

RESEARCH

Open Access



# A two-sequence motif-based method for the inventory of gene families in fragmented and poorly annotated genome sequences

Anton Frisgaard Nørrevang<sup>1</sup>, Sergey Shabala<sup>2,3</sup> and Michael Palmgren<sup>1\*</sup>

## Abstract

Databases of genome sequences are growing exponentially, but, in some cases, assembly is incomplete and genes are poorly annotated. For evolutionary studies, it is important to identify all members of a given gene family in a genome. We developed a method for identifying most, if not all, members of a gene family from raw genomes in which assembly is of low quality, using the P-type ATPase superfamily as an example. The method is based on the translation of an entire genome in all six reading frames and the co-occurrence of two family-specific sequence motifs that are in close proximity to each other. To test the method's usability, we first used it to identify P-type ATPase members in the high-quality annotated genome of barley (*Hordeum vulgare*). Subsequently, after successfully identifying plasma membrane H<sup>+</sup>-ATPase family members (P3A ATPases) in various plant genomes of varying quality, we tested the hypothesis that the number of P3A ATPases correlates with the ability of the plant to tolerate saline conditions. In 19 genomes of glycophytes and halophytes, the total number of P3A ATPase genes was found to vary from 7 to 22, but no significant difference was found between the two groups. The method successfully identified P-type ATPase family members in raw genomes that are poorly assembled.

**Keywords** Fragmented genomes, Gene families, Halophytes, P-type ATPase, Plasma membrane H<sup>+</sup>-ATPase, Two-motif

## Introduction

Sequencing technology has progressed rapidly, and associated costs have declined, which has resulted in increased throughput in genome sequencing. Consequently, databases of sequenced genomes are now growing exponentially (RefSeq growth statistics [1];

GenBank and WGS Statistics [2]), and raw sequencing data are accumulating at an even greater rate (e.g., NCBI's Sequence Read Archive [SRA] database [3]) [4, 5]. It has therefore become challenging to assemble and curate the genomes at the same pace as sequencing takes place [6]. For plants, even though published genomes are mostly complete with respect to the sequencing level, many of the assemblies are fragmented, with a scaffold N50 (the statistical value that defines assembly quality in terms of contiguity) that is below the standard 1 Mb [7]. A fragmented genome assembly will impact how well genes can be annotated, both in regards to entirely missing protein-coding genes and structural errors in the coding sequence (CDS) [8, 9]. Thus, it is challenging to identify all members of a gene family of interest in the predicted proteome of a published genome. This is also true in data

\*Correspondence:

Michael Palmgren  
palmgren@plen.ku.dk

<sup>1</sup> NovoCrops Center, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C DK-1871, Denmark

<sup>2</sup> School of Biological Sciences, University of Western Australia, Crawley WA6009, Australia

<sup>3</sup> International Research Centre for Environmental Membrane Biology, Foshan University, Foshan 528000, China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

sets where both sequence similarity and ab initio gene prediction have been used for predictions [10]. A limitation of using transcriptomic data to support the annotation is that gene family members often have a differential expression with respect to development, cell type, and changes in the environment, and the transcriptome may not be representative of all stages. For example, several members of the P3A plasma membrane (PM) H<sup>+</sup>-ATPase family from *Arabidopsis thaliana* are almost exclusively expressed in the gametophyte and not in the sporophyte [11].

Sequence alignment of related genes has identified sequence motifs that characterize members of specific gene families (e.g., as can be found in the Pfam database of protein families and domains [12]), and such motifs can be used as baits when searching new genomes for regions in genes that are similar to the identified sequence motifs. To derive such sequence profiles, or Hidden Markov Models (HMMs), from sequence alignments, methods such as HMMER [13] have been developed [14]. Some families or domains are defined not just by one motif but by the co-occurrence of two or more motifs. Whereas the presence of just one of these motifs may be insufficient to assign a protein to a particular family or domain, the simultaneous occurrence of linked motifs improves the confidence that the sequence belongs to the gene family being considered [15]. Alignment-free analytical methods have also been developed [16].

A problem with fragmented genomes is that they hinder the automatic annotation of genes. In this work, we developed a tool to identify members of gene families in fragmented genomes of plants that have not been fully annotated yet. The method is based on the translation of an entire genome in all six reading frames and the co-occurrence of two sequence motifs. We tested our method on P3A ATPases, a major family of PM H<sup>+</sup>-ATPases in plants. An advantage of our method is that it avoids using the predicted proteome/transcriptome for a given organism, which can potentially be misleading.

## Materials and methods

### Annotation of P-type ATPases in *Hordeum vulgare*

The inventory of P-type ATPase genes in the *H. vulgare* genome was based on identified homology to the *A. thaliana* and *Oryza sativa* inventories. Sequences for *A. thaliana* and *O. sativa* were retrieved from refs. [17, 18], respectively. The *H. vulgare* genome was retrieved from the IPK website together with high-confidence CDS and protein predictions.

To identify P-type ATPases in the *H. vulgare* cultivar Morex, the functional annotation of the high-confidence

protein predictions on the IPK database [19] was first searched for proteins belonging to the PFAM00122 family (E1-E2 ATPase) [20]. The identified sequences were inspected for the DKTGT motif, and all sequences not containing this motif were discarded. Subsequently, the sequences were used for preliminary phylogenetic analysis to determine the closest homologs in *O. sativa* and *A. thaliana*. An alignment for each of the *H. vulgare* CDSs, the corresponding genomic sequence, and the CDS for the closest homolog from *O. sativa* was created to correct and check the identified ATPases (Supplementary files 1, 2, 3, 4, 5 and 6). This alignment was used to correct the CDSs by hand, compared to the *O. sativa* CDS, and by inspection of intron–exon splice sites.

### Phylogenetic analysis

Sequences were aligned in MEGA6 using multiple sequence comparison by log expectation (MUSCLE) [21]. The phylogenetic analysis was performed with the CIPRES Science Gateway [22]. For Maximum likelihood analysis, the RAxML-HPC2 workflow implemented in the XSEDE tool was used with the following parameters: bootstraps=1,000; data type=protein; Protein Substitution Matrix=LG; and the rest of the settings were default. For Bayesian inference analysis, the MrBayes algorithm implemented in XSEDE 3.2.6 was used with the following parameters: number of generations=100,000; data type=protein; Protein Substitution Matrix=LG; number of runs=2; likelihood model=inverse gamma distribution; number of chains=8; temperature parameter=0.05; and the rest of the settings were default.

### Building the query sequences

The query sequences were created for each family (P1B, P2A, P2B, P3A, P4, and P5) using the ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare* (Supplementary file 7) and were built by aligning all protein sequences for a family in Genius Prime using MUSCLE with standard parameters. From the alignment, two query sequences were constructed. The first query sequence consisted only of the family-specific motif, the superfamily-specific motif, and the distance between the two. The second query sequence contained the consensus surrounding the family- and superfamily-specific motifs.

### Genomes used for identifying P3A ATPases

All genomes used for determining the number of P3A PM H<sup>+</sup>-ATPases were reference genomes as of October 2021 and were retrieved from NCBI with the exception of *H. vulgare* and *Aeluropus littoralis*, for which the genomes were retrieved from IPK in October 2021. Links to all genomes and assemblies used are given in Supplementary Table 1.

### Translation of genomes and building and searching databases

To build the databases for interrogation of the genomes, all genomes were imported into CLC Main Workbench 20.0.4. Following this, the genomes were translated in all six reading frames and basic local alignment search tool (BLAST) databases were created in CLC Main Workbench for each of the now translated genomes. All BLAST searches were also carried out using CLC Main Workbench.

