Science Signaling
AAAS

stke.sciencemag.org/cgi/content/full/12/578/eaav3810/DC1

Supplementary Materials for

Tracing the origin and evolution of pseudokinases across the tree of life

Annie Kwon, Steven Scott, Rahil Taujale, Wayland Yeung, Krys J. Kochut, Patrick A. Eyers, Natarajan Kannan*

*Corresponding author. Email: nkannan@uga.edu

Published 23 April 2019, *Sci. Signal.* **12**, eaav3810 (2019)
DOI: 10.1126/scisignal.aav3810

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at stke.sciencemag.org/cgi/content/full/12/578/eaav3810/DC1)

Table S1 (Microsoft Excel format). Kinase and pseudokinase sequence counts detected in 10,092 archaeal, bacterial, and eukaryotic proteomes.
Table S2 (Microsoft Excel format). Catalog and annotation of pseudokinase families.
Table S3 (Microsoft Excel format). Counts of canonical sequences classified into pseudokinase families.
Table S4 (Microsoft Excel format). Distribution of plant IRAK pseudokinase families across diverse plant species.
Table S5 (Microsoft Excel format). Known plant IRAK pseudokinases and their classifications.
Data file S1 (FASTA file format). Alignments of model organism kinomes and pseudokinomes.
Data file S2 (FASTA file format). Sequences and alignments of pseudokinase families.
Data file S3 (FASTA file format). Alignments of canonical sequences classified into pseudokinase families.

