UC Riverside
UC Riverside Previously Published Works

Title

Transcription factor binding specificities of the oomycete Phytophthora infestans reflect conserved and divergent evolutionary patterns and predict function.

Permalink

https://escholarship.org/uc/item/6nk536qf

Journal

BMC Genomics, 25(1)

Authors

Vo, Nguyen
Yang, Ally
Leesutthiphonchai, Wiphawee
et al.

Publication Date

2024-07-23

DOI

10.1186/s12864-024-10630-6

Peer reviewed

Transcription factor binding specificities of the oomycete *Phytophthora infestans* reflect conserved and divergent evolutionary patterns and predict function

Nguyen N. T. Vo[1], Ally Yang[2], Wiphawee Leesutthiphonchai[1,3], Yulong Liu[2], Timothy R. Hughes[2] and Howard S. Judelson[1*]

Abstract

Background Identifying the DNA-binding specificities of transcription factors (TF) is central to understanding gene networks that regulate growth and development. Such knowledge is lacking in oomycetes, a microbial eukaryotic lineage within the stramenopile group. Oomycetes include many important plant and animal pathogens such as the potato and tomato blight agent *Phytophthora infestans*, which is a tractable model for studying life-stage differentiation within the group.

Results Mining of the *P. infestans* genome identified 197 genes encoding proteins belonging to 22 TF families. Their chromosomal distribution was consistent with family expansions through unequal crossing-over, which were likely ancient since each family had similar sizes in most oomycetes. Most TFs exhibited dynamic changes in RNA levels through the *P. infestans* life cycle. The DNA-binding preferences of 123 proteins were assayed using protein-binding oligonucleotide microarrays, which succeeded with 73 proteins from 14 families. Binding sites predicted for representatives of the families were validated by electrophoretic mobility shift or chromatin immunoprecipitation assays. Consistent with the substantial evolutionary distance of oomycetes from traditional model organisms, only a subset of the DNA-binding preferences resembled those of human or plant orthologs. Phylogenetic analyses of the TF families within *P. infestans* often discriminated clades with canonical and novel DNA targets. Paralogs with similar binding preferences frequently had distinct patterns of expression suggestive of functional divergence. TFs were predicted to either drive life stage-specific expression or serve as general activators based on the representation of their binding sites within total or developmentally-regulated promoters. This projection was confirmed for one TF using synthetic and mutated promoters fused to reporter genes *in vivo*.

Conclusions We established a large dataset of binding specificities for *P. infestans* TFs, representing the first in the stramenopile group. This resource provides a basis for understanding transcriptional regulation by linking TFs with their targets, which should help delineate the molecular components of processes such as sporulation and host

*Correspondence:
Howard S. Judelson
howard.judelson@ucr.edu

Full list of author information is available at the end of the article

Vo *et al. BMC Genomics*   (2024) 25:710

Page 2 of 18

infection. Our work also yielded insight into TF evolution during the eukaryotic radiation, revealing both functional conservation as well as diversification across kingdoms.

Keywords Transcription factor binding site, Protein-binding oligonucleotide microarray, Oomycete, Promoter, Phytophthora infestans, Gene regulation, DNA-binding protein

Background

Transcription factors (TFs) establish patterns of gene expression by binding specific sequences in DNA which are usually 5' of the target gene and 6 to 12 nt in size [1]. Over 70 families of eukaryotic TFs have been identified and classified based on the structure of their DNA-binding domains. Some families are restricted to specific taxonomic groups while others occur across kingdoms [2, 3]. Changes in TF binding specificities underlie important evolutionary processes [4, 5] since many of the proteins are master regulators, capable of activating or repressing expression in a tissue or condition-specific manner. Other TFs serve as general activators or direct RNA polymerase to a specific start site. Many TFs act in concert with cofactor proteins or other partners [6]. For example, Basic Leucine Zipper Domain TFs (bZIP) TFs usually act as homo- or heterodimers, while canonical Heat Shock Factors (HSFs) form homo- or heterotrimers [7, 8]. The binding preferences of TFs are typically described as motifs containing a mix of invariable and degenerate sites [9]. Identifying these motifs is an important step towards understanding the function of a TF.

Studies of TFs are relatively limited in the filamentous microbe *Phytophthora infestans*, an oomycete member of the stramenopile lineage of eukaryotes [10]. A few TFs have been shown to regulate its life cycle or those of close relatives [11–14]. *P. infestans* is a devastating pathogen of potato and tomato and is notorious as a cause of the Irish Famine of the 1840's [15]. It is a useful model for oomycetes since it can be cultured on artificial media or plants, technologies exist for manipulating genes [16–18], and a chromosome-scale genome assembly is available [19]. The asexual life cycle of *P. infestans* involves the growth of branched vegetative hyphae which extract nutrients from media or a plant host [20]. As cultures or plant lesions age, the hyphae produce multinucleate sporangia capable of traveling in wind or water to new hosts. Upon chilling in a moist environment, the cytoplasm of each sporangium cleaves into individual zoospores which swim, encyst, and produce germ tubes able to penetrate host tissues. There is also a sexual cycle that occurs when the hyphae of opposite mating types (A1 and A2) interact, causing gametangia to differentiate which unite to generate oospores [21].

Changes in gene expression during the *P. infestans* life cycle are extensive. RNA-seq analyses have shown that of the approximately 17,000 genes expressed from the 219 Mb genome, as many as 49% show more than a 5-fold change in mRNA abundance between life-stages with 8% showing greater than a 100-fold change [22–24]. Examples of important genes that vary during these transitions include those encoding regulators of mitotic dormancy in spores [25], structural components of zoospore flagella [26], or effectors that suppress host defenses during early plant infection or cause host cell death during late infection [27]. Despite the obvious importance of TFs in such processes, the DNA target of only one oomycete TF has been identified [13]. Being able to describe the DNA-binding specificities of *P. infestans* TFs will help illuminate the transcriptional networks of these important but understudied microbes.

Functional binding sites for most TFs in *P. infestans* are thought to reside close to the gene since intergenic distances average only 430 nt and 5' untranslated regions are typically smaller than 50 nt [28, 29]. Distant enhancers seem unlikely to play a major role in transcriptional control since adjacent genes usually have distinct patterns of expression [28]. Also, studies with reporter genes have shown that about 250 nt of DNA upstream of the transcription start site is sufficient to drive normal expression [18, 30]. Consequently, bioinformatic approaches for identifying TF binding sites have generally focused on the 500 nt upstream of the translation start site. Several motifs have been predicted bioinformatically but remain unlinked to a specific TF [28, 31, 32].

In the present study, we used protein binding microarrays (PBMs; [33, 34]) to successfully define the DNA binding preferences of 73 sequence-specific TFs representing the major families in *P. infestans.* Several targets were confirmed using electrophoretic mobility shift assays (EMSA) or chromatin immunoprecipitation (ChIP). We observed that paralogs bearing similar DNA-binding domains often bound related targets and clustered in the genome but displayed distinct patterns of expression consistent with neo- and subfunctionalization models of gene duplication [35]. Motif enrichment analysis combined with studies using a reporter gene suggested which TFs may serve as general activators or stage-specific regulators. About half of the *P. infestans* TFs bound sequences resembling the targets of related proteins from human and plants, reflecting both functional conservation as well as diversification across kingdoms.

Methods

Identification of TFs and tree construction

Two parallel methods were used to identify TFs from *P. infestans*, using gene models from strain T30-4 [36]. One approach, implemented in the Cis-BP pipeline (http://cisbp.ccbr.utoronto.ca), relied mostly on PFAM domains [37]. The second method searched for INTERPRO domains [38]. The two lists were compared and manually curated. As part of this process, predicted genes lacking expression in 16 developmental stages and growth conditions based on RNA-seq data [22–24, 39] were considered to be pseudogenes and removed from the list of TF candidates. Also eliminated were several genes resulting from apparent false duplications in the T30-4 assembly.

For species other than *P. infestans*, domain searches were performed using genome data in Fungidb or Ensembl Protists (http://protists.ensembl.org) or Fungidb [40]. For *P. infestans* and several other species, putative interaction domains not defined by PFAM or INTERPRO were identified using Waggawagga and Deepcoil [41, 42]. Phylogenetic trees were constructed from the DNA-binding domains using MUSCLE and PhyML as implemented in SEAVIEW [43] with 1000 bootstrap replicates.

RNA-seq

Expression analyses were performed using tissues and fastq files from prior studies using an average of three biological replicates [22–24, 39]. In brief, RNA was isolated using kits from Sigma or Agilent. RNA-seq was performed using indexed libraries prepared using the Illumina Truseq kit, and sequenced to produce 75-nt single-end reads. Reads passing the quality filter were aligned to the *P. infestans* T30-4 genome using Bowtie 2.2.5 and Tophat version 2.0.14, allowing for one mismatch [44]. Expression and differential expression calls were made with edgeR [45]. Data for the heatmaps were from isolate 1306 except that the mating data reflects the average of three crosses: 88069×618, 8811×E13, and 88069×E13. Heatmaps were generated using Seaborn [46] using per-gene normalized data. The latter were obtained by dividing FPKM values for each gene against the mean of that gene across all conditions, such that the mean value across the heatmap for that gene would equal 1.0. Thus, the reads in each gene are scaled similarly which allows gene expression across conditions in the heatmap to be compared.

Protein Binding Microarrays (PBMs)

For each TF, parallel arrays were analyzed using recombinant protein produced in *E. coli* and by in vitro transcription-translation. These assays included the PFAM-defined DNA-binding domain, 50 flanking N-terminal and 50 flanking C-terminal residues (or until the end of the protein), and a 6×His tag. The protein

sequences used are shown in Table S1. These included oligomerization domains such as leucine zippers and other coiled-coil domains since they were either included in the PFAM domain (e.g., bZIP and HLH TFs) or resided within the 50 amino acids adjacent to that domain (e.g., HSF and HTH TFs). After optimization for expression in *E. coli*, the sequences were synthesized and cloned into pTH6838 or pTH7069, which contain a T7 promoter and add an N or C-terminal glutathione S-transferase (GST) tag, respectively [33]. In vitro transcription-translation was performed using the PURExpress In Vitro Protein Synthesis Kit from New England BioLabs. For proteins produced in *E. coli*, soluble proteins were purified using nickel resin and eluted in phosphate-buffered saline (PBS) containing 10 to 20% glycerol. Proteins in inclusion bodies were first solubilized in 2 M urea in PBS and then refolding was achieved in 0.5 M arginine containing 10 to 20% glycerol, which was adjusted through dialysis to phosphate-buffered saline with 10% glycerol. Gel analysis indicated that the proteins were 85 to 90% pure and of the expected size.

The methods for analyzing the arrays were as described [33, 34, 47]. Each TF was analyzed in duplicate on two different arrays with differing probe sequences (HK and ME), from which positive 8-mers were identified and E- and Z-scores calculated as described [48]. Experiments were judged as successful if at least one 8-mer had an E-score above 0.45 on both arrays, if the two arrays yielded correlated E- and Z-scores, and if the arrays defined similar motifs based on alignments of the 8-mers using Top10AlignZ [37].

Electrophoretic mobility shift assays

These were performed as described [49] using ca. 35-nt double-stranded DNAs and a Cy5 label. The oligonucleotide sequences, DNA concentrations, and protein concentrations employed are shown in Table S2. Preliminary titrations established conditions where DNA was in excess. After incubation for 15 min at room temperature, the mixtures were separated on a 5% acrylamide nondenaturing gel and imaged with a Typhoon laser scanner. Dissociation constants were calculated based on band intensities measured with ImageJ [50].

Reporter gene analysis

Vectors for *P. infestans* transformation were based in either the promoter-less reporter plasmid pNPGUS or a derivative containing the 74-nt minimal *NifS* promoter [28]. The sequences shown in Table S3 were generated by polymerase chain reaction or synthesized and then cloned upstream of the GUS reporter. Transformants of strain 1306 were obtained by the protoplast method using G418 selection [51]. Transformant tissue was disrupted by grinding in liquid nitrogen and assayed using

Vo *et al. BMC Genomics*   (2024) 25:710

Page 4 of 18

bromochloroindoyl-β-glucuronide or 4-methylumbelliferyl glucuronide [31].

Binding site enrichment analysis

Lists of genes upregulated in a developmental or infection stage were identified using RNA-seq data [22–24, 39]. This involved comparisons to nonsporulating mycelia grown on rye-sucrose, except for nonsporulating mycelia and mating cultures which were compared to sporangia and single cultures of the parents, respectively. For each gene list as well as control lists (i.e., the remaining genes), 500-nt putative promoters were extracted and scanned for the motifs using FIMO using a *p*-value cutoff of 10^{-4} [52]. The statistical significance of over- or under-representation was measured by Chi-Square tests.

Chromatin immunoprecipitation

Triplicate samples of sporangia and sporulating hyphae were collected in 10 mM Tris pH 8.0, 10 mM $MgCl_2$, 0.4 M sucrose, and 1 mM phenylmethylsulfonyl fluoride (PMSF). After adding 1.5 or 2% formaldehyde, the mixture was shaken for 15 min at 50 rpm and glycine was added to 0.125 M. After 5 min of further shaking, the tissue was pelleted at 700×*g* for 4 min, washed three times in PBS pH 7.4, frozen in liquid nitrogen, and ground using a mortar and pestle. After resuspension in buffer A (10 mM Tris pH 8.0, 0.25 M sucrose, 10 mM $MgCl_2$, 1% Triton X-100, 1% protease inhibitor cocktail, 1% PMSF), the material was vortexed for 30 s, subjected to 20 strokes in a tight-fitting Dounce homogenizer, incubated on ice for 1 h, and passed through 15 μm mesh. Nuclei were pelleted from the flow-through at 1,800×*g* at 4ºC for 10 min. After shearing the chromatin to 100–300 nt in a Covaris S220 sonicator, samples were incubated overnight with gentle rocking at 4ºC with mouse IgG or a custom MADS antibody. Protein A magnetic beads (Surebeads, Bio-Rad) were then added and gently mixed for 3 h at 4ºC. Using a magnetic stand, the beads were then washed twice with Buffer A lacking protease inhibitors, twice with Buffer B (100 mM Tris pH 8.0, 500 mM LiCl, 1% Triton X-100, 1% deoxycholic acid), and once with Buffer C (Buffer B plus 150 mM NaCl). The beads were eluted by shaking at 800 rpm for 30 min at room temperature with 0.1 ml of Buffer D (100 mM $NaHCO_3$, 1% SDS). A small portion (5 μl) was saved for immunoblot analysis to confirm the presence of MADS protein, with the rest (95 μl) used for DNA extraction. The latter entailed adding NaCl to 0.54 M followed by overnight incubation at 65ºC to reverse formaldehyde crosslinks. Then, 2 μl of 20 mg/ml Proteinase K was added and incubated for 2 h at 45ºC. The material was mixed with an equal volume of 25:24:1 phenol: chloroform: isoamyl alcohol for 5 min followed by separation at 10,000×*g* for 10 min, and then mixed for 2 min with 24:1 chloroform: isoamyl alcohol followed by

2 min of centrifugation at 10,000×*g*. To the aqueous phase was added sodium acetate pH 5.2 to 0.3 M, glycogen to 1 μg/μl, and 2.5 volumes of cold ethanol. After overnight incubation at -20ºC, the DNA was pelleted at 15,000×*g* for 20 min, washed with 95% ethanol, dried for 10 min, and dissolved in 10 μl of 10 mM Tris pH 8.0. The DNA was then subjected to paired-end Illumina sequencing. Each sample yielded an average of 4.7 million 75-nt reads, which were trimmed using Cutadapt and mapped to the reference genome using Bowtie2 [53]. Peaks were detected using Homer [54]. Motifs enriched in peaks unique to the anti-MADS samples were identified using STREME [55].

Results

Genome-wide identification of TFs in *P. infestans* and relatives

Of 325 *P. infestans* genes annotated as encoding proteins with DNA-binding activity (GO:0003700), 197 were selected for further study since they belonged to TF families known to exhibit sequence-specific DNA binding. This was trimmed to 190 by eliminating genes (including potential pseudogenes) that lacked expression based on RNA-seq data from nonsporulating mycelia from rye-sucrose and minimal media, sporangia, sporangia undergoing zoosporogenesis, motile zoospores, germinated zoospore cysts, 10-day mating cultures, and early and late stages of potato and tomato infection [22–24, 39]. The 190 genes belonged to 22 families with the largest encoding bZIP, Myb, Heat Shock Factor (HSF), and C2H2 (Cys_2His_2) zinc finger proteins (Fig. 1). The figure also indicates the number of proteins in each family that were tested in the PBMs, as will be described in more detail in later sections. The similarity of proteins within each family can be surmised from the phylogenetic trees in Fig. S1.

The chromosomal distribution of the genes was consistent with growth of many of the families through unequal crossing-over. For example, many genes encoding bZIP and C2H2 proteins were clustered (Fig. S2). Such expansions appeared to have been ancient since similar numbers of TFs were detected in most oomycetes including other members of *Phytophthora* and representatives of *Globisporangium*, *Pythium*, and *Saprolegnia* (Table S4). However, none of the enlargements were as extensive as those described for certain families in plants and animals [3]. Some families were smaller in *Hyaloperonospora arabidopsidis* and *Albugo laibachii*, which are obligately pathogenic species having streamlined genomes [57]. Several small families such as GCR1 were not detected in oomycetes besides *Phytophthora.* Most TF families had similar numbers in other stramenopiles, except for the C2H2 group which was about one-fifth the size in diatoms.
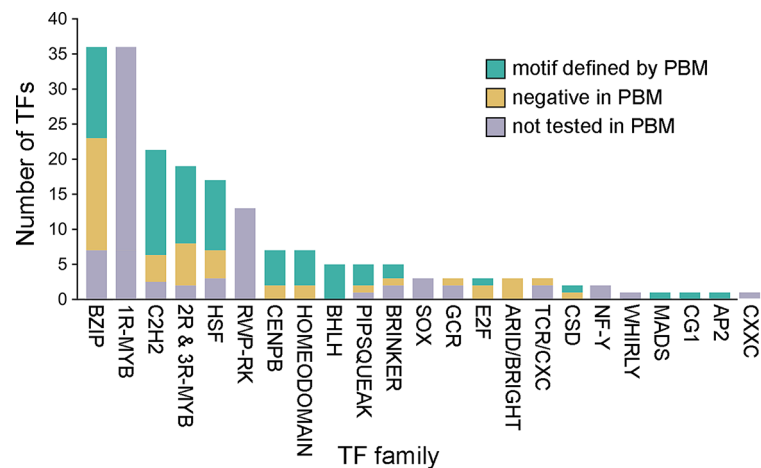
Fig. 1 Predicted sequence-specific transcription factors (TFs) from *P. infestans*. Each bar is colored to indicate the number that yielded a DNA-binding motif in the PBMs, as described in Methods. Myb proteins with single DNA-binding domains (1R-Myb) or two two or three such domains (2R and 3R-Mybs) as listed separately since many 1R-Mybs may be telomere-binding factors rather than TFs [56]. Predicted pseudogenes are not included

Determination of binding specificities

Protein binding microarrays (PBMs) were used to assay 123 of the *P. infestans* TFs for their DNA-binding specificities. This was limited to the 19 families that had yielded positive results in PBM studies of other species [37, 58]. In brief, this involved expressing their DNA-binding domains and adjacent dimerization domains both in *E. coli* and by coupled in vitro transcription-translation, with GST tags. After incubating the protein fragments with PBMs, fluorophore-conjugated GST antibodies identified the bound oligonucleotides from which TF binding sequences were extracted.

After filtering out low-quality data, binding motifs were generated for 73 TFs by aligning top-scoring 8-mers extracted from the PBMs (Fig. S3). Position-specific frequency matrices (PFMs) based on those alignments are supplied in Fig. S4. These PFMs are also represented by sequence logos that will be presented in the following sections, and the number of proteins within each TF family that successfully yielded a PFM are summarized in Fig. 1. Thirteen of the motifs resembled those associated with *P. infestans* promoters in a prior study [31].

Heat Shock Factor (HSF) family

DNA-binding specificities were determined for 10 of the 17 expressed HSFs (Fig. 2A). To help interpret their evolution, in the figure their DNA targets are overlaid on a phylogenetic tree based on the DNA-binding domains. Also shown are expression patterns of the TFs, presented in the same order as in the tree (Fig. 2B). Motifs bound by selected HSFs from human and *Arabidopsis thaliana* are displayed to assess the conservation of the sites across kingdoms (Fig. 2C).

Nearly all characterized HSFs from other eukaryotes including human HSF1 and *A. thaliana* HSFC1

bind sites that contain one or more units of nTTCn [59, 60]. These usually occur in a head-to-head orientation, forming repeats of TTC and GAA separated by a 2-nt gap although HSFs binding head-to-tail arrays are also described [61]. Only a few HSFs, such as human HSFY2, are reported to bind ungapped arrays [62]. Both forms of binding preferences were observed in *P. infestans*. Most common were HSFs that bound ungapped motifs, such as PITG_08199 and PITG_11760 which recognized GAATTC (Fig. 2A). In contrast, the predicted site for PITG_04701 was the gapped motif TTCTAGAA. This target was confirmed by fluorescent EMSA (Fig. 2D).

Whether the *P. infestans* HSFs bound gapped or ungapped sites was incongruent with relationships in the tree (Fig. 2A). For example, the upper-most clade includes HSFs which bound ungapped and gapped dimers, as does the clade in the middle of the tree (e.g., TTCNGAA for PITG_03306 and TTCGAATTC for PITG_22459).

The logos for the three proteins at the base of the tree displayed the TTC motif but with complex flanking positions. Examination of the 8-mers from their PBMs suggested that this was due to flexibility in binding. For example, for both PITG_04694 and PITG_20387 the 8-mers included both ungapped and 2-nt gapped arrays (e.g., TTCGAA and TTCnnGAA) and solo TTC motifs (Fig. S3). Some yeast HSFs have also been shown to bind targets with varying gaps [63].

Adding the human and *A. thaliana* HSFs to phylograms based on the *P. infestans* DNA-binding domains did not support a relationship between binding preference and domain sequence (Fig. S1). In particular, *P. infestans* HSFs that bound gapped or ungapped arrays both clustered with HSFs having gapped targets such as HsHSF1 and AtHSFC1. HsHSFY2, which binds an ungapped target, appeared as an outgroup consistent with a separate
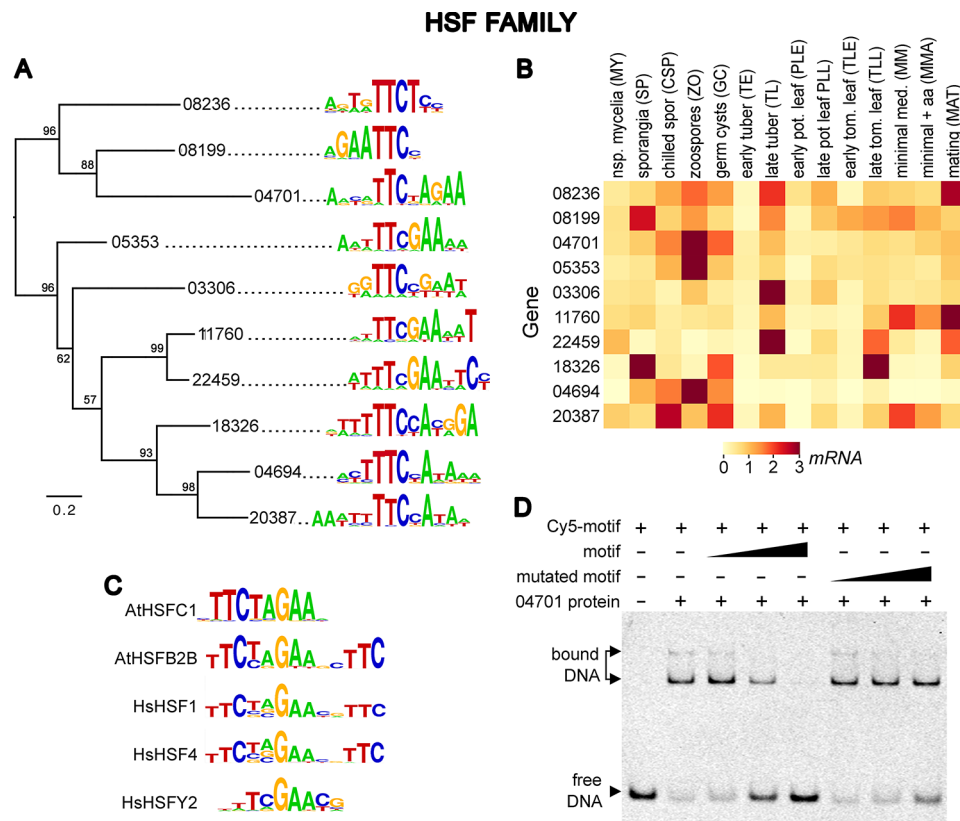
HSF FAMILY

Fig. 2 DNA-binding preferences of HSF family TFs. (**A**) Phylogenetic tree based on amino acid sequences of DNA-binding domains, aligned with logo plots representing the predicted DNA targets of each TF. Bootstrap values above 50% are shown at nodes. (**B**) mRNA levels of the TFs in different stages of development including nonsporulating mycelia on rye-sucrose or minimal media, sporangia, sporangia chilled to initiate zoosporogenesis, swimming zoospores, germinating zoospore cysts, infected potato and tomato leaves or potato tuber, and mating cultures. The latter were obtained from 10-day mixed cultures of A1 and A2 hyphae, containing young gametangia and oospores. Expression levels in the heat map are shown as per-gene normalized FPKM values such that the mean of each gene across all samples is 1.0, as described in Methods. A subset of the data was reported previously [22]. (**C**) Binding specificity of HSF TFs from *A. thaliana* (At prefix) or human (Hs prefix) as determined from prior SELEX, DAP-seq, or PBM studies [62, 65–67]. In this and later figures, the human and plant motifs represent those with binding specificities closest to those of the *P. infestans* proteins. Otherwise, representative human and/or plant proteins are shown. (**D**) Validation of PITG_04701 binding site by EMSA using a Cy5-labeled double-stranded oligonucleotide containing TTCTAGAA. The large triangle represents competitors added at 1, 10, and 100-fold the concentration of labeled probe. The results demonstrate specific binding since the same sequence added as an unlabeled competitor reduced the intensity of the retarded bands, unlike DNA containing a mutated motif. A full-length gel is presented in Figure S5

evolutionary history. As noted previously, its DNA-binding domain diverges substantially from those of other HSFs [64].

Although HSF proteins in other kingdoms often form oligomers that can bind three or more of units of TTC or its reverse complement [59, 68, 69], our motifs usually contained only one or two units. This likely reflects the tendency of the PBM approach to underestimate the width of binding sites [65]. We addressed this by using EMSA to measure the binding of PITG_04701 to one, two, or three copies of TTC (Fig. 3). Dissociation constants calculated from those assays (Fig. 3, right) indicated that binding to the trimeric DNA site was stronger than to the dimer or monomer. Binding to the trimer was not as strong as that reported for human HSF1 [68], but this might be explained by variation in the methods employed (EMSA versus fluorescence polarization) or

our use of a higher pH (7.9 versus 7.5) which may suppress oligomerization [70]. Nevertheless, the ability of PITG_04701 to bind diverse sites may enable graded control of transcription. Consistent with this, we detected both dimeric, trimeric, gapped, and ungapped target motifs in the promoters of *P. infestans* genes encoding Hsp70, which in other species are often regulated by HSF factors [71].

PITG_04701 contains a coiled-coil domain (i.e., hydrophobic heptad repeats) near its DNA binding-domain. This configuration occurs in animal, yeast, and plant HSFs where the repeats enable oligomerization [72–74]. The occurrence of a coiled-coil domain in PITG_04701 is consistent with the multiple large, retarded bands observed in EMSA (Fig. 3), although some of the bands might be due to individual proteins binding separate motifs. The presence of this putative oligomerization
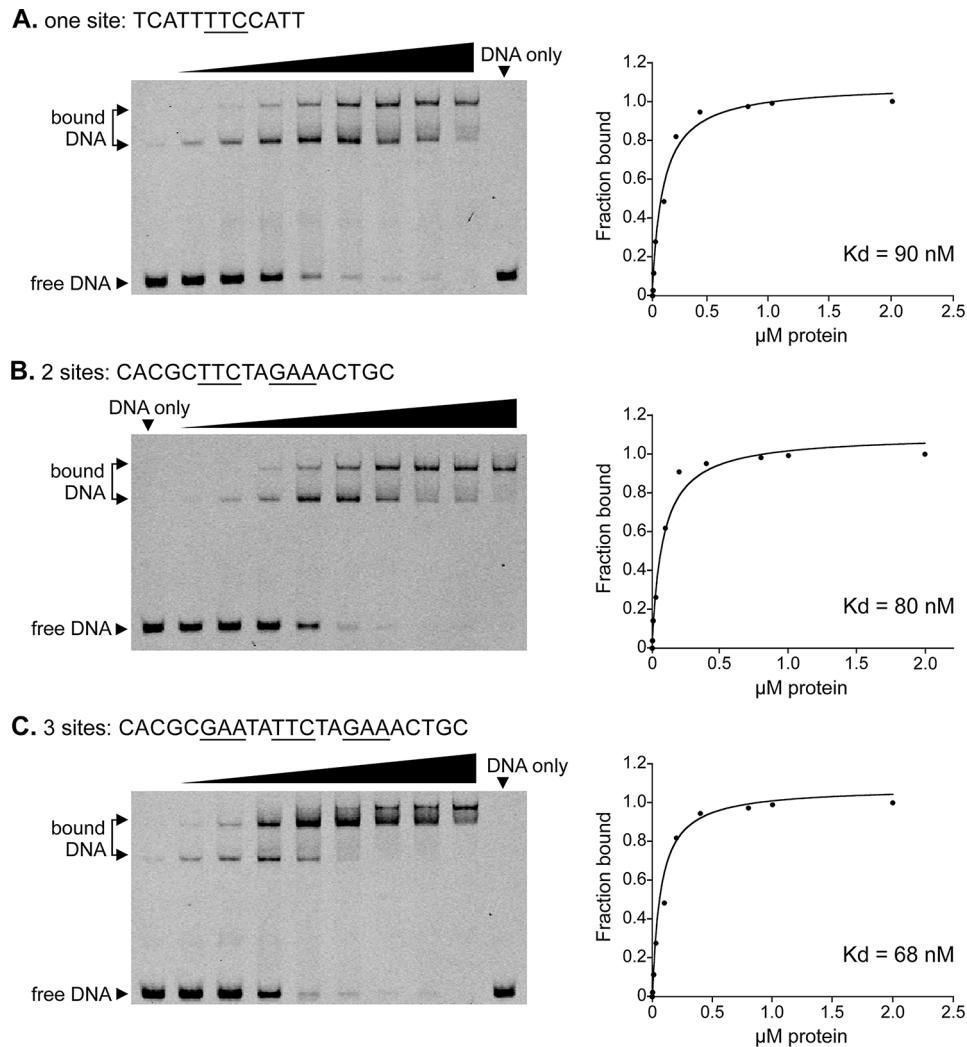
Fig. 3 Binding of HSF protein PITG_04701 to targets with one, two, or three TCC/GAA units. These EMSA assays used 20 nM of Cy5-labeled double-stranded oligonucleotides bearing the indicated sequences flanked by 14–17 nt of random DNA, and 0 to 2000 nM of PITG_04701 (triangle). The fraction of bound DNA was determined by densitometry of the gels and used to calculate K_d. The portion of the protein tested here as well as on the PBMs (Table S1) includes the coiled-coil domain which is believed to enable oligomerization. Full-length gels are presented in Figure S5

domain is also consistent with the protein's optimal trimeric binding site. Due to the universal presence and functional importance of coiled-coil domains in plant, fungal, and animal HSF proteins, we were surprised to discover that only 7 of the 17 *P. infestans* proteins (PITG_03306, 04700, 04701, 05353, 06935, 08199, 15654) had coiled-coil domains based on the six prediction tools within Waggawagga and a newer neural network method, DeepCoil [41, 42]. The absence of coiled-coils did not appear to be due to errors in gene models based on the alignment of the genes to RNA reads and since orthologs mined from other oomycete genomes had similar structures.

RNA-seq data revealed diverse patterns of expression of the HSF genes, with those binding similar motifs often displaying distinct transcriptional profiles and vice versa

(Fig. 2B). All 10 genes were upregulated in at least one life-stage or growth condition, especially in one of the spore stages. For example, PITG_04694 and PITG_04701 were expressed primarily in zoospores. Even though their genes reside within 50 kb which suggests evolution from a common ancestor, their protein sequences did not cluster in the phylogram and their DNA binding motifs are distinct. Similarly, while PITG_03306 and PITG_22459 were transcribed primarily during late tuber infection, they did not cluster in the tree and had distinct binding sites. In contrast, PITG_11760 and PITG_22459 clustered and bound similar motifs but had divergent expression patterns although both were upregulated during mating.

