

RESEARCH

Open Access



AcademH, a lineage of *Academ* DNA transposons encoding helicase found in animals and fungi

Kenji K. Kojima

Abstract

Background: DNA transposons are ubiquitous components of eukaryotic genomes. *Academ* superfamily of DNA transposons is one of the least characterized DNA transposon superfamilies in eukaryotes. DNA transposons belonging to the *Academ* superfamily have been reported from various animals, one red algal species *Chondrus crispus*, and one fungal species *Puccinia graminis*. Six *Academ* families from *P. graminis* encode a helicase in addition to putative transposase, while some other families encode a single protein which contains a putative transposase and an XPG nuclease.

Results: Systematic searches on Rebase and BLAST searches against publicly available genome sequences revealed that several species of fungi and animals contain multiple *Academ* transposon families encoding a helicase. These *AcademH* families generate 9 or 10-bp target site duplications (TSDs) while *Academ* families lacking helicase generate 3 or 4-bp TSDs. Phylogenetic analysis clearly shows two lineages inside of *Academ*, designated here as *AcademH* and *AcademX* for encoding helicase or XPG nuclease, respectively. One sublineage of *AcademH* in animals encodes plant homeodomain (PHD) finger in its transposase, and its remnants are found in several fish genomes.

Conclusions: The *AcademH* lineage of TEs is widely distributed in animals and fungi, and originated early in the evolution of *Academ* DNA transposons. This analysis highlights the structural diversity in one less studied superfamily of eukaryotic DNA transposons.

Keywords: DNA transposon, *Academ*, Helicase, Target site duplication, Transposase, Plant homeodomain

Introduction

Transposable elements (TEs), or transposons are ubiquitous components of genomes in all three domains of life [1, 2]. TEs are traditionally classified into 2 classes: Class I retrotransposons and Class II DNA transposons [3]. Autonomous retrotransposons encode a reverse transcriptase and during the transposition, the information of RNA is transformed into DNA by reverse transcription. DNA transposons do not have a process of reverse transcription in their transposition cycle. At least 5

independent DNA-cleaving/recombining enzymes (DDE transposase or DDD/E transposase, tyrosine recombinase, serine recombinase, HUH nuclease, and Cas1 endonuclease) have been incorporated into TEs and related mobile genetic elements [4, 5]. DDE transposase or integrase is the most ubiquitous enzyme that functions as transposase of DNA transposons, as well as of long terminal repeat (LTR) retrotransposons and of retroviruses [6]. Eukaryotic DNA transposons are now classified into around 20 superfamilies [1]. Most of these superfamilies, such as *Mariner/Tc1* and *Harbinger/PIF1*, are known to encode a DDE transposase.

Correspondence: kojima@girinst.org

Genetic Information Research Institute, Cupertino, CA 95014, USA



© The Author(s). 2020 **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.

DDE transposase is topologically a member of RNaseH-like fold [6]. The conserved core of the transposase domain is $\beta 1$ - $\beta 2$ - $\beta 3$ - $\alpha 1$ - $\beta 4$ - $\alpha 2/3$ - $\beta 5$ - $\alpha 4$ - $\alpha 5$. Three acidic residues, DDD or DDE play the central role in the transposition. The first D is located on $\beta 1$ and the second D on or just after $\beta 4$. The last D or E is on or just before $\alpha 4$. In the case of the integrase encoded by human immunodeficiency virus type 1 (HIV-1), the distance between the second D and the last E is 35 residues. In some DNA transposons, the catalytic core domain between $\beta 5$ and $\alpha 4$ is extended by “insertion domain.” In the case of RAG1, recombination activating gene 1, which originated from an eukaryotic DNA transposon superfamily *Transib* [7], the insertion domain is 264 residues in length and entirely α -helical [6]. The transposase encoded by *Hermes*, a member of eukaryotic DNA transposon superfamily *hAT*, contains a 288-aa-long insertion domain [6].

The *Academ* superfamily of eukaryotic DNA transposons was first described by Kapitonov and Jurka [8] from various animals. To date, *Academ* has been found from animals, fungi, and plants [1]. In animals, *Academ* is widely distributed and found from genomes of 7 phyla: Chordata, Hemichordata, Echinodermata, Annelida, Mollusca, Arthropoda, and Cnidaria. In contrast, in fungi and in plants, only one species of each group is reported to have *Academ* transposons: a red alga *Chondrus crispus* [9] and a pathogenic fungus *Puccinia graminis* [10], while the wide distribution of *Academ* in fungi was suggested [11]. The transposase domain of *Academ* is predicted to be a DDE transposase [12]. An entirely α -helical insertion

domain was predicted between $\beta 5$ and $\alpha 4$, as are the cases of RAG1 and *Hermes*. Another insertion domain was predicted between $\beta 2$ and $\beta 3$, unlike any other transposases. Many *Academ* families encode a large protein that contains three recognizable domains, a transposase, an XPG nuclease, and a putative Cys8 zinc finger [8] (Fig. 1). These three domains can be recognized among *Academ* families from animals and *C. crispus*. The *Academ* families from *P. graminis* do not encode an XPG nuclease. Instead, they encode a superfamily II helicase as a separate protein [10] (Fig. 1). This lineage was designated as *AcademH*. It is not yet known whether the presence of helicase is a recently acquired characteristic specific for *Academ* families from *P. graminis*, or it is an ancient trait shared by various *Academ* families from diverse organisms.

In this study, many families of *AcademH* from other fungi and animals were characterized. No intact *AcademH* transposons were found from vertebrates, but some fish genomes still contain remnants of *AcademH* transposons. *AcademH* shows 9 or 10-bp target site duplications (TSDs), although other *Academ* shows 3 or 4-bp TSDs. The sequence comparison and phylogenetic analysis revealed two independent lineages with different TSD length and protein composition inside of the *Academ* DNA transposons.

Results

Academ families encoding a superfamily II helicase

Manual inspection of Repbase entries revealed that besides 6 *AcademH* families from *P. graminis*, *Academ-1_ADi*, *Academ-2_ADi*, *Academ-3_ADi* from the coral *Acropora digitifera*, and *Academ-2_CGi* from the Pacific

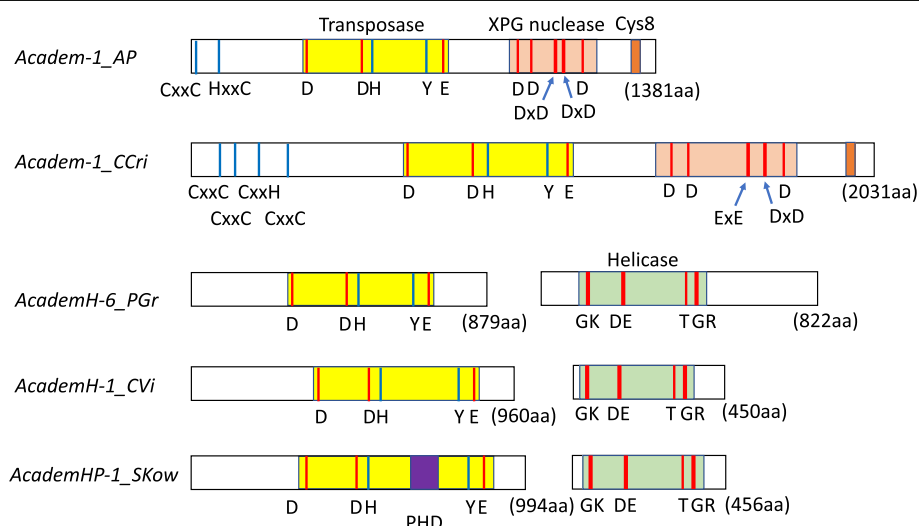


Fig. 1 Schematic structures of proteins encoded by *Academ* transposons. Encoded proteins are shown as open boxes, and inside them, protein domains are shown as colored boxes. Conserved residues are shown as bars in red and blue. Protein lengths are shown in parentheses. Transposase and helicase proteins of *AcademH* families are encoded in opposite directions

oyster *Crassostrea gigas* also encode a superfamily II helicase protein. Using the helicase protein sequences from these families as queries, Censor search [13] against published genome sequences was performed. It led to the characterization of *AcademH* families encoding a helicase protein from 7 species of basidiomycetes fungi (*Laccaria bicolor*, *Puccinia coronata*, *Puccinia horiana*, *Puccinia striiformis*, *Puccinia triticina*, *Serpula lacrymans*), one species of fungi in Mucoromycotina (*Lobosporangium transversale*), and another oyster *Crassostrea virginica*, in addition to more families from the three species above (Table 1 and Supplementary Dataset S1). Non-autonomous DNA transposons showing similarity in terminal regions with *AcademH* families were also found from three cnidarians (*Exaiptasia pallida*, *Orbicella faveolata*, and *Stylophora pistillata*) and the Yesso scallop *Mizuhopecten yessoensis* (Table 1 and Supplementary Dataset S1).

With two of these characterized *AcademH* protein sequences (*AcademH-6_PGr* and *AcademH-1_CVi*) as queries, BLASTP search against the non-redundant protein sequences (nr) at NCBI BLAST website hits many proteins from diverse fungi and animals (Supplementary Table S1). In fungi, proteins related to *AcademH* transposases were found from three subdivisions (Agaricomycotina, Pucciniomycotina, Ustilaginomycotina) within Basidiomycota, one subdivision (Pezizomycotina) within Ascomycota, and one subdivision (Mortierellomycotina) within Mucoromycota. Despite the report that *Academ* transposons are widely distributed in fungi [11], no other fungal group was revealed to contain *AcademH* transposons in this analysis. In animals, genomes from 9 phyla (Porifera, Cnidaria, Mollusca, Annelida, Brachiopoda, Priapulida, Chordata, Hemichordata, and Echinodermata) encode proteins related to *AcademH* transposases. Most of these protein sequences were encoded by single-copy, non-repetitive sequences. Basidiomycetes fungi with more than 5 protein hits and all other species were further analyzed. If their terminal inverted repeats (TIRs) longer than 10 bp and TSDs adjacent to TIRs could be detected in flanking 10,000-bp sequences, they were considered as full-length *Academ* transposons (Table 1 and Supplementary Dataset S2). Most of these single-copy *Academ* transposons encode a helicase protein. The sequence lengths, numbers of uninterrupted full-length copies, and the sequence identities to the consensus sequences are shown in Supplementary Table S2.

Secondary structure-based protein homology search HHpred was performed with helicase proteins encoded by *AcademH* DNA transposons. The top hit was RecQ DNA helicase from *Escherichia coli*, followed by U5 small nuclear ribonucleoprotein 200 and RNA helicase Vasa. The pairwise alignment generated by HHpred and multiple protein alignment generated by MAFFT were combined. It revealed that *AcademH* helicases conserve

all motifs important for catalytic reactions, nucleic acid binding, and ATP binding (Fig. 2a). Censor search using helicase proteins encoded by *AcademH* against Repbase hit some families of *KolobokH*, a lineage of *Kolobok* DNA transposons encoding a helicase [14]. However, helicases encoded by *AcademH* and *KolobokH* are not so closely related to each other and are likely acquired independently in these two lineages of DNA transposons (data not shown). Helicases encoded by *Helitron* DNA transposons are Superfamily I helicases related to PIF1 helicase [15], and thus, there is little sequence similarity between helicases encoded by *AcademH* and *Helitron*.

Academ families without helicase often, but not always, contain 1 long open reading frame for a large protein containing three recognizable domains: a transposase, an XPG nuclease, and a putative Cys8 zinc finger (Figs. 1 and 2). Here, *Academ* families with XPG nuclease are designated as *AcademX*. In contrast, *AcademH* usually contain introns and encodes two proteins in opposite directions. These two proteins are encoded without overlapping. None of *AcademH* families encode an XPG nuclease or a Cys8 zinc finger.

Longer TSDs generated by *AcademH* than *AcademX* families

It is reported that *AcademX* DNA transposons generate 3-bp or 4-bp TSDs [8, 16]. In contrast, *AcademH* generates relatively long TSDs. Fungal *AcademH* families generate 9-bp TSDs with some exceptions (Fig. 3, and Supplementary Fig. S1). Animal *AcademH* families generate 9 or 10-bp TSDs (Fig. 4 and Supplementary Fig. S2). In the genome of coral *A. digitifera*, both lineages of *Academ* DNA transposons (*AcademH* and *AcademX*) are present. *AcademH* is usually inserted with 9-bp TSDs. *AcademX* generates 3-bp TSDs the same as previously reported *AcademX* DNA transposons from animals.

Sequence comparison against reported non-autonomous TEs deposited in Repbase revealed that some of non-autonomous DNA transposons whose classification has not yet been determined are either *AcademX* or *AcademH* (Supplementary Table S3). DNA transposons with 8-bp or 9-bp TSDs show sequence similarity to *AcademH* termini while DNA transposons with 3-bp TSDs show sequence similarity to *AcademX* termini. One fungal species *Melampsora larici-populina*, closely related to *Puccinia*, and *Nematostella vectensis*, similarly to other cnidarian species, contain non-autonomous *AcademH* families (Table 1 and Supplementary Tables S2 and S3).

The presence of a pyrimidine (C or T) at the 5' terminus and a purine (G or A) at the 3' terminus is shared among almost all *Academ* families (Figs. 3 and 4). Some *Academ* families contain > 100-bp TIRs, represented by 526-bp TIRs of *AcademH-1_LoTr* and 575-bp TIRs of *AcademH-16_CVi*, while some have shorter than 10-bp TIRs; for

Table 1 *AcademH* distribution

| Classification | Organism | <i>AcademH</i> families |
|------------------------------------|---|--|
| Fungi/Basidiomycota | | |
| -Pucciniomycetes/Pucciniales | <i>Puccinia coronata</i> | <i>AcademH-1_PCor</i> to <i>17_PCor</i> , <i>AcademH-14N1_PCor</i> , <i>AcademH-N1_PCor</i> to <i>N13_PCor</i> |
| | <i>Puccinia graminis</i> | <i>AcademH-1_PG</i> to <i>6_PG</i> , <i>AcademH-N2_PG</i> to <i>N6_PG</i> , <i>AcademH-2B_PG</i> , <i>AcademH-N3B_PG</i> |
| | <i>Puccinia horiana</i> | <i>AcademH-1_PHor</i> , <i>AcademH-N1_PHor</i> to <i>N25_PHor</i> |
| | <i>Puccinia sorghi</i> | <i>AcademH-N1_PSor</i> |
| | <i>Puccinia striiformis</i> | <i>AcademH-1_PSt</i> to <i>18_PSt</i> , <i>AcademH-N1_PSt</i> to <i>N12_PSt</i> |
| | <i>Puccinia triticina</i> 1–1 BBBB Race 1 | <i>AcademH-1_PTrit</i> to <i>6_PTrit</i> , <i>AcademH-1B_PTrit</i> , <i>AcademH-N1_PTrit</i> to <i>N20_PTrit</i> |
| | <i>Melampsora larici-populina</i> | <i>AcademH-N1_MLP</i> |
| -Agaricomycetes/Auriculariales | <i>Exidia glandulosa</i> HHB12029 | <i>AcademH-1_ExGI</i> |
| -Agaricomycetes/Atheliales | <i>Fibularhizoctonia</i> sp. CBS 109695 | <i>AcademH-1_FiCBS</i> |
| -Agaricomycetes/Agaricales | <i>Gymnopilus dilepis</i> | <i>AcademH-1_GyDi</i> |
| | <i>Hebeloma cylindrosporum</i> h7 | <i>AcademH-1_HeCy</i> |
| | <i>Laccaria bicolor</i> S238N-H82 | <i>AcademH-1_LB</i> , <i>AcademH-2_LB</i> |
| | <i>Panaeolus cyanescens</i> | <i>AcademH-1_PaCy</i> , <i>AcademH-2_PaCy</i> |
| -Agaricomycetes/Amylocorticiales | <i>Plicaturopsis crispa</i> FD-325 SS-3 | <i>AcademH-1_PICr</i> to <i>3_PICr</i> |
| -Agaricomycetes/Boletales | <i>Serpula lacrymans</i> | <i>AcademH-1_SLL</i> |
| -Agaricomycetes/Hymenochaetales | <i>Schizopora paradoxa</i> | <i>AcademH-1_ScPa</i> , <i>AcademH-2_ScPa</i> |
| -Agaricomycetes/Polyporales | <i>Dichomitus squalens</i> LYAD-421 SS1 | <i>AcademH-1_DiSq</i> |
| | <i>Trametes cinnabarina</i> | <i>AcademH-1_TrCi</i> , <i>AcademH-2_TrCi</i> |
| | <i>Trametes pubescens</i> | <i>AcademH-1_TrPu</i> |
| -Exobasidiomycetes/Tilletiales | <i>Tilletia caries</i> | <i>AcademH-1_TiCa</i> , <i>AcademH-2_TiCa</i> |
| | <i>Tilletia indica</i> | <i>AcademH-1_TiIn</i> |
| Fungi/Ascomycota | | |
| -Pezizomycetes/Pezizales | <i>Ascobolus immersus</i> RN42 | <i>AcademH-1_Aslm</i> , <i>AcademH-2_Aslm</i> |
| Fungi/Mucoromycota | | |
| -Mortierellomycetes/Mortierellales | <i>Lobosporangium transversale</i> | <i>AcademH-1_LoTr</i> |
| | <i>Mortierella verticillata</i> NRRL 6337 | <i>AcademH-1_MoVe</i> , <i>AcademH-2_MoVe</i> |
| Metazoa/Porifera | | |
| -Demospongiae/Haplosclerida | <i>Amphimedon queenslandica</i> | <i>AcademH-1_AQ</i> to <i>3_AQ</i> , <i>AcademH-N1_AQ</i> to <i>N2_AQ</i> |
| Metazoa/Cnidaria | | |
| -Anthozoa/Scleractinia | <i>Acropora digitifera</i> | <i>Academ-1_ADi</i> to <i>3_ADi</i> , <i>AcademH-4_ADi</i> to <i>7_ADi</i> , <i>AcademH-N1_ADi</i> to <i>N14_ADi</i> |
| | <i>Orbicella faveolata</i> | <i>AcademH-N1_OrFa</i> to <i>N6_OrFa</i> |
| | <i>Stylophora pistillata</i> | <i>AcademH-1_StPi</i> , <i>AcademH-N1_StPi</i> to <i>N3_StPi</i> |
| -Anthozoa/Actiniaria | <i>Exaiptasia pallida</i> | <i>AcademH-1_ExPa</i> to <i>2_ExPa</i> , <i>AcademH-N1_ExPa</i> to <i>N3_ExPa</i> |
| | <i>Nematostella vectensis</i> | <i>AcademH-N1_NV</i> , <i>N1A_NV</i> , <i>N2_NV</i> , <i>N2A_NV</i> |
| Metazoa/Priapulida | | |
| -Priapulimorpha/Priapulimorphida | <i>Priapulius caudatus</i> | <i>AcademHP-1_PrCa</i> |
| Metazoa/Mollusca | | |
| -Bivalva/Ostreoida | <i>Crassostrea gigas</i> | <i>Academ-2_CGi</i> , <i>AcademH-1_CGi</i> to <i>3_CGi</i> , <i>AcademH-2N1_CGi</i> , <i>AcademH-N1_CGi</i> to |

Table 1 *AcademH* distribution (Continued)

| Classification | Organism | <i>AcademH</i> families |
|-----------------------------|--------------------------------------|--|
| | | <i>N3_CGi</i> , <i>AcademH-2B_CGi</i> |
| | <i>Crassostrea virginica</i> | <i>AcademH-1_CVi</i> to <i>16_CVi</i> , <i>AcademH-2N1_CVi</i> , <i>AcademH-7N1_CVi</i> , <i>AcademH-N1_CVi</i> to <i>N2_CVi</i> , <i>AcademHP-1_CVi</i> |
| -Bivalva/Pectinoida | <i>Mizuhopecten yessoensis</i> | <i>AcademH-N1_MiYe</i> to <i>N3_MiYe</i> |
| Metazoa/Hemichordata | | |
| -Euteropneusta | <i>Saccoglossus kowalevskii</i> | <i>AcademHP-1_SKow</i> to <i>2_SKow</i> , <i>AcademH-N1_SKow</i> , <i>AcademH-N1B_SKow</i> |
| Metazoa/Echinodermata | | |
| -Echinoidea/Echinoidea | <i>Strongylocentrotus purpuratus</i> | <i>AcademHP-1_SP</i> |
| Metazoa/Chordata | | |
| -Leptocardii/Amphioxiformes | <i>Branchiostoma floridae</i> | <i>AcademH-N1_BF</i> |
| | <i>Branchiostoma belcheri</i> | <i>AcademH-1_BBe</i> , <i>AcademH-2_BBe</i> |

example, *AcademH-2_PSt* and *AcademH-N13_PHor* have 8-bp TIRs.

