 8.3 million bp. Supposing that a target sequence is accessible and that it is located at a suitable position within a gene control region, binding of the zinc finger artificial TF is likely to affect the expression pattern of the gene. In this manner, a relatively simple zinc finger artificial TF can act as an ectopic master switch that modulates the regulation of a multitude of target genes. These primary effects probably result in a plethora of changes in transcription patterns of genes further downstream. These changes in the transcriptome will obviously be accompanied by changes in the proteome and the metabolome.

The ability to cause massive changes in gene expression by just expressing a single artificial TF with lower target site-specificity enables researchers to seek for novel phenotypes, possibly never observed when down regulating or overexpressing single genes. Mutagenesis by means of introducing a library of gene constructs each encoding a unique artificial TF with a rather simple DNA binding domain has been termed 'genome interrogation.' Since the target cells contain the same genome (except for different artificial TF-encoding transgenes), an overview of the potential of a genome is achieved in this way by expressing it differently and/or differentially.

Thus far, genome interrogation has only been accomplished using libraries of zinc finger artificial transcription factors. It is relatively easy to have essentially all members of a large
library of different artificial TF-encoding gene constructs represented in single celled organisms or in cell cultures [60-64]. When the libraries consist of tens of thousands or even millions of different artificial TFs, it is simply not realistic to study the full spectrum of mutant phenotypes that are induced. However, when the introduction of artificial TFs is combined with a single or multiple rounds of selective screening for phenotypes of interest, particular phenotypes might easily be found, even when the causal artificial TFs are relatively rare. Successful isolation of mutants with a phenotype of interest and the subsequent delivery of the proof of principle that the respective isolated zinc finger artificial TF-encoding construct induces this phenotype has led to the notion that genome interrogation is indeed a powerful tool for generating novel phenotypes [60-65]. Interestingly, genome interrogation has previously yielded bacterial phenotypes that could not be mimicked by overexpression of single or two combined candidate genes that were identified as the most likely target genes of the respective zinc finger artificial TF. This can be regarded as a strong indication that the causal artificial TF was able to induce the phenotype by affecting the expression of a combination of genes, precisely as the principle of genome interrogation predicts [62]. Obviously, genome interrogation is not restricted to the zinc finger technology. Both dTALE- and CRISPR/Cas9-based technology should also allow for the generation of libraries of artificial TFs that are suitable for genome interrogation. These alternative DNA binding domains seem not to have been applied in the context of genome interrogation. The first proofs of principle will probably be published soon or have been so at the time of publication of this review.

Applying genome interrogation at the level of multicellular organisms requires generating and culturing transgenic organisms in sufficient quantities to obtain meaningful experimental depth. The first proof of principle of genome interrogation in multicellular organisms was delivered for the model plant species Arabidopsis by using a relatively small collection of about 4000 GNN based three fingered artificial TF-encoding genes. In this study, a specific artificial TF inducing very high levels of somatic homologous DNA recombination [65] was identified. Further experimental evidence indicated that this three finger ATF acts as an ectopic master switch orchestrating the timely expression of a set of endogenous genes. This then leads to enhanced somatic recombination. The resulting enhancement is much greater than the one that is accomplished by the overexpression of each of the individual genes [66].

4.3. Comparing genome interrogation to activation tagging

Thus far, the use of genome interrogation for finding novel phenotypes of interest in plants has not yet been reported by other groups. This could be due to various reasons, the most prosaic one being that the method is relatively new and unknown. However, more fundamental questions might relate to its performance and utility in comparison to other methods. Since genome interrogation involves raising transgenic individuals with dominant phenotypes, it
is very different from chemical mutagenesis or radiation mutagenesis which yields mostly recessive mutations in a dose-dependent manner. Moreover, these more traditional mutagenic techniques are based on introducing genetic variation at the DNA sequence level. The phenotypic changes induced by genome interrogation are in principle due to changes in the expression patterns of particular sets of genes. A comparison to activation tagging, however, is appropriate. This method and its derivatives are also based on the introduction of foreign DNA and are also aimed at recovering dominant mutations [67,68]. The essential difference between the two techniques is that genome interrogation is based upon the *in trans* activity of an artificial TF-encoding gene (with its precise location being irrelevant), while activation tagging is based upon the integration of a transcriptional enhancer sequence that can only affect the expression of a gene nearby (*in cis*), mostly within a few kb of the insertion site [68].

