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 (theseed.uchicago.edu/FIG/index.cgi) when the BMC genomics paper that is in preparation is published.
<table>
<thead>
<tr>
<th>Case</th>
<th>Hypothesis</th>
<th>TAIR ID</th>
<th>COG, gene name/</th>
<th>Subsystem in SEED</th>
<th>Experimental verification status</th>
<th>PubMed ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pterin carbinolamine dehydratase with role in Moco metabolism</td>
<td>At1g29610, At1g51110</td>
<td>COG2154, phnB</td>
<td>Pterin_carbinolamine_dehydratase</td>
<td>Validated in 7 euukaryotes and 6 prokaryotes</td>
<td>18245455</td>
</tr>
<tr>
<td>2</td>
<td>tSA biosynthesis</td>
<td>At5g00590</td>
<td>COG0009, YrdC</td>
<td>YrdC-YcO</td>
<td>Validated in yeast, archaeara and two bacteria, Arabidopsis, Pelthana in progress</td>
<td>19287007</td>
</tr>
<tr>
<td>3</td>
<td>PTPS family protein replacing the F0B step in folate synthesis</td>
<td>-</td>
<td>COG0720</td>
<td>Experimental-PTPS</td>
<td>Validated in 1 euukaryote and 3 prokaryotes</td>
<td>19395485, 18805734</td>
</tr>
<tr>
<td>4</td>
<td>Metal chaperone-Znc homeostasis</td>
<td>At1g15730, At1g26520, At1g80480</td>
<td>COG0523</td>
<td>COG0523</td>
<td>Validated in several bacteria</td>
<td>19622009</td>
</tr>
<tr>
<td>5</td>
<td>Folate-dependent, Fe/S cluster synthesis or repair protein</td>
<td>At4g12130, At1g60990</td>
<td>COG0354, ygFZ</td>
<td>YgfZ-Fe-S</td>
<td>Validated in E. coli, Bartonella hensellae, Halofex xanthan, Arabidopsis, Leishmania in yeast, mouse</td>
<td>20489182</td>
</tr>
<tr>
<td>6</td>
<td>Alternative route for 5-formyl/tetrahydrofolate disposal</td>
<td>At2g20830</td>
<td>COG3843</td>
<td>Experimental_Histidine_Degradation</td>
<td>Verified in 5 prokaryotes</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>7</td>
<td>5-Formyl/tetrahydrofolate cycloisomerase paralog</td>
<td>At1g76730</td>
<td>COG0212</td>
<td>5-FCL-like_protein</td>
<td>Predicted role in thiamine recycling (testing in progress)</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>8</td>
<td>tSA biosynthesis</td>
<td>At2g45270, At4g22720</td>
<td>COG0533, YgP</td>
<td>YrdC-YcO</td>
<td>Validated in yeast</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>9</td>
<td>Niacin and/or choline transporter</td>
<td>NaP homolog</td>
<td>MFS superfamil</td>
<td>Choline transport and metabolism</td>
<td>In progress in Bacillus subtilis, Ralstonia solanacearum, Burkholderia xenovorans, plants, mouse</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>10</td>
<td>Phytol phosphate kinase</td>
<td>At1g76820</td>
<td>COG1836, ari1612</td>
<td>COG1836</td>
<td>In progress in Arabidopsis, Synchocystis</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>11</td>
<td>NAD-dependent ribosomal protein modification</td>
<td>At1g12930, At1g67620</td>
<td>COG0709, ari1616</td>
<td>lolo</td>
<td>In progress in E. coli</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>12</td>
<td>Glycosylphosphatidylinositol involved in plant cell wall breakdown</td>
<td>At5g12950, At5g12960</td>
<td>COG3533, SAV1144</td>
<td>COG3533</td>
<td>In progress in X. campestris</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>13</td>
<td>mrRNA in small rRNA</td>
<td>At4g28330</td>
<td>COG2253</td>
<td>rRNA_modification_Arachae</td>
<td>In Progress in H. volcanii</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>14</td>
<td>Ribosome assembly</td>
<td>At1g09150</td>
<td>COG2016</td>
<td>rRNA_modification_Arachae</td>
<td>In progress in yeast and H. volcanii</td>
<td>Manuscript in prep</td>
</tr>
<tr>
<td>15</td>
<td>Pyridoxal phosphate enzyme in amino acid metabolism.</td>
<td>At4g26860, At1g11939</td>
<td>COG0325, