BMC Genomics

Research article

Open Access

Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome

Claudia C Englbrecht[1], Heiko Schoof[2] and Siegfried Böhm*[3]

Address: [1]GSF-Forschungszentrum für Umwelt und Gesundheit, Institut für Bioinformatik, Ingolstädter Landstrasse1, D-85764 Neuherberg, Germany, [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85354 Freising-Weihenstephan, Germany and [3]Max-Delbrück-Centrum für Molekulare Medizin, Department of Genetics, Bioinformatics and Structural Biology, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany

Email: Claudia C Englbrecht - c.englbrecht@gsf.de; Heiko Schoof - h.schoof@gsf.de; Siegfried Böhm* - boehm@mdc-berlin.de

* Corresponding author

Published: 05 July 2004

BMC Genomics 2004, **5**:39 doi:10.1186/1471-2164-5-39

Received: 16 March 2004

Accepted: 05 July 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/39

Background: The classical C2H2 zinc finger domain is involved in a wide range of functions and can bind to DNA, RNA and proteins. The comparison of zinc finger proteins in several eukaryotes has shown that there is a lot of lineage specific diversification and expansion. Although the number of characterized plant proteins that carry the classical C2H2 zinc finger motifs is growing, a systematic classification and analysis of a plant genome zinc finger gene set is lacking.

Results: We found through *in silico* analysis 176 zinc finger proteins in *Arabidopsis thaliana* that hence constitute the most abundant family of putative transcriptional regulators in this plant. Only a minority of 33 *A. thaliana* zinc finger proteins are conserved in other eukaryotes. In contrast, the majority of these proteins (81%) are plant specific. They are derived from extensive duplication events and form expanded families. We assigned the proteins to different subgroups and families and focused specifically on the two largest and evolutionarily youngest families (A1 and C1) that are suggested to be primarily involved in transcriptional regulation. The newly defined family A1 (24 members) comprises proteins with tandemly arranged zinc finger domains. Family C1 (64 members), earlier described as the EPF-family in *Petunia*, comprises proteins with one isolated or two to five dispersed fingers and a mostly invariant QALGGH motif in the zinc finger helices. Based on the amino acid pattern in these helices we could describe five different signature sequences prevalent in C1 zinc finger domains. We also found a number of non-finger domains that are conserved in these families.

Conclusions: Our analysis of the few evolutionarily conserved zinc finger proteins of *A. thaliana* suggests that most of them could be involved in ancient biological processes like RNA metabolism and chromatin-remodeling. In contrast, the majority of the unique *A. thaliana* zinc finger proteins are known or suggested to be involved in transcriptional regulation. They exhibit remarkable differences in the features of their zinc finger sequences and zinc finger arrangements compared to animal zinc finger proteins. The different zinc finger helix signatures we found in family C1 may have important implications for the sequence specific DNA recognition and allow inferences about the evolution of the members in this family.

Background

C2H2 zinc finger proteins (ZFPs) constitute an abundant family of nucleic acid binding proteins in the genomes of higher and lower eukaryotes. The number of ZFPs identified by *in silico* analysis corresponds to ~2.3 and ~3% of all genes in diptera and mammalia, respectively [1,2]. Approximately 0.8% of the proteins in *Saccharomyces cerevisiae* [3] have C2H2 zinc finger domains and about 0.7% in *Arabidopsis thaliana* (this paper). C2H2 zinc fingers (ZF) display a wide range of functions, from DNA or RNA binding to the involvement in protein-protein interactions. Therefore ZFPs not only act in transcriptional regulation, either directly or through site-specific modification and/or regulation of chromatin, but also participate in RNA metabolism and in other cellular functions that probably require specific protein contacts of the ZF domain. In addition, the comparison of the whole ZFP sets in major eukaryotic lineages has revealed a remarkable level of complexity through lineage specific diversification and expansion. These expansions often include ZFPs that contain conserved lineage specific non-finger domains like the vertebrate specific KRAB domain (reviewed in [4]) or the ZAD domain specific to diptera [1]. These domains are protein interaction domains with known or suggested repressor functions. Several ZFPs in plants, e.g. *Arabidopsis* and *Petunia*, have already been functionally characterized. They are involved in a variety of processes such as the regulation of floral organogenesis, leaf initiation, lateral shoot initiation, gametogenesis and stress response. Former reviews on plant-ZFPs [5] have been limited to approximately 30 proteins. But the systematic analysis of a complete ZFP set of a plant genome with the aim to predict some basic molecular functions is lacking. The genome annotation of the model plant *Arabidopsis thaliana* has reached high quality and allows comprehensive computational analyses. Here we describe the classification of the full set of ZFPs in the *Arabidopsis* genome including a genome-wide comparative analysis based on the *in silico* analysis of the whole proteome of this plant.

Results

General classification and characterization of ZFPs in the Arabidopsis *genome*

In *A. thaliana* we found altogether 176 proteins that contain one or more ZF domains (Table S1 [see Additional File 1]) which by far exceeds the previously reported numbers by Riechmann et al. [6,7]. Therefore, according to our estimate, *A. thaliana* ZFPs (AT-ZFPs) constitute the most abundant family of putative transcriptional regulators in *A. thaliana*. So far most studies on the DNA recognition of ZFPs have been carried out on proteins with tandem arrays of fingers (reviewed in [8]). In the genomes of animals, these types of ZFPs (classified as sets A and B, see Methods) constitute the majority (about two thirds) of all ZFPs (S.B. unpublished data). In striking contrast to the animal kingdom, only a minority of 33 AT-ZFPs (about 20%) contain tandem ZF arrays with 32 proteins in set A, containing up to five ZF, and one protein (TF3A) in set B (Table 1), containing nine ZF in more than one array. The vast majority of AT-ZFPs (about 80%) contain a single ZF or several dispersed ZFs (classified as set C). Set C can be further classified into three clearly distinguishable subsets, C1, C2 and C3 (Table 1). These subsets are characterized by ZF types that differ in their spacing between the two invariant zinc coordinating histidine residues by three (C1), four (C2) or five (C3) amino acid residues. A complete list of all classified 176 AT-ZFPs is shown in the supplementary Table S1 [see Additional File 1] including additional data and links to database information (MatDB, TIGR). Pairwise sequence comparisons of all ZF domains found showed that pairwise distances (PAM Dayhoff matrix, see Methods) among domains of subsets A1, C1, C2 and C3 varied and were lowest in A1 and C1 (1,02 and 1,27) and twice as high in C2 and C3 (2,35 and 2,57). This suggests that ZF domains of subsets A1 and C1

Table 1: Overview of conserved and unique *Arabidopsis* ZFPs in different subsets. Maximum values are in boldface. Unique ZFPs are further classified into families, pairs and singletons. C1S and C2S belong to one family but contain ZFs both with HX3H and HX4H spacing, an asterisk marks a pair with the exceptional combination of ZFs with HX4H and HX3H spacing in the first and second finger, respectively. The most expanded families, A1 and C1 are in italics.

Classified set	Set A+B	SubsetC1	SubsetC2	SubsetC3	∑ all Sets
Conserved AT-ZFPs	6 (18%)	2 (3%)	9 (20%)	**16 (73%)**	33 (19%)
unique AT-ZFPs	**27 (82%)**	**75 (97%)**	**35 (80%)**	6 (27%)	**143 (81%)**
- ZFPs in families	*A1 24*	*C1 64* C1Q 4 C1S 3	C2A 6 C2B 17 C2S 2	C3A 3 C3B 3	126
- ZFPs in pairs	2	2*	8	-	12
- Singletons	1	2	2	-	5
∑ all ZFPs	33	77	44	22	176

are younger than those of C2 and C3 and result from a more recent expansion.

Conserved AT-ZFPs

BLAST analysis against the non-redundant database of NCBI resulted in 33 AT-ZFPs that are conserved in other taxa. The proportion of conserved AT-ZFPs varies remarkably between the different subsets (Table 1), with minor proportions in sets A/B (6 of 33), C1 (2 of 77) and C2 (9 of 44), but a major proportion in set C3 (16 of 22). The 33 conserved AT-ZFPs and their assignment to 27 evolutionarily conserved ZFP families are highlighted in Table S1 [see Additional File 1]. The family names given in the table have either been published already (Table 2) or are proposed by us. About half of the ZFP families are new (SF1-SF13) and do not contain any functionally characterized members in any species. Only six of the 33 conserved AT-ZFPs have so far been functionally characterized in *Arabidopsis* or other plants. They will be discussed below together with our functional predictions for some of the conserved AT-ZFPs that have so far not been experimentally investigated.

Sets A and B

The first TF3A homolog in a plant (At1g72050) has recently been cloned and characterized [9]. AT-TF3A is highly diverged in sequence from TF3As in vertebrates, but it binds specifically to 5S rRNA and 5S rDNA as shown in [9]. Interestingly, the nine fingers of AT-TF3A are arranged differently from animal TF3A. Instead of having the nine fingers in a single array, the first finger is isolated and fingers 2–4 and 5–9 are arranged in two separate tandem arrays. We found two other plant orthologs of AT-TF3A in *Oryza sativa* and *Medicago truncatula* (data not shown). One additional conserved AT-ZFP has tandem ZFs with the typical characteristics of a DNA binding ZFP of animals. This ZFP, At4g06634, contains four tandem ZFs and is, like its putative ortholog TRM1 from *Zea maize*, a member of the YY1-family (Table 2). TRM1 was recently cloned and characterized as suppressor of *rbcS-m3* [10], a gene involved in photosynthetic CO_2 fixation. It was shown in [10] that TRM1 binds to an YY1-like DNA site and to two other regions with no homology to the YY1 site. Since TRM1 and At4g06634 are very conserved in sequence, we suggest that At4g06634 probably has similar DNA target sites. No functions have been described so far for any of the remaining four conserved AT-ZFPs (Table S1 [see Additional File 1], newSF1 and 2, ZP207-SF) in set A or for their homologs in other eukaryotes. They contain tandem fingers with rare HX4H spacing and have unusually short linkers (one or no residue), features that are conserved in the eukaryotic homologs.

Subset C1

Subset C1 comprises 77 ZFPs containing ZFs with HX3H or HX3C spacing. Only two of them, At5g09740 and At5g64610, are also conserved in other kingdoms. They arose through duplication and belong to the SAS-MOZ family (Table 2). Both ZFPs contain the conserved combination of a single ZF with CHROMO and SAS domains (Table 2). Therefore we suggest that they have a function in histone acetylation (HAT), a key process in chromatin-remodeling (reviewed in [11]).