### cDNA cloning

Seeds of *H. vulgare* cv. RGT Planet were surface sterilized by removing the husk, submerging the seeds in a bleach solution (80% primo-bleach, 19% water, and 0.1% Triton X-100), and then placing the submerged seeds on a pivoting table for 20 min. After sterilization, the seeds were washed 10 times in sterilized Mili-Q water. Seeds were germinated on sterile wet filter paper for 3 days at 20 °C with 16 h of light. After germination, the seeds were moved to full MS plates and left to grow for 7 days at 20 °C with 16 h of light. Sterile seedlings were homogenized into a fine powder using a clean mortar submerged in liquid nitrogen, from which RNA was extracted with the Qiagen Plant RNeasy Kit following the manufacturer's protocol. gDNA was removed with the TURBO DNA-free Kit from Ambion following the manufacturer's instructions. cDNA was produced using the iScript kit from Bio-Rad according to the provided protocol.

### Identification and cloning of HvHMA7 and HvHMA1-like

The *HvHMA7* transcript was validated by partially cloning it from cDNA of 10-day-old whole seedlings of the cultivar RGT Planet using the following primer pairs: HvHMA7Fp2; 5'-TCCGCAACTGTCAATCT-3' X HvHMA7Rp10; 5'-TCCTCAAAAATGTCTTC-3' and HvHMA7Fp8\_BLAST; 5'-AGGCTGCCAACTGCATCAAT-3' X HvHMA7Rp8\_BLAST; 5'-CAAC GGGTAGCCCAAC AATG-3'. *HvHMA1B* transcript expression was confirmed by cloning the first 1850 bp of the gene using the following primer pairs: HvHMA1BFp1\_BLAST; 5'-ATG CGGCTTGACTCC-3' X HvHMA1BRp1\_BLAST; 5'-CATGCCCGCCTTTCACAAA-3' and HvHMA1BFp9; 5'-CAGGTGGTGCTAGGAACCTG-3' X HvHMA1BRp9; 5'-CACTGGCACGTTGAGCTAGA-3'. All PCR products were sequenced by Eurofins Genomics.

### Statistics

Statistical analysis was performed using GraphPad Prism 9.4.0 using Student's *t*-tests.

## Results

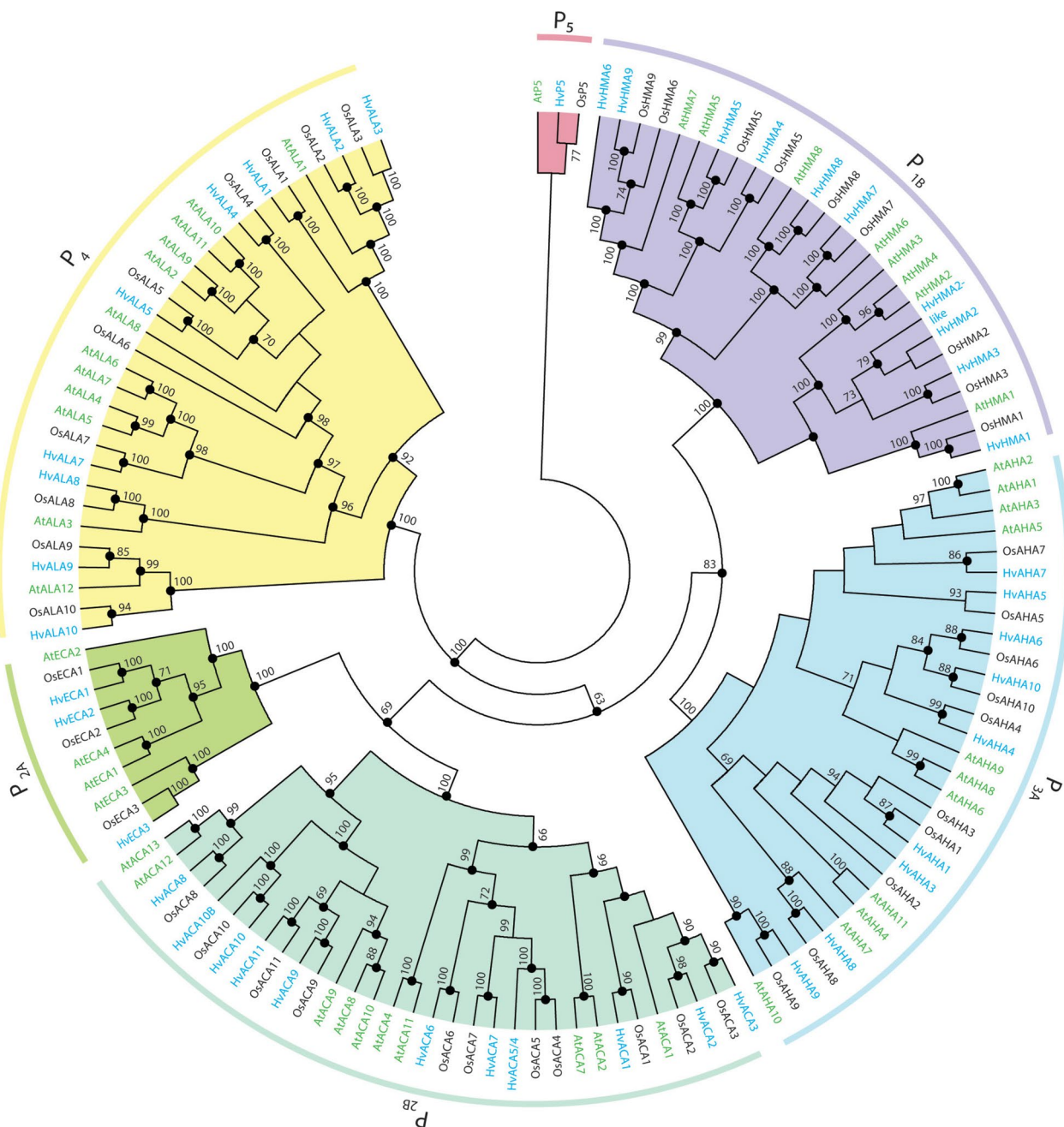
### Generating query sequences for different P-type ATPase families

P-type ATPases (also called E1-E2 ATPases) are a large superfamily of primary active pumps comprising several families in plants (P1B, P2A, P2B, P3A, P4, and P5). All P-type ATPases have the Asp-Lys-Thr-Gly-Thr (DKTGT) motif in the phosphorylation (P) domain where the phosphorylatable aspartate residue resides [23, 24]. Upstream of the DKTGT motif, each family also has a specific motif surrounding the proline helix breaker in transmembrane helix four, namely, P1B: CPC/SPC; P2A/B: PEXL; P3A: PIA; P4: PIS; and P5: PPXXP [24–27]. We thus considered that the two motifs in close proximity to each other could be used in combination to search the database of translated genomes. A search string for each family only containing the family- and ATPase-specific motifs with a fixed distance between the two was created (see Additional file 14). We called this approach the two-motif search.

### A reference inventory of P-type ATPases in *H. vulgare*

Several P-type ATPases have been identified in *H. vulgare* [28–31], but an exhaustive inventory of this superfamily has yet to be constructed. To test our method, we started by using the high-quality assembled genome of *H. vulgare* (current N50 = 69.6 Mb). First, we identified and annotated all P-type ATPases using conventional methods. Subsequently, we searched through the *H. vulgare* genome using the above-described method.