***AcademHP*, a sublineage of *AcademH* with PHD zinc fingers**

Although no proteins from vertebrates were hit in the first iteration of PSI-BLAST search with the transposase of *AcademH-1_CVi* or *AcademH-6_PGr* as a query, the protein sequences from the four teleost fishes were hit in the second iteration. They are from the climbing perch *Anabas testudineus* (XP_026195931, XP_026196227, XP_026196228, XP_026196229), the California yellowtail *Seriola lalandi dorsalis* (XP_023286175, XP_023286176), the spiny chromis damselfish *Acanthochromis polyacanthus* (XP_022063315, XP_022063316, XP_022063317, XP_022063318), and the rohu *Labeo rohita* (RXN19178, RXN19557). Besides these species, the genomes from a species of thornfishes *Cottoperca gobio*, the Siamese fighting fish *Betta splendens*, the bicolor damselfish *Stegastes partitus*, and the spotted seabass *Lateolabrax maculatus* contain related sequences (Supplementary Table S4). These proteins do not have all residues conserved among *AcademH* transposases (Fig. 2b, XP_026196227; and data not shown). Further investigation revealed that apparently intact *AcademH* transposons related to these proteins are present in the genomes of two deuterostomes: *AcademHP-1_SP* from the purple sea urchin *Strongylocentrotus purpuratus* and *AcademHP-1_SKow* from the acorn worm *Saccoglossus kowalevskii* (Fig. 2). These families encode 2 plant homeodomain (PHD) fingers between the second D and the last E catalytic residues (Figs. 1 and 2). One PHD finger contains 1 histidine residue sandwiched by 4 and 3 cysteine residues (Cys₄-His-Cys₃). PHD fingers share an ability to bind to tri-methylated lysines on histones [17], and thus, it is expected that the PHD fingers in the

transposases of *AcademHP* families also bind to histones. Several copies of *AcademHP* families show 9-bp TSDs similarly to other *AcademH* families (Supplementary Fig. S3). One *AcademHP* sequence was also found as a single-copy sequence from the genome of *Priapulus caudatus*, although it encodes only one PHD finger (Fig. 2). Another protein encoded in the genome of *P. caudatus* (XP_014663285.1) contains 2 PHD fingers, although no TIRs flanked with recognizable TSDs were detected around the sequence encoding this protein. Thorough investigation revealed that other *AcademH* families from animals also contain a zinc finger motif between the second D and the last E catalytic residues, but they are CCHH-type (Fig. 2B2).

***AcademH* and *AcademX*, two distant lineages inside of *Academ* superfamily**

HHpred analysis with *Academ* transposases did not indicate any specific relationships with other transposases. The transposase domains of *Academ* are considered to belong to the DDE transposases, and thus to the RNaseH fold, based on Yuan and Wessler [12] which reported the conserved motifs and residues among *Academ* transposases. With more divergent transposases included in this analysis, fewer conserved residues are recognized (Fig. 2b). Only 7 residues, including the proposed DDE triad, are conserved among diverse *Academ* transposases. Compared with other DDE transposases, the first catalytic D and the second catalytic D are very distant (138–192 residues apart) in *Academ* transposases. The conserved G/A/E/QxxH motif following the second catalytic D residue might correspond to C/DxxH motif in *MuDR*, *P*, *hAT*, *Kolobok* and *Dada*, predicted to be located at the beginning of insertion domain [12].

The phylogenetic analysis revealed that *Academ* superfamily can be classified into two large groups, *AcademH*

(a) Superfamily II Helicase

Table of protein sequences for Superfamily II Helicase. Includes entries for Vasa (2DB3 B), SNRNP200 (6AH0_X), RecQ (10YW A), and various AcademH-1 to AcademH-11 Pst (Puccinia) and other species like Acropora, Crassostrea, Saccoglossus, and Rhizophagus. Motif labels Q, I, Ia, Ic, II are shown at the bottom.

(b) Academ transposase

Table of protein sequences for Academ transposase. Lists various Academ-1 to Academ-8 species including Acyrthosiphon, Branchiostoma, Ciona, Nematostella, Chondrus, Chondrus, Puccinia, Puccinia, Puccinia, Puccinia, Puccinia, TrPua (Trametes), Ascobolus, Panaeolus, Plicaturoopsis, Lobosporangium, Fibularhizoctonia, Laccaria, Gymnopilus, Mortierella, Tilletia, Schizopora, Amphimedon, Branchiostoma, Acropora, Crassostrea, Crassostrea, Anabas. Motif labels DN, H, I, Y, D, E, *, ZnF are shown at the bottom.

(B1) PHD fingers in insertion domain of AcademH transposase

Table showing PHD1 and PHD2 motifs for AcademH-1 SP (Strongylocentrotus), AcademHP-1 SKow (Saccoglossus), and AcademHP-1 PrCa (Priapulius). Includes amino acid sequences for each motif.

(B2) Zinc finger in insertion domain of AcademH transposase

Table showing Zinc finger motifs for Academ-3 Adi (Acropora), AcademH-1 Cvi (Crassostrea), and AcademH-8 Cvi (Crassostrea). Includes amino acid sequences for each motif.

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Motifs of protein domains conserved in the *Academ* superfamily. Conserved residues are shown in red. The lengths between motifs are shown in parentheses. **a** Superfamily II helicase. The motif names are shown below alignment. **b** *Academ* transposase. Conserved residues are shown below alignment. Asterisks indicate three catalytic residues. ZnF indicates the location where animal transposases encode PHD fingers or another zinc finger motif, which are shown in (B1) and (B2), respectively

and *AcademX*, corresponding to the protein coding ability (Fig. 5). *AcademX* can be further divided into two lineages, consistent with the difference in TSD length and distribution. *AcademX* with 3-bp TSDs are distributed among animals. *AcademX* with 4-bp TSDs has been found only from the red alga *C. crispus*. Two clusters for *AcademH* correspond to the *AcademH* from fungi and animals. The three *AcademHP* families with the *AcademHP* transposase-like protein encoded on the genome of *A. testudineus* (XP_026196227.1) clustered together inside of animal *AcademH*. *AcademH* transposons from closely related organisms are often clustered together, for examples, three families from Mucoromycote fungi (*AcademH-1_LoTr*, *AcademH-1_MoVe* and *AcademH-2_MoVe*) or five families from the oysters in the genus *Crassostrea* (*AcademH-8_CVi*, *AcademH-1_CVi*, *AcademH-4_CVi*, *Academ-2_CGi*, and *AcademH-2_CVi*). All *AcademH* families from the genus *Puccinia* are very closely related. However, deeper phylogeny of *AcademH* transposases is not consistent with their host phylogeny. Considering the small number of genomes from which *AcademH* families were characterized, and low bootstrap supports for deeper nodes, the contribution of horizontal transfer to the *AcademH* evolution remains to be investigated.

