Title: Phylogenomics-Guided Validation of Function for Conserved Unknown Genes

Project ID 0013832
Prog Mgr Marvin Stodolsky Phone: 301-903-4475 Division: SC-23.2
PI: Valérie de Crécy-Lagard; co-PI: Andrew Hanson
Award Register: ER64498

Summary of Results

Identifying functions for all gene products in all sequenced organisms is a central challenge of the post-genomic era. However, at least 30-50% of the proteins encoded by any given genome are of unknown function, or wrongly or vaguely annotated. Many of these 'unknown' proteins are common to prokaryotes and plants. We accordingly set out to predict and experimentally test the functions of such proteins.

Our approach to functional prediction is integrative, coupling the extensive post-genomic resources available for plants with comparative genomics based on hundreds of microbial genomes, and functional genomic datasets from model microorganisms. The early phase is computer-assisted; later phases incorporate intellectual input from expert plant and microbial biochemists. The approach thus bridges the gap between automated homology-based annotations and the classical gene discovery efforts of experimentalists, and is much more powerful than purely computational approaches to identifying gene-function associations.

Among Arabidopsis genes, we focused on those (2,325 in total) that (i) are unique or belong to families with no more than three members, (ii) are conserved between plants and prokaryotes, and (iii) have unknown or poorly known functions. Computer-assisted selection of promising targets for deeper analysis was based on homology-independent characteristics associated in the SEED database with the prokaryotic members of each family, specifically gene clustering and phyletic spread, as well as availability of functional genomics data, and publications that could link candidate families to general metabolic areas, or to specific functions. In-depth comparative genomic analysis was then performed for about 500 top candidate families, which connected ~55 of them to general areas of metabolism and led to specific functional predictions for a subset of ~25 more. Twenty predicted functions were experimentally tested in at least one prokaryotic organism via reverse genetics, metabolic profiling, functional complementation, and recombinant protein biochemistry. Our approach predicted and validated functions for 10 formerly uncharacterized protein families common to plants and prokaryotes; none of these functions had previously been correctly predicted by computational methods (In orange in Table 1 below). The functions of five more are currently being validated (in yellow in Table 1 below). Experimental testing of diverse representatives of these families combined with in silico analysis allowed accurate projection of the annotations to hundreds more sequenced genomes.

This work has already led to 8 published papers and reviews, plus 5 more in preparation, and to 18 communications at state, national, and international conferences. Both the general and experimentally validated predictions will be implemented in the publically available SEED database (*the**seed.uchicago**.edu/**FIG**/index.cgi*) when the BMC genomics paper that is in preparation is published.

Table 1. Protein families that have been experimentally validated or are in the validation process

Case	Hypothesis	TAIR ID	COG, gene name	Subsystem in SEED	Experimental verification status	PubMed ID
1	Pterin carbinolamine dehydratase with role in Moco metabolism	At1g29810 At5g51110	COG2154, phhB	Pterin_carbinolamine_dehydratase	Validated in 7 eukaryotes and 8 prokaryotes	18245455
2	t6A biosynthesis	At5g60590	COG0009, YrdC	YrdC-YciO	Validated in yeast, archaea and two bacteria; A. thaliana in progress	19287007
3	PTPS family protein replacing the FolB step in folate synthesis	-	COG0720	Experimental-PTPS	Validated in in 1 eukaryote and 8 prokaryotes	19395485, 18805734
4	Metal chaperone-Zinc homeostasis	At1g15730 , At1g26520 , At1g80480	COG0523	COG0523	Validated in several bacteria	19822009
5	Folate-dependent Fe/S cluster synthesis or repair protein	At4g12130 At1g60990	COG0354, ygfZ	YgfZ-Fe-S	Validated in E. coli, Bartonella hensellae, Haloferax volcanii, Arabidopsis, Leishmania, yeast, mouse	20489182
6	Alternative route for 5-formyltetrahydrofolate disposal	At2g20830	COG3643	Experimental_Histidine_Degradation	Verified in 5 prokaryotes	Manuscript in prep
7	5-Formyltetrahydrofolate cycloligase paralog	At1g76730	COG0212	5-FCL-like_protein	Predicted role in thiamine recycling (testing in progress)	Manuscript in prep
8	t6A biosynthesis	At2g45270 , At4g22720	COG0533, YgjD	YrdC-YciO	Validated in yeast	Manuscript in prep
9	Niacin and/or choline transporter	NiaP homolog At1g13050	MFS superfamily	Choline transport and metabolism	In progress in Bacillus subtilis, Ralstonia solanacearum, Burkholderia xenovorans, plants, mouse	Manuscript in prep
10	Phytol phosphate kinase	At1g78620	COG1836, alr1612	COG1836	In progress in Synechocystis , Arabidopsis	Manuscript in prep
11	NAD-dependent ribosomal protein modification	At3g12930 , At1g67620	COG0799, alr4169	lojap	In progress E.coli	
12	Glycosylhydrolase involved in plant cell wall breakdown	At5g12950 , At5g12960	COG3533, SAV1144	COG3533	In progress in X. campestris	
13	m6A in small rRNA	At4g28830	COG2263	rRNA_modification_Archaea	in Progress in *H. volcanii*	
14	Ribosome assembly	At1g09150	COG2016	rRNA_modification_Archaea	In progress in yeast and H. volcanii	
15	Pyridoxal phosphate enzyme in amino acid metabolism.,	At4g26860 , At1g11930	COG0325, yggS	PROSC	In progress in E. coli	

Papers and other products delivered

Papers (5)

1. Waller J.C., Alvarez S., Naponelli V., Lara-Nuñez A., Blaby I.K., Da Silva V., Ziemak M.J., Vickers T.J., Beverley S.M., Edison A.S., Rocca, J.R., Gregory J.F. 3rd, de Crécy-Lagard V. and Hanson, A.D. (2010). A role for tetrahydrofolates in the metabolism of iron-sulfur clusters in all domains of life. *Proc Natl Acad Sci U S A* (in press). (PMID: 20489182).
2. Haas C. A., Rodionov D. A., Kropat J., Malasarn D., Merchant S. S. and de Crécy-Lagard V. (2009) A subset of the diverse COG0523 family of putative metal chaperones is linked to zinc homeostasis in all kingdoms of life. *BMC Genomics* **10**, 470 (PMID: 19822009).
3. El Yacoubi,B., Lyons B., Cruz Y., Reddy R., Nordin B., Agnelli F., Williamson,J.R., Schimmel P., Swairjo M.A. and de Crécy-Lagard, V. (2009) The universal YrdC/Sua5 family is required for the formation of treonylcarbamoyladenosine in tRNA. *Nucleic Acids Res* **37**, 2894-2909 (PMID: 19287007).
4. Pribat A., Jeanguenin L., Lara-Núñez A., Ziemak M.J., Hyde J.E., de Crécy-Lagard,V. and Hanson A.D. (2009) 6-Pyruvoyltetrahydropterin synthase paralogs replace the folate synthesis enzyme dihydroneopterin aldolase in diverse bacteria. *J Bacteriol* **191**, 4158-4165 (PMID: 19395485).
5. Naponelli V., Noiriel A., Ziemakv M. J., Beverley S. M., Lye L. F., Plume A. M., Botella J. R., Loizeau K., Ravane S., Rébeillé F., de Crécy-Lagard V. and Hanson A.D. (2008) Phylogenomic and functional analysis of pterin-4a-carbinolamine dehydratase family (COG2154) proteins in plants and microorganisms. *Plant Physiol* 1**46**,1515-1527(PMID: 18245455).

Reviews (3)

6. Hanson A.D., Pribat A., Waller J.C. and de Crécy-Lagard V. (2009) 'Unknown' proteins and 'orphan' enzymes: The missing half of the engineering parts list – and how to find it. *Biochem J* **425**, 1-11 (PMID: 20001958).
7. Hyde J.H., Dittrich S., Wand P., Sims P. F. G., de Crécy-Lagard V. and Hanson A.D. (2008) *Plasmodium falciparum*: a paradigm for alternative folate biosynthesis in diverse organisms. *Trends Parasitol.* **24**, 502-508 (PMID: 18805734).
8. de Crécy-Lagard V. and Hanson A.D. (2007) Finding novel metabolic genes through plant-prokaryote phylogenomics. *Trends Microbiol* **15**, 563-570.

Manuscripts in preparation (5)

9. Jeanguenin L., Pribat A., Hamner Mageroy S., Lara-Núñez A., Gregory J.F. , Blaby I.K., de Crécy-Lagard V. and Hanson A.D. (2010). 5-Formyltetrahydrofolate metabolism: An alternative to the cycloligase reaction in bacteria and archaea. *J. Biol Chem* (in preparation)
10. Pribat A., Blaby I.K., Lara-Núñez A., Jeanguenin L., Northen T, Bowen B., Giuliani S., Collart F., Begley T., Gregory J.F. 3rd, de Crécy-Lagard V. and Hanson A.D. (2010) Comparative genomic and functional analysis of a 5-formyltetrahydrofolate cycloligase paralog family from archaea, bacteria, and eukaryote. *Funct Integ Geno* (in preparation).
11. Lara-Núñez A., Ziemak M., Rodionov D.A., Gregory J.F. 3rd and Hanson A.D. (2010) Comparative genomic and functional analysis of the NiaP family of membrane proteins. *J Bacteriol* (in preparation).
12. El-Yacoubi B., Deutsch C., Bailly M. Thiaville P., Murzin A., Dirk Iwata-Reuyl D. and de Crécy-Lagard V. (2010). The universal Kae1/YgjD (COG0533) family is required for the formation of threonylcarbamoyladenosine in tRNA *Proc Natl Acad Sci U S A* (in preparation).
13. Gerdes S., El Yacoubi B., Bailly M., Blaby I.K., Haas C.2, Jeanguenin L., Lara-Núñez A., Waller J.C., Overbeek R., Hanson A.D. and de Crécy-Lagard V. (2010). Synergistic use of plant-prokaryote comparative genomics for functional annotations *BMC genomics* (in preparation).

Conference presentations (18)

Posters

1. Waller J.C., Shen G., Alvarez S., Loizeau K., Blaby I.K., Edison A.S., Rocca J.R., Golbeck J.H., Ravanel S., de Crécy-Lagard V. and Hanson A. D. Plants have two COG0354 proteins with non-redundant, folate-dependent functions in iron sulfur cluster protein metabolism. Plant Biology 2010 Joint Annual Meeting of the American Society of Plant Biologists & the Canadian Society of Plant Physiologists, Montreal, QC Canada, August 2010.
2. de Crécy-Lagard V., El Yacoubi B. , Bailly M., Pribat A., Lara-Nunez A. and Hanson A. D.. Phylogenomics guided validation of function for conserved unknown genes. Genomics Science Awardee Workshop VIII, Arlington, VA, February 2010.
3. Waller J.C., Alvarez S., Naponelli V., Lara-Nuñez A., Loizeau K., Blaby I.K., Da Silva V., Ziemak M.J., Vickers T.J., Beverley S.M., Edison A.S., Rocca J.R., Gregory JF 3rd, de Crécy-Lagard V., Ravanel S. and Hanson A.D. Unprecedented role for a folate-dependent protein in iron-sulfur cluster metabolism. Annual Meeting of the Canadian Society of Plant Physiologists, Burnaby, BC, Canada, June 2009.
4. Haas C.E., Rodionov D.A., Kropat K., Malasarn M, Merchant S.S. and de Crécy-Lagard V. 2009. A subset of the diverse COG0523 family of putative metal chaperones is linked to zinc homeostasis in all kingdoms of life. Gordon Research Conference: Cell Biology of Metals. Newport, Rhode Island. August 2009.
5. El Yacoubi B. and de Crécy-Lagard V. Comparative genomics and experimental validation to find universal, globally missing genes: the universal families COG009 and COG0533. Genomics:GTL Awardee Workshop VII, Bethesda, MD, February 2009.
6. Pribat A., Jeanguenin L., de Crécy-Lagard V. and Hanson A.D.. Genome annotation: Coupling the power of plant-prokaryote comparative genomics to experimental validation, COG0720 and COG3404. Genomics:GTL Awardee Workshop VII, Bethesda, MD, February 2009.
7. El Yacoubi B., Lyons B., Cruz Y., Reddy R., Nordin B., Agnelli F., Williamson J.R., Schimmel P, Swairjo M. and V. de Crécy-Lagard. The YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA: bioinformatic identification and experimental validation 16th Annual conference on *Microbial Genomics,* Lake Arrowhead, CA, September 2008.
8. Waller J.C., Da Silva V., Lyons, B.J., Ziemak M.J., Vickers T.J., Loizeau K., Beverley S.M., de Crécy-Lagard V, Gregory J.F. 3rd, Ravanel S. and Hanson A.D. A role for tetrahydrofolate in the metabolism of iron-sulfur clusters in all kingdoms of life. 50th Annual Meeting of the Canadian Society of Plant Physiologists, University of Ottawa, Ottawa, ON, Canada,June 2008.
9. de Crécy-Lagard V., El Yacoubi B., Haas C., Naponelli V., Noiriel A., Waller F.C., and Hanson A.D. Phylogenomics-guided validation of function for conserved unknown genes. Genomics:GTL Awardee Workshop VI, Bethesda MD, February 2008.
10. El Yacoubi B., Agnelli F., Williamson J., Takacs J., Lorsch J., de Crécy-Lagard V. and Swairjo M. A.. Biosynthesis of tRNA anticodon loop modification threonylcarbamoyl adenosine and role in ribosome function. 22nd International tRNA workshop, Uppsala, Sweden, November 2007.
11. El Yacoubi B., Swairjo M. and de Crécy-Lagard V. Identification of a universal gene family involved in threonylcarbamoyl adenosine (t^6A_{37}) biosynthesis. FEBS-CNRS workshop DNA and RNA Modification enzymes: Comparative Structure Mechanism Function and Evolution, Aussois, France, September 2007.