P-type ATPase genes were annotated by retrieving all predictions with a PFAM00122 motif from the high-confidence predictions, which were curated guided by intron–exon splice sites and homology to their orthologues in *O. sativa*. The *HvHMA7* assembled transcript was only identified in Morex V3 and was difficult to curate properly. Therefore, the transcript was validated by partially cloning it from 10-day-old *H. vulgare* seedlings and searching the transcriptomic shotgun assembly (TSA) database at NCBI. All sequences are presented in Supplementary files 1, 2, 3, 4, 5 and 6. Phylogenetic analysis was performed to compare the ATPases found in *H. vulgare* with those of *O. sativa* and *A. thaliana*. *H. vulgare* sequences were named by adopting the name from the closest homolog in *O. sativa* (Fig. 1). HvPAA1 was renamed to HvHMA3 for consistency. A few differences could be found when comparing the number of P-type ATPase genes in *O. sativa* and *H. vulgare*. Two orthologues, *HvHMA2* and *HvHMA2-like*, were identified for *OsHMA2* in the P1B family. There also appeared to be two orthologues of *OsHMA1* in *H. vulgare*, *HvHMA1* and



**Fig. 1** Phylogenetic analysis of P-type ATPases from *A. thaliana*, *H. vulgare*, and *O. sativa*. Full-length protein sequences were aligned in MEGA6 using MUSCLE with standard parameters. The alignment was then subjected to maximum likelihood analysis and Bayesian inference. For maximum likelihood analysis, the RAXML-HPC2 workflow implemented in the XSEDE tool was used on the CIPRES Science Gateway with the following parameters: bootstraps = 1000; data type = protein; Protein Substitution Matrix = LG; and the rest of the settings were default. For Bayesian inference analysis, the MrBayes v. 3.2.7 tool was used on the CIPRES Science Gateway with the following parameters: aamodelpr = fixed(lg); prset statefreqpr = fixed(empirical); lset rates = invgamma; mcmc ngen = 100,000; mcmc nruns = 2; mcmc nchains = 8; mcmc temp = 0.05; mcmc mcmcdiagn = yes; and mcmc diagnfreq = 10,000. Numbers at nodes show likelihood from maximum likelihood analysis, and filled circles indicate full support in Bayesian inference analysis. In the phylogenetic analysis, the orthologue of OsALA6 was excluded as this gene is partial

*HvHMA1-like*. The *HvHMA1-like* gene has a premature stop codon after codon number 80 and is therefore characterized as a pseudogene. The premature stop codon

was confirmed by partially cloning *HvHMA1-like* from cDNA of 10-day-old seedlings of the cultivar RGT Planet and was also identified in the cultivar Price by searching

the TSA database with the CDS of *HvHMA1-like*. In the P2B family, a second homolog of *OsACA10* was identified in *H. vulgare*, and a homolog of *OsACA4/OsACA5* was missing. Furthermore, in the P3A family no ortholog of *OsAHA2* could be identified in *H. vulgare*. For a complete list of the genes, see Table 1.

#### Evaluating the two-motif search in *A. thaliana*, *O. sativa*, and *H. vulgare*

Three genomes in which all P-type ATPases have been identified are now available, which allowed us to test the two-motif method. We searched the translated genomes of *A. thaliana*, *H. vulgare*, and *O. sativa* with the two-motif string and, with the exception of the P4 family, identified all P-type ATPases of a specific family that did not have an intron between the two motifs (Table 2). To identify all members of the P4 family, the search string was made more flexible to accept a distance between the two motifs that varied by 1 amino acid (Supplementary Fig. 5). As P2A and P2B have the same family motif, the two groups are indistinguishable and need to be added together.

#### Modifying the two-motif method to include the area surrounding the family- and P-Type ATPase-specific motifs

As the two-motif method alone falls short regarding the identification of all ATPases within a group, we aimed to modify it by including the consensus sequence surrounding the family- and ATPase-specific motifs. To identify genes with an intron between the two motifs, it is necessary to build a bait sequence with additional information between the motifs. We modified the two-motif search and built an enlarged bait sequence for each family (P1B, P2A, P2B, P3A, P4, and P5) using the entire inventory of P-type ATPases from *A. thaliana*, *H. vulgare*, and *O. sativa*. We aligned sequences from the different families and produced a consensus sequence surrounding the DKTGT motif and family-specific motif for each family. The consensus sequences were some 55 to 70 amino acid residues in length (Supplementary Figs. 1–6). Each consensus sequence was used as a short query sequence to identify the number of genes in a specific family.

#### Testing the enlarged query sequence on the annotated genomes of *A. thaliana*, *O. sativa*, and *H. vulgare*

To test the enlarged query sequence, we used the full inventories of P-type ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*. To avoid including the data used to build the query sequence, we rotated organisms out of the query sequence. Thus, for testing against *H. vulgare*, the query sequence would be created with *A. thaliana* and *O. sativa*. After construction, we used the

sequence for a BLAST search against the translated genome of *H. vulgare*. The assumption was that if the DKTGT and the family-specific motifs were located in close proximity to each other and in the correct orientation relative to each other, the hit would be assessed as constituting a P-type ATPase of that family. This way, a gene can be identified even when the motifs are located in different exons.

The testing showed that all genes belonging to the P3A and P5 ATPase families could be scored (Table 2). These families have high overall conservation, which facilitated their identification (Supplementary Table 2 and Supplementary Figs. 4 and 6). Also, these families have the query sequence in two exons. The method identified fewer P1B, P2B, and P4 ATPases than what is present in the inventory. This could have been due to low overall conservation in the families or because several of the genes have the query sequence distributed onto three exons. Lastly, for P2A, the number is overestimated, most likely due to the close similarity between P2A and P2B (see Supplementary Figs. 2 and 3). If P2A and P2B are considered to be one family, all genes where the query sequence is found in one or two exons can be identified. A complete list of the number of exons the query sequences are located in is provided in Supplementary Table 3. Overall, it appears that the number of genes in a family can be relatively well predicted with the exception of genes where the query sequences are spread out onto three exons. To ensure that the result obtained for *H. vulgare* was consistent, we further tested *O. sativa* and *A. thaliana*, with results similar to those from *H. vulgare* (Table 2). Thus, we assumed that the method could be used to determine the number of P3A and P5 genes in a plant genome and estimate the number of P1B, P2A/B, and P4 genes.

#### Identification of P3A ATPases in complete assemblies

To further evaluate the ability of our approach to correctly determine the number of P3A ATPases, several different well-sequenced and annotated genomes were retrieved and translated in all six reading frames, and a database was created for each. Using the constructed consensus sequence, the databases were interrogated for the number of P3A ATPases (Supplementary Table 4). The number obtained was compared to the number annotated in the genome found by performing BLAST searches using the protein sequence of AtAHA2 against the protein database for each species in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [32]. From this analysis, it appears that the method can predict the number of P3A ATPases in a genome relatively well, and this applies to both very large genomes such as that of *Triticum aestivum* (genome size = 14.6 Gb) and

**Table 1** Inventory of P-type ATPases in *H. vulgare*

Gene name	Morex V3 gene ID	Chromosome	Length (aa)	Status	Accession to the transcript on NCBI	Ref
<b>P1B</b>						
HvHMA1	HORVU.MOREX.r3.7HG0730400	7	828	Complete	XM_045100963.1	[28]
HvHMA2	HORVU.MOREX.r3.7HG0727750	7	1009	Complete	XM_045106273.1	[29]
HvHMA2-like	HORVU.MOREX.r3.7HG0727780	7	946	Complete	XM_045106438.1	
HvHMA3	HORVU.MOREX.r3.5HG0509930	5	837	Complete	XM_045128791.1	[30]
HvHMA4	HORVU.MOREX.r3.5HG0509930	5	980	Complete	XM_045095517.1	[29]
HvHMA5	HORVU.MOREX.r3.2HG0188860	2	995	Complete	XM_045111619.1	
HvHMA6	HORVU.MOREX.r3.6HG0571380	6	1001	Complete	XM_045095317.1	[29]
HvHMA7	HORVU.MOREX.r3.7HG0677890	7	891	Complete	XM_045099939.1 <sup>c</sup>	
HvHMA8	HORVU.MOREX.r3.4HG0404710	4	904	Complete	XM_045126876.1	
HvHMA9	HORVU.MOREX.r3.7HG0738190	7	1002	Complete	XM_045102148.1	[29]
<b>P2A</b>						
HvECA1	HORVU.MOREX.r3.4HG0386480	4	1062	Complete	XM_045126035.1	
HvECA2	HORVU.MOREX.r3.1HG0007600	1	1049	Complete	XM_045127204.1	
HvECA3	HORVU.MOREX.r3.4HG0339640	4	1000	Complete	XM_045124294.1	
<b>P2B</b>						
HvACA1	HORVU.MOREX.r3.4HG0386480	4	1020	Complete	XM_045126605.1	
HvACA2	HORVU.MOREX.r3.5HG0438560	5	1020	Complete	XM_045092435.1	
HvACA3	HORVU.MOREX.r3.4HG0342980	4	1031	Complete	XM_045124470.1 <sup>c</sup>	
HvACA4/5	HORVU.MOREX.r3.5HG0461290	5	1044	Complete	XM_045090434.1	
HvACA6	HORVU.MOREX.r3.3HG0317250	3	1035	Complete	XM_045117438.1	
HvACA7	HORVU.MOREX.r3.1HG0069030	1	1044	Complete	XM_045120351.1	
HvACA8	HORVU.MOREX.r3.1HG0077790	1	1022	Complete	XM_045106798.1	
HvACA9	HORVU.MOREX.r3.6HG0570250	6	1093	Complete	XM_045099510.1	
HvACA10	HORVU.MOREX.r3.7HG0682170	7	1083	Complete	XM_045100081.1 <sup>c</sup>	
HvACA10-like	HORVU.MOREX.r3.5HG0492640	5	1079	Complete	XM_045094011.1	
HvACA11 <sup>b</sup>	HORVU.MOREX.r3.2HG0192750	2	1083	Complete	XM_045111798.1	
<b>P3A</b>						
HvAHA1	HORVU.MOREX.r3.4HG0334970	4	962	Complete	XM_045123783.1	
HvAHA3	HORVU.MOREX.r3.5HG0420210	5	958	Complete	XM_045128668.1	
HvAHA4	HORVU.MOREX.r3.7HG0743890	7	952	Complete	XM_045104844.1	
HvAHA5	HORVU.MOREX.r3.1HG0065040	1	957	Complete	XM_045118633.1	
HvAHA6	HORVU.MOREX.r3.6HG0624610	6	950	Complete	XM_045096904.1	
HvAHA7	HORVU.MOREX.r3.2HG0202670	2	951	Complete	XM_045112242.1	
HvAHA8	HORVU.MOREX.r3.4HG0419450	4	970	Complete	XM_045127541.1 <sup>c</sup>	
HvAHA9	HORVU.MOREX.r3.4HG0405790	4	908	Complete	XM_045126914.1	
HvAHA10	HORVU.MOREX.r3.7HG0658610	7	946	Complete	XM_045102180.1	
<b>P4</b>						
HvALA1	HORVU.MOREX.r3.4HG0378760	4	1247	Complete	XM_045125679.1	
HvALA2	HORVU.MOREX.r3.3HG0257460	3	1162	Complete	XM_045119932.1	
HvALA3	HORVU.MOREX.r3.4HG0379790	4	1121	Complete	XM_045125739.1	
HvALA4	HORVU.MOREX.r3.1HG0000110	1	1230	Complete	XM_045113689.1	
HvALA5	HORVU.MOREX.r3.7HG0676610	7	1205	Complete	XM_045105875.1	
HvALA6	HORVU.MOREX.r3.7HG0717230	7	1227	Partially	XM_045106260.1 <sup>c</sup>	
HvALA7	HORVU.MOREX.r3.7HG0717230	7	1217	Complete	XM_045106260.1	
HvALA8	HORVU.MOREX.r3.1HG0034380	1	1238	Complete	XM_045112594.1	
HvALA9	HORVU.MOREX.r3.2HG0134710	2	1110	Complete	XM_045109393.1	
HvALA10	HORVU.MOREX.r3.2HG0160620	2	1172	Complete	XM_045110317.1	

**Table 1** (continued)

Gene name	Morex V3 gene ID	Chromosome	Length (aa)	Status	Accession to the transcript on NCBI	Ref
<b>P5</b>						
P5	HORVU.MOREX.r3.1HG0063350	1	1175	Complete	XM_045117991.1	[31]
<b>Pseudogenes</b>						
	HORVU4Hr1G086810 <sup>a</sup>	4		Pseudo gene		
HvHMA1-like	HORVU5Hr1G066370 <sup>a</sup>	5		Pseudo gene		
	HORVU7Hr1G091830 <sup>a</sup>	7		Pseudo gene		

<sup>a</sup> These genes do not have a gene ID in later versions of the *H. vulgare* genome, and the V1 ID is therefore adopted

<sup>b</sup> HvACA11 could potentially be duplicated in the genomic locations Nr 1: 605591154-605603057 and Nr 2: 605666642-605678539 on chromosome 7. However, the two genomic areas are highly identical and they are therefore viewed as one gene here

<sup>c</sup> The transcript was manually curated

**Table 2** Comparison of ATPases identified by searching translated genomes with the two-motif method and the expanded two-motif method. The inventory of the respective species and the number of exons in which the query sequence is present

	Inventory	One exon	Two exons	Three exons	Two-motif	Expanded two-motif
<b>P1B</b>						
<i>A. thaliana</i>	8	4	3	1	4	6
<i>H. vulgare</i>	10	5	3	1	5	7
<i>O. sativa</i>	9	4	4	1	4	8
<b>P2A</b>						
<i>A. thaliana</i>	4	3	1	0	10	10
<i>H. vulgare</i>	3	2	1	0	9	9
<i>O. sativa</i>	3	2	1	0	10	10
<b>P2B</b>						
<i>A. thaliana</i>	10	7	0	3	10	11
<i>H. vulgare</i>	10	7	0	4	9	9
<i>O. sativa</i>	11	8	0	3	10	11
<b>P3A</b>						
<i>A. thaliana</i>	11	8	3	0	8	11
<i>H. vulgare</i>	9	5	4	0	5	9
<i>O. sativa</i>	10	5	5	0	5	10
<b>P4</b>						
<i>A. thaliana</i>	12	10	0	2	10	10
<i>H. vulgare</i>	10	7	0	3	6	6
<i>O. sativa</i>	10	7	0	3	7	7
<b>P5</b>						
<i>A. thaliana</i>	1	0	1	0	0	1
<i>H. vulgare</i>	1	0	1	0	1	1
<i>O. sativa</i>	1	0	1	0	0	1

smaller genomes such as that of *Beta vulgaris* (genome size = 540.5 Mb).

#### Testing the method for the P2 family

We further tested the method in the *Glycine max* assembly, where P2 family ATPases have already been

annotated [33]. The *G. max* genome assembly used here has a scaffold N50 of 48 Mb and a contig N50 of 419 kb. Twelve P2 Ca<sup>2+</sup> ATPases were identified [33]. This is a surprisingly low number considering that *G. max* is a tetraploid [34] and that the diploid *A. thaliana* contains 14 P2 ATPases [18]. When searching the translated

*G. max* assembly with the bait made from *O. sativa*, *A. thaliana*, and *H. vulgare*, 23 potential  $\text{Ca}^{2+}$  ATPases were identified (Supplementary Table 5), demonstrating that care must be taken when using protein prediction models.