Subset C2

In subset C2, there are more ZFPs that are involved in chromatin-remodeling processes and are conserved between plants and animals. These are VERNALIZATION 2 (VRN2), EMBRYONIC FLOWER2 (EMF2) and FERTILIZATION-INDEPENDENT SEED (FIS2) [12-15]. They belong to the Polycomb group (PcG) and were given the name VEF family. As first described in [16] the ZF domains and other non-finger parts of these three AT-ZFPs are conserved in the Su(z)12 proteins of *Drosophila* and human. PcG proteins are required to maintain the transcriptionally repressed state of homeotic genes throughout development. The molecular function(s) of their single ZF are unknown, but data from *Drosophila* Su(z)12 suggest their involvement in specific protein contacts, but not DNA binding. FIS2 [12,13], VRN2 [14] and EMF2 [15] act as repressors in different developmental stages of *Arabidopsis* (reviewed in [11]). Another conserved AT-ZFP in subset C2 is the protein SERRATE (SE, At2g27100) [17]. The phenotype of the SE mutant reveals a role of the affected protein in the early steps of organ elaboration and a role in the regulation of gene expression via chromatin modification was also suggested [17]. We assigned the proteins At5g01160 and At3g12270 to the evolutionarily conserved E7 and PRMT families (Table 2), respectively, based on their conserved combination of a single ZF with a RING and a PRMT3 domain.

Subset C3

16 out of the 22 ZFPs in subset C3 are conserved in other eukaryotes. There is no information available regarding the function of these proteins. Interestingly, we found that eleven of them are predicted to have a U1 type ZF (Table S1 [see Additional File 1], Table 2) that contains conserved extensions on both sides of the ZF which has a HX5H spacing. They are known or suggested to be involved in RNA binding and therefore we suggest that these conserved AT-ZFPs with U1 type ZFs could be involved in RNA metabolic processes, e.g. splicing like the splicosome associated proteins of the SAP62 family (Table 2), corresponding to At2g32600 in *Arabidopsis.* According to our prediction we found, that other conserved AT-ZFPs with U1 type ZFs are combined with domains, like DNAJ (At1g74250), KOW (At1g55460) or

Table 2: Description of acronyms: Description and Pfam/SMART accession numbers (if available) of acronyms used throughout the text and in Table S1 are given. Column "conserved" refers to classification in conserved ZF-families (ZF-fam) and/or non-finger domains (NFD).

Acronym	Description	Pfam	SMART	conserved
ANK	Ankyrin repeat	PF00023	SM00248	NFD
C3H	C3H1 zinc finger	PF00642	SM00356	NFD
CHROMO	CHRomatin Organisation MOdifier	PF00385	SM00298	NFD domain
DNAJ	Dnaj molecular chaperone domain	PF00226	SM00271	NFD/ZF-fam
DUF537	Plant family of unknown function	PF04396	-	NFD
DUF627	Plant family of unknown function	PF04781	-	NFD
DUF629	Plant family of unknown function	PF04780	-	NFD
E7	E7cadherine binding protein family	-	-	ZF-fam
EXO3	Exonuclease domain	PF00929	SM00479	NFD
FZFnew	New ZFP family FZF	-	-	ZF-fam
G-patch	G-patch domain	PF01585	SM00443	NFD
HLH	Helix-loop-helix domain	PF00010	SM00353	NFD
IPPT	IPP transferase family	PF01715	-	NFD/ZF-fam
JmjC	Jumonji domain, C-terminal part	PF02373	SM00558	NFD
JmjN	Jumonji domain, N-terminal part	PF02375	SM00545	NFD
KIN17	KIN17 family	-	-	ZF-fam
KOW	Kyprides, Ouzounis, Woese motif	PF00467	SM00739	NFD
MBD	Methyl-CpG binding domain	PF01429	SM00391	NFD
OTU	OTU-like cysteine protease	PF02338	-	NFD
PHD	PHD zinc finger	PF00628	SM00249	NFD
PreSet	N-terminal to some SET domains	-	SM00468	NFD
PRMT3	Ribosomal protein L11 methyl-transferase (PrmA)family	PF06325	-	NFD/ZF-fam
PUG	Domain in protein kinases	-	SM00580	NFD
RING	zf-C3HC4 or RING finger	PF00097	SM00184	NFD
RRM	RNA recognition motif	PF00076	SM00360	NFD
SAP61	Splicosome associated protein 61/PRP9 family	-	-	ZF-fam
SAP62	Splicosome associated protein 62/PRP11 family	-	-	ZF-fam
SAS	MOZ/SAS family	PF01853	-	NFD/ZF-fam
SE	Serrate family	-	-	ZF-fam
SET	SET domain	PF00856	SM00317	NFD
TF3A	Transcription factor IIIA family	-	-	ZF-fam
U1-ZF	U1-like zinc finger	PF06220	SM00451	NFD
UBA	UBA/TS-N domain (ubiquitin associated domain)	PF00627	SM00165	NFD
UBX	UBX domain	PF00789	SM00166	NFD
UPF0023	Uncharacterized protein family	PF01172	-	NFD
VEF	VRN2-EMF2-FIS2 family	-	-	ZF-fam
YY1	Ying Yang 1 family	-	-	ZF-fam
ZF-AN1	AN1-like zinc finger	PF01428	SM00154	NFD
ZP207SF	ZFP207 family	-	-	ZF-fam

G-patch (At5g26610) (Table 2), which are also known to be involved in different RNA metabolic processes (reviewed in [18]). Like chromatin-remodeling, many pathways of the RNA metabolism are ancient, conserved processes in eukaryotes, which is reflected by our finding that several ZFPs that are described above are evolutionarily conserved in all eukaryotic taxa from Protozoa to Mammalia (data not shown).

AT-specific ZFPs and their families

BLAST analysis resulted in the assignment of the 143 AT-specific ZFPs to eight families, six pairs and to five single occurrences. This distribution reflects the high incidence of duplication events in the *Arabidopsis* genome. The two largest families, named A1 and C1, contain 24 and 64 members, respectively. Together they constitute about 60% of all AT-specific ZFPs. Additional data on other AT-specific ZFP families and pairs with uncharacterized members are given in Table S1 [see Additional File 1]. The C1 family is part of subset C1. We investigated the two biggest families in more detail.

The A1 family

General description

The A1 family, with 24 members, represents by far the most expanded AT-ZFP family of set A. All ZFPs of the A1 family contain four similar ZFs with a conserved arrangement as shown in Figures 1 and 3. They are classified by us into four different subgroups named A1a, A1b, A1c and A1d with 13, three, two and six members, respectively (Figures 1 and 2). These assignments are new and now group plant ZFPs described in the past as belonging to different groups or families. Among them are StPCP1 from *Solanum tuberosum* [19] and ZmID1 from *Zea mays* [20,21], which are members of the A1a subgroup. Also part of this group are ZFPs of the recently described WIP-subfamily [22] which is identical to the subgroup A1d that contains six members. So far three members of family A1 have been studied in more detail and shown to be quite variable in function. StPCP1 was identified by its ability to confer growth on sucrose as the sole carbon source upon a sucrose uptake-deficient yeast strain [19]. ZmId1, the indeterminate1 (*id1*) gene, was the first example of a gene other than photo-receptors that is involved in the production or transmission of a flowering signal [20,21]. In the *id1* maize mutant the terminal shoot meristem continues to display vegetative (i.e. indeterminate) growth. One characterized member of the WIP-subfamily or subgroup A1d is transparent testa 1 (*tt1*, At1g34790), a gene that has been found to be involved in seed coat development in *Arabidopsis* [22]. All three studied representatives of family A1 have been suggested to be transcriptional regulators. However, no regulatory DNA sites or target genes have been reported for any of them.

ZF arrays and subgroups

All 24 members of the A1 family contain four ZFs with the conserved arrangement F1 isolated, F2–F3–F4 in tandem (Figure 3). In former reports [19,20] only the first and the third ZF, F1 and F3, have been considered, but in [22] the existence of two possible additional fingers (F2 and F4) was mentioned though not discussed in detail. Both ZFs reported earlier match the consensus ZF pattern X2CX2CX12HX3H/C and have high scores in Pfam searches. In contrast, the "new" fingers F2 and F4, have unusual ZF patterns (Figure 3) resulting in low scores. Pairwise comparisons (not shown) indicate highest similarities between members of the same subgroup with identities of 77–96 % for the subgroups A1a, A1b and A1d and 62.5 % for the two members of subgroup A1c. Comparisons between subgroups showed that similarities are about 30–40%. From the neighbor-joining tree (Figure 2) and the domain architecture (Figure 3) one can infer that A1a and A1b are more similar to each other than any of them is to A1c or d. Also, A1c and A1d are more closely related to each other than either is to A1a or A1b. Remarkably, amino acid residues in the ZF helix positions -1, 3 and 6, known as primary DNA recognition positions, are among the most conserved positions. Moreover, about half of the conserved amino acids are residues with high specificity to particular DNA bases according to a proposed ZF-DNA "recognition code" [8] which allows for the prediction of possible core bases of a DNA site (Figure 4). We emphasize that this prediction is in part speculative because the ZF-DNA recognition rules are more complex than previously assumed.

Conserved non-finger regions in the A1 family

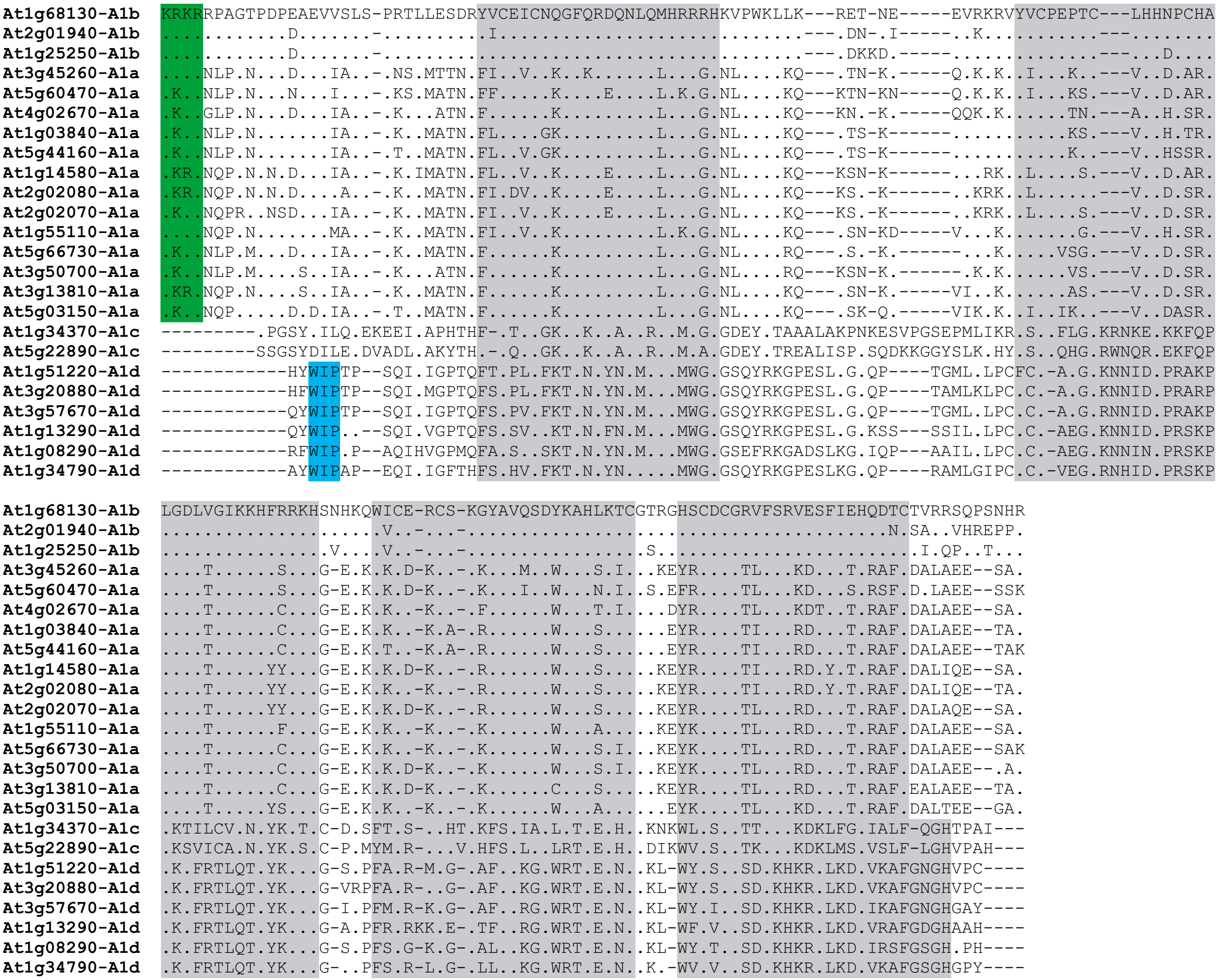
Using the expectation maximization search tool MEME (see Methods) we found additional conserved sequence parts in the regions outside of the ZF domains (Figure 3). These conserved regions vary in length and are not shared among the four subgroups, with the exception of a conserved N-terminal region of 29 residues that starts with an R/K rich stretch that is common to subgroups A1a and A1b. The basic amino acids could represent a nuclear localization signal [23]. In addition, all members of subgroup A1a contain two other conserved sequences at their C-terminus. The consensus of the first is 'SATALLQKAAQMGS', the second is characterized by the pattern 'T [R/L]DFLG [L/V]' (Figure 3). These patterns could be necessary for protein interactions or localization.

The C1 family

The C1 family represents with 64 members by far the most expanded AT-ZFP family and includes about 85% of all ZFPs in subset C1. ZFPs in the C1 family are characterized by either a single finger or a varying number (2–5) of dispersed ZFs, most of them with the conserved QALGGH sequence in their alpha-helix positions 2–7. A part of this conserved plant family was investigated in petunia and named the EPF family [24]. Based on 21 petunia-ZFPs with two, three or four dispersed ZFs, a first systematic classification of their ZF types was described in [24]. About 20 members of the *Arabidopsis* C1 family have also been described regarding their biological functions and expression characteristics [25-28]. We have subdivided family C1 according to the varying numbers of fingers as subclasses C1-1i (N = 33), C1-2i (N = 20), C1-3i (N = 8), C1-4i (N = 2) and C1-5i (N = 1).

Representatives of the different C1 subclasses

Subclass C1-1i has 33 members exclusively found in plants or specifically *A. thaliana* (Table S1 [see Additional File 1]). 28 members show the invariant motif QALGGH in the alpha-helix (Figure 5a). The best studied member of this subclass is the SUPERMAN protein (SUP, At3g23130, AtZFP9) [28]. One suggested function of SUP is the coordination of stamen- and carpel-specific meristematic cells and hence the maintenance of the boundary between whorls three and four of the floral organ [29]. The *sup* gene encodes a transcription factor with specific DNA



```
At1g68130-A1b  KRKRRPAGTPDPEAEVVSLS-PRTLLESDRYVCEICNQGFQRDQNLQMHRRRHKVPWKLLK---RET-NE-----EVRKRVYVCPEPTC---LHHNPCHA
At2g01940-A1b  ............D.......-..........I.............................---.DN-.I-----..K...........---........
At1g25250-A1b  ............D.......-........................................---.DKKD.-----..................---...D....
At3g45260-A1a  ....NLP.N...D...IA..-.NS.MTTN.FI..V..K..K......L...G.NL....KQ---.TN-K.-----Q.K.K..I...K..---V..D.AR.
At5g60470-A1a  .K..NLP.N...N...I...-.KS.MATN.FF.....K....E....L.K.G.NL....KQ---KTN-KN-----Q.K.K..I...KS.---V..D.AR.
At4g02670-A1a  .K..GLP.N...D...IA..-.K...ATN.F......K.........L...G.NL....KQ---KN.-K.-----QQK.K......TN.---A..H.SR.
At1g03840-A1a  .K..NLP.N.......IA..-.K..MATN.FL....GK.........L...G.NL....KQ---.TS-K------...........KS.---V..H.TR.
At5g44160-A1a  .K..NLP.N.......IA..-.T..MATN.FL..V.GK.........L...G.NL....KQ---.TS-K------...........K..---V..HSSR.
At1g14580-A1a  .KR.NQP.N.N.D...IA..-.K.IMATN.FL..V..K....E....L...G.NL....KQ---KSN-K------...RK..L....S.---V..D.AR.
At2g02080-A1a  .KR.NQP.N.N.D....A..-.K..MATN.FI.DV..K....E....L...G.NL....KQ---KS.-K------..KRK..L......---V..D.SR.
At2g02070-A1a  .K..NQPR..NSD...IA..-.K..MATN.FI..V..K....E....L...G.NL....KQ---KS.-K------..KRK..L....S.---V..D.SR.
At1g55110-A1a  ....NQP.N.......MA..-.K..MATN.FI..V..K.........L.K.G.NL....KQ---.SN-KD-----V...K.......G.---V..H.SR.
At5g66730-A1a  .K..NLP.M...D...IA..-.K..MATN.F......K.........L...G.NL....RQ---.S.-K.------...K.....VSG.---V..D.SR.
At3g50700-A1a  .K..NLP.M....S..IA..-.K...ATN.F......K.........L...G.NL....RQ---KSN-K.------.K.K......VS.---V..D.SR.
At3g13810-A1a  .KR.NQP.N....S..IA..-.K..MATN.F......K.........L...G.NL....KQ---.SN-K.-----VI..K......AS.---V..D.SR.
At5g03150-A1a  .K..NQP.....D.D.IA..-.T..MATN.F......K.........L...G.NL....KQ---.SK-Q.-----VIK.K..I..IK..---V..DASR.
At1g34370-A1c  ---------.PGSY.ILQ.EKEEI.APHTHF-.T..GK..K..A..R..M.G.GDEY.TAAALAKPNKESVPGSEPMLIKR.S..FLG.KRNKE.KKFQP
At5g22890-A1c  ---------SSGSYDILE.DVADL.AKYTH.-.Q..GK..K..A..R..M.A.GDEY.TREALISP.SQDKKGGYSLK.HY.S..QHG.RWNQR.EKFQP
At1g51220-A1d  ------------HYWIPTP--SQI.IGPTQFT.PL.FKT.N.YN.M...MWG.GSQYRKGPESL.G.QP----TGML.LPCFC.-A.G.KNNID.PRAKP
At3g20880-A1d  ------------HFWIPTP--SQI.MGPTQFS.PL.FKT.N.YN.M...MWG.GSQYRKGPESL.G.QP----TAMLKLPC.C.-A.G.KNNID.PRARP
At3g57670-A1d  ------------QYWIPTP--SQI.IGPTQFS.PV.FKT.N.YN.M...MWG.GSQYRKGPESL.G.QP----TGML.LPC.C.-A.G.RNNID.PRAKP
At1g13290-A1d  ------------QYWIP..--SQI.VGPTQFS.SV..KT.N.FN.M...MWG.GSQYRKGPESL.G.KSS---SSIL.LPC.C.-AEG.KNNID.PRSKP
At1g08290-A1d  ------------RFWIP.P--AQIHVGPMQFA.S..SKT.N.YN.M...MWG.GSEFRKGADSLKG.IQP---AAIL.LPC.C.-AEG.KNNIN.PRSKP
At1g34790-A1d  ------------AYWIPAP--EQI.IGFTHFS.HV.FKT.N.YN.....MWG.GSQYRKGPESLKG.QP----RAMLGIPC.C.-VEG.RNHID.PRSKP

At1g68130-A1b  LGDLVGIKKHFRRKHSNHKQWICE-RCS-KGYAVQSDYKAHLKTCGTRGHSCDCGRVFSRVESFIEHQDTCTVRRSQPSNHR
At2g01940-A1b  .....................V..-...-............................................N.SA..VHREPP.
At1g25250-A1b  ................V....V..-...-.................S.........................I.QP..T...
At3g45260-A1a  ....T......S...G-E.K.K.D-K..-.K...M..W...S.I...KEYR....TL...KD...T.RAF.DALAEE--SA.
At5g60470-A1a  ....T......S...G-E.K.K.D-K..-.K...I..W...N.I..S.EFR....TL...KD...S.RSF.D.LAEE--SSK
At4g02670-A1a  ....T......C...G-E.K.K..-K..-.F......W...T.I....DYR....TL...KDT..T.RAF.DALAEE--SA.
At1g03840-A1a  ....T......C...G-E.K.K..-K.A-.R......W...S......EYR....TI...RD...T.RAF.DALAEE--TA.
At5g44160-A1a  ....T......C...G-E.K.T..-K.A-.R......W...S......EYR....TI...RD...T.RAF.DALAEE--TAK
At1g14580-A1a  ....T.....YY...G-E.K.K.D-K..-.R......W...S.....KEYR....TI...RD.Y.T.RAF.DALIQE--SA.
At2g02080-A1a  ....T.....YY...G-E.K.K..-K..-.R......W...S.....KEYR....TI...RD.Y.T.RAF.DALIQE--TA.
At2g02070-A1a  ....T.....YY...G-E.K.K.D-K..-.R......W...S.....KEYR....TL...RD...T.RAF.DALAQE--SA.
At1g55110-A1a  ....T......F...G-E.K.K..-K..-.K......W...A.....KEYK....TL...RD...T.RAF.DALAEE--SA.
At5g66730-A1a  ....T......C...G-E.K.K..-K..-.K......W...S.I...KEYK....TL...RD...T.RAF.DALAEE--SAK
At3g50700-A1a  ....T......C...G-E.K.K.D-K..-.K......W...S.I...KEYK....TL...RD...T.RAF.DALAEE--.A.
At3g13810-A1a  ....T......C...G-E.K.K.D-K..-.K......C...S.....KEYR....TL...RD...T.RAF.EALAEE--TA.
At5g03150-A1a  ....T.....YS...G-E.K.K..-K..-.K......W...A......EYK....TL...KD...T.RAF.DALTEE--GA.
At1g34370-A1c  .KTILCV.N.YK.T.C-D.SFT.S-..HT.KFS.IA.L.T.E.H..KNKWL.S..TT...KDKLFG.IALF-QGHTPAI---
At5g22890-A1c  .KSVICA.N.YK.S.C-P.MYM.R-...V.HFS.L..LRT.E.H..DIKWV.S..TK...KDKLMS.VSLF-LGHVPAH---
At1g51220-A1d  .K.FRTLQT.YK...G-S.PFA.R-M.G-.AF..KG.WRT.E.N..KL-WY.S..SD.KHKR.LKD.VKAFGNGHVPC----
At3g20880-A1d  .K.FRTLQT.YK...G-VRPFA.R-..G-.AF..KG.WRT.E.N..KL-WY.S..SD.KHKR.LKD.VKAFGNGHVPC----
At3g57670-A1d  .K.FRTLQT.YK...G-I.PFM.R-K.G-.AF..RG.WRT.E.N..KL-WY.I..SD.KHKR.LKD.IKAFGNGHGAY----
At1g13290-A1d  .K.FRTLQT.YK...G-A.PFR.RKK.E-.TF..RG.WRT.E.N..KL-WF.V..SD.KHKR.LKD.VRAFGDGHAAH----
At1g08290-A1d  .K.FRTLQT.YK...G-S.PFS.G-K.G-.AL..KG.WRT.E.N..KL-WY.T..SD.KHKR.LKD.IRSFGSGH.PH----
At1g34790-A1d  .K.FRTLQT.YK...G-..PFS.R-L.G-.LL..KG.WRT.E.N..K.-WV.V..SD.KHKR.LKD.VKAFGSGHGPY----
```