It is important to have at least some idea about the size of the mutant population that is required for finding a particular phenotype of interest when considering large scale forward mutagenesis as a tool for genetics. A typical three finger domain might, on average, find about 1000 loci that contain its cognate 9 bp recognition sequence in Arabidopsis, with a haploid genome size of approximately 135 Mbp. Hence, screening a library of only a few thousand transgenic Arabidopsis plants can already be regarded as a potent brute force approach to discover genes or sets of genes where differential expression results in novel and dominant traits. About 750,000 nucleosomes and their cognate 180 bp DNA stretches exist in Arabidopsis when simply thinking of chromatin-embedded DNA as a string of nucleosome structures occurring every 180 bp of DNA. Suppose that the expression level of a particular three finger artificial TF and accessibility of its cognate 9 bp recognition sites are not limiting factors, then one out of each 750 nucleosomes might already be targeted by this 3F-ATF. Having a total of a few thousand different zinc finger artificial TFs represented in about 3000 transgenic plants would in theory be enough to probe every nucleosome structure on average about four times. A large collection of transgenic plants is required to find an *in cis* effect of activation tagging on a particular gene. Up to 100,000 insertion mutants would be required for Arabidopsis to be able to state that each endogenous gene has had a fair chance of being affected by the introduced activation tagging construct. The number of transgenic plants required increases in a linear fashion to perform a saturating screen for a given phenotype of interest in plant species with larger genomes. The number of genome interrogation mutant plants will stay the same; larger genomes contain correspondingly more binding sites for any given artificial TF that can all be targeted *in trans*, provided that the expression level of the artificial TF encoding transgene is sufficiently high. Therefore, screening a relatively small mutant population enables the discovery of novel dominant phenotypes. However, this very strength of genome interrogation could also be considered a weakness when one is interested in elucidating the gene expression patterns that are induced by the artificial TF
and induce the phenotype of interest. Each ATF is a potential trigger of changes in genome wide transcription patterns. Each of the changes in primary target gene expression can have their subsequent pleiotropic effects, apart from the primary effects at the genomic loci having an accessible recognition site for any given artificial TF. Only after having gathered sufficient experimental evidence that reintroduction of a particular artificial TF indeed induces the phenotype, it can be concluded that the particular artificial TF is the master switch triggering all of these primary and pleotropic effects. However, finding the initial artificial TF-induced changes among the plethora of changes that abide in plants exhibiting a phenotype of interest might seem like an impossible task, especially when considering that any given artificial TF might bind at hundreds of primary recognition sites. In that respect, activation tagging has a clear advantage over genome interrogation. Simply identifying the genomic locus of integration of the activation tagging construct construct, for instance by TAIL-PCR [69], is in fact a direct lead to the primary candidate gene(s) in the vicinity of the insertion site. This is provided that the genome is adequately annotated. Linking a single gene with a phenotype of interest can then proceed by more or less standard protocols, as it will just require the rather straightforward experimental proof of causality that the ectopic overexpression of the candidate gene also induces the phenotype of interest. An overview of this comparison between genome interrogation and activation tagging is presented in Fig. 2.

4.4. Practical considerations for the application of genome interrogation

Genome interrogation in plants has thus far only been performed with zinc finger artificial TFs [65,66]. When using three finger domains, it is to be expected that a considerable fraction of the zinc finger artificial TF library can only bind to DNA with low affinity. The potential target sites can at best only transiently be occupied by the artificial TF, particularly when the expression levels of the artificial TF-encoding transgene are low. The chances of a given target site being occupied at a given time will increase when expressing artificial TFs with low binding affinity at higher levels. However, one could wonder whether inherently weak DNA-binding affinities will allow for building a molecular platform affecting transcriptional regulation. Polydactyl zinc fingers must have a very high affinity for their target sites to access DNA in a region with an actively repressed chromatin state, and require at least 5F domains for transactivation of gene expression [13]. Surprisingly, access to DNA within typical heterochromatic regions was already possible using 3F domains [70].
Fig. 2 A comparison of genome interrogation (a) and activation tagging (b). For this example, expression cassettes are assumed to have been introduced as T-DNAs with a left and a right border sequence (LB and RB, respectively) via *Agrobacterium tumefaciens*-mediated transformation. The genome interrogation construct in (a) encodes an effector domain (ED) fused to an array of three different zinc fingers (ZFs). Expression of this gene is controlled by the promoter of the *RPS5a* gene from *Arabidopsis* and the *NOS* terminator sequence. The N-terminus of the protein contains a Nuclear Localization Signal (NLS) to send it to the nucleus. Note that –N and –C refer to the N and C terminal ends of the fusion proteins represented in the figure, but not necessarily to an optimal position for the effector domain. The activation tagging construct in (b) harbors just an element that is able to enhance transcription.