### Comparing the number of P3A ATPases in glycophytes and halophytes

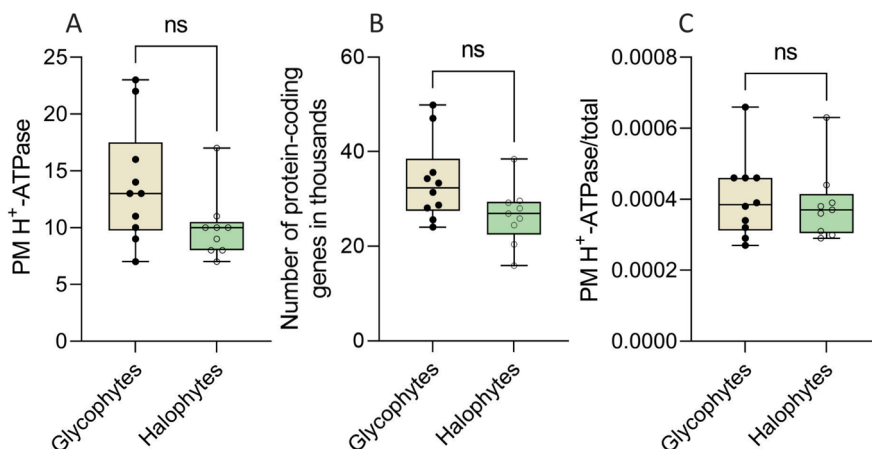
Under salt stress, plants extrude  $\text{Na}^+$  back into the soil to lower the  $\text{K}^+/\text{Na}^+$  ratio. This extrusion is done in roots by Salt Overly Sensitive 1 (SOS1), which is a  $\text{Na}^+/\text{H}^+$  antiporter, driven by the electrochemical  $\text{H}^+$  gradient generated by the PM  $\text{H}^+$ -ATPase [35–38]. The uptake of  $\text{K}^+$  from the soil into roots, as well as  $\text{K}^+$  retention in the cytosol [39], through channel proteins and symporters is also dependent on the activity of the PM  $\text{H}^+$ -ATPase [40]. Thus, we speculated that the number of PM  $\text{H}^+$ -ATPases could contribute to the difference between halophytes (salt-tolerant plants) and glycophytes (salt-sensitive plants) with respect to salt tolerance as an expansion of genes and their subsequent diversification could possibly enable a more fine-tuned regulation at the tissue expression level.

To determine if there is a difference in the number of P3A ATPases between halophytes and glycophytes, we only used diploid genomes for the analysis, as the ploidy could influence the results. Halophytic species were selected according to the The Food and Agriculture Organization of the United Nations (FAO) [41]. We first investigated if there was a general expansion in the number of genes between the two groups. No significant difference between the number of protein-coding genes in the two groups could be observed (Fig. 2A). Searching through the translated genomes with the consensus sequence for P3A ATPases earlier identified, the number of PM  $\text{H}^+$ -ATPases was determined for the two groups (Table 3). On average, there were 10–12 P3A ATPase genes per plant genome (Fig. 2B), which could be normalized to about 0.00035% of protein-coding genes (Fig. 2C). The mean numbers were slightly lower in the halophyte group ( $10.3 \pm 0.9$  vs.  $13.8 \pm 1.6$ ), but due to the sample size being limited by the number of species available for analysis it could not be ascertained whether the difference was significant.

**Table 3** The number of genes and PM  $\text{H}^+$ -ATPases and the ratio between PM  $\text{H}^+$ -ATPases and the number of genes in glycophytes and halophytes. Additional references for numbers of protein-coding genes are given in Supplementary Table 6

Plant	Ploidy	Protein-coding genes	PM $\text{H}^+$ -ATPase	PM $\text{H}^+$ -ATPase/total	Scaffold N50	Citation for ploidy
<b>Glycophytes</b>						
<i>Allium cepa</i>	diploid	47066	22	0,00046	460,7 kb	[42]
<i>Eucalyptus grandis</i>	diploid	33352	11	0,00032	58,5 Mb	[43]
<i>Gossypium raimondii</i>	diploid	35609	14	0,00039	62,2 Mb	[44]
<i>Hordeum vulgare</i>	diploid	31448	9	0,00029	610,3 Mb	[45]
<i>Oryza sativa</i>	diploid	28738	10	0,00034	30 Mb	[46]
<i>Phaseolus vulgaris</i>	diploid	28134	13	0,00046	50,4 Mb	[47]
<i>Raphanus sativus</i>	diploid	49855	23	0,00046	19,9 kb	[48]
<i>Sesamum indicum</i>	diploid	24075	16	0,00066	2,1 Mb	[49]
<i>Solanum lycopersicum</i>	diploid	25613	7	0,00027	66,7 Mb	[50]
<i>Zea mays</i>	diploid	34337	13	0,00038	226,4 Mb	[51]
<b>Halophytes</b>						
<i>Aeluropus litoralis</i>	diploid	15916	10	0,00063	3,6 kb	[52]
<i>Beta vulgaris</i>	diploid	24491	7	0,00029	2 Mb	[53]
<i>Eutrema salsugineum</i>	diploid	26943	10	0,00037	13,4 Mb	[54]
<i>Limonium bicolor</i>	diploid	38444	17	0,00044	340,4 Mb	[55]
<i>Phoenix dactylifera</i>	diploid	29239	11	0,00038	4,7 Mb	[56]
<i>Setaria viridis</i>	diploid	28032	10	0,00036	11,2 Mb	[57]
<i>Solanum chilense</i>	diploid	25885	8	0,00031	70,7 kb	[58]
<i>Suaeda aralocaspica</i>	diploid	29604	9	0,00030	69,5 kb	[59]
<i>Zostera marina</i>	diploid	20450	8	0,00039	485,6 kb	[60]





**Fig. 2** Comparison of the number of plasma membrane (PM)  $H^+$ -ATPases (P3A ATPases) in halophytes and glycophytes. **A** The total numbers of protein-coding genes in the genomes of halophytes and glycophytes are listed in Table 3. **B** Comparison of the number of PM  $H^+$ -ATPases in the listed halophytes and glycophytes. **C** Comparison of PM  $H^+$ -ATPases in halophytes and glycophytes after normalization to the total number of protein-coding genes. Whiskers are minimum and maximum, the box is upper and lower quartile and median. ns, no significance

## Discussion

### An expanded two-motif method for determining the number of genes in a family

We have described a simple and unbiased method for determining the number of genes within a gene family that circumvents the use of a transcriptome and/or predicted proteome (Fig. 3). The genome of interest is translated in all six reading frames to produce six protein sequences for each scaffold or chromosome, be it complete or fragmented. The protein sequences are used to build a BLAST database that can be searched with a short query sequence for the gene family in question. To apply the method, a gene family in question must contain two motifs in close proximity to each other. The query sequence can then be produced by aligning the protein sequences, and the consensus sequence adjacent to the two motifs can then be used to search the translated genome database. Once a sequence containing the motifs has been identified, it must be verified that (1) the motifs are located in the correct orientation relative to each

other and (2) in close proximity to each other. If the hit fulfills these criteria, it can be considered to constitute a gene in the family. The method described can be used to identify genes of specific families in the P-type ATPase superfamily and most likely also in other gene families.