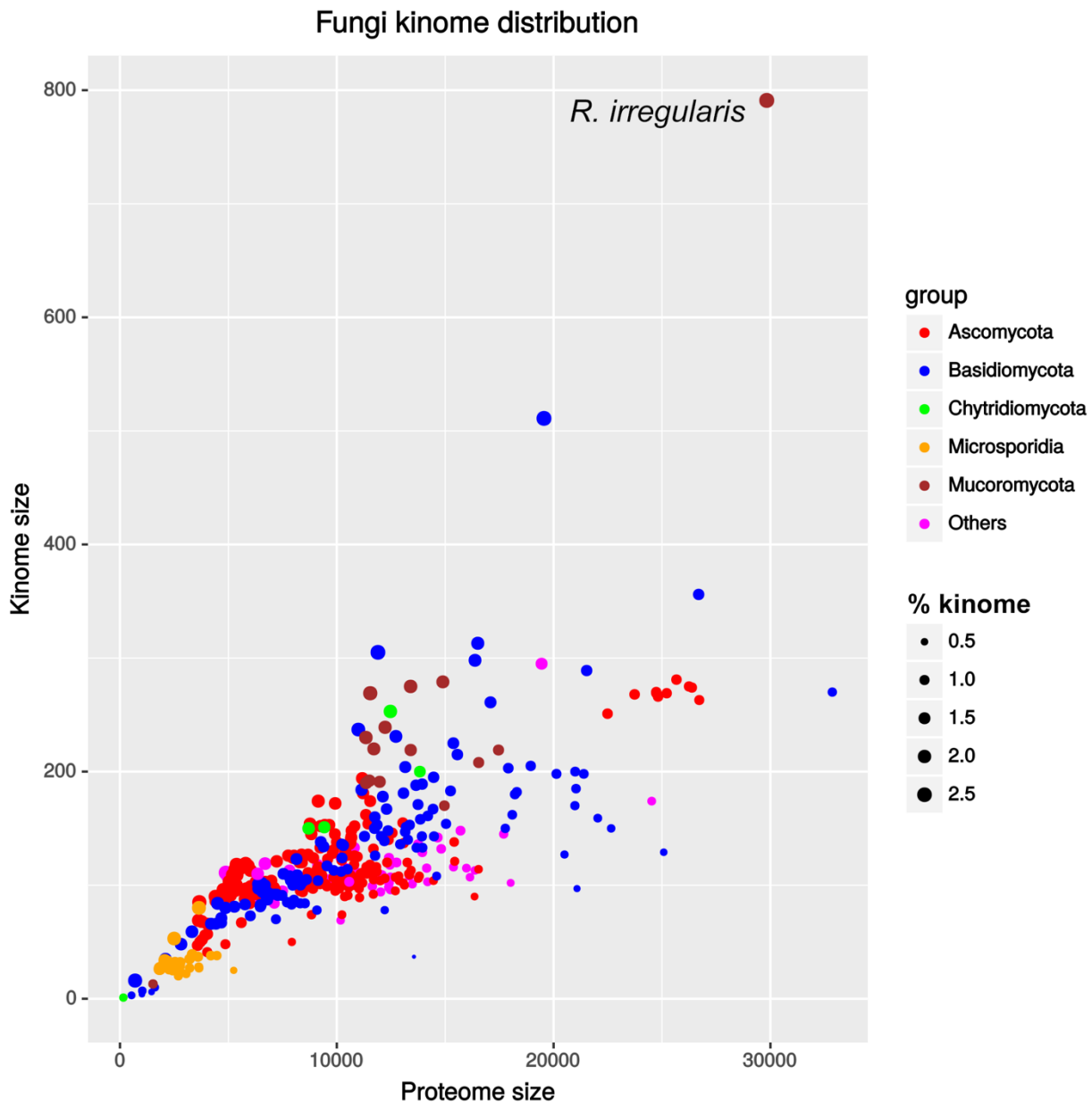

Fig. S1. Expansions of fungal proteomes and kinomes. The proteome and kinome sizes for 449 fungal species/strains are shown, with each species colored by the fungal phylum to which it belongs. Dot sizes represent the proportion of the proteome that is comprised by kinases. The point for *R. irregularis* is labeled.

Table S1. Kinase and pseudokinase sequence counts detected in 10,092 archaeal, bacterial, and eukaryotic proteomes. The counts of protein kinases and pseudokinases detectable in all archaeal, bacterial, and eukaryotic reference proteomes available in the UniProt database (release 2018_9) (*59*) are provided. Percentages indicate the fraction of total kinases from each proteome that were predicted to be pseudokinases based on the lack of at least one residue in the catalytic triad. Protein kinase sequences were extracted and aligned from each reference proteome using diverse ePK sequence profiles. Table is provided in Microsoft Excel format.

Table S2. Catalog and annotation of pseudokinase families. Detailed sequence and taxonomic information is provided for each pseudokinase family. The classifications of each pseudokinase family into the traditional kinase groups and families are noted. The size of each pseudokinase family in the original omcBPPS cluster is provided, as well as the size of the final pseudokinase family determined from the UniProt pseudokinase sequence set (see methods). The major taxonomic groups represented in each pseudokinase family is provided. Taxonomic groups are reported if greater than 95% of sequences in the pseudokinase family belongs to a taxonomic group. The numbers of species/strains from each major taxonomic group are also provided. Percentages of sequences that contain the canonical catalytic triad motifs at each of the three positions were calculated and are noted. Alignments of pseudokinase sequences at and near the catalytic triad motifs were manually examined, and the confidence of the alignment at each position was annotated for each family. The alignment confidence was annotated for each of the three catalytic positions for each pseudokinase family ('H' for 'high confidence' and 'L' for low confidence). Pseudokinase families were also annotated based on alignment confidence and their conservation across taxonomic groups ('high confidence' families are marked in green and 'low confidence' families are marked in yellow). Table is provided in Microsoft Excel format.

Table S3. Counts of canonical sequences classified into pseudokinase families. The numbers of canonical protein kinase sequences that were classified into each pseudokinase family are provided. The percentages of canonical sequences that comprise each pseudokinase family is also provided. Pseudokinase families are highlighted white if they contain no canonical members, yellow if fewer than 10% of sequences are canonical, and orange if greater than 10% of sequence are canonical. Detailed notes regarding canonical sequences are provided for pseudokinase families where canonical sequences were found. Table is provided in Microsoft Excel format.

Table S4. Distribution of plant IRAK pseudokinase families across diverse plant species. The counts of pseudokinases in each plant IRAK pseudokinase family across diverse representative plant species are shown. The total number of kinases and pseudokinases detected in each plant reference proteome from UniProt are also provided. Plant species are organized into major plant phyla. Table is provided in Microsoft Excel format.

Table S5. Known plant IRAK pseudokinases and their classifications. Previously identified plant IRAK pseudokinases are noted. Their placements in the pseudokinase classification, as

well as their placements in the IRAK subclassification (*52*) are provided. Table is provided in Microsoft Excel format.

Data file S1. Alignments of model organism kinomes and pseudokinomes. FASTA alignments of pseudokinase sequences are provided for representative archaeal, bacterial, and eukaryotic reference proteomes shown in Figs. 1 and 2. The first sequence in each alignment represents an ePK consensus sequence. Data are provided in FASTA file format.

Data file S2. Sequences and alignments of pseudokinase families. Full-length pseudokinase sequences and FASTA alignments of pseudokinase domains are provided for each pseudokinase family. The first sequence in each alignment represents an ePK consensus sequence. Data are provided in SEQ and FASTA file format.

Data file S3. Alignments of canonical sequences classified into pseudokinase families. FASTA alignments of canonical sequences that cluster into pseudokinase families are provided when applicable. The first sequence in each alignment represents an ePK consensus sequence. Data are provided in FASTA file format.