In other species, subsets of HSF genes are transcribed constitutively, developmentally-regulated, or induced
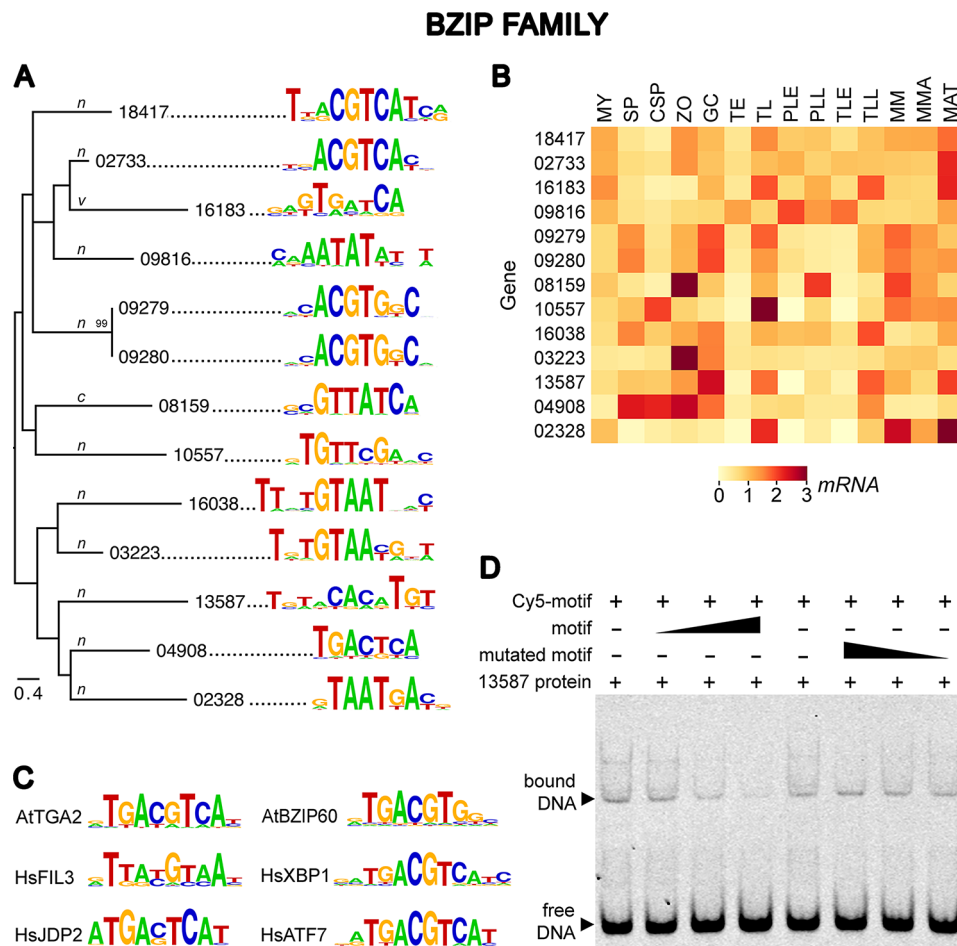
BZIP FAMILY

Fig. 4 DNA binding specificities of bZIP family. The organization of panels **A**-**D** are the same as in Fig. 2 except that the unlabeled competitors used in EMSA were at 1, 10, 50-fold the concentration of Cy5-labeled probe. The single-letter code on the branches of the phylogenetic tree indicates whether the TFs were the Asn (n), Cys (c), or Val (v) types. The codes for the RNA samples in panel B are shown in Fig. 2. A full-length gel is presented in Figure S5

by stresses including heat, starvation, and reactive oxygen [75]. It was therefore notable that PITG_11760 and PITG_20378 were induced in our minimal media, a near-starvation condition which supports very poor growth. We also discovered that PITG_05353, PITG_08236, and PITG_11760 are upregulated by hydrogen peroxide by mining data from a prior study [76].

bZIP family

This group includes members with a canonical DNA-binding domain and those in which an evolutionarily conserved Asn corresponding to residue 235 of *Saccharomyces cerevisiae* GCN4 is substituted by Cys, Val, or Tyr [11, 77]. Thirty bZIPs with detectable transcription were identified and used for PBM analysis including 17, 7, 5, and 1 in the Asn, Cys, Val, and Tyr categories, respectively. DNA motifs were predicted for 11 of the 17 Asn bZIPs (Fig. 4A). In contrast, motifs were identified for only one of the seven Cys types and one of the five Val forms. The predicted motif for PITG_13587 was validated by EMSA (Fig. 4D). bZIP proteins bind DNA

as dimers; we speculate that some proteins that did not yield motifs may be obligate heterodimers.

The position of a bZIP on the phylogenetic tree often correlated with its DNA-binding specificity. Four of the six bZIPs in the largest clade, ranging from PITG_18417 to PITG_09280, bound sequences with a palindromic ACGT core. This motif also occurs within many targets of bZIPs from *A. thaliana* and humans (Fig. 4C). However, binding preferences lacking an ACGT core were predicted for about half of the *P. infestans* bZIPs, representing the lower clades of the tree.

About 85% of the bZIPs appeared to bind palindromes, which is expected since this family typically acts as dimers [7]. For example, the palindrome recognized by PITG_09816 was ATATAT. Others bound gapped palindromes, such as PITG_04908 (TGACTCA). Although palindromes were not evident in the consensus logos of several other bZIPs, individual 8-mers from their PBM data were often palindromic (Fig. S3). Examples are PITG_16038 and PITG_02323 (GTAATTA
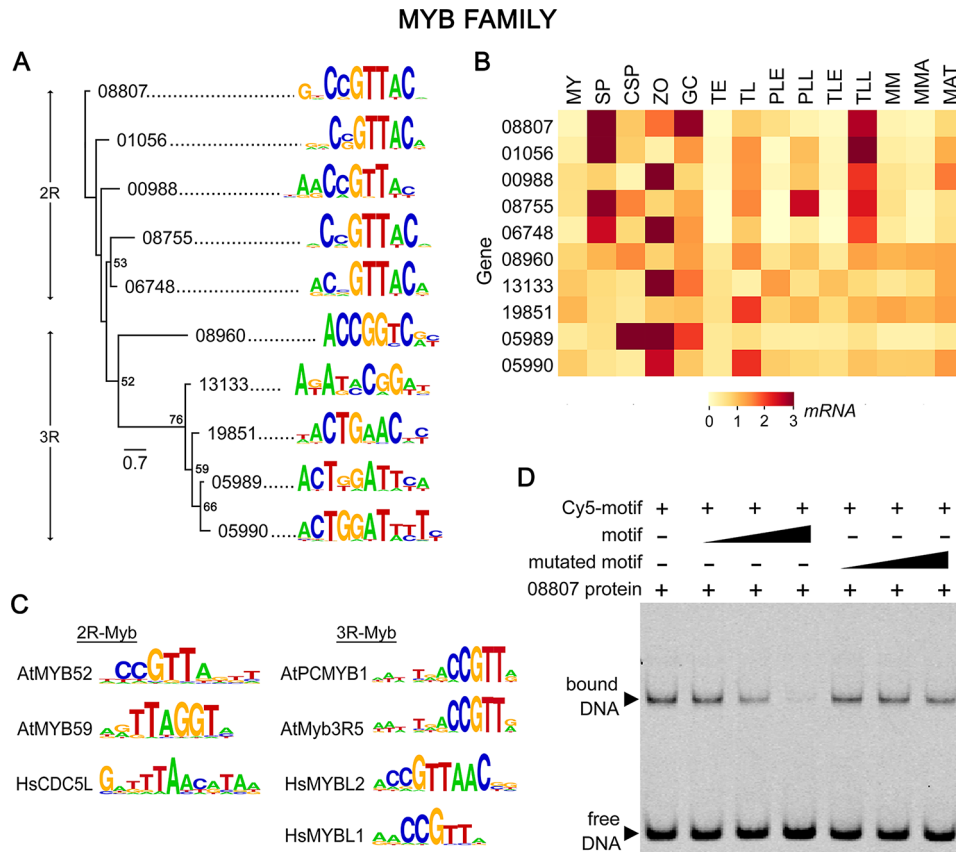
Vo *et al. BMC Genomics*   (2024) 25:710

Page 9 of 18

MYB FAMILY

Fig. 5 DNA binding specificities of MYB family. The organization of panels **A**-**D** are the same as in Fig. 2. The clades in the phylogenetic tree are marked to denote the 2R-Myb and 3R-Myb types. A full-length gel is presented in Figure S5
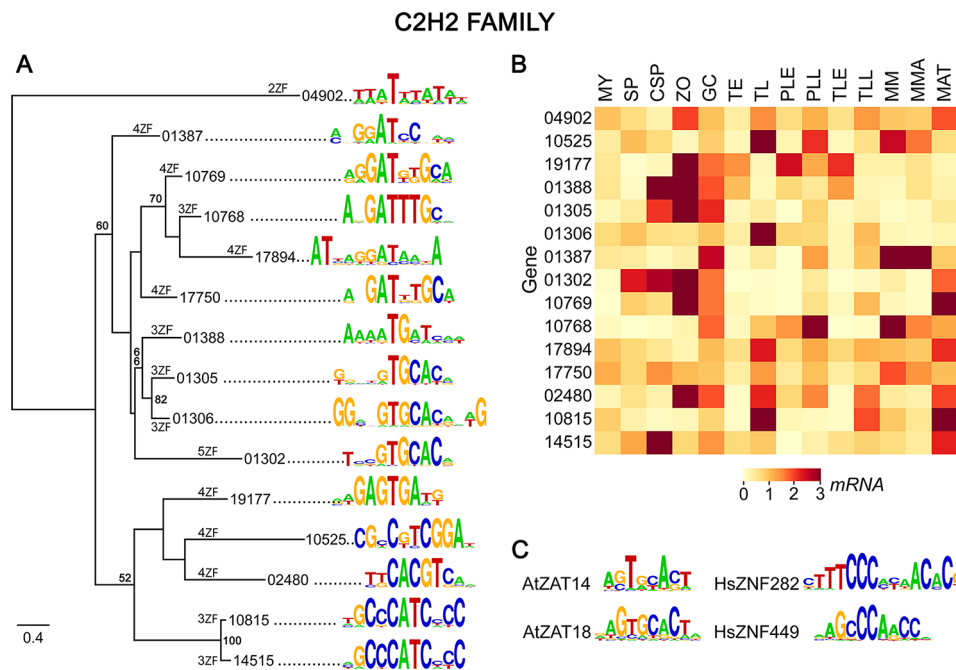
C2H2 FAMILY

Fig. 6 DNA binding specificities of C2H2 zinc-finger family. The organization of panels **A**-**C** are the same as in Fig. 2. The number of zinc fingers in each protein is marked on the branches in panel A

C), PITG_10557 (GTTCGAAC), and PITG_16183 (CAT CGATG).

Transcripts of about three-quarters of the bZIPs were induced in at least one stage of the life or disease cycles (Fig. 4B). However, fewer bZIPs than HSFs had dynamic changes in mRNA levels during mycelial-spore transitions while more bZIPs were upregulated in hyphae from minimal media compared to rye-sucrose media. Most bZIP exhibited low relative mRNA levels in the early (biotrophic) stages of tuber and leaf infection. Interestingly, most other TF families including HSFs showed this same pattern.

Continuing another trend seen with the HSFs, the DNA-binding preference of a bZIP and its transcription profile were not always correlated. For example, PITG_18417 and PITG_02733 bound similar motifs and had similar expression profiles. In contrast, PITG_16038 and PITG_03223 had different patterns of expression despite having similar binding preferences.