When good transformation or transfection protocols exist, such as for unicellular organisms or (plant) cell cultures, very large numbers of different artificial TFs can be expressed and assessed for their ability to induce a phenotype of interest. If needed, one or more additional rounds of selective screening can be employed to further enrich for the cells that harbor an artificial TF-encoding gene construct inducing such a phenotype. After their recovery, candidate gene constructs can then be tested individually for their phenotype-inducing properties. Obviously, generating very large numbers of genome interrogation mutants at the plant level can never be a practical option. However, as further described below, a wealth of
different phenotypes can already be observed in small collections of plants, representing at most a few thousand artificial TF-encoding constructs. In our opinion, the most attractive application of artificial TF-encoding genes therefore lies in their use as factors for the induction of novel, complex phenotypes of interest.

4.5. Preliminary results with genome interrogation in plants
We have established collections of seeds of the primary transformants obtained through floral dip transformation of Arabidopsis plants with libraries of zinc finger artificial TF constructs. These seed collections are valuable tools for the discovery of (novel) mutant phenotypes. Amongst others, plant lines with phenotypes of interest such as increased vigor, increased tolerance to salinity, and changes in flowering time were easily isolated from a seed pool representing a maximum of approximately 3500 VP16-based three finger artificial TFs (Van Tol and Van der Zaal, unpublished data). An example of the variety of phenotypes observed among primary transformants expressing a fusion of an EAR repressor domain to the C-terminus of three fingers is shown in Fig 3. These plants were transferred to soil in a random fashion from medium containing kanamycin for selection for the presence of the transgene. It is evident that this particular set of plants is very rich in different leaf phenotypes, but one cannot immediately attribute all phenotypes in primary transformants to expression of artificial TFs. In our experience however, the most conspicuous phenotypes will reappear as dominant traits in the next generation. For a large fraction of the phenotypes that were experimentally addressed, a zinc finger artificial TF-encoding construct was recovered that could reestablish the respective phenotype when introduced into wild type plants. Obviously, a particular artificial TF-induced phenotype will be linked to transgene activity and might disappear when the gene gets inactivated. Up till now, we have no evidence that silencing occurs. The artificial TF-induced somatic recombination phenotype [65] has been stable over the next six generations that have been studied thus far. We are now particularly interested in screening for phenotypes associated with more complex traits, such as growth and photosynthesis. Orchestrating novel gene expression patterns by means of genome interrogation may prove valuable in aiding breakthroughs in these fields of research.

Artificial transcription factor-mediated regulation of gene expression | 33
Fig. 3 An example of the variety of leaf phenotypes observed in randomly picked primary transformants plants expressing 3F-EAR fusion proteins.