Figure 1
Multiple sequence alignment of the zinc finger regions and their conserved flanks in 24 AT-ZFPs of family A1: ZF domains are highlighted in grey, conserved stretches of basic amino acids in green and WIP-stretches in blue. Positions identical to the first sequences are dots, '-' indicate gaps.

binding properties [30]. It was shown in [30] that the minimal region required for specific DNA binding includes the single zinc finger and two basic regions on either side of the ZF domain. Based on these DNA binding studies and on the NMR structure of the ZF domain of SUP [31] a peculiar DNA recognition mechanism was proposed by the authors and will be discussed below. Recently two further SUPERMAN-like representatives of subclass C1-1i were described for *A. thaliana*. RABBIT EARS (RBE, At5g06070) regulates petal development and is probably required for the early development of the organ primordia of the second whorl [32]. Additionally, the expression of At2g42410 (ZFP11) was investigated [33]. It has very low expression levels in flowers, axillary meristems, roots and stems. Interestingly, the deletion of the R/K rich stretch at the C-terminus of the ZF, that was shown to be important for DNA binding in SUP, revealed that this stretch is necessary for the nuclear localization of the protein.

Subclass C1-2i comprises 20 members (Table S1 [see Additional File 1]). Among them are a few with known biological functions. In [27] the expression of four proteins, namely STZ/ZAT10 (At1g27730), AZF1 (At5g67450), AZF2 (At3g19580) and AZF3 (At5g43170) was investigated. The authors showed that all four genes are involved in the plant's water-stress response. Our analysis assigned them into one subgroup (C1-2iD) as shown

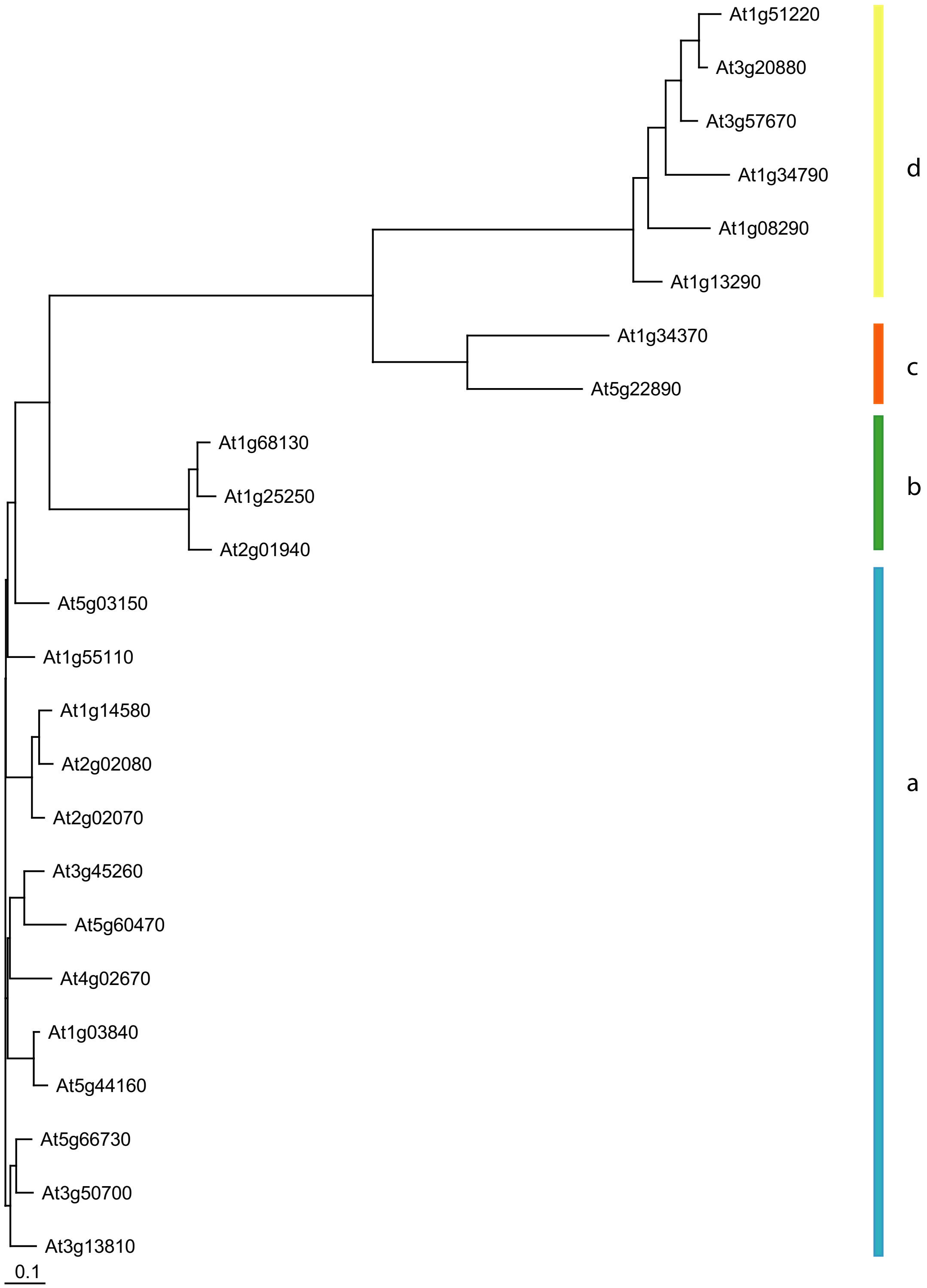


Figure 2

Neighbor-joining tree of the multiple sequence alignment as shown in Figure 1. Relationships between members of the four subgroups A1a to A1d are illustrated.

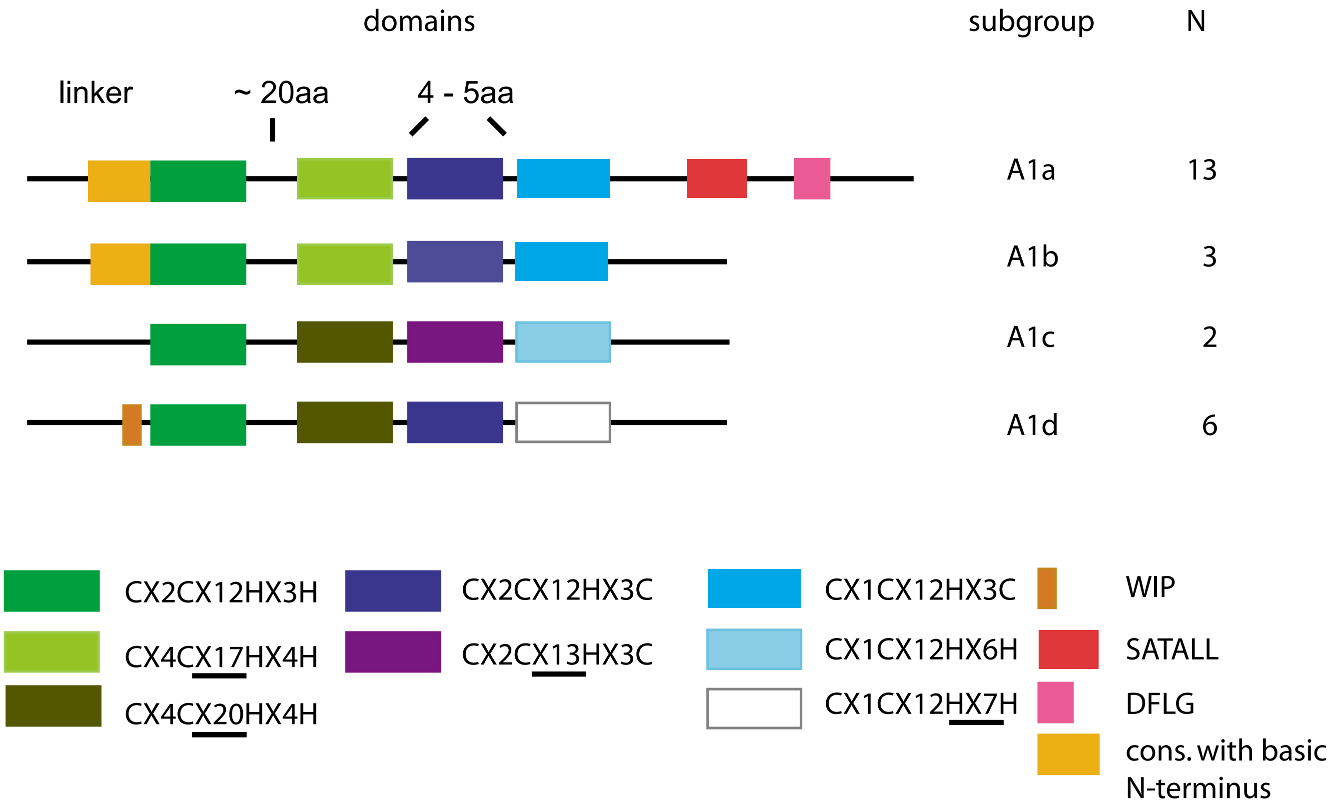
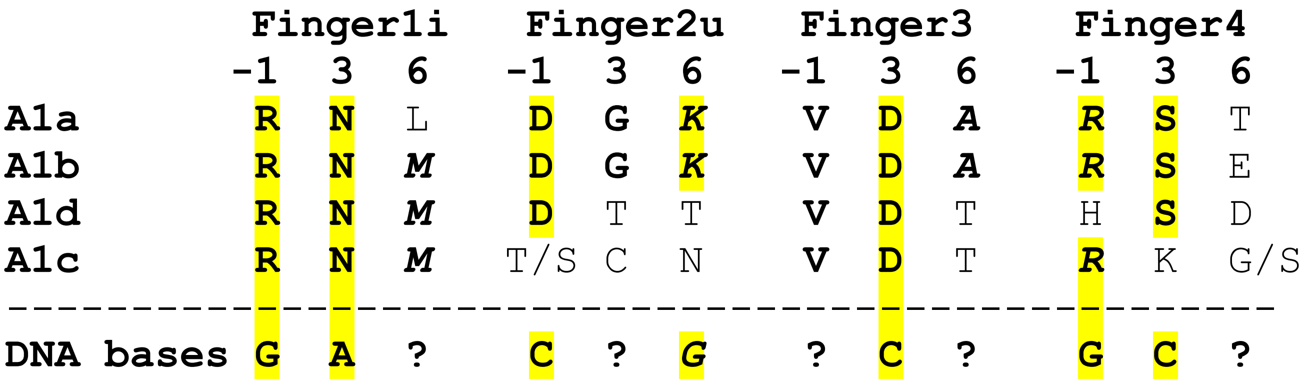


Figure 3
Schematic presentation of zinc finger domain distribution and other conserved motifs of groups A1a to A1d. Length of linkers between ZF domains is given and patterns of the different ZF types are given. Unusual spacing is underlined.



	Finger1i			**Finger2u**			**Finger3**			**Finger4**		
	-1	**3**	**6**	**-1**	**3**	**6**	**-1**	**3**	**6**	**-1**	**3**	**6**
A1a	**R**	**N**	L	**D**	**G**	***K***	**V**	**D**	***A***	***R***	**S**	T
A1b	**R**	**N**	***M***	**D**	**G**	***K***	**V**	**D**	***A***	***R***	**S**	E
A1d	**R**	**N**	***M***	**D**	T	T	**V**	**D**	T	H	**S**	D
A1c	**R**	**N**	***M***	T/S	C	N	**V**	**D**	T	***R***	K	G/S
DNA bases	**G**	**A**	**?**	**C**	**?**	***G***	**?**	**C**	**?**	**G**	**C**	**?**

Figure 4
Amino acid residues in helix positions -1, 3 and 6 in ZF domains of groups A1a to A1d. Conserved amino acid residues with high specificity to particular DNA bases according to a proposed ZF-DNA "recognition code" and corresponding putative DNA bases are highlighted in yellow.

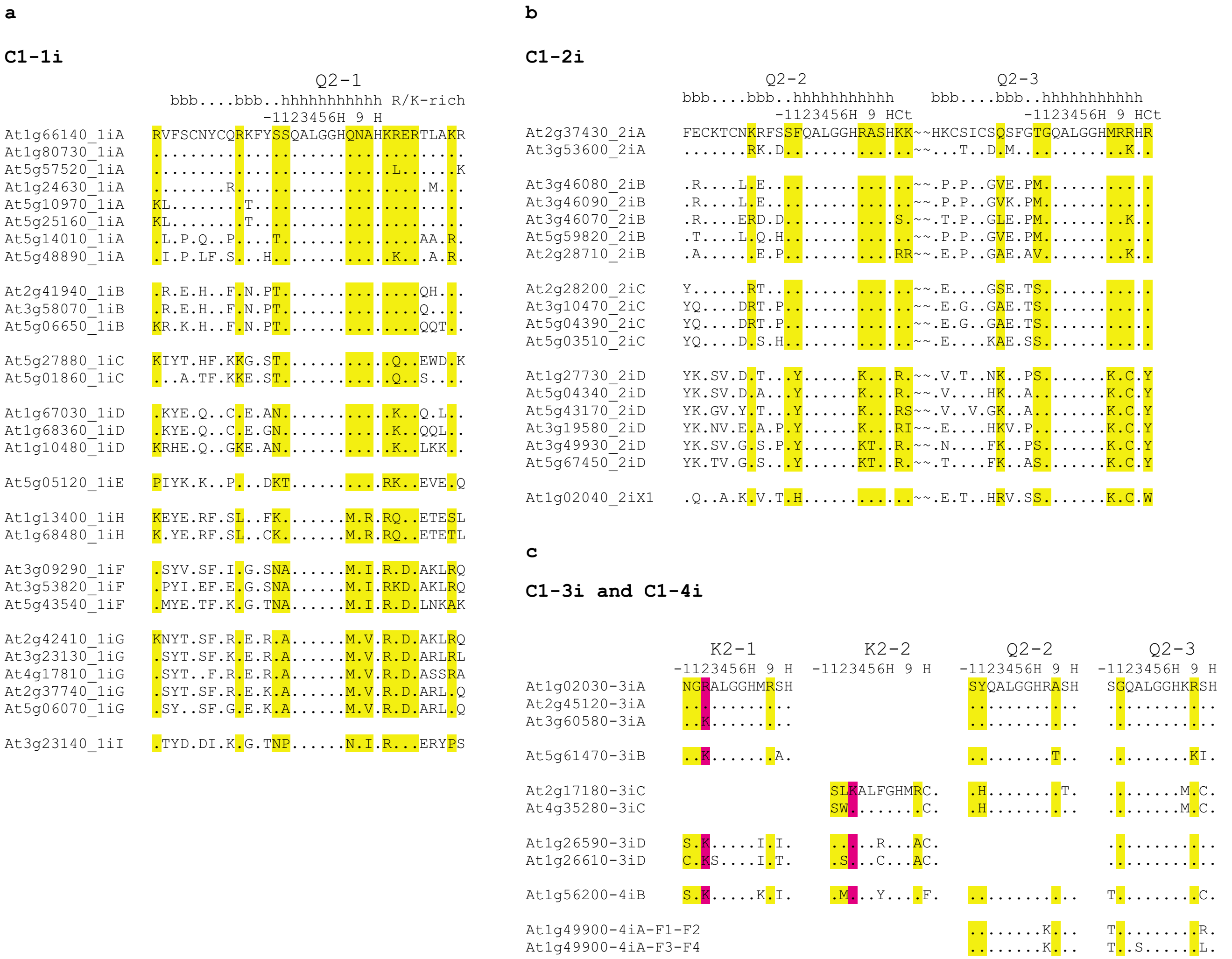


Figure 5

Alignment of ZF domains of family C1 and classification of the four subclasses into different subgroups. (a) C1-1iA-I including the conserved C-terminal R/K-rich ZF flank. (b) C1-2iA-D, X1, also shown are positions C-terminal (Ct) of the ZF. (c) C1-3iA-D, C1-4iA, B, only the helical sequence parts are shown. Signatures and other important positions are highlighted in yellow. The plant specific sequence QALGGH in helix (h) position 2–7 is invariant in all ZF domains except for K2-1 and K2-2 where the K/R is highlighted in pink.

in Figure 5b. This subgroup contains two additional members, At3g49930 and At5g04340 that could be functionally related to ZAT10/AZF1-3. Other representatives of the C1-2i subclass are ZAT5, 7, 11 and 12 that were studied in [26]. As shown in Figure 5b these ZFPs were assigned to three different subgroups C1-2iA (At2g37430, ZAT11), C1-2iB (At3g46070, ZAT7 and At5g59820, ZAT12) and C1-2iC (At2g28200, ZAT5). These three subgroups contain other yet uncharacterized members. Most of them are the result of duplications for example the segmentally duplicated ZAT11 and At3g53600 or the tandemly duplicated ZAT7 (At3g46090), At3g46070 and At3g46080 (Table S1 [see Additional File 1]).

Subclass C1-3i contains eight ZFPs with three dispersed ZFs (Table S1 [see Additional File 1]). They are assigned to four different subgroups C1-3iA to C1-3iD as illustrated in Figure 5c. This figure is for clarity restricted to the alignment of their ZF helices in positions -1 to 11. The only characterized ZFP of this subclass is ZAT1 (At1g02030), a member of subgroup C1-3iA [26]. Its sequence is very similar to the segmentally duplicated At2g45120 and At3g60580.

The only two proteins we found with four dispersed ZFs (C1-4i), At1g49900 and At5g56200, are very different in length and in their sequences. The sequences of the ZF helices are also shown in Figure 5c. Only one protein with five dispersed zinc fingers (C1-5i), At3g29340, occurs in the *Arabidopsis* genome. Nothing is known about the function of the four and five fingered proteins.

Classification of C1 ZFPs and evaluation of different ZF helix types

Based on multiple sequence alignments and tree analyses of the complete sequences of the C1 family members, we further assigned ZFPs of different subclasses to several subgroups as illustrated in Table S1 [see Additional File 1] and Figure 5. These assignments remain the same if we only compare the corresponding ZF sequences and their flanks (Figure 5). The comparison of the ZF helices only (position -1 to 10) resulted in five main signature types (Figure 5 and 6).

28 ZFPs of subclass C1-1i contain an invariant QALGGH sequence in their ZF-helix. The alignment shown in Figure 5a reveals several sequence features of the ZFs and their flanks which are unique to these subclasses, e.g. the invariant N residue in helix position 9 and the conservation of the C-terminal R/K rich flank in five positions. We will refer to significant and apparently conserved positions as signature positions (e.g. the helix positions -1,1,8, 9 and 10). We call the most prevalent signature found in group C1-1i Q2-1 (Q2 refers to the glutamine in position 2 of the helix) (Figure 6). Comparison of the other two and more fingered ZFPs showed that Q2-1 is unique for subgroup C1-1i. 18 ZFPs of the C1-2i subclass contain in both ZFs the invariant QALGGH (Figure 5b). The signature of the two fingers is different from Q2-1 and we refer to finger number one and two as Q2-2 and Q2-3, respectively (Figure 6). They contain in contrast to the Q2-1 type, an A/T (Q2-2) and R (Q2-3) residue at helix position 9. Furthermore, Q2-2 has an aromatic amino acid (F, Y or H) in helix position 1 whereas Q2-3 shows a conserved G. This observation is identical to the ZF domains of two fingered petunia ZFPs [24] where Q2-2 and Q2-3 are named ZF types A and B, respectively. The signatures of the three fingered ZFPs (C1-3i) are more varied (Figure 6). Based on the signature and arrangement of the three fingers we could subdivide the eight proteins into four groups C1-3iA to C1-3iD (Figure 5c). We found signatures in the first finger of all eight proteins and in the second of subgroup D that we encountered previously in a few representatives of C1-1i (C1-1iCa and C1-1iCb) and C1-2i (C1-2iX2). This signature has as the most prominent feature the Q of QALGGH replaced with K or R. We classified the new signatures again according to their residues in helix position 1 and 9. Signature K2-1 has a G in position 1 and a R/K in position 9 whereas K2-2 has L, W, M or S in position 1 and R or A in position 9. We also investigated the signature of the four and five fingered ZFPs (C1-4i, C1-5i). The four fingered At1g49900 has the signature Q2-2/Q2-3/Q2-2/Q2-3 (Figure 5c). The second member of subclass C1-4i, At5g56200, shows the combination K2-1/K2-2/Q2-2/Q2-3 (Figure 5c). None of the signatures we found to this point match any of the ZF domains of the only five fingered ZFPs (At3g29340). The signatures we found are summarized in Figure 6 and all domain arrangements in class C1 are presented schematically in Figure 7. The classification of the signatures is clearly reflected in the neighbor-joining tree of all C1 domains (Figure 8).