Discussion

The diversity and distribution of *Academ*

The *Academ* superfamily of DNA transposons has been found from three different groups of eukaryotes: animals, fungi and red algae. With a relatively small number of sequences, the phylogeny and structural characteristics of *Academ* are straightforward. The *AcademX* lineage encodes one large protein containing a transposase, an XPG nuclease and a putative zinc finger. It is distributed in animals and red algae. *AcademX* generates relatively short (3 or 4-bp) TSDs upon integration. The *AcademH* lineage encodes two proteins, one of which is a transposase and the other of which is a superfamily II helicase. *AcademH* generates relatively long (9 or 10-bp) TSDs upon integration. *AcademH* is distributed in animals and fungi. *AcademHP* is a sublineage inside of *AcademH* and this lineage encodes one or two PHD fingers between the second D and the last E catalytic residues. In vertebrates, the genomes of some teleost fishes keep remnants of *AcademHP* copies.

Functional implications for helicase in the life cycle of *AcademH*

The length of TSDs is one of the hallmarks of superfamilies of DNA transposons. In general, inside of the

| 5' junction | Transposon family | 3' junction |
|----------------------------------|-------------------------|----------------------------------|
| <i>Puccinia coronata</i> | | |
| GGACGCACATTTTCCGAACTTAACCAAG | <i>AcademH-1_PCor</i> | GCTTTGTTAAGCTCGACATTTTCCACGCTC |
| CAGGCCATTTGATCCCTAACTTAACGCAAG | <i>AcademH-2_PCor</i> | GCTTAGTAAAGCTAGATTTGATCCCTAGGGGA |
| CCTTGGGACGTTTGTCTAGCTTAACACAAG | <i>AcademH-3_PCor</i> | ATTGTGTTAAGCTAGGACGTTTGTGAGGAA |
| GACCATGTACAGTGGCTAGCTTAACCCAAG | <i>AcademH-4_PCor</i> | GCTTAGTAAAGCTAGGTACAGTGGCTGGAA |
| GAAAAGGCTTGAACGCTAGCTTAACATAAGC | <i>AcademH-5_PCor</i> | GCTTAGTAAAGCTAGGCTTGAACGCTGGTA |
| GGCTGAGGTCACTTTCTAACTTAACATAAGT | <i>AcademH-6_PCor</i> | ACTTAGTAAAGTAAAGTGGTCACTTTGCTCCT |
| GGGATAGGCTCAACGCTAACTTAACACAAG | <i>AcademH-7_PCor</i> | GCTTAGTAAAGCTAGGCTCAACGCTATATG |
| TTAAGGCACCTACGGCGAACTTAACACAAGT | <i>AcademH-8_PCor</i> | ACTTAGTAAAGTTGCACTTACGGTAGAATA |
| CATATGTGCGGATTCGAACCTTAACACAAG | <i>AcademH-9_PCor</i> | GCTTTGTTAAGCTCGTGGCGCATTTGCTGCT |
| GTGGAGCCCAAGGAGCTAGCTTAACACAAG | <i>AcademH-10_PCor</i> | ATTGTGTTAAGCTAGAGGAGAGATGAAC |
| GAATGGATTTCTTCAAACTTATCAAAGT | <i>AcademH-11_PCor</i> | GCTTTGTTAAGCTCGATTTCTTCAATC |
| ATCGTACGTTGAGCGAACTTAACACAAGT | <i>AcademH-12_PCor</i> | GCTTAGTAAAGCTTGGTGGAGCTACAGT |
| AAATGGGTGTAATGCTAACTTAACACAAG | <i>AcademH-13_PCor</i> | GGCTTTGTTAAGCTTGTGTAATGGAGGAC |
| TAGCCCTTTAAGCTGCTAACTTAACACAAG | <i>AcademH-14_PCor</i> | CFTATGTTAAGCTAGTTAAGCTAGTGGT |
| GAGCTGGGAATACTCAACTTAACACAAG | <i>AcademH-15_PCor</i> | GCTTTGTTAAGCTCGGGAATACTGGGTT |
| AAATGATGCTGCTCAACTTAACACAAG | <i>AcademH-16_PCor</i> | GCTTAGTAAAGTTGATGCTGCTAGTCAA |
| AGGCACATGGGGGGCTAACTTAACACAAGT | <i>AcademH-17_PCor</i> | ACTTTGTTAAGTTAGATGGGGGGCCATTTT |
| TTGTAAGTGGCATCTCTAACAATAAGTAAAGT | <i>AcademH-N1_PCor</i> | CTTGTGTTAAGTTAGTGGCATCTCAAAGT |
| TGGACGAGTGGCGAGCTAGCTTAACCCAAT | <i>AcademH-N2_PCor</i> | GCTTAGTAAAGCTAGAGTGGCGAGGTGGCA |
| ACGGGTAGGGGTGATGCTAACTTAACACAAG | <i>AcademH-N3_PCor</i> | GCTTTGTTAAGCTAGAGGGTATGCTCCTT |
| AAATCTAAGAAAGTCTAACTTAACACAAG | <i>AcademH-N4_PCor</i> | ACTTAGTAAAGTTAAGAAAGGTGTAATA |
| GTACAAATAGACCTTTCTAACTTAACATAAGT | <i>AcademH-N5_PCor</i> | ACTTAGTAAAGTTAGTAGACCTTTGGGGTC |
| TGGCTGGTTGCAAGCTAACTTAACACAAG | <i>AcademH-N6_PCor</i> | CFTATGTTAAGCTAGGTTGCAAGCGACCT |
| CACCCACATCAGTCAACGAACCTTAACACAAG | <i>AcademH-N7_PCor</i> | GCTTAGTAAAGCTCGCATCAGTCAAGGGTT |
| CATACCGAGTAATACCGAACTTAACGCAAG | <i>AcademH-N8_PCor</i> | GCTTTGTAGGCTCGCAGTAATACCAATCG |
| CACACCTGTGAGTCAAACTTAACGCAAG | <i>AcademH-N9_PCor</i> | GCTTTGTTAAGCTCGTCACTCATTTC |
| AACAGGATTCATACCCCTAACTTAACACAAGT | <i>AcademH-N10_PCor</i> | ACGTTGTTAAGTTAGATTCATACCCCTTGT |
| TACATGACTGGGGACTAACTTGACACAAG | <i>AcademH-N11_PCor</i> | CFTATGTTAAGCTAGGACTGGGGAGGGGTA |
| TAATGTGGTCTGCTAGCTTATCACAAG | <i>AcademH-N12_PCor</i> | GCTTAGTAAAGCTTGGTCTGCTTATGTA |
| ATACTATACACACCCGAACTTAACGCAAG | <i>AcademH-N13_PCor</i> | GCTTTGTTAAGCTTGTACACACCCACCAC |