Myb family

As in other eukaryotes, *P. infestans* encodes proteins with one, two, or three Myb DNA-binding domains which are named 1R, 2R, and 3R-Myb proteins, respectively. Since many proteins with single Myb domains lack
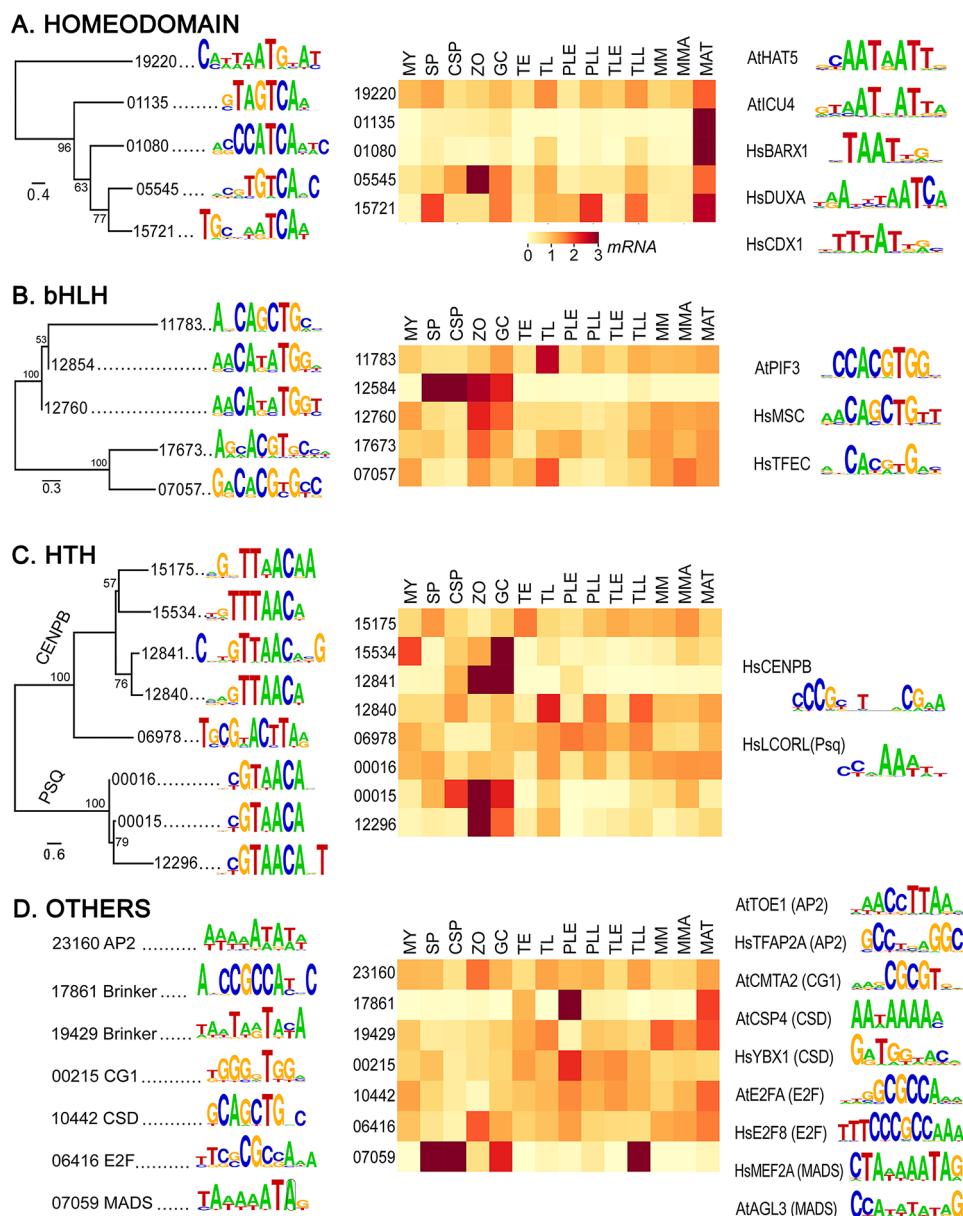

Fig. 7 DNA binding specificities of homeodomain, bHLH, HTH, and other TF groups, showing phylogenetic trees and predicted motifs (left), RNA level heatmaps (center), and motifs bound by *A. thaliana* and human orthologs (right, with At and Hs prefixes)

sequence-specific TF activity [56, 78], we focused our PBM studies on the 2R and 3R proteins. DNA binding specificities for six of seven 2R and five of nine 3R types were obtained (Fig. 5A).

The DNA motifs for the 2R and 3R groups were distinct (Fig. 5A). The 2R members all bound sequences sharing CCGTTAC, which resembles the targets of human and *A. thaliana* 2R-Mybs (Fig. 5C). This target was confirmed for PITG_08807 by EMSA although a tendency for nonspecific binding was evident since the unlabeled oligonucleotide was only about five times more effective as a competitor than one with a mutated motif (Fig. 5D). Non-specific binding of Myb proteins from other kingdoms has also been described [79]. In contrast to the 2R-Mybs, the five 3R types bound more divergent sequences although three shared an ACTG motif. None of the targets resembled those bound by human or *A. thaliana* 3R-Mybs (Fig. 5C).

Our RNA-seq data revealed distinct patterns of expression between and within the 2R and 3R-Myb subfamilies (Fig. 5B). One difference was that four of the five 2R-Mybs, but no 3R form, were upregulated strongly in sporangia. Continuing a trend seen with other TF families, members within both the 2R and 3R groups that had similar DNA-binding domains and target motifs often had different transcriptional profiles. For example, PITG_05989 and PITG_05990 bound similar DNA motifs, but mRNA levels of the former were upregulated in sporangia chilled to initiate zoosporogenesis (CSP) while the latter was not upregulated until after zoospores were released (ZO). Also, only PITG_05990 was induced during late tuber infection. Another continuing trend was that transcript levels were typically low during early plant infection, especially for the 2R-Mybs.

C2H2 zinc finger family

Twenty-two proteins were identified with high-confidence as C2H2 TFs from *P. infestans* based on the presence of two or more C2H2 DNA-binding domains. An additional 50 proteins were defined as low-confidence hits since they either contained sequences seemingly inconsistent with TF activity such as retroelement domains or bore only a single C2H2 domain, which is normally insufficient for binding DNA [80, 81]. It is possible that some of the latter might have affinity for DNA in the presence of other structural elements as seen with plant Superman proteins [81]. No protein resembled the related DOF single-domain TFs of plants [82].

Our PBM studies were limited to the high-confidence hits, and resulted in the definition of binding preferences for 15 proteins (Fig. 6A). These exhibited more variation in binding preference than seen within the other TF families. Nevertheless, there was some congruence between the motifs bound by some C2H2 proteins and

their position in the phylogenetic tree. For example, the clade containing PITG_01388 and PITG_01306 all bound sequences containing GTGCAC, while the motifs for clustered proteins PITG_10815 and PITG_14515 both had a GCCCATC core. The latter resembles the sites bound by human ZNF282 and ZNF449 but considering the diversity of the human family this may not imply an evolutionary relationship since there was little similarity between the amino acid sequences of the oomycete and human DNA-binding domains. Similarly, the motif predicted for PITG_01305 resembled the targets of *A. thaliana* ZAT6 and ZAT18.

Binding sites within the plant and animal families are also diverse due to variation in the sequence, number, or spacing of their zinc fingers [65, 83]. Each *P. infestans* C2H2 protein contained an average of 3.5 zinc fingers, ranging between two and five. This is similar to the number in other oomycetes, but less than the nine found in the average human C2H2 protein [83] and more than the average of two fingers in *A. thaliana*, fungi, and members of the SAR (stramenopile-alveolate-rhizaria) supergroup.

As with other TF families, many C2H2 genes had mRNA levels that increased in zoospores and germinated cysts (Fig. 6B). However, compared to other families, fewer genes were upregulated in sporangia and more rose during mating, such as PITG_10815 and PITG_14515. Interestingly, although these two clustered in the tree and bound similar motifs, only PITG_14515 was expressed strongly in chilled sporangia and only PITG_10815 in late tubers, providing another example of possible subfunctionalization. Slightly more C2H2 genes were expressed during early infection than seen for other TF families.

Homeodomain family

Binding preferences were determined for all five members of this family (Fig. 7A, left). Their motifs were largely dissimilar although four of the five predicted targets contained TCA. While most human and *A. thaliana* homeodomain TFs bind AT-rich sequences, this was observed only for PITG_19220. That protein contains three homeodomains which were tested separately, but only the N-terminal domain yielded a positive result.

A striking feature of the expression profiles of the *P. infestans* genes was that all were expressed highly during mating (Fig. 7A, center). This was especially true for PITG_01080 and PITG_01135. Despite this similar pattern, their predicted DNA targets only shared a TCA motif. None of the top twenty 8-mers bound by each protein on the arrays were in common.

Basic Helix-Loop-Helix (bHLH) family

DNA targets were identified for all five proteins in this group, with similarities in their motifs congruent with locations of the proteins in the tree. Consistent with data

from other species that bHLHs bind as dimers, the DNA targets all contained the gapped palindrome CANNTG (Fig. 7B, left). This is also within the targets of the representative bHLHs from *A. thaliana* and human, which have been described as E- (CANNTG) and G- (CAC-GTG) boxes, respectively (Fig. 7B, right) [65, 84].

The *P. infestans* genes displayed diverse patterns of expression despite having similar DNA-binding preferences (Fig. 7B, center). For example, PITG_11783 was transcribed at similar levels in most tissues except for tubers where its mRNA levels were high. In contrast, PITG_12584 was expressed primarily in the asexual spores.

Helix-Turn-Helix (HTH) families

The HTH domain occurs in two families of DNA-binding proteins, CENPB and Pipsqueak (Psq). While Psq proteins regulate protein-coding genes, human CENPB participates in assembling centromeres and repressing transcription of non-coding sequences [85, 86]. A relative of CENPB (but not Psq) may exist in *A. thaliana*, but its binding sites are unknown [87].

PBM analysis identified binding motifs for five of seven CENPB and three of four Psq proteins from *P. infestans* (Fig. 7C, left). Despite functional differences between CENPB and Psq, nearly all of their predicted binding sites shared a TAACA motif. This does not occur in the targets of human CENPB or Psq (Fig. 7C right). Despite the similar binding preferences of the *P. infestans* proteins, each group displayed distinct and diverse transcriptional profiles (Fig. 7C, center). For example, PITG_12296 and PITG_00015 exhibited spore-associated expression patterns while PITG_00016 was more constitutive.

While human CENPB proteins bind a 17 bp centromeric motif called the CENPB box (Fig. 7C, right), we determined that the GTTTAAC and GTTAAC motifs bound by the proteins from *P. infestans* are not enriched at its centromeres. Unlike the human CENPB box, the *P. infestans* motifs are palindromes which suggests that its proteins have a distinct function.

Other families

DNA targets were also determined for the small Brinker, E2F, CSD, AP2, CG1, CSD, E2F, and MADS-box families which have five, three, two, two, one, one, and one members, respectively (Fig. 7D, left). In several cases the targets resembled those of *A. thaliana* and human proteins (Fig. 7D, right). For example, MADS-box proteins from all three species bound AT-rich sequences, while E2F proteins all targeted motifs bearing CGCCA. In contrast, the binding sites of the *P. infestans* AP2, CG1, and CSD proteins were dissimilar to those from the human and plant. We are unaware of a known binding site for a human Brinker, but the binding specificity of

PITG_17861 resembles the GC-rich targets of Brk from *Drosophila melanogaster* [88].

A range of expression patterns were observed for the *P. infestans* TFs in this section. Most showed moderate mRNA levels in most tissue samples, while PITG_17861 (Brinker) and PITG_07059 (MADS) were strongly stage-specific. For example, PITG_07059 was upregulated strongly in sporangia.

Chromatin immunoprecipitation confirms the predicted MADS-box target

Although a prior study in a mammalian system concluded that TF binding preferences identified by ChIP-seq matched in vitro binding results [37], we chose to also test this with an oomycete TF. We examined PITG_07059 since a satisfactory antibody had been generated as part of a prior study [12]. Due to the gene's sporulation-associated expression pattern, samples for ChIP-seq were prepared from both sporulating mycelia and purified sporangia. This led to the identification of 259 and 367 peaks across chromosomes for the two tissues, respectively (Fig. 8A). These reside predominantly in promoters. The sequence over-represented in those peaks matched that obtained from the PBM (Fig. 8B).

Enrichment analysis links DNA targets to expression profiles

To better understand the roles of the TFs, we calculated whether their motifs showed biased representation in promoters with specific patterns of activity. To accomplish this, we developed lists of genes having mRNA levels that were 10-fold higher in each tissue compared to nonsporulating mycelia in rye-sucrose broth; genes higher in nonsporulating mycelia were identified by comparison to sporangia. Then, matches to the motifs were identified using FIMO with the default *P*-value threshold of 10^{-4} [52]. We hypothesized that targets of TFs that upregulate genes in a specific tissue would be over-represented in the relevant promoter set, while under-represented sites might bind repressors or stimulate transcription in other stages. A second hypothesis is that sites binding general activators would show little bias across the promoter sets.