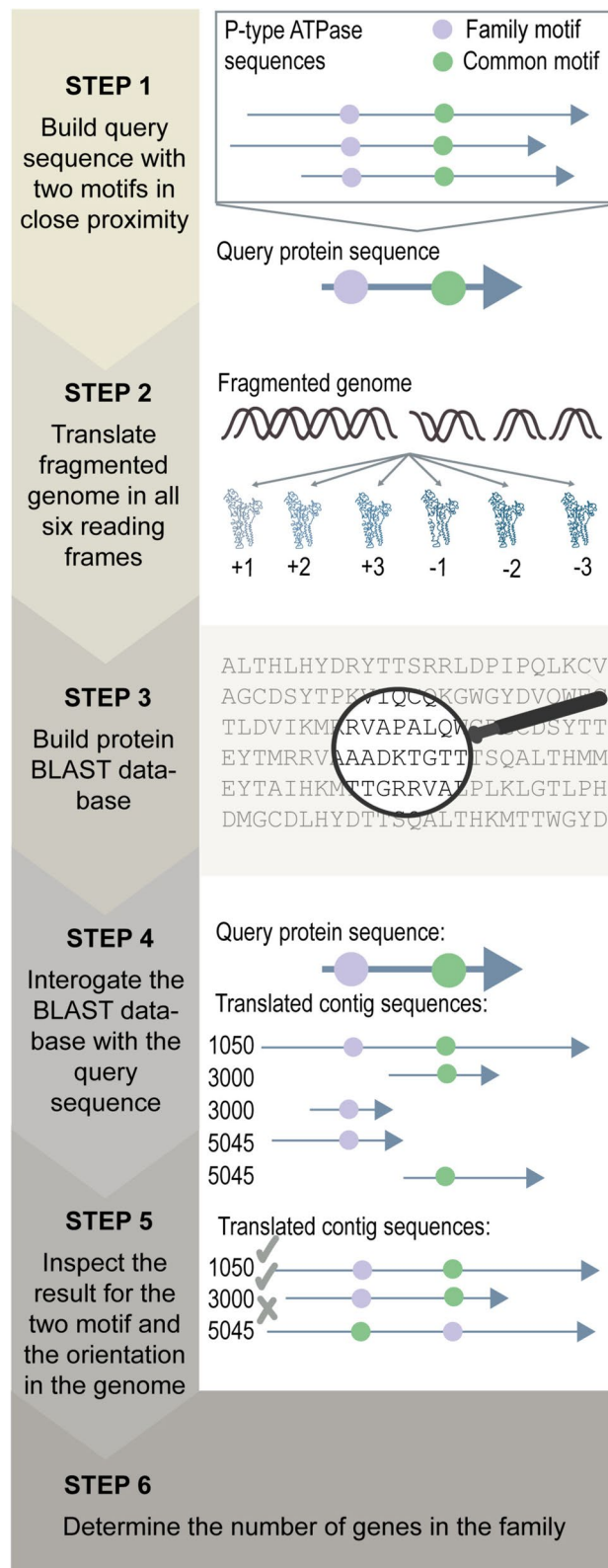
### Comparison with other methods

Tools are available for such analysis of novel genomes, as the Pfam database is based on sequence alignments and identification of common sequence motifs in gene family members. P-type ATPases form a large superfamily of primary active pumps divided into five major families. A unique motif for all members of the superfamily is the sequence DKTGT, but Pfam entries for the families remain to be identified.

An HMM has been used to identify soybean P2-type  $Ca^{2+}$  ATPase genes in the *G. max* genome [33]. Using the automatic annotation according to these criteria, 12 P2-type ATPases were identified, which is a number far below the

(See figure on next page.)

**Fig. 3** Flow chart of the procedure used to predict the number of genes for a family in a species of an incompletely assembled genome using the expanded two-motif method. An overview of the workflow for identifying all members of a specific family. Step 1: A query sequence is built by aligning sequences from the desired family. The query sequence should contain two motifs, with at least one being unique to the family. Step 2: The genome for the organism of interest is translated in all six reading frames. Step 3: A BLAST database is built using the translated genome. Step 4: The BLAST database is searched with the query sequence. Step 5: The BLAST result is inspected; the orientation and distance (if there are few or no gaps between the motifs in the family) of the two motifs is used to determine if a hit can constitute a member of the gene family. In our hypothetical example, inspection revealed hits on three hypothetical contigs: 1050, 3000, and 5045. For Contig 1050, the two motifs are close and correctly oriented. For 3000, the two motifs are located on each side of an intron; thus, the position in the genome is used to verify the orientation and establish if the hits are located relatively close to each other. 5045 is a negative hit. In this contig, there are two hits, but inspection revealed that the motifs are incorrectly oriented and therefore is deemed not to constitute a member of the family



**Fig. 3** (See legend on previous page.)

23 P2A/B ATPases we identified in this genome using our method based on the expanded two-motif method.

#### Inventory of P-type ATPases in *H. vulgare* and identification of new family members

Using conventional methods, we identified all P-type ATPases presently annotated in the *H. vulgare* cultivar Morex V3 HC genome. *HvHMA7*, *HvACA3*, *HvACA10*, *HvAHA8*, and *HvALA6* were found to be incorrectly annotated when compared to the homologs from *A. thaliana* and *O. sativa* but could be corrected based on homology, intron–exon splice sites, and cloning.

We identified two novel P-type ATPases, namely, *HvHMA1-like* and *HvHMA2-like*. *HvHMA2* and *OsHMA2* are both involved in the root-to-shoot transport of zinc and belong to the same phylogenetic sub-clade of P1B ATPases. Among monocots, only *T. aestivum* has recently been reported to have two genes belonging to the same clade as *OsHMA2* [61]. *HvHMA2-like* was found to be a paralogue of *HvHMA2* and is located approximately 50 kb upstream of *HvHMA2*. Therefore, this gene could be the result of a gene duplication event by unequal crossover.

*HvHMA1-like* was found to have high homology to *HvHMA1* and *OsHMA1*, but *HvHMA1-like* has a premature stop codon after codon number 80, and neither *HvHMA1* nor *OsHMA1* has a premature stop codon. *HvHMA1-like* is expressed and can be cloned from cDNA from 10-day-old whole-seedling cDNA. This suggests that the N-terminal region of the protein is expressed, but its role, if any, for example, in regulating heavy metal transport, remains unknown.

#### Evaluation of methods

We first surmised that the presence of two distinct motifs near each other with a fixed distance between them could be used to identify the genes in a specific family. As we tested this hypothesis, we found that even though the two motifs were relatively close together, a large part of the sought-after genes had an intron between the two motifs and could not be identified using this method. This shortcoming limits the effectiveness of the approach; however, though it can identify genes without an intron extremely well.

To address this shortcoming, we devised a method that expanded on the two-motif method to include more information between and surrounding the two motifs. The expanded two-motif method can successfully identify genes in which the two motifs are separated by an intron but fails to identify genes when the bait is distributed across three exons. This could potentially be negated by including more sequence information on either side of the motifs. However, if this is done, it

might not be a feasible method for fragmented genomes where continuity is low.

#### P3A ATPases in glycophytes and halophytes

We used the expanded two-motif approach to approximate the number of P3A ATPases in different species divided into two groups, glycophytes and halophytes. We aimed to determine if there is a genetic difference between the two groups in regards to P3A ATPases. We found no significant difference using the species selected here when the data were normalized to the number of protein-coding genes in each species. However, it may be too early to establish whether P3A ATPases have been reduced or expanded in halophytes as part of their salt tolerance mechanism. The number of species used for analysis was rather limited (as we opted to only use diploids), and the variability within each group made the difference not significant. Future studies involving a larger number of species in each group will provide an explicit answer about the possible causal link between salinity tolerance and the number of PM H<sup>+</sup>-ATPase copies.

#### Advantages and limitations of the methods

The advantage of the expanded two-motif method described here is that it can be used on poorly annotated and poorly assembled genomes and on any gene family in which two unique sequence motifs occur in close proximity to each other.

When using the two-motif method when the distance between the two motifs is fixed, the rate of false positives will only be impacted by the probability of the two motifs randomly occurring with the specific distance between them. This was only observed once during the testing for *HvP5*. The larger the distance between the two motifs, the higher the possibility of them being split onto two scaffolds/contigs and vice versa or being separated by an intron. The distance should therefore be kept relatively short.