Fig. 3 Termini and TSDs of *Academ* superfamily of DNA transposons from the fungus *Puccinia coronata*. Only one representative insertion is shown for each family. TSDs are colored in red. All termini and TSDs characterized in the study are shown in Supplementary Fig. S1

| 5' junction | Transposon family | 3' junction |
|----------------------------------|------------------------|---------------------------------|
| <i>Crassostrea virginica</i> | | |
| GTTGCTGTAACGACCCCAATCTAATTAATAA | <i>AcademH-1_CVi</i> | ATTTTAATAGATTGTGTAACGACCCAAAGG |
| AATAATCCATTACATACCAATCTGATTAATAA | <i>AcademH-2_CVi</i> | ATTTTAATCAGATTGCCATTACATATACAT |
| AAACACATTGCCATGCACTCTGATTAATAAC | <i>AcademH-3_CVi</i> | GTTTTAATCAGACTGATGCCATGAAAATA |
| ATATTGTTGATTGCAATCTGATTAATAA | <i>AcademH-4_CVi</i> | ATTTTAATCAGATTGGTGGATTGAATTCG |
| GTGTCACTGCCGATGCAATCTAATTAATAA | <i>AcademH-5_CVi</i> | ATTTTAATAGATTGGTCCGATGTTTAT |
| TTGTCACTCCCAAGCAATCTCGGAACCCA | <i>AcademH-6_CVi</i> | TGGGTCCGACGATGCTACCCAAACCTACAT |
| CAGAGACTTTTCCCTCAATCTAATTAATAA | <i>AcademH-7_CVi</i> | ATTTTAATAGATTGCTTTTCTCTGAATCC |
| GAGTTTACCTAAATCAATCTCGGAACCCA | <i>AcademH-8_CVi</i> | TGGGTCCGACGATGACTAAATCTGTCTCA |
| TTGGGTGTTTCAACAGCTCTGATTAATAA | <i>AcademH-9_CVi</i> | ATTTTAATAGACTGGTTTCAACGCACTC |
| AGTGCACAAATGTGCAATCTGACTAAAGT | <i>AcademH-N1_CVi</i> | ACTTTACTCAGATTGAAAATATGTCTGAA |
| ATTAGTAAATACAGTCAATCTAATTAATAA | <i>AcademH-N2_CVi</i> | ATTTTAATTAATTTGAAAATATAGTATATC |
| <i>Acropora digitifera</i> | | |
| CGTTCAATCACAGCTTAGCCTGCGAAAGCA | <i>Academ-1_ADi</i> | TGCTTTCGACAGGCTAGTGTATGTTGCCG |
| TCAGCCAAATTCGGTCCAACTCGTTCCCGAG | <i>Academ-2_ADi</i> | CTGGGAACGAGGTTGAATTCGGTCAACATG |
| CAGGGTCTCCGAGGATAGGCTCCTCCGAG | <i>Academ-3_ADi</i> | CTGGGAAGGAGACTACTCCGAGGATGGCTG |
| AAGTCTTACTGGGATCAATCTCGTTCCCGAG | <i>Academ-4_ADi</i> | CTGGGAACGAGATTGACTGGGATTAAGTCT |
| GCTGATACCAAGGGTCAATCTCGTTCCCGAG | <i>AcademH-5_ADi</i> | CTGGGAACGAGATTGACCAAGGGTCAATTTG |
| AGTGCCATGTTGAGATAGCCTGTGTCCAG | <i>AcademH-6_ADi</i> | CTGTGACACAGGCTAGATTTTATTGAAAAC |
| GCGCAGCCATAAGTCAATCTCGTTCCCGAG | <i>AcademH-N1_ADi</i> | CTGGGAACGAGATTGCCATAACATCCGAGG |
| TCTCATAGTATGTCACAACTCGTTCCCGAG | <i>AcademH-N2_ADi</i> | CTGGGAACGAGGTTGAGTATGCTATGATAT |
| TAGCGAGCTAAACGGTACTCTCCTCCGAG | <i>AcademH-N3_ADi</i> | CTGGGAAGGAGACTAGCTAAACGCGGCGCA |
| TGACGAGAGAAGTCTCAACTCGTTCCCGAG | <i>AcademH-N4_ADi</i> | CTGGGAACGAGGTTGGAGAAGCTTTGGTT |
| TCAATGTTGTTCTATCAACTCGATTCCGAG | <i>AcademH-N6_ADi</i> | CTGGGAACGAGGTTGTTGTTCTATGAGAAA |
| TTACCTCATCGAGCCAACTCGTTTCAAG | <i>AcademH-N7_ADi</i> | CCGGGAACGAGGTTGCTATCGAGACGAAAC |
| TAGACTAACTGGCGTTAGCCTGCGAGCAGG | <i>AcademH-N8_ADi</i> | CCTGCTCGCAGGCTAAACTGGCGTGGGTT |
| AAGAGGCTATATAGTCAATCTCGTTCCCGAG | <i>AcademH-N9_ADi</i> | TTGGGAACGAGATTGCTATATAGTGGCCTT |
| GGATAGTCCCGGTTCAACTCGTTCCCGAG | <i>AcademH-N10_ADi</i> | CTGGGAACGAGGTTGGTCCCGGTTAAGGTT |
| GAGTTCACTCCGGTACCTGAGTCCCGAG | <i>AcademH-N11_ADi</i> | CTGGCACCCAGGTTAAACTCCCGGCGGACC |
| GCAATGCGGTTAGTATAGCCTGCGAGCAAG | <i>AcademH-N12_ADi</i> | CTTGCTCGCAGGCTACCGTTTATGTTCCACA |
| TATCCCATATTTGTTATAGACTCGGTGCACT | <i>AcademH-N14_ADi</i> | ACTGCTCGCAGTCTACATTTGTTAAGTGCC |
| GAAAGCTCGTTATAGTCTCGGAGCAGAA | <i>Academ-4_ADi</i> | CTTGCTCCTGACTATACGTTCCATTTCTC |
| AAGTATAGTCTAATAGTCTGGCAGCATG | <i>Academ-5_ADi</i> | AATGCTCCAGACTAATGAAAGGCGGCTCA |
| ATTCCGTAAGGGTTATAGTCCAGTCATTTT | <i>Academ-6_ADi</i> | AAAAATGACTGGACTATTAACAGAGGCGAGT |
| GAAAACGGCAATTAATAGTCCAGGCATTTT | <i>Academ-7_ADi</i> | AAAATGCCTGGACTAATAAACAGACAAAT |
| TAATATTCTTATAGTATAGTCCAGGGCCAG | <i>Academ-8_ADi</i> | CTGGCCCCCTGAATAGTAGCGCATGTGAAC |
| ACACGCCACCATTATAGTCCATGGGCCAG | <i>Academ-9_ADi</i> | CTGGCCCCATGGACTATTAACAGGCTTTCTT |
| TTCTTTCCCGTGTATAGACTGGGACAAAT | <i>Academ-N1_ADi</i> | AAATGCTCCAGTCTATTTTGAACATTAATA |
| ATAAATAAATCAATAGTCCAGGCGTTAT | <i>Academ-N2_ADi</i> | ATAACCGCTAACTAATAAGGAAATGGGCC |
| ACAAATTGACAAGGATAGTGTGGTCGCTTA | <i>Academ-N4_ADi</i> | TAAGCGCCAGACTAGGATGCTTATGTGAA |

Fig. 4 Termini and TSDs of *Academ* superfamily of DNA transposons from two animal species, *Crassostrea virginica* and *Acropora digitifera*. Only one representative insertion is shown for each family. TSDs are colored in red. No clear TSDs were observed for *Academ-1_ADi* and *AcademH-6_ADi*, while a short derivative of *Academ-1_ADi* is flanked by 9-bp TSDs (data not shown). All termini and TSDs characterized in the study are shown in Supplementary Fig. S2

superfamily of DNA transposons, the lengths of TSDs are not so divergent [4, 18]. Almost all of superfamilies show strict restriction of TSD lengths, which allows only 1-bp difference. As rare exceptions, the *hAT* superfamily shows TSDs of 5, 6 or 8-bp, and the *EnSpm* superfamily shows TSDs of 2-bp, 3-bp, or 4-bp. In contrast, inside of the *Academ* superfamily, *AcademX* generates 3-bp or 4-bp TSDs, while *AcademH* generates TSDs of 9 or 10 bps in length.

AcademH families encode a superfamily II helicase related to RecQ, while *AcademX* families encode an XPG nuclease. Mutually exclusive presence of helicase or nuclease in *Academ* transposons implies the functional similarity of these two enzymes in the life cycle of *Academ* transposons. RecQ helicase family works for various DNA repair pathways including homologous recombination and non-homologous end joining [19]. XPG nuclease families are needed to repair DNA damages by a process called nucleotide excision repair [20]. It can be speculated that helicase and nuclease encoded by *Academ* transposons are coupled with cellular proteins in DNA repair pathway during the transposition of *Academ* transposons. DDE transposases cleave DNA at both termini of DNA transposons [21]. The difference in how DNA repair pathway is recruited to resolve

the transposition intermediate might dictate the junction structures different between *AcademX* and *AcademH*.

Conclusions

The *Academ* superfamily of DNA transposons has 2 deep-branching lineages: *AcademX* and *AcademH*. Besides its transposase, *AcademH* encodes a superfamily II helicase, which may contribute to the generation of long TSDs.

Methods

Characterization of *Academ* DNA transposons

All genome sequences used in this study were downloaded from either of three websites: NCBI Assembly database (<https://www.ncbi.nlm.nih.gov/assembly>), UCSC Genome Browser (<https://genome.ucsc.edu/>), and OIST Marine Genomics Unit (http://marinegenomics.oist.jp/lingula/viewer/download?project_id=47), and listed in Supplementary Table S5.

Cursor searches [13] using reported *Academ* sequences as queries against genomes were performed. Sequences showing similarity to *Academ* were clustered by BLASTCLUST in the NCBI Blast package. Cursor searches were done with consensus sequence of each cluster and the hits with flanking sequences were

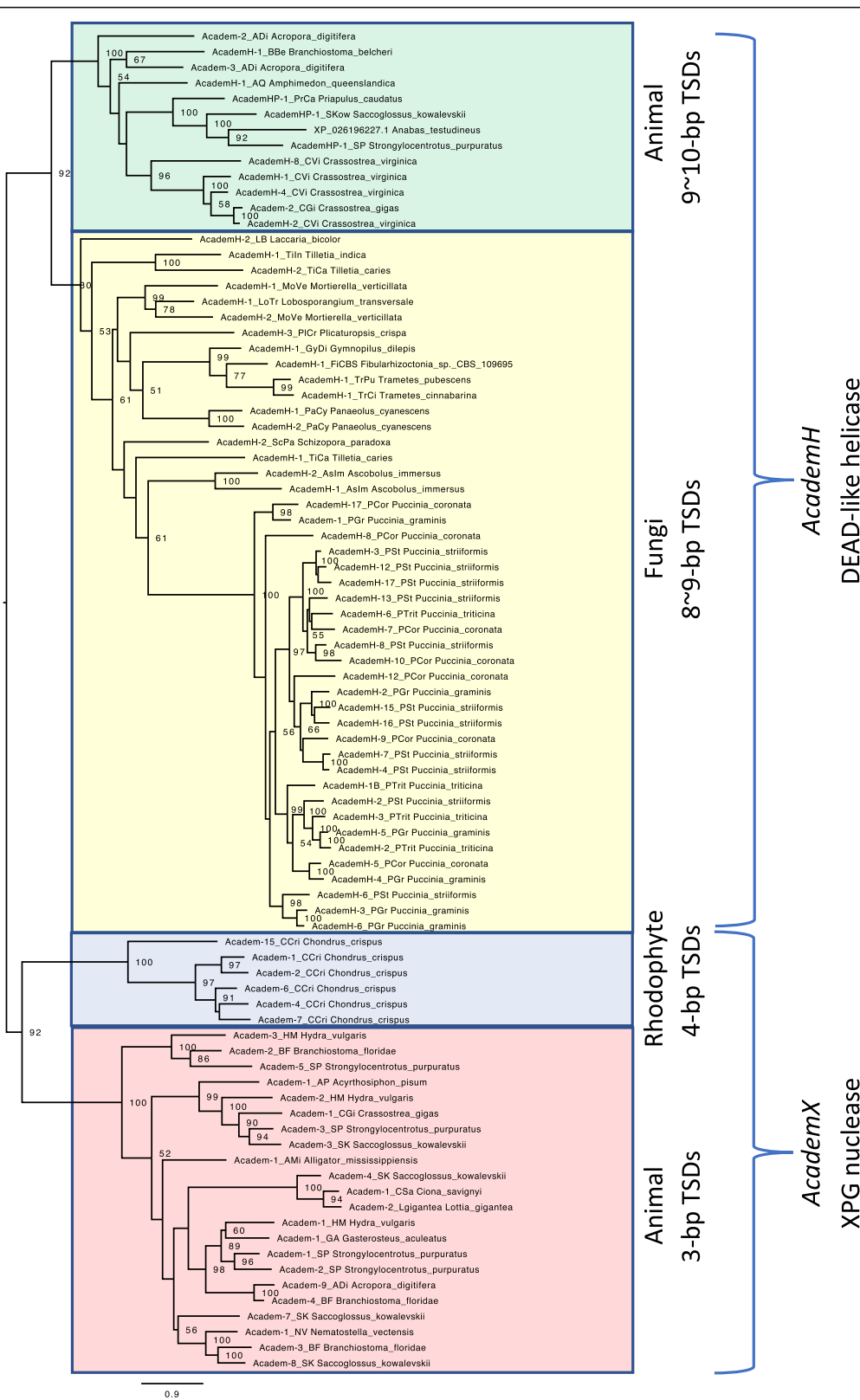


Fig. 5 Phylogeny of *Academ* transposases. Four lineages recognized are highlighted in different colors. Bootstrap values of 100 replicates are shown at nodes when they are > 50. Host organisms of TEs are shown after TE names

extracted to characterize the complete repeat unit until TSDs were detected.