Other conserved patterns

The motif search tool MEME revealed a number of other conserved patterns in family C1 (Figure 7). Most remarkable is a leucine rich stretch at the C-terminus of almost all ZFPs in this group. This was also described for homologs in petunia [24]. This stretch is called ERF-associated amphiphilic repression motif and is essential for repression activity [34,35]. We found a basic stretch adjacent to all Q2-1 ZF domains of subclass C1-1i. For SUP, RBE and ZFP11 this basic stretch was already described and suggested to be either involved in DNA binding or nuclear localization (see above). We suggest that it could have a dual function and may be important for both, as it was shown for other proteins [36,37]. Furthermore we found the motifs 'CLMLL' and 'KRKSTKR' N-terminal of Q2-2 in C1-2i and in varying places in C1-3i. The conserved stretches of basic amino acids found in different positions in ZFPs of subclasses C1-2i, 3i and 4i may also serve as nuclear localization signal. The location of all conserved patterns is shown in Figure 7.

Discussion

Evolution

The majority of ZFPs in *A. thaliana* are plant specific and not conserved in other eukaryotes. Comparisons of pairwise distances revealed that ZF domains of subsets C2 and C3 show greater pairwise distances than those of the families C1 and A1. Therefore we can conclude that ZFPs of C2 and C3 are evolutionarily older than A1 and C1 which is supported by our finding that the proportion of conserved proteins is highest in subset C3 followed by subset C2 and that many of them are involved in ancient processes such as RNA metabolism and chromatin-remodeling. Families A1 and C1 are probably the result of a recent expansion. Both families almost exclusively contain plant and AT-specific proteins which supports the notion that they are younger families.

type	**—112345678910**	**N**	**Pos.**	**subgroup**
Q2-1				
	SS**QALGGH**QNA	7	1/1	1iA
	T..........	6	1/1	1iA, B, C
	N..........	3	1/1	1iD
	NA......M.I	3	1/1	1iF
	NP......N.I	1	1/1	1iI
	.A......M.V	5	1/1	1iG
	K.......M.R	2	1/1	1iH
	KT.........	1	1/1	1iE
	sum	28		
Q2-2				
	SF**QALGGH**RAS	11	1/2	2iA, B, C
	.Y.........	4	2/3, 3/4	3iA, 4iB
	.Y......K..	6	1/2, 1+3/4	2iD, 4iA
	.Y......KT.	2	1/2	2iD
	.Y.......T.	1	2/3	3iB
	TY.T....QTF	1	2/2	2iX
	.Y......QTF	1	2/2	2iX
	CY.......RL	1	1/2	2iX3
	.H........T	1	2/3	3iC
	.H.........	1	2/3	3iC
	sum	30		
Q2-3				
	TG**QALGGH**MRR	2	2/2	2iA
	..........K	2	2/2	2iA
	V.........K	1	2/2	2iB
	M.........K	1	2/2	2iB
	M..........	3	2/2	2iB
	S..........	4	2/2	2iC
	S.......K.C	4	2/2	2iD, X1
	........K.C	4	2/2, 4/4	2iD, 4iB
	S.........C	2	3/3	3iC
	S.......K.S	5	3/3	3iA, 3iD
	S.......KKI	1	3/3	3iB
	...S....K.L	1	4/4	4iA
	........K..	1	2/4	4iA
	sum	29		
K2-1				
	NG**RALGGH**MRS	2	1/3	3iA
	..K........	1	1/3	3iA
	.........K.	1	1/1	1iCb
	..K.......A	1	1/3	3iB
	C.KS....I.T	1	1/3	3iD
	S.K.....I.I	1	1/3	3iD
	S.K.....K.I	1	1/4	4iB
	S.K.Y...V.R	1	1/2	2iX2
	sum	9		
K2-2				
	SW**KALFGH**MRC	1	1/3	3iC
	.L.........	1	1/3	3iC
	.L...R...A.	1	2/3	3iD
	.S...C...A.	1	2/3	3iD
	.M...Y....F	1	2/4	4iB
	....V.....A	1	1/1	1iCa
	....V...L.Q	1	1/1	1iCa
	sum	7		

Figure 6

Five main helix signature types: Types are given, amino acids in helix positions -1 to 10, number of ZFPs with identical helices (N) total sum of the helix signature type, ZF position versus total number of ZFs in the protein (Pos.) and assignments to subgroups. Residues in helix positions 1 and 9 are in yellow, the QALGGH motif or variations thereof are in bold.

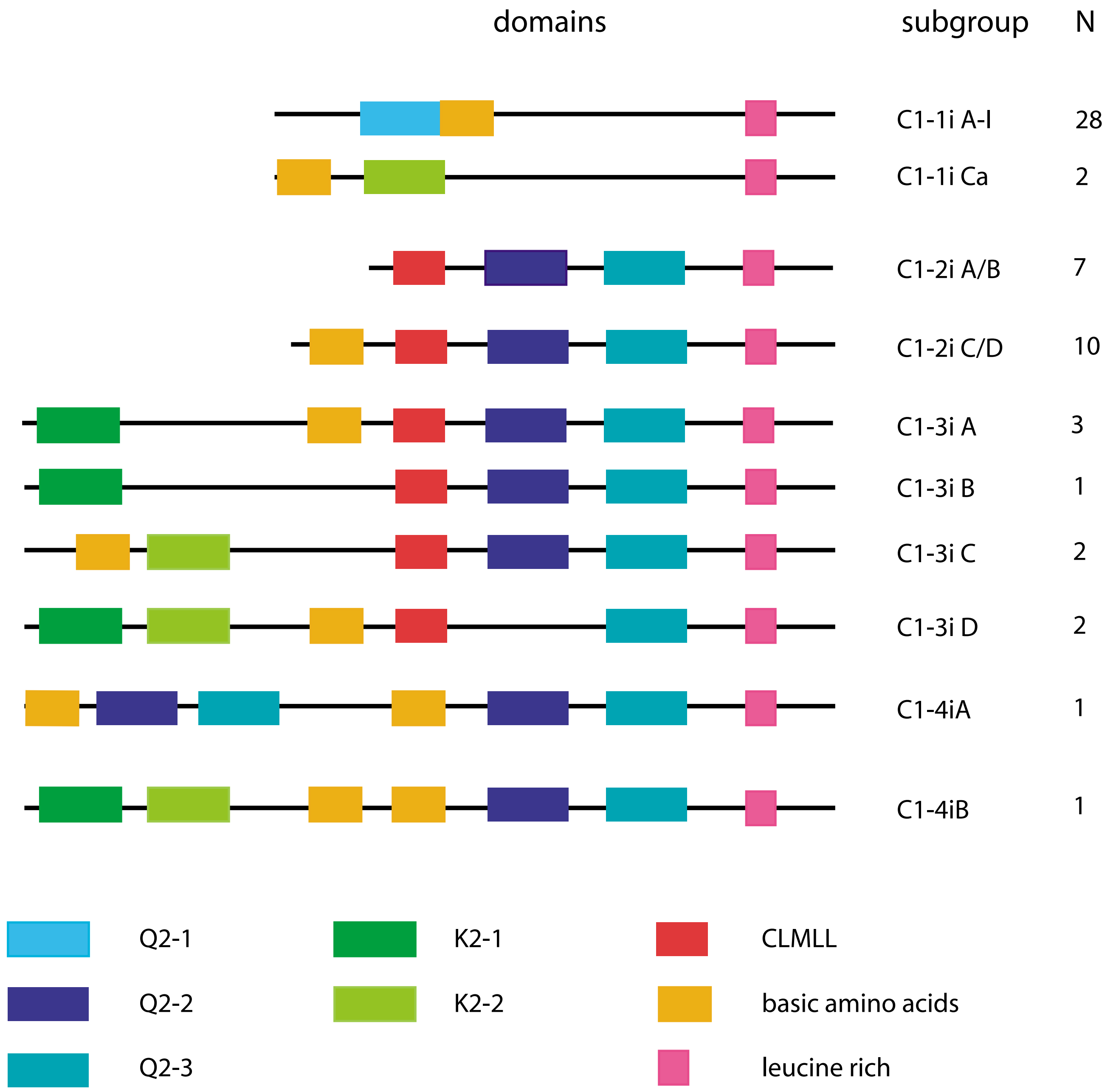


Figure 7
Schematic presentation of the distribution of zinc finger domains and other conserved motifs in family C1. The five main ZF helix types are presented by different colors.

Kubo and coworkers [24] investigated members of the family C1 with two, three and four fingers and suggested, based on the distribution of domains, that multi-fingered proteins in petunia are probably older than those with two fingers. Based on our more comprehensive analysis of family C1 that includes also ZFPs with a single finger, we favor the alternative hypothesis that the single and two fingered proteins are older and the three and four fingered are derived. Q2-2 and Q2-3 are conserved between *Arabidopsis* and petunia (where so far only one single fingered

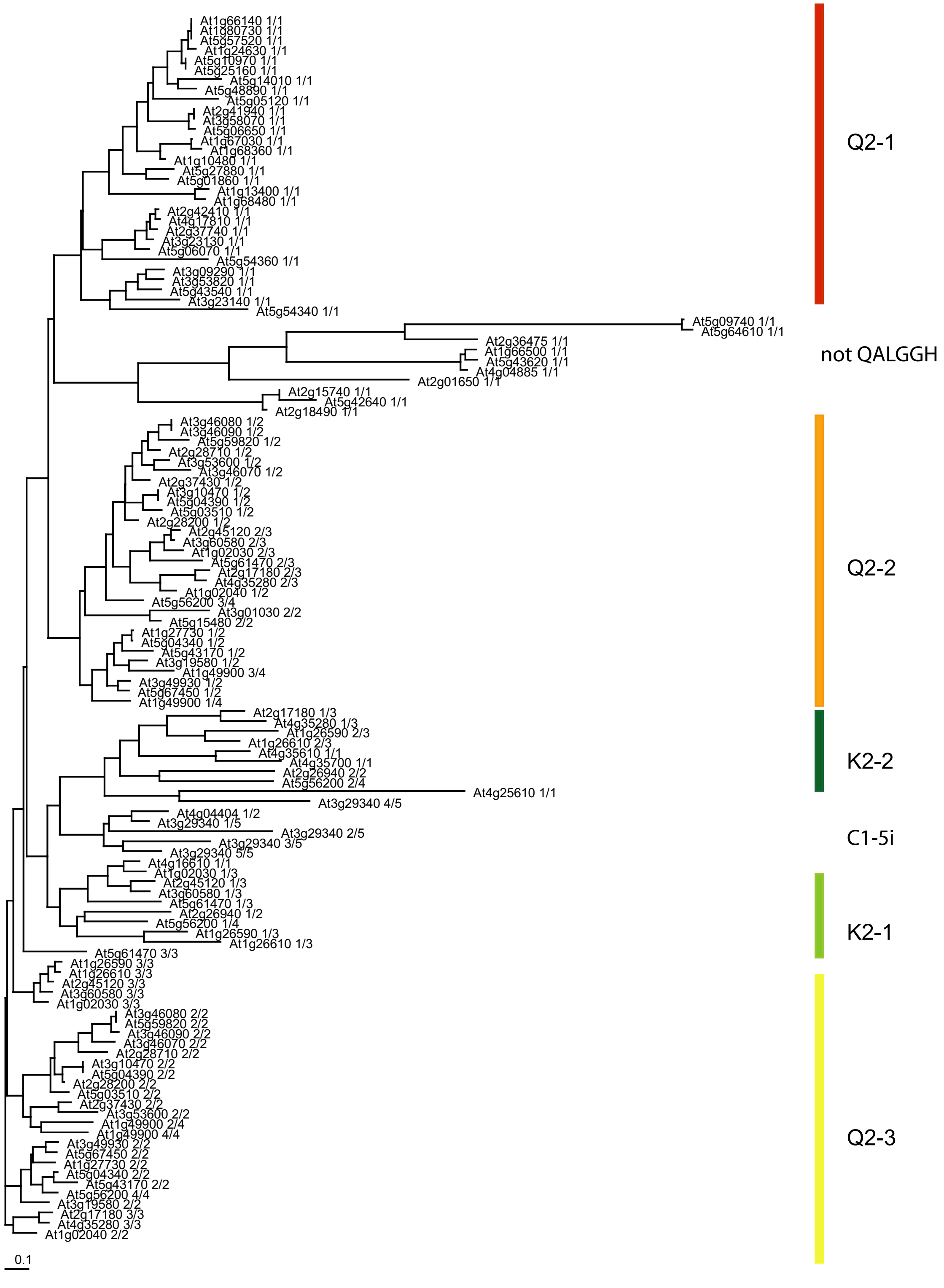


Figure 8

Neighbor-joining tree of all domains found in subset C1. ZF domains assigned to the same signature (Q2-1, Q2-2, Q2-3, K2-1, K2-2) cluster together.

```
Second.str.          ..hhhhhhhhhhh C-term.flank
Helix pos.           -11234567 9   12 4   8
Superman              SAQALGGHMNVH RRDRARLRL
Put. DNA contacts     + ++     +   ++ + + +
Sup R/A mutants                       $ $ $

ZPT2-2/F1             SYQALGGHKASH RK
Put. DNA contacts     + ++    +    ++
ZPT2-2/F2             TGQALGGHKRRH YEG
Put. DNA contacts     + ++    +++
```

Figure 9

Comparison of putative DNA contacting amino acids in SUP and ZPT2-2. The ZF helix sequences, their conserved C-terminal flanks, suggested DNA contacting amino acid residues (+) and positions mutated in DNA-binding studies and abolishing binding activity (§). Signature positions are highlighted in yellow.

protein has been reported) and Q2-1, Q2-2 and Q2-3 are highly conserved between single and two fingered ZFPs of rice and *Arabidopsis*. The K-types vary between rice and *Arabidopsis* which could indicate that there is less selective pressure on this type of zinc finger. We suggest the following scenario: the ancestor domains evolved to Q2-1 domains and duplicated to evolve into Q2-2 and Q2-3 domains, respectively, leading to the C1-2i ZFPs. Another duplication (probably Q2-3) led to three fingered ZFPs (C1-3i) and the domains K2-1 and K2-2. Recombination and also loss of domains could have led to the different three and four fingered types we see today, but can also explain the rare occurrence of one and two fingered proteins with K-type or similar domains (Figure 7). Based on the signatures of At1g49900 (Figure 5c and 6) we conclude that it arose from the duplication of a C1-2i protein. The second four fingered protein At5g56200 probably arose from recombining proteins of the subset C1-3i. The only five fingered ZFP we found is too diverged in sequence to allow inferences about its evolution. We think that the number of the members of the respective subgroups, the distribution of Q2-2 and Q2-3 as well as the distribution of non-finger conserved motifs (Figure 7) favor the assumption of evolution from a low number to a higher number of domains and not vice versa. All main signatures we found seem to be conserved in the plant kingdom. The conservation of the signatures, especially the Q-type implies that ZF types with the same signature may recognize similar DNA sites.

DNA recognition by C1 family zinc fingers

We found five main signatures that are prevalent in the ZF helices of the C1 family which suggests variability in the DNA recognition sites. DNA binding assays determined binding sites with an AGT core sequence for SUP (C1-1i) [30], like in earlier reports for ZPT2-1 (or EPF1) and ZPT2-2 of petunia (C1-2i) [38]. However, the ZF-DNA recognition mechanism of this family is not entirely understood. Experimental data published so far for SUP [30,31] and for the two petunia proteins [24,38,39] revealed a peculiar DNA recognition mechanism that is only partially in line with the canonical binding mode of tandem ZFs [8]. It was suggested for SUP [30,31] that all or some of the amino acid residues in positions -1, 2 and 3 of the alpha-helix and/or residues at the C-terminus of the helix are responsible for the base specific DNA recognition. Similar conclusions were drawn from amino acid mutation studies of ZPT2-2 [24]. These results emphasized the importance of both the invariant QALGGH sequence of the helix and the C-terminal flanking residues for DNA binding. In extension of their previous reports Yoshioka and coworkers showed recently [40] that the optimal binding sites of petunia ZPT2-2 are slightly different with AGC(T) and CAGT for the first and second ZF, respectively, which is in accord with the observation of different signature sequences (Q2-2 and Q2-3 or A and B) for the two fingers [24]. We suggest that differences outside the invariant QALGGH could be responsible for the slightly different optimal DNA binding sites of the first and second finger of ZPT2-2. Additionally, the optimal DNA binding site of SUP might also vary in comparison with the two DNA core sites of ZPT2-2. This is supported by an extended basic C-terminal flank of the ZF of SUP (Q2-1) which is characteristic for many other Q2-1 type fingers (Figure 5a), but is shorter or lacking in the two ZFs of ZPT2-2. In the case of SUP the flank contains five arginine residues and it was shown in [30] that mutations from R to A in three of them abolishes the DNA binding activity. In Figure 9 we compare the ZF helices and the C-terminal flanks of SUP and ZPT2-2 and suggest positions that may contact DNA. We emphasize that detailed conclusions concerning DNA recognition mechanisms of the C1-family ZFs, i.e. which amino acid residues make direct base contacts and which make stabilizing interactions with the phosphate backbone, cannot be drawn until the structure of a ZF/DNA complex is solved by NMR or X-ray analysis.

Conclusions

We showed that the minority of AT-ZFPs is evolutionarily conserved and our analysis further suggests that most of them could be involved in ancient biological processes like RNA metabolism and chromatin-remodeling. The majority of AT (plant)-specific ZFPs are known or suggested to be involved in transcriptional regulation and exhibit remarkable differences in the features of their ZF sequences and ZF arrangements compared to animal ZFPs. In *A. thaliana* we found two major families with recent expansions, one with zinc fingers arranged in tandem (A1), the other with a varying number of dispersed zinc fingers and the plant-specific invariant QALGGH motif in the alpha-helix (C1). However, our studies showed that most ZFPs of *A. thaliana* have their domains arranged in a dispersed manner and not in tandem. Additionally, novel plant specific ZFP-associated domains were detected that may be involved in DNA binding or repressor functions. Our results reflect the diversity of the transcriptional regulation guided by ZFPs in plants compared to animals. Our findings on signatures in zinc finger domains of the largest family C1, and on conserved non-finger motifs give insight into the evolution of the ZFPs and will help to understand their DNA binding function.

Methods

Identification of ZFPs and of conserved ZFP-associated motifs

For the identification of the ZFPs we searched the *Arabidopsis* proteome (MatDB_v110103) using the HMMer package 2.1.1 [41] and the Pfam domain ZF-C2H2 (PF00096) [2]. The minimal cut-off for the search was chosen at a score of 0. The choice of this rather low threshold permits the detection of all ZFs/ZFPs, but also results in the detection of many false-positives. Therefore all identified ZFs/ZFPs subsequently were checked for overlaps with other protein motifs by manual inspection with Pfam and SMART [42] search tools and by BLAST search [43]. Putative C2H2 hits that overlapped with more significant hits of other motifs were eliminated. Usually, questionable C2H2 hits have very low scores and do not exactly fit the spacing of the Pfam C2H2 pattern. Pfam and SMART were also used for the identification of conserved non-finger domains in the AT-ZFPs. In addition, we have applied the program "Multiple Expectation Maximization for Motif Elicitation" (MEME) [44] for the detection of short conserved sequence parts that have not been described yet as Pfam and/or SMART motif. The program MEME detects conserved domains (with unknown sequence) in unaligned sequences. It starts with an initial alignment which provides an estimate of the amino acid composition at each position of the respective conserved stretch that is found. The two steps that follow, the expectation and maximization steps, are applied repeatedly to finally converge to a solution that offers the best location of the motif in each sequence and an estimate of the amino acid composition of each position of the motif.

Classification of AT-ZFPs into families and subgroups

The identified AT-ZFPs were compared to the NR protein database of the NCBI in order to find evolutionarily conserved proteins. Furthermore, all against all BLAST searches of the AT-ZFPs were performed to define families and subgroups and the number of their members. ZFPs in any genome can be classified first into a few main sets based on the number, types and arrangements of their fingers as proposed earlier by us for ZFPs of the yeast genome [3]. All ZFPs containing tandem ZFs in one array or in more than one array are assigned accordingly to sets A and B, respectively, and all ZFPs containing a single ZF or dispersed ZFs are assigned to set C. Based on the results of our statistical analysis of linker lengths in ZFPs (S.B. unpublished data) we have defined tandem ZFs as fingers linked by zero to ten amino acid residues, with five residues as the most frequent (consensus) linker length. ZFs separated by longer spacers of eleven or more residues are considered as dispersed ZFs. Our choice of the upper and lower limits of ten and eleven linker/spacer residues for tandem and dispersed ZFs may seem somewhat arbitrary. However it reflects experimental data on DNA binding ZFPs from literature, where a range of two to seven residues for the linker is given, but most frequently a consensus linker with five residues and the conserved sequence 'TGEK/RP'. ZF domains of large subfamilies were also subjected to phylogenetic analyses using ClustalX [45] for alignments and the PHYLIP package [46] for pairwise sequence distance (PAM Dayhoff matrix) and neighbor-joining analyses.