As shown in Fig. 9, 68 of the 73 TF motifs were over-represented (blue) or under-represented (red) in one or more promoter sets, based on a *P*-value threshold of 0.05. The most common patterns included over-representation in genes upregulated during plant infection, mycelial growth, or spore development. However, the trends varied between and within the TF families. For example, while about half of the C2H2 binding sites were enriched in genes expressed more in nonsporulating mycelia, this was true for only one bZIP and one homeodomain target.
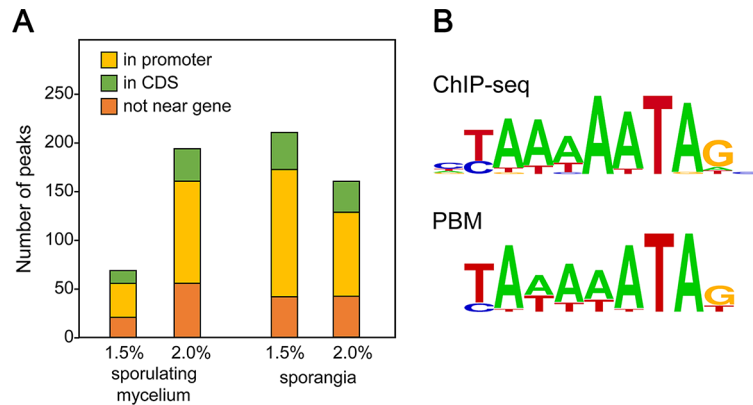
Fig. 8 Targets of MADS-box protein determined by ChIP-seq. (**A**) Number and genomic context of peaks in chromatin from sporulating mycelia and sporangia crosslinked with 1.5% and 2% formaldehyde. Locations include coding sequences (CDS), promoters, and sites distant from genes. (**B**) Logo plots representing motifs enriched in promoter region based on the ChIP-seq and PBM data
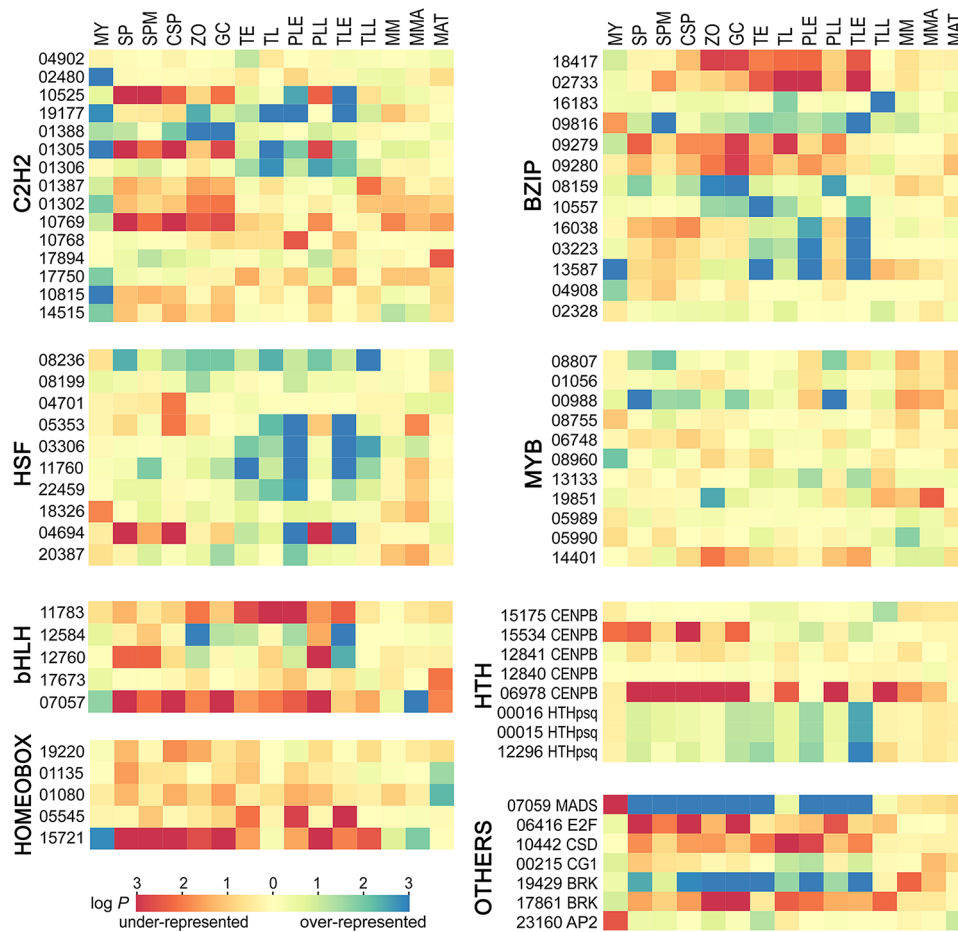

Fig. 9 Enrichment analysis of TF binding sites in promoters displaying different patterns of activity. The heatmaps display $\log_{10} P$-values associated with over- or under-representation (blue and red, respectively) in promoters upregulated > 10-fold in the indicated tissues when compared to nonsporulating mycelia, except that nonsporulating mycelia (MY) was compared to sporangia (SP) and the mating culture (MAT) was compared to the mean of the parents grown separately. The tissues are the same as in Fig. 2 with the addition of SPM, which represents the entire tissue (sporangia and mycelia) from 5-day cultures

Supporting our hypothesis that over-represented motifs would be linked to stage-specific transcription, the binding site for the MADS-box TF was overly abundant in promoters of genes upregulated in sporulating hyphae and sporangia (Fig. 9, lower right panel). This protein was proven earlier to regulate many sporulation-induced genes [12]. Its binding site was also abundant in promoters upregulated in many of the plant infection samples. In the late infection timepoints, this might be attributed to the occurrence of sporulation. However, over-representation of the motif in the early timepoints might instead reflect the involvement of a different TF with a similar binding site, since sporulation initiates only near the end of the disease cycle.

We extended our search for motifs in co-regulated gene sets to search for potential *cis*-regulatory modules (CRMs). These are defined as clusters of binding sites for distinct TFs that combine to dictate a pattern of expression, and are common in the promoters of plants and animals [89]. CRMs are typically identified by searching for over-represented combinations of TF binding sites within a genome [90]. We therefore searched total *P. infestans* promoters and subsets upregulated at each stage of the life cycle using MCAST [91] and custom scripts using outputs from FIMO [52]. After eliminating gene families in which promoter and coding sequences of members

were nearly identical, no over-represented motif combinations (potential CRMs) were identified.

Functional analysis of a motif confirms its predicted function

In the analysis shown in Fig. 9 several motifs lacked a strong association with any stage-specific pattern of expression. We hypothesized that such motifs bind general activators. Prior studies have shown that adding the DNA target of a general activator to a minimal promoter often stimulates transcription, while its elimination from a full promoter reduces expression [92]. This was tested using the motif for HSF PITG_04701 which was not over-represented significantly in any promoter set. As shown in prior studies [28] and repeated here, the *NifS* minimal promoter does not drive the β-glucuronidase (GUS) reporter when transformed into *P. infestans* (Fig. 10A). However, expression resulted when the motif bound by PITG_04701 was added 5' to the *NifS* sequences. Also as hypothesized, transcription was impaired by mutating the motif in a full promoter. While the intact promoter drove robust expression of the reporter, a weak signal resulted when the motif was mutated. This was shown initially using histochemical staining and then confirmed by a quantitative assay (Fig. 10B). As might be expected for a general activator, PITG_04701 is well-expressed in
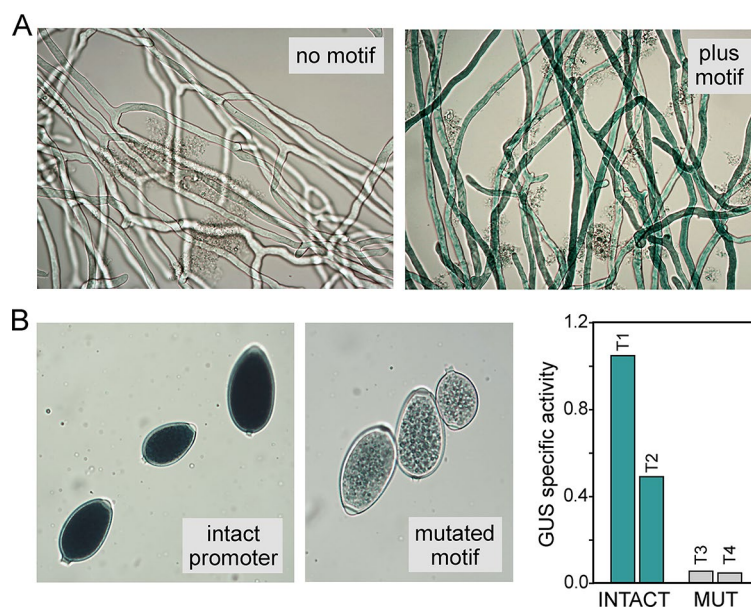

Fig. 10 Evaluation of the PITG_04701 binding site in *P. infestans* transformants. (**A**) Tests using the minimal *NifS* promoter fused to GUS. The left panel shows a representative transformant with the unmodified minimal promoter, subjected to histochemical staining. As described previously [28], this partial promoter does not drive expression in any life stage. However, expression was detected when the PITG_04701 motif was inserted upstream of the minimal promoter. Similar results were obtained with four independent transformants. (**B**) Assays using the GUS gene fused to the promoter of PITG_20221, which contains the PITG_04701 motif. Indicated are a representative transformant containing the intact promoter (left) and one in which the PITG_04701 binding site was mutated (right). Similar results were obtained with four independent transformants. While deleting the PITG_04701 motif did not prevent induction of the gene, the mutation strongly repressed expression. The graph shows quantitative data from a fluorometric assay using sporulating hyphae from two transformants with the intact promoter (T1, T2) and two with the mutation (T3, T4. Due to position and transgene copy number effects, transformants obtained with the same plasmid (e.g., T1 and T2) are not expected to yield the same level of expression

all stages, ranking for example in the 77% percentile of genes in mycelia. The gene is upregulated in zoospores, which might relate to the increase in expression of general housekeeping genes that occurs when their cysts germinate [22].

Discussion

We have defined the DNA binding preferences of 73 TFs from *P. infestans* representing 14 of its 22 families and associated many of those TFs with patterns of gene expression during the life cycle. Several binding sites were verified by EMSA or ChIP-seq. The latter represents its first application to an oomycete TF. Combined with RNA-seq resources [22–24, 39], methods for blocking gene activity through homology-based gene silencing or editing [51, 93], a chromosome-scale genome assembly [19], and emerging data on nucleosome occupancy [94] the prospect of revealing the networks that regulate growth, development, and pathogenesis are now much improved.

Our success rate of 59% for proteins tested on the PBMs compares favorably with results from other organisms [37, 58]. The failures might be due to the absence of a cofactor or post-translational modification, a requirement for heterodimerization [95], or poor folding of the proteins. Some failures might be related to the fact that we did not test the whole protein, although prior studies indicated that full-length TFs and their isolated DNA-binding domains nearly always bound similar sequences [62]. In both the current and prior studies, a strong correlation was observed between binding data from in vitro studies using DNA-binding domains alone and ChIP-seq data targeting the native protein in vivo [37].

Reflecting on some causes for failures on the PBMs may provide insight into the biology of *P. infestans.* For example, yielding poor results were the bZIPs with novel amino acids (Cys, Val, or Tyr) in their DNA-binding domains. It is possible that these only associate with DNA as heterodimers. Alternatively, rather than binding DNA directly their cellular function might entail blocking the attachment of the canonical Asn bZIPs to DNA by forming heterodimers. Another possibility is that DNA binding by the Cys types may depend on that residue's oxidation state, which would be consistent with our discovery that those bZIPs help defend against oxidative stress [11, 96].