Invited Oral presentations

12. Hanson A.D., Waller J.C., Alvarez S., Naponelli V., Lara-Nuñez A., Blaby I.K., Da Silva V., Ziemak M.J., Vickers T.J., Beverley S.M., Edison A.S., Rocca, J.R., Gregory J.F. 3rd and de Crécy-Lagard V. A role for tetrahydrofolates in the metabolism of iron-sulfur clusters in all domains of life. FASEB Summer Research Conference on Folic Acid, Vitamin B12, and One Carbon Metabolism, Carefree, AZ, August 2010
13. de Crécy-Lagard V. and Hanson A.D. Phylogenomics-guided validation of function for

conserved unknown genes Co-presented at the Genomics:GTL Awardee Workshop VII, Bethesda, MD, February 2009.

14. de Crécy-Lagard V. Making sense of genomes: Linking gene and function by comparative genomics. 16th Annual Conference on Microbial Genomics, Lake Arrowhead, CA, September 2008.

Contributed oral presentation

15. Jeanguenin L., Lara-Núñez A., Pribat A., Hamner Mageroy M., Gregory III J. F., de Crécy-Lagard V. and Hanson A.D. An alternative pathway for 5-formyltetrahydrofolate metabolism. Society of Experimental Biology (SEB) Annual Main Meeting, Prague, Czech Republic, July 2010.

16. El Yacoubi B., McGuirk H., Hatin, I,, Iwata-Reuyl D, Murzin A. and V. de Crécy-Lagard. Function of the YrdC/ygjD conserved protein network the t^6A lead. 23rd International tRNA workshop, Aveiro, Portugal, January 2010.

17. Jeanguenin L., Pribat A., Hamner M., Ziemak M.J. and Hanson A.D. Identification of a new degradation pathway for 5-formyl-THF by comparative genomics. PMCB Annual Workshop Crystal River, FL, May 2009.

18. Haas C. and V. de Crécy-Lagard. Characterization of the role of COG0523 in Zinc trafficking. Joint Annual Meeting of Southeastern Branch and Florida Branch of the American Society for Microbiology, Jacksonville, FL, November, 2008.