List of abbreviations

ZFPs C2H2 zinc finger proteins

ZF C2H2 zinc finger

AT-ZFPs C2H2 zinc finger proteins found in *A. thaliana*

Authors' contributions

CE and SB carried out bioinformatics analyses and drafted the manuscript. SB performed the classification of the ZFPs and CE provided information on function and expression of proteins. HS carried out Pfam analyses and validated gene models. All authors read an approved the final manuscript.

Additional material

Additional File 1

Table S1. Complete set of classified C2H2 zinc finger proteins in A. thaliana. *Listed are ZFP subsets and subfamilies as classified in the text.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-39-S1.pdf]

Acknowledgements

We would like to thank K. FX. Mayer, A. Facius and P. Pagel for technical support and A. D. Greenwood for helpful comments on the manuscript.

References

1. Chung H-R, Schäfer U, Jäckle H, Böhm S: **Genomic expansion and clustering of ZAD-containing C2H2 zinc-finger genes in Drosophila.** *EMBO reports* 2002, **3(12):**1158-1162.
2. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32 Database issue:**D138-D141.
3. Böhm S, Frishman D, Mewes HW: **Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins.** *Nucleic Acids Res* 1997, **25(12):**2464-2469.
4. Urrutia R: **KRAB-containing zinc-finger repressor proteins.** *Genome Biol* 2003, **4:**231.
5. Takatsuji H: **Zinc-finger proteins: the classical zinc finger emerges in contemporary plant science.** *Plant Mol Biol* 1999, **39:**1073-1078.
6. **BioOne: The Arabidopsis Book** *Transcriptional Regulation: a Genomic Overview* [http://www.bioone.org/].
7. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G-L: **Arabidopsis Transcription Factors: Genome-Wide Analysis Among Eukaryotes.** *Science* 2000, **290:**2105-2110.
8. Wolfe SA, Nekludova L, Pabo CO: **DNA recognition by Cys$_2$His$_2$ zinc finger proteins.** *Annu Rev Biophys Biomol Struct* 2000, **29:**183-212.
9. Mathieu O, Yukawa Y, Prieto JL, Vaillant I, Sugiura M, Tourmente S: **Identification and characterization of transcription factor IIIA and ribosomal protein L5 from Arabidopsis thaliana.** *Nucleic Acids Res* 2003, **31(9):**2424-22433.
10. Xu T, Purcell M, Zucchi P, Helentjaris T, Bogorad L: **TRM1, a YY1-like suppressor of rbcS-m3 expression in maize mesophyll.** *Proc Natl Acad Sci USA* 2001, **98:**2295-2300.
11. Reyes JC, Hennig L, Gruissem W: **Chromatin-remodeling and memory factors. New regulators of plant development.** *Plant Physiol* 2002, **130(3):**1090-1101.
12. Luo M, Bilodeau P, Dennis ES, Peacock WJ, Chaudhury A: **Expression and parent-of-origin effects for FIS2, MEA, and FIE in the endosperm and embryo of developing Arabidopsis seeds.** *Proc Natl Acad Sci USA* 2000, **97(19):**10637-10642.
13. Luo M, Bilodeau P, Kolturnow A, Dennis ES, Peacock WJ, Chaudhury AM: **Genes controlling fertilization-independent seed development in Arabidopsis thaliana.** *Proc Natl Acad Sci USA* 1999, **96(1):**296-301.
14. Gendall AR, Levy YY, Wilson A, Dean C: **The VERNALIZATION 2 Gene Mediates the Epigenetic Regulation of Vernalization in Arabidopsis.** *Cell* 2001, **107:**525-535.
15. Yoshida N, Yanai Y, Chen L, Kato Y, Hiratsuka J, Miwa T, Sung ZR, Takahashi S: **EMBRYONIC FLOWER2, a Novel Polycomb Group Protein Homolog, Mediates Shoot Development and Flowering in Arabidopsis.** *Plant Cell* 2001, **13:**2471-2481.
16. Birve A, Sengupta AK, Beuchle D, Larsson J, Kenninson J, Rasmuson-Lestander A, Muller J: **Su(z)12, a novel Dorosophila Polycomb group gene that is conserved in vertebrates and plants.** *Development* 2001, **128(17):**3371-3379.
17. Prigge MJ, Wagner DR: **The Arabidopsis SERRATE Gene Encodes a Zinc-Finger Protein Required for Normal Shoot Development.** *Plant Cell* 2001, **13:**1263-1279.
18. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30:**1427-1464.
19. Kuhn C, Frommer WB: **A novel zinc finger protein encoded by a couch potato homologue from Solanum tuberosum enables a sucrose transport-deficiency.** *Mol Gen Genet* 1995, **247(6):**759-763.
20. Colasanti J, Yuan Z, Sundaresan V: **The indeterminate Gene Encodes a Zinc Finger Protein and Regulates a Leaf-Generated Signal Required for the Transition to Flowering in Maize.** *Cell* 1998, **93:**593-603.
21. Colasanti J, Sundaresan V: **'Florigen' enters the molecular age: long-distance signals that cause plants to flower.** *Trends Biochem Sci* 2000, **25:**236-240.
22. Sagasser M, Lu G-H, Hahlbrock K, Weisshaar B: **A. thaliana TRANSPARENT TESTA 1 is involved in seed coat development and defines the WIP subfamily of plant zinc finger proteins.** *Gene & Development* 2002, **16:**138-149.
23. Liu L, White MJ, MacRae TH: **Transcription factors and their genes in higher plants functional domains, evolution and regulation.** *Eur J Biochem* 1999, **262:**247-257.
24. Kubo K, Sakamoto A, Kobayashi A, Rybka Z, Kanno Y, Nakagawa H, Nishino T, Takatsuji H: **Cys$_2$/His$_2$ zinc-finger protein family of petunia: evolution and general mechanism of target-sequence recognition.** *Nucleic Acids Res* 1998, **26(2):**608-615.
25. Tague BW, Goodman HM: **Characterization of a family of Arabidopsis zinc finger protein cDNAs.** *Plant Mol Biol* 1995, **28(2):**267-279.
26. Meissner R, Michael AJ: **Isolation and characterisation of a diverse family of Arabidopsis two and three-fingered C$_2$H$_2$ zinc finger protein genes and cDNAs.** *Plant Mol Biol* 1997, **33:**615-634.
27. Sakamoto H, Araki T, Meshi T, Iwabuchi M: **Expression of a subset of the Arabidopsis Cys$_2$/His$_2$-type zinc-finger protein gene family under water stress.** *Gene* 2000, **248:**23-32.
28. Sakai H, Medrano LJ, Meyerowitz EM: **Role of SUPERMAN in maintaining Arabidopsis floral whorl boundaries.** *Nature* 1995, **378:**199-203.
29. Sakai H, Krizek BA, Jacobsen SE, Meyerowitz EM: **Regulation of SUP Expression Identifies Multiple Regulators Involved in Arabidopsis Floral Meristem Development.** *Plant Cell* 2000, **12:**1607-1618.
30. Dathan N, Zaccaro L, Esposito S, Isernia C, Omichinski JG, Rccio A, Pedone C, Blasio BD, Fattorusso R, Pedone PV: **The Arabidopsis SUPERMAN protein is able to specifically bind DNA through its single Cys$_2$-His$_2$ zinc finger motif.** *Nucleic Acids Res* 2002, **30(22):**4945-4951.
31. Isernia C, Bucci E, Leone M, Zaccaro L, DiLello P, Digilio G, Esposito S, Saviano M, DiBlasio B, Pedone C, Pedone P, Fattorusso R: **NMR structure of the single QALGGH zinc finger domain from the Arabidopsis thaliana SUPERMAN protein.** *Chembiochem* 2003, **4(2–3):**171-180.
32. Takeda S, Matsumoto N, Okada K: **RABBIT EARS, encoding a SUPERMAN-like zinc finger protein, regulates petal development in Arabidopsis thaliana.** *Development* 2004, **131:**425-434.
33. Dinkins RD, Pflipsen C, Collins GB: **Expression and deletion analysis of an Arabidopsis SUPERMAN-like zinc finger gene.** *Plant Science* 2003, **165:**33-41.

34. Hiratsu K, Ohta M, Matsui K, Ohme-Takagi M: **The SUPERMAN protein is an active repressor whose carboxy-terminal repression domain is required for the development of normal flowers.** *FEBS Lett* 2002, **514:**351-354.
35. Ohta M, Matsui K, Hiratsu K, Shinshi H, Ohme-Takagi M: **Repression Domains of Class II ERF Transcriptional Repressors Share an Essential Motif for Active Repression.** *Plant Cell* 2001, **13:**1959-1968.
36. Fernández-Martínez J, Brown CV, Díez E, Tilburn J, Jr HNA, Penalva MÁ, Espeso EA: **Overlap of Nuclear Localisation Signal and Specific DNA-binding Residues Within the Zinc Finger Domain of PacC.** *J Mol Biol* 2003, **334:**667-684.
37. Cokol M, Nair R, Rost B: **Finding nuclear localization signals.** *EMBO reports* 2000, **1(5):**411-415.
38. Takatsuji H, Matsumoto T: **Target-sequence recognition by separate-type Cys2/His2 zinc finger proteins in plants.** *J Biol Chem* 1996, **271(38):**23368-23673.
39. Takatsuji H: **A single amino acid determines the specificity for the target sequence of two zinc-finger proteins in plants.** *Biochem Biophys Res Comm* 1996, **224(1):**219-223.
40. Yoshioka K-i, Fukushima S, Yamazaki T, Yoshida M, Takatsuji H: **The Plant Zinc Finger Protein ZPT2-2 Has a Unique Mode of DNA Interaction.** *The Journal of Biological Chemistry* 2001, **276(38):**35802-35807.
41. Eddy SR: **Profile Hidden Markov Models.** *Bioinformatics* 1998, **14(9):**755-763.
42. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucleic Acids Res* 2004, **32 Database issue:**D142-D144.
43. Altschul SF, Madden TL, Schäfer AA, Zhang J, Miller W, Lippman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
44. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* AAAI Press; 1994:28-36.
45. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
46. Felsenstein J: **PHYLIP (Phylogeny Interference Package), version 3.5c.** *University of Washington, Departement of Genetics, Seattle* 1993.
47. Dinneny J, Yadegari R, Fischer R, Yanofsky M, Weigel D: **The role of JAGGED in shaping lateral organs.** *Development* 2004, **131:**1101-1110.
48. Ohno C, Reddy G, Heisler M, Meyerowitz E: **The *Arabidopsis* JAGGED gene encodes a zinc finger protein that promotes leaf tissue development.** *Development* 2004, **131:**1111-1122.
49. Wu K, Tian L, Zhou C, Brown D, Miki B: **Repression of gene expression by Arabidopsis HD2 histone deacetylase.** *Plant J* 2003, **34:**241-247.