One could very well wonder whether artificial TFs in genome interrogation experiments should best be expressed in a constitutive, inducible, or in a tissue specific manner. We have used the promoter of the RPS5A gene (pRPS5a) from Arabidopsis for driving the expression of zinc finger artificial TFs [65]. We were motivated by the fact that this promoter is highly active in zygotes, early embryos, and meristematic tissue [71]. Hence, all cells in an Arabidopsis rosette have expressed the zinc finger artificial TF-encoding transgene during their juvenile stages, but are not experiencing unnecessary transcription factor-induced stress when they have matured and/or differentiated. It might be expected, however, that pRPS5A controlled expression of artificial TFs in the early embryo shortly after floral dip transformation induces lethal phenotypes. Expression under control of pRPS5a in dividing cells might hamper cell division in tissue culture-based transformation protocols, thereby abolishing the formation of any regenerated plant material. Hence, unless early lethality is eliminated by a particular choice of expression cassette, it might result in a final mutant population that is biased for expressing only those artificial TFs that induce relatively mild phenotypes. Also, drastic changes in gene expression patterns invoked by artificial TFs might themselves be lethal in early stages of development independently of the expression cassette that is chosen. Indeed, most of the primary transformants harboring pRPS5A expression cassettes appeared to have rather mild phenotypes. Sterility (either male or female) was often found, strongly suggesting that expression of the transgene in gametes can have a substantial effect on fertility. Remarkably,
the majority of surviving primary transformants also lacked an obvious phenotype when using the constitutive CaMV 35S promoter instead of pRPS5A. We did not observe a very conspicuous drop in the efficiency of floral dip transformation. Thus, we tend to believe that indeed most artificial TF-induced phenotypes in our mutant libraries are relatively mild. This does not mean, however, that transcription patterns are also only marginally affected. In plants that were macroscopically normal there was differential expression of more than 1500 genes [66]. Apparently, plants are able to thrive whilst experiencing drastic changes in genome wide transcription patterns.

Different artificial TFs should be found that trigger a similar phenotype when working with high complexity artificial TF libraries and when also being able to generate the required number of transgenic individuals. This approach has led to the identification of gene expression signatures related to the development of drug resistance in human cancer cell lines [64]. However, it might be a problem to find truly comparable phenotypes in relatively small mutant populations when screening at the plant level. A way out of this dilemma is to make a dedicated collection of artificial TFs that are very similar to the one that triggers a uniquely found phenotype of interest. Such an approach was feasible for three finger artificial TFs. By keeping two zinc finger domains constant and alternating the third one, alternative zinc finger artificial TFs were obtained that triggered the same phenotype of interest. Comparative transcriptome analyses were performed between plants expressing phenotype-inducing three finger artificial TFs (Nearest Active Neighbors) and plants expressing highly similar three finger artificial TFs that did not induce the phenotype (Nearest Inactive Neighbors). A core set of genes was identified that specifically contributes to the phenotype of interest [66]. An alternative approach might be to use samples taken from pools of plants expressing related zinc finger artificial TFs that do not induce a phenotype of interest as negative controls for RNA sequencing experiments. When it is of particular interest to identify the genes that cause to an artificial TF-induced phenotype, several approaches could be used. One could use an inducible artificial TF expression system and investigate gene expression changes soon after induction. Target gene identification could then further be corroborated when cognate artificial TF binding sites can be found in their respective gene control regions, if these have functionally been delineated. However, considering the lack of knowledge on the true extent of the gene control region, and the possibly low binding specificity, we do not think that an in silico screening for biologically relevant artificial TF target sites among the large number of target sites that exist in a given genome can yield more than just tentative clues. This is especially true in the case of genomes that have not completely or incorrectly been annotated.
5. Concluding remarks

The availability of modular DNA binding domains allows for a wealth of possibilities for artificial regulation of gene expression. Still, alternative methods for gene regulation are available as well. When looking for the specific merits of artificial TF technology for plant biotechnology, we think its niche should be to help us assess which novel traits and phenotypes can be generated by differentially expressing essentially the same genome. This might pave the way for finding novel phenotypes that have not yet been or cannot be made available by single gene-based methods. In this way, it might enable exploring the potential of a genome in terms of the phenotypes it can produce, and to get to know the hidden properties of a selected plant line or cultivar. In this way, artificial TF-mediated genome interrogation will allow us to place these traits under experimental control. It should be realized however, that the genome interrogation technology greatly depends on genetic modification. It should be noted that when there would be the requirement of introducing complex, multilocus properties into a selected plant line or cultivar, genetic modification is almost certainly going to be a part of such a procedure. In any case, genome interrogation allows for an intriguing complementary approach to methods based on learning from extant genetic variation in plant species.

Acknowledgements

This work was carried out within the research programme of BioSolar Cells, co-financed by the Dutch Ministry of Economic Affairs, Agriculture and Innovation, and was cofinanced by the Consortium for Improving Plant Yield (CIPY) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research. We are indebted to Dr. Martijn Rolloos for experimental data on genome interrogation with ZF-EAR constructs, and to Prof. Dr. PJJ. Hooykaas and Dr. B.S. de Pater for critically reading the manuscript.
References


Artificial transcription factor-mediated regulation of gene expression