Our phylogenetic analyses revealed that TFs within a cluster often had similar binding specificities, in line with findings from other taxa [37, 97]. Cases where clustered TFs bound distinct sequences could often be explained by small variations in their DNA-binding regions. PITG_05989 and PITG_19851, for example, differ in one of the eleven residues in each of their R1 and R2 Myb domains that are thought to bind DNA [98].

TFs in a family that bound the same motif often had distinct patterns of expression consistent with subfunctionalization after gene duplication. This transcriptional divergence may have resulted from incomplete duplication of the promoter, small mutations, or acquisition of a new regulatory site [99, 100]. For functionally equivalent paralogs, a new expression profile might serve to fine-tune mRNA levels through the life cycle, giving a basis for retaining the duplicated gene. The new pattern might also be beneficial if changes outside the DNA-binding domain conferred a new function. For example, while C2H2 proteins PITG_10815 and PITG_14515 target the same DNA sequence and have similar DNA binding domains, their 200-amino acid C-termini lack similarity. These might associate with different proteins or cofactors as described for paralogs in other systems [101, 102].

Non-paralogous TFs also often had similar predicted binding specificities or targets with staggered overlaps. One such example involves the MADS-box protein PITG_07059 and Brinker protein PITG_19429. Such proteins may compete for binding, switching the transcription pattern of a target gene [103]. The interaction may also be synergistic, enhancing transcription through an assisted loading model [104].

Prior studies showed that TF orthologs within a taxonomic group such as plants, animals, or fungi often have similar binding specificities [37, 88, 105]. However, few other than HSFs have been shown to bind similar targets across kingdoms [106]. Our data have thus extended knowledge of binding site conservation across long evolutionary distances by examining oomycetes, which lack taxonomic affinity with traditional model organisms [10]. Members of five *P. infestans* families (bZIP, E2F, HSF, Myb, MADS-box) targeted motifs resembling those of relatives from plants and/or humans. Sometimes the sequences were nearly identical, as with the human and *P. infestans* MADS-box proteins. More often there was only partial overlap as with a subset of the *P. infestans* bZIPs where only an ACGT core was shared across kingdoms. Interestingly, while the *P. infestans* HSF proteins bound the canonical TTC motif, a majority lacked an obvious coiled-coil domain that is central to their function in other kingdoms [72–74, 107].

Besides assigning binding sites to the *P. infestans* TFs, many were linked to specific patterns of transcription based on their relative representation in promoters. While similar approaches have been employed in other species [65, 108], the method can be challenging. Statistical significance may be hard to achieve due to noise from binding site degeneracy or if a TF regulates only a few genes. Another complication is that an expression pattern may be determined by several TFs operating independently, multiple regulators working in concert through heterodimerization or as part of a TF cascade,

or TFs acting *in trans* by co-opting cofactors. Thus, additional strategies will be needed to link many TFs to their cellular targets.

Conclusions

The databases of transcription factors and their binding specificities yielded by this study are foundations for future explorations of the evolution of TFs specificity and function across diverse eukaryotic groups. While some features have been well-conserved during the eukaryotic radiation such as binding sites of MADS proteins, others have varied through the emergence of new binding sites or structures such as the *P. infestans* bZIPs containing Cys in their DNA-binding domain and HSF factors lacking canonical oligomerization regions. Our data will also help connect *P. infestans* TFs with their genic targets, providing insight into the regulation of transcription and life-stage transitions. This will help reveal the molecular components of processes such as sporulation, germination, and host colonization that are central to plant disease and potentially defeatable by targeted inhibitors.

Abbreviations

bHLH	basic Helix-Loop-Helix
bZIP	basic Leucine Zipper Domain
ChIP-seq	Chromatin Immunoprecipitation sequencing
RM ANOVA	Repeated measures analysis of variance
CPM	Counts Per Million
CRM	*Cis*-Regulatory Module
EMSA	Electrophoretic Mobility Shift Assay
FPKM	Fragments Per Kilobase of Transcript Per Million Mapped Reads
HSF	Heat Shock Factor
HTH	Helix-Turn-Helix
PBM	Protein-Binding Microarray
TF	Transcription Factor
TMM	Trimmed Mean of M values

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10630-6.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Author contributions
Conceptualization: NV, HJ, and TH. Experimentation: NV, AY, WL, YL, HJ. Original manuscript draft: NV. Manuscript revisions: NV, AY, TH, HJ. All authors approved the final manuscript.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
¹Department of Microbiology and Plant Pathology, University of California, Riverside, CA 92521, USA
²Department of Molecular Genetics and Donnelly Center, University of Toronto, Toronto, ON M5S 3E1, Canada
³Current address: Department of Plant Pathology, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand

References
1. Latchman DS. Eukaryotic transcription factors. 5th ed. San Diego: Academic; 2008.
2. Lambert SA, Yang AWH, Sasse A, Cowley G, Albu M, Caddick MX, Morris QD, Weirauch MT, Hughes TR. Similarity regression predicts evolution of transcription factor sequence specificity. Nat Genet. 2019;51:981–9.
3. de Mendoza A, Sebe-Pedros A, Sestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci USA. 2013;110:E4858–66.
4. Krieger G, Lupo O, Wittkopp P, Barkai N. Evolution of transcription factor binding through sequence variations and turnover of binding sites. Genome Res. 2022;32:1099–111.
5. Gera T, Jonas F, More R, Barkai N. Evolution of binding preferences among whole-genome duplicated transcription factors. Elife. 2022;11:e73225.
6. Reiter F, Wienerroither S, Stark A. Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev. 2017;43:73–81.
7. Llorca CM, Berendzen KW, Malik WA, Mahn S, Piepho HP, Zentgraf U. The elucidation of the interactome of 16 *Arabidopsis* bZIP factors reveals three independent functional networks. PLoS ONE. 2015;10:e0139884.
8. Andrasi N, Pettko-Szandtner A, Szabados L. Diversity of plant heat shock factors: regulation, interactions, and functions. J Exp Bot. 2021;72:1558–75.
9. Tognon M, Giugno R, Pinello L. A survey on algorithms to characterize transcription factor binding sites. Brief Bioinform. 2023;24:1–16.
10. Adl SM, Simpson AG, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 2005;52:399–451.
11. Gamboa-Melendez H, Huerta AI, Judelson HS. bZIP transcription factors in the oomycete *Phytophthora infestans* with novel DNA-binding domains are involved in defense against oxidative stress. Eukaryot Cell. 2013;12:1403–12.
12. Leesutthiphonchai W, Judelson HS. A MADS-box transcription factor regulates a central step in sporulation of the oomycete *Phytophthora infestans*. Mol Microbiol. 2018;110:562–75.
13. Xiang Q, Judelson HS. Myb transcription factors and light regulate sporulation in the oomycete *Phytophthora infestans*. PLoS ONE. 2014;9:e92086.
14. Lin L, Ye WW, Wu JW, Xuan MR, Li YF, Gao J, Wang YL, Wang Y, Dong SM, Wang YC. The MADS-box transcription factor PsMAD1 is involved in zoosporogenesis and pathogenesis of *Phytophthora sojae*. Front Microbiol. 2018;9:2259.
15. Yuen J. Pathogens which threaten food security: *Phytophthora infestans*, the potato late blight pathogen. Food Secur. 2021;13:247–53.

Vo *et al. BMC Genomics*   (2024) 25:710

Page 17 of 18

16. Ah-Fong A, Boyd AM, Matson MEH, Judelson HS. A Cas12a-based gene editing system for *Phytophthora infestans* reveals monoallelic expression of an elicitor. Molec Plant Pathol. 2021;22:737–52.

17. Ah-Fong AM, Judelson HS. Vectors for fluorescent protein tagging in *Phytophthora*: tools for functional genomics and cell biology. Fungal Biol. 2011;115:882–90.

18. Ah Fong A, Xiang Q, Judelson HS. Architecture of the sporulation-specific Cdc14 promoter from the oomycete *Phytophthora infestans*. Eukaryot Cell. 2007;6:2222–30.

19. Matson MEH, Liang Q, Lonardi S, Judelson HS. Karyotype variation, spontaneous genome rearrangements affecting chemical insensitivity, and expression level polymorphisms in the plant pathogen *Phytophthora infestans* revealed using its first chromosome-scale assembly. PLoS Pathog. 2022;18:e1010869.

20. Leesutthiphonchai W, Vu AL, Ah-Fong AMV, Judelson HS. How does *Phytophthora infestans* evade control efforts? Modern insight into the late blight disease. Phytopathology. 2018;108:916–24.

21. Judelson HS. Expression and inheritance of sexual preference and selfing potential in *Phytophthora infestans*. Fungal Genet Biol. 1997;21:188–97.

22. Ah-Fong AM, Kim KS, Judelson HS. RNA-seq of life stages of the oomycete *Phytophthora infestans* reveals dynamic changes in metabolic, signal transduction, and pathogenesis genes and a major role for calcium signaling in development. BMC Genomics. 2017;18:198.

23. Ah-Fong AM, Shrivastava J, Judelson HS. Lifestyle, gene gain and loss, and transcriptional remodeling cause divergence in the transcriptomes of *Phytophthora infestans* and *Pythium ultimum* during potato tuber colonization. BMC Genomics. 2017;18:764.

24. Niu X, Ah-Fong AMV, Lopez LA, Judelson HS. Transcriptomic and proteomic analysis reveals wall-associated and glucan-degrading proteins with potential roles in *Phytophthora infestans* sexual spore development. PLoS ONE. 2018;13:e0198186.

25. Ah Fong AM, Judelson HS. Cell cycle regulator Cdc14 is expressed during sporulation but not hyphal growth in the fungus-like oomycete *Phytophthora infestans*. Molec Microbiol. 2003;50:487–94.

26. Judelson HS, Shrivastava J, Manson J. Decay of genes encoding the oomycete flagellar proteome in the downy mildew *Hyaloperonospora Arabidopsidis*. PLoS ONE. 2012;7:e47624.

27. Fabro G. Oomycete intracellular effectors: specialised weapons targeting strategic plant processes. New Phytol. 2022;233:1074–82.

28. Roy S, Kagda M, Judelson HS. Genome-wide prediction and functional validation of promoter motifs regulating gene expression in spore and infection stages of *Phytophthora infestans*. PLoS Pathog. 2013;9:e1003182.

29. Win J, Kanneganti T-D, Torto-Alalibo T, Kamoun S. Computational and comparative analyses of 150 near full-length cDNA sequences from the oomycete plant pathogen *Phytophthora infestans*. Fungal Genet Biol. 2006;43:20–33.

30. Tani S, Judelson HS. Activation of zoosporogenesis-specific genes in *Phytophthora infestans* involves a 7-nucleotide promoter motif and cold-induced membrane rigidity. Eukaryot Cell. 2006;5:745–52.

31. Roy S, Poidevin L, Jiang T, Judelson HS. Novel core promoter elements in the oomycete pathogen *Phytophthora infestans* and their influence on expression detected by genome-wide analysis. BMC Genomics. 2013;14:106.

32. Seidl MF, Wang R-P, Van den Ackerveken G, Govers F, Snel B. Bioinformatic inference of specific and general transcription factor binding sites in the plant pathogen *Phytophthora infestans*. PLoS ONE. 2013;7:e51295.

33. Narasimhan K, Lambert SA, Yang AWH, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi HS, Reece-Hoyes JS, Bass JIF, et al. Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities. Elife. 2015;4:e06967.

34. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol. 2013;31:126–34.

35. Birchler JA, Yang H. The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell. 2022;34:2466–74.

36. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature. 2009;461:393–8.

37. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.

38. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, Bileschi ML, Bork P, Bridge A, Colwell L, et al. InterPro in 2022. Nucleic Acids Res. 2022;51:D418–27.

39. Abrahamian M, Ah-Fong AM, Davis C, Andreeva K, Judelson HS. Gene expression and silencing studies in *Phytophthora infestans* reveal infection-specific nutrient transporters and a role for the nitrate reductase pathway in plant pathogenesis. PLoS Path. 2016;12:e1006097.

40. Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert CJ, Kissinger JC, et al. FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi. 2018;4:39.

41. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics. 2019;35:2790–5.

42. Simm D, Hatje K, Kollmar M. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α-helices (SAH domains). Bioinformatics. 2015;31:767–9.

43. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27:221–4.

44. Backman TWH, Girke T, systemPipeR. NGS workflow and report generation environment. BMC Bioinformatics. 2016;17:388.

45. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

46. Waskom ML. Seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021.

47. Lam KN, van Bakel H, Cote AG, van der Ven A, Hughes TR. Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucleic Acids Res. 2011;39:4680–90.

48. Berger MF, Philippakis AA, Qureshi AM, He FXS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006;24:1429–35.

49. Heffler MA, Walters RD, Kugel JF. Using electrophoretic mobility shift assays to measure equilibrium dissociation constants: GAL4-p53 binding DNA as a model system. Biochem Mol Biol Edu. 2012;40:383–7.

50. Schindelin J, Rueden CT, Hiner MC, Eliceiri KW. The ImageJ ecosystem: an open platform for biomedical image analysis. Mol Reprod Dev. 2015;82:518–29.

51. Ah-Fong AM, Bormann-Chung CA, Judelson HS. Optimization of transgene-mediated silencing in *Phytophthora infestans* and its association with small-interfering RNAs. Fungal Genet Biol. 2008;45:1197–205.

52. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.

53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–U354.

54. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molec Cell. 2010;38:576–89.

55. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37:2834–40.

56. Zhang KX, Tarczykowska A, Gupta DK, Pendlebury DF, Zuckerman C, Nandakumar J, Shibuya H. The TERB1 MYB domain suppresses telomere erosion in meiotic prophase I. Cell Rep. 2022;38:110289.

57. Judelson HS. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. Eukaryot Cell. 2012;11:1304–12.

58. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009;4:393–411.

59. Perisic O, Xiao H, Lis JT. Stable binding of *Drosophila* Heat-Shock factor to head-to-head and tail-to-tail repeats of a conserved 5-bp recognition unit. Cell. 1989;59:797–806.

60. Santoro N, Johansson N, Thiele DJ. Heat shock element architecture is an important determinant in the temperature and transactivation domain requirements for heat shock transcription factor. Mol Cell Biol. 1998;18:6340–52.

61. Yamamoto N, Takemori Y, Sakurai M, Sugiyama K, Sakurai H. Differential recognition of heat shock elements by members of the heat shock transcription factor family. FEBS J. 2009;276:1962–74.

62. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei GH, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
63. Yamamoto A, Mizukami Y, Sakurai H. Identification of a novel class of target genes and a novel type of binding sequence of heat shock transcription factor in *Saccharomyces cerevisiae*. J Biol Chem. 2005;280:11911–9.
64. Fujimoto M, Nakai A. The heat shock factor family and adaptation to proteotoxic stress. FEBS J. 2010;277:4112–25.
65. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci USA. 2014;111:2367–72.
66. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Perez NM, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022;50:D165–73.
67. O'Malley RC, Huang SSC, 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;165:1280–92.
68. Jaeger AM, Makley LN, Gestwicki JE, Thiele DJ. Genomic heat shock element sequences drive cooperative human heat shock factor 1 DNA binding and selectivity. J Biol Chem. 2014;289:30459–69.
69. Xiao H, Perisic O, Lis JT. Cooperative binding of *Drosophila* Heat-Shock factor to arrays of a conserved 5 Bp unit. Cell. 1991;64:585–93.
70. Zhong M, Kim SJ, Wu C. Sensitivity of *Drosophila* heat shock transcription factor to low pH. J Biol Chem. 1999;274:3135–40.
71. Yu EM, Yoshinaga T, Jalufka FL, Ehsan H, Welch DBM, Kaneko G. The complex evolution of the metazoan HSP70 gene family. Sci Rep. 2021;11:17794.
72. Scharf KD, Berberich T, Ebersberger I, Nover L. The plant heat stress transcription factor (hsf) family: structure, function and evolution. Bba-Gene Regul Mech. 2012;1819:104–19.
73. Sorger PK, Nelson HCM. Trimerization of a yeast transcriptional activator via a coiled-coil motif. Cell. 1989;59:807–13.
74. Rabindran SK, Haroun RI, Clos J, Wisniewski J, Wu C. Regulation of heat-shock factor trimer formation - role of a conserved leucine zipper. Science. 1993;259:230–4.
75. Guo M, Liu JH, Ma X, Luo DX, Gong ZH, Lu MH. The plant heat stress transcription factors (HSFs): structure, regulation, and function in response to abiotic stresses. Front Plant Sci. 2016;7:114.
76. Luo XM, Tian TT, Bonnave M, Tan X, Huang XQ, Li ZG, Ren MZ. The molecular mechanisms of *Phytophthora infestans* in response to reactive oxygen species stress. Phytopathology. 2021;111:2067–79.
77. Suckow M, Schwamborn K, Kisterswoike B, Vonwilckenbergmann B, Mullerhill B. Replacement of invariant bzip residues within the basic region of the yeast transcriptional activator Gcn4 can change its DNA-binding specificity. Nucleic Acids Res. 1994;22:4395–404.
78. Fiore A, Liang Y, Lin YH, Tung J, Wang HC, Langlais D, Nijnik A. Deubiquitinase MYSM1 in the hematopoietic system and beyond: a current review. Int J Mol Sci. 2020;21:3007.
79. Jiang MQ, Sun LF, Isupov MN, Littlechild JA, Wu XL, Wang QC, Wang Q, Yang WD, Wu YK. Structural basis for the target DNA recognition and binding by the MYB domain of phosphate starvation response 1. FEBS J. 2019;286:2809–21.
80. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000;29:183–212.
81. Dathan N, Zaccaro L, Esposito S, Isernia C, Omichinski JG, Riccio A, Pedone C, Di Blasio B, Fattorusso R, Pedone PV. The *Arabidopsis* SUPERMAN protein is able to specifically bind DNA through its single Cys2-His2 zinc finger motif. Nucleic Acids Res. 2002;30:4945–51.
82. Noguero M, Atif RM, Ochatt S, Thompson RD. The role of the DNA-binding one zinc finger (DOF) transcription factor family in plants. Plant Sci. 2013;209:32–45.
83. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, Albu M, Weirauch MT, Radovani E, Kim PM, et al. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotechnol. 2015;33:555–U181.
84. Apone S, Hauschka SD. Muscle gene E-box control elements - evidence for quantitatively different transcriptional activities and the binding of distinct regulatory factors. J Biol Chem. 1995;270:21420–7.
85. Gamba R, Fachinetti D. From evolution to function: two sides of the same CENP-B coin? Exp Cell Res. 2020;390:111959.
86. Siegmund T, Lehmann M. The Drosophila pipsqueak protein defines a new family of helix-turn-helix DNA-binding proteins. Dev Genes Evol. 2002;212:152–7.
87. Hall SE, Kettler G, Preuss D. Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. Genome Res. 2003;13:195–205.
88. Nitta KR, Jolma A, Yin YM, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Elife. 2015;4:e04837.
89. Marand AP, Eveland AL, Kaufmann K, Springer NM. cis-Regulatory elements in plant development, adaptation, and evolution. Annu Rev Plant Biol. 2023;74:111–37.
90. Ni P, Wilson D, Su Z. A map of cis-regulatory modules and constituent transcription factor binding sites in 80% of the mouse genome. BMC Genomics. 2022;23:714.
91. Grant CE, Johnson J, Bailey TL, Noble WS. MCAST: scanning for cis-regulatory motif clusters. Bioinformatics. 2016;32:1217–9.
92. Coustry F, Maity SN, Decrombrugghe B. Studies on transcription activation by the multimeric CCAAT-binding factor Cbf. J Biol Chem. 1995;270:468–75.
93. Mendoza CS, Findlay AC, Judelson H. An LbCas12a variant and elevated incubation temperatures enhance the rate of gene editing in the oomycete *Phytophthora infestans*. Mol Plant Microbe Interact. 2023;in press.
94. Chen H, Shu H, Fang Y-F, Song W, Zhi L, Fang Y-J, Wang Y, Dong S. Genome-wide analyses of histone modifications and chromatin accessibility reveal the distinct genomic compartments in the Irish potato famine pathogen *Phytophthora infestans*. Biorxiv. 2022. https://doi.org/10.1101/2022.02.18.480484.
95. Zwiers LH, De Waard MA. Characterization of the ABC transporter genes MgAtr1 and MgAtr2 from the wheat pathogen Mycosphaerella graminicola. Fungal Genet Biol. 2000;30:115–25.
96. Yin Z, Machius M, Nestler EJ, Rudenko G. Activator Protein-1: redox switch controlling structure and DNA-binding. Nucleic Acids Res. 2017;45:11425–36.
97. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu Rev Cell Dev Biol. 2019;35:357–79.
98. Wang BH, Luo Q, Li YP, Yin LF, Zhou NN, Li XN, Gan JH, Dong AW. Structural insights into target DNA recognition by R2R3-MYB transcription factors. Nucleic Acids Res. 2020;48:460–71.
99. Huminiecki L, Wolfe KH. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. 2004;14:1870–9.
100. Bailon-Zambrano R, Sucharov J, Mumme-Monheit A, Murry M, Stenzel A, Pulvino AT, Mitchell JM, Colborn KL, Nichols JT. Variable paralog expression underlies phenotype variation. Elife. 2022;11:e79247.
101. Feng S, Rastogig C, Loker R, Glassford WJ, Rube HT, Bussemaker HJ, Mann RS. Transcription factor paralogs orchestrate alternative gene regulatory networks by context-dependent cooperation with multiple cofactors. Nat Commun. 2022;13:3808.
102. Reece-Hoyes JS, Pons C, Diallo A, Mori A, Shrestha S, Kadreppa S, Nelson J, DiPrima S, Dricot A, Lajoie BR, et al. Extensive rewiring and complex evolutionary dynamics in a multiparameter transcription factor network. Mol Cell. 2013;51:116–27.
103. Liu N, Xu SQ, Yao QM, Zhu Q, Kai Y, Hsu JY, Sakon P, Pinello L, Yuan GC, Bauer DE, et al. Transcription factor competition at the gamma-globin promoters controls hemoglobin switching. Nat Genet. 2021;53:511–20.
104. Goldberg D, Charni-Natan M, Buchshtab N, Bar-Shimon M, Goldstein I. Hormone-controlled cooperative binding of transcription factors drives synergistic induction of fasting-regulated genes. Nucleic Acids Res. 2022;50:5528–44.
105. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol. 2004;2:2202–19.
106. Sakurai H, Enoki Y. Novel aspects of heat shock factors: DNA recognition, chromatin modulation and gene expression. FEBS J. 2010;277:4140–9.
107. Kmiecik SW, Le Breton L, Mayer MP. Feedback regulation of heat shock factor 1 (Hsf1) activity by Hsp70-mediated trimer unzipping and dissociation from DNA. EMBO J. 2020;39.
108. Rubin JD, Stanley JT, Sigauke RF, Levandowski CB, Maas ZL, Westfall J, Taatjes DJ, Dowell RD. Transcription factor enrichment analysis (TFEA) quantifies the activity of multiple transcription factors from a single experiment. Commun Biol. 2021;4:661.

Publisher's Note