For the expanded two-motif method, if there is an intron between the motifs, it does not necessarily prevent identification but requires more work. The BLAST output will show two hits from the same scaffold/contig, and it will be necessary to investigate the position of the two hits, both regarding the orientation of the two compared to each other and the distance between them. Furthermore, as is evident from the testing, when the query sequence is distributed onto three exons, it is impossible to identify the family member using this approach. The method described will identify the location of the two motifs and thus will not identify the whole gene. Therefore, the method is prone to error

because it cannot differentiate between true genes and pseudogenes. The expanded two-motif method proved useful for identifying gene members of the superfamily of P-type ATPases. Using the expanded two-motif method to screen well annotated genomes, we successfully identified all members of P3A and P5 ATPases even though the two motifs in some members of these families are separated by an intron.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-023-09859-4>.

**Additional file 1: Supplementary file 1.** *Hordeum vulgare* P1B ATPase coding sequences.

**Additional file 2: Supplementary file 2.** *Hordeum vulgare* P2A ATPase coding sequences.

**Additional file 3: Supplementary file 3.** *Hordeum vulgare* P2B ATPase coding sequences.

**Additional file 4: Supplementary file 4.** *Hordeum vulgare* P3A ATPase coding sequences.

**Additional file 5: Supplementary file 5.** *Hordeum vulgare* P4 ATPase coding sequences.

**Additional file 6: Supplementary file 6.** *Hordeum vulgare* P5 ATPase coding sequences.

**Additional file 7: Supplementary file 7.** P-type ATPase protein sequences from *A. thaliana*, *O. sativa*, and *H. vulgare*.

**Additional file 8: Supplementary Table 1.** Genomes and assemblies used for the different species.

**Additional file 9: Supplementary Table 2.** Percent amino acid identity between P3A-ATPases (AHAs) from *A. thaliana*, *O. sativa* and *H. vulgare*.

**Additional file 10: Supplementary Table 3.** Number of introns in P-type ATPases from *A. thaliana*, *O. sativa* and *H. vulgare*.

**Additional file 11: Supplementary Table 4.** Annotated and estimated number of plasma membrane H<sup>+</sup>-ATPases in annotated plant genomes.

**Additional file 12: Supplementary Table 5.** Overview of location, sequence hit, and chromosome location for Ca<sup>2+</sup>-ATPases in the G. max assembly GCA\_022114995.1

**Additional file 13: Supplementary Table 6.** Number of protein-coding genes used for generation of Fig. 2.

**Additional file 14: Supplementary Figure 1.** Alignment of Heavy Metal ATPases (P1B ATPases) from *A. thaliana*, *O. sativa*, and *H. vulgare*. **Supplementary Figure 2.** Alignment of P2A ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*. **Supplementary Figure 3.** Alignment of P2B ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*. **Supplementary Figure 4.** Alignment of P3A ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*. **Supplementary Figure 5.** Alignment of P4 ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*. **Supplementary Figure 6.** Alignment of P5 ATPases from *A. thaliana*, *O. sativa*, and *H. vulgare*.

## Acknowledgements

We thank Anett Stéger for help with preparing Fig. 3.

## Authors' contributions

M.P. and S.S. conceptualised the research idea. A.F.N. and M.P. designed all experiments. A.F.N. performed all experiments. A.F.N. and M.P. wrote the first draft of the manuscript. All authors contributed to analysing the data. All authors contributed to the writing of the final manuscript.

## Funding

Open access funding provided by Copenhagen University The research was supported by Novo Nordisk Fonden (NovoCrops; M.P.), Carlsbergfondet (RaisingQuinoa; M.P.), Australian Research Council (S.S.), and the National Natural Science Foundation of China (S.S.).

## Availability of data and materials

Sequence data used in this article are available either at the KEGG webpage (<https://www.genome.jp/kegg/>), the National Library of Medicine (<https://www.ncbi.nlm.nih.gov/>), or the The Leibniz Institute of Plant Genetics and Crop Plant Research (<https://galaxy-web.ipk-gatersleben.de/>). Links to the genome assemblies used can be found in Supplementary Tables 1 and codes for organisms in KEGG can be found in Supplementary Table 4. P-type ATPase CDSs from *A. thaliana* and *O. sativa* can be found in refs. [17, 18]. P-type ATPase CDSs from *H. vulgare* can be found in Supplementary files 1, 2, 3, 4, 5 and 6, with all accession numbers given in Table 1.

## Declarations

### Ethics approval and consent to participate

Research on plants was conducted in accordance with institutional, national, and international guidelines and legislations.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

Received: 1 September 2023 Accepted: 29 November 2023

Published online: 03 January 2024

## References

- RefSeq. growth statistics. <http://www.ncbi.nlm.nih.gov/genbank/statistics/>.
- GenBank, Statistics WGS. <http://www.ncbi.nlm.nih.gov/genbank/>.
- NCBI's Sequence Read Archive SRA. database growth <http://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/>.
- O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44:D733-745.
- Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011;39:D19-21.
- Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data: Astronomical or Genomical? *PLoS Biol.* 2015;13: e1002195.
- Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, Liu X, Soltis PS. Green plant genomes: what we know in an era of rapidly expanding opportunities. *Proc Natl Acad Sci U S A.* 2022;119: e2115640118.
- Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS ONE.* 2011;6: e21400.
- Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol.* 2014;10: e1003998.
- Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. *BMC Genomics.* 2020;21:293.
- Hoffmann RD, Portes MT, Olsen LI, Damineli DSC, Hayashi M, Nunes CO, Pedersen JT, Lima PT, Campos C, Feijó JA, et al. Plasma membrane H<sup>+</sup>-ATPases sustain pollen tube growth and fertilization. *Nat Commun.* 2020;11:2395.
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 2021;49:D412-d419.
- Eddy SR. Profile hidden Markov models. *Bioinformatics.* 1998;14(9):755-63.

14. Sánchez-Pulido L, Yuan YP, Andrade MA, Bork P. NAIL-Network Analysis Interface for linking HMMER results. *Bioinformatics*. 2000;16:656–7.
15. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*. 2002;3:265–74.
16. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol*. 2019;20:232.
17. Baxter I, Tchieu J, Sussman M, Boutry M. Genomic comparison of P-type ATPase ion pumps in Arabidopsis and rice. *Plant Phys*. 2003;132:618–28.
18. Axelsen KB, Palmgren MG. Inventory of the superfamily of P-type ion pumps in Arabidopsis. *Plant Phys*. 2001;126:696–706.
19. IPK database. [https://webblast.ipk-gatersleben.de/barley\\_ibsc/downloads/](https://webblast.ipk-gatersleben.de/barley_ibsc/downloads/).
20. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44:D279–285.
21. Edgar Robert C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004;5: 113.
22. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 gateway computing environments workshop (GCE). 2010;1–8.
23. Palmgren MG, Nissen P. P-type ATPases. *Annu Rev Biophys*. 2011;40:243–66.
24. Pedersen CN, Axelsen KB, Harper JF, Palmgren MG. Evolution of plant P-type ATPases. *Front Plant Sci*. 2012;3:31.
25. Arguello JM. Identification of ion-selectivity determinants in heavy-metal transport P1B-type ATPases. *J Membr Biol*. 2003;195:93–108.
26. Palmgren M, Østerberg JT, Nintemann SJ, Poulsen LR, López-Marqués RL. Evolution and a revised nomenclature of P4 ATPases, a eukaryotic family of lipid flippases. *Biochim Biophys Acta Biomembr*. 2019;1861:1135–51.
27. Møller AB, Asp T, Holm PB, Palmgren MG. Phylogenetic analysis of P5 P-type ATPases, a eukaryotic lineage of secretory pathway pumps. *Mol Phylogenet Evol*. 2008;46:619–34.
28. Mikkelsen MD, Pedas P, Schiller M, Vincze E, Mills RF, Borg S, Møller A, Schjoerring JK, Williams LE, Baekgaard L, et al. Barley HvHMA1 is a heavy metal pump involved in mobilizing organellar Zn and Cu and plays a role in metal loading into grains. *PLoS ONE*. 2012;7: e49027.
29. Mills RF, Peaston KA, Runions J, Williams LE. HvHMA2, a P<sub>1B</sub>-ATPase from barley, is highly conserved among cereals and functions in Zn and Cd transport. *PLoS ONE*. 2012;7: e42640.
30. Wang X-K, Gong X, Cao F, Wang Y, Zhang G, Wu F. HvPAA1 encodes a P-type ATPase, a novel gene for cadmium accumulation and tolerance in barley (*Hordeum vulgare* L.). *Int J Mol Sci*. 2019;20: 1732.
31. Sørensen DM, Holen HW, Hølemans T, Vangheluwe P, Palmgren MG. Towards defining the substrate of orphan P5A-ATPases. *Biochim Biophys Acta Biomembr*. 2015;1850:524–35.
32. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30.
33. Wang J, Fu X, Zhang S, Chen G, Li S, Shangguan T, Zheng Y, Xu F, Chen ZH, Xu S. Evolutionary and regulatory pattern analysis of soybean Ca<sup>2+</sup> ATPases for abiotic stress tolerance. *Front Plant Sci*. 2022;13: 898256.
34. Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, et al. Genome duplication in soybean (*Glycine subgenus soja*). *Genetics*. 1996;144:329–38.
35. Maughan PJ, Turner TB, Coleman CE, Elzinga DB, Jellen EN, Morales JA, Udall JA, Fairbanks DJ, Bonifacio A. Characterization of Salt overly sensitive 1 (SOS1) gene homologs in quinoa (*Chenopodium quinoa* Willd.). *Genome*. 2009;52:647–57.
36. Shi H, Ishitani M, Kim C, Zhu JK. The Arabidopsis thaliana salt tolerance gene SOS1 encodes a putative Na<sup>+</sup>/H<sup>+</sup> antiporter. *Proc Natl Acad Sci U S A*. 2000;97:6896–901.
37. Shi H, Lee B-h, Wu S-J, Zhu J-K. Overexpression of a plasma membrane Na<sup>+</sup>/H<sup>+</sup> antiporter gene improves salt tolerance in Arabidopsis thaliana. *Nat Biotechnol*. 2002;21:81–5.
38. Shi H, Quintero FJ, Pardo JM, Zhu J-K. The putative plasma membrane Na<sup>+</sup>/H<sup>+</sup> antiporter SOS1 controls long-distance Na<sup>+</sup> transport in plants. *Plant Cell*. 2002;14:465–77.
39. Shabala S, Bose J, Fuglsang AT, Pottosin I. On a quest for stress tolerance genes: membrane transporters in sensing and adapting to hostile soils. *J Exp Bot*. 2016;67:1015–31.
40. Shabala S, Cui TA. Potassium transport and plant salt tolerance. *Physiol Plant*. 2008;133:651–69.
41. Food and Agriculture Organization of the United Nations. FAO irrigation and drainage paper 61. Annex. 1. Crop salt tolerance data. 2002. [www.fao.org/3/y4263e/y4263e0e.htm](http://www.fao.org/3/y4263e/y4263e0e.htm).
42. Finkers R, van Kaaunen M, Ament K, Burger-Meijer K, Egging R, Huits H, Kodde L, Kroon L, Shigyo M, Sato S, et al. Insights from the first genome assembly of Onion (*Allium cepa*). G3 (Bethesda). 2021;11:jkab243.
43. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. The genome of *Eucalyptus grandis*. *Nature*. 2014;510:356–62.
44. Udall JA, Long E, Hanson C, Yuan D, Ramaraj T, Conover JL, Gong L, Arick MA, Grover CE, Peterson DG, et al. De novo genome sequence assemblies of *Gossypium Raimondii* and *Gossypium turneri*. G3 (Bethesda). 2019;9:3079–85.
45. Sato K. History and future perspectives of barley genomics. *DNA Res*. 2020;27: dsaa023.
46. Eckardt NA. Grass genome evolution. *Plant Cell*. 2008;20:3–4.
47. McClean PE, Lavin M, Gepts P, Jackson SA. *Phaseolus vulgaris*: a diploid model for soybean. In: New York, NY: Springer New York; 2008. p. 55–76.
48. Mitsui Y, Shimomura M, Komatsu K, Namiki N, Shibata-Hatta M, Imai M, Katayose Y, Mukai Y, Kanamori H, Kurita K, et al. The radish genome and comprehensive gene expression profile of tuberous root formation and development. *Sci Rep*. 2015;5:10835–10835.
49. Wang X, Wang S, Lin Q, Lu J, Lv S, Zhang Y, Wang X, Fan W, Liu W, Zhang L, et al. The wild allotetraploid sesame genome provides novel insights into evolution and lignan biosynthesis. *J Adv Res*. 2023;50:13–24.
50. Barone A, Chiusano ML, Ercolano MR, Giuliano G, Grandillo S, Frusciante L. Structural and functional genomics of tomato. *Int J Plant Genomics*. 2008;2008:820274.
51. Gaut BS, Le Maud AS, Sawkins MC. Maize as a model for the evolution of plant nuclear genomes. *Proc Natl Acad Sci U S A*. 2000;97:7008–15.
52. Hashemi-Petroudi SH, Arab M, Dolatabadi B, Kuo Y-T, Baez MA, Himmelmelbach A, Nematzadeh G, Maibody SAMM, Schmutz T, Mälzer M, et al. Initial description of the genome of *Aeluropus litoralis*, a halophile grass. *Front Plant Sci*. 2022;13:906462–906462.
53. Galewski P, McGrath JM. Genetic diversity among cultivated beets (*Beta vulgaris*) assessed via population-based whole genome sequences. *BMC Genomics*. 2020;21:189–189.
54. Xiao M, Hao G, Guo X, Feng L, Lin H, Yang W, Chen Y, Zhao K, Xiang L, Jiang X, et al. A high-quality chromosome-level *Eutrema salsugineum* genome, an extremophile plant model. *BMC Genomics*. 2023;24:174–174.
55. Yuan F, Wang X, Zhao B, Xu X, Shi M, Leng B, Dong X, Lu C, Feng Z, Guo J, et al. The genome of the recretalhalophyte *Limonium bicolor* provides insights into salt gland development and salinity adaptation during terrestrial evolution. *Mol Plant*. 2022;15:1024–44.
56. Chao CT, Krueger RR. The date palm (*Phoenix dactylifera* L.): overview of biology, uses, and cultivation. *HortScience*. 2007;42:1077–82.
57. Pinghua LI, Brutnell TP. *Setaria viridis* and *Setaria italica*, model genetic systems for the Panicoid grasses: exploiting the engine of C4 photosynthesis. *J Exp Bot*. 2011;62:3031–7.
58. Böndel KB, Nosenko T, Stephan W. Signatures of natural selection in abiotic stress-responsive genes of *Solanum chilense*. *R Soc Open Sci*. 2018;5:171198–171198.
59. Cheng Y, Yang P, Zhao L, Priyadarshani SVGN, Zhou Q, Li Z, Li W, Xiong J, Lin Z, Li L, et al. Studies on genome size estimation, chromosome number, gametophyte development and plant morphology of salt-tolerant halophyte *Suaeda salsa*. *BMC Plant Biol*. 2019;19:473–473.
60. Olsen J, Rouzé P, Verhelst B, Lin Y-C, Bayer T, Collen J, Dattolo E, De Paoli E, Dittami S, Maumus F, et al. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature*. 2016;530:331–5.
61. Amini S, Arsova B, Hanikenne M. The molecular basis of zinc homeostasis in cereals. *Plant Cell Environ*. 2022;45:1339–61.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.