In parallel, RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) and Repbase [1] were used for the initial screening of repetitive families with default parameters for all animal genomes used here except for *C. gigas*, *C. virginica*, *M. yessoensis*, *C. teleta*, *P. caudatus*, and *B. floridae*. Consensus sequences generated by RepeatModeler output with the annotation as *Academ* were chosen to reconstruct the second consensus sequences using the top 10 hits with the 1000-bp flanking sequences at both sides in the Censor search.

Single-copy sequences similar to *AcademH* families were annotated as *AcademH* transposons if > 10-bp TIRs and adjacent > 8-bp TSDs were detected within their 10,000-bp flanking sequences.

The consensus or single-copy representative sequences for all TE families reported here have been submitted to Repbase [1], and are also available in Supplementary Datasets S1 and S2.

Protein structure and phylogenetic analyses

Protein coding regions were predicted from consensus sequences and representative single-copy sequences with Softberry FGENESH (<http://www.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind>) [22], followed by manual curation with reference to predicted mRNA sequences available at NCBI website. NCBI CD-Search (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) [23] was done to detect protein domains. HHpred (<https://toolkit.tuebingen.mpg.de/tools/hhpred>) [24] was used to find similar structures of respective proteins.

Multiple sequence alignment was done with MAFFT with linsi option [25]. *Academ* transposase domains were extracted following the definition in [12, 16]. Protein sequences with truncation or internal deletion inside of transposase domain were excluded from the analysis. The final dataset used for the phylogenetic analysis contains 86 sequences which are 319 to 541 residues in length (Supplementary Dataset S3). Maximum likelihood trees with bootstrap values of 100 replicates were constructed using PhyML [26] with the amino acid substitution model LG + G + I + F, which was chosen based on the best Akaike Information Criterion score. The phylogenetic trees were drawn with the aid of FigTree 1.3.1 (<http://tree.bio.ed.ac.uk/software/figtree/>).

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13100-020-00211-1>.

Additional file 1: Figure S1. Termini and TSDs of newly characterized families of *Academ* from the fungus *Puccinia coronata*. **Figure S2.**

Termini and TSDs of newly characterized families of *Academ* from two animal species, *Crassostrea virginica* and *Acropora digitifera*. **Figure S3.** Termini and TSDs of *AcademHP* families from animals. **Table S3.** Non-autonomous DNA transposons newly classified as *Academ*. **Table S4.** *AcademHP* remnants found in teleost.

Additional file 2 : Table S1. Protein sequences showing similarity to *AcademH* transposases. **Table S2.** Characteristics of *AcademH* families. **Table S5.** Genome assembly sequences used in this study.

Additional file 3 : Data S1. Consensus sequences of multicopy *Academ* transposons characterized in this study. **Data S2.** Representative sequences of single-copy *Academ* transposons characterized in this study.

Data S3. Protein multiple alignment of *Academ* transposase domains used for the phylogenetic analysis.

Abbreviations

TSD: Target site duplication; TIR: Terminal inverted repeat; PHD: Plant homeodomain; TE: Transposable element; LTR: Long terminal repeat; RAG1: Recombination activating gene 1

Acknowledgements

Not applicable.

Author's contributions

KKK performed experiments and analysis, and wrote the manuscript. The author read and approved the final manuscript.

Funding

The author received no specific funding for this work.

Availability of data and materials

All data generated or analyzed in this study are included in this published article and its supplementary information files. Consensus and single-copy representative sequences of TEs are also submitted to Repbase (<http://www.girinst.org/repbase/>).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declare that he has no competing interests.

Received: 29 January 2020 Accepted: 6 April 2020

Published online: 18 April 2020

References

- Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11.
- Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's Guide to Bacterial Insertion Sequences. *Microbiol Spectr*. 2015;3(2):MDNA3-0030-2014.
- Finnegan DJ. Eukaryotic transposable elements and genome evolution. *Trends Genet*. 1989;5(4):103-7.
- Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. *Genes Genet Syst*. 2019;94:233-52.
- Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA*. 2017;8:19.
- Hickman AB, Chandler M, Dyda F. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Crit Rev Biochem Mol Biol*. 2010; 45(1):50-69.
- Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol*. 2005;3(6):e181.
- Kapitonov VV, Jurka J. *Academ* - a novel superfamily of eukaryotic DNA transposons. *Repbases Rep*. 2010;10(4):643-7.
- Bao W, Jurka J. DNA transposons from the red seaweed. *Repbases Rep*. 2013; 13(10):2271-85.

10. Kojima KK, Jurka J. DNA transposons from the *Puccinia graminis* genome. *Rebase Rep.* 2015;15(8):2495–508.
11. Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. Cut-and-paste transposons in Fungi with diverse lifestyles. *Genome Biol Evol.* 2017;9(12):3463–77.
12. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci U S A.* 2011;108(19):7884–9.
13. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Rebase: RebaseSubmitter and censor. *BMC Bioinformatics.* 2006;7:474.
14. Kapitonov VV, Jurka J. Kolobok transposons in the Glomeromycota fungus. *Rebase Rep.* 2014;14(7):1925–9.
15. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A.* 2001;98(15):8714–9.
16. Zhang HH, Shen YH, Xiong XM, Han MJ, Qi DW, Zhang XG. Evidence for horizontal transfer of a recently active *Academ* transposon. *Insect Mol Biol.* 2016;25(3):338–46.
17. Musselman CA, Kutateladze TG. Handpicking epigenetic marks with PHD fingers. *Nucleic Acids Res.* 2011;39(21):9061–71.
18. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Rebase. *Nat Rev Genet.* 2008;9(5):411–2 author reply 4.
19. Croteau DL, Popuri V, Opresko PL, Bohr VA. Human RecQ helicases in DNA repair, recombination, and replication. *Annu Rev Biochem.* 2014;83:519–52.
20. Schärer OD. XPG: its products and biological roles. *Adv Exp Med Biol.* 2008; 637:83–92.
21. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol.* 2003;4(11):865–77.
22. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 2006;7(Suppl 1):S10 1–2.
23. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 2015; 43(Database issue):D222–6.
24. Zimmermann L, Stephens A, Nam SZ, Rau D, Kubler J, Lozajic M, et al. A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J Mol Biol.* 2018;430(15):2237–43.
25. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 2005;33(2):511–8.
26. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 2010;59(3):307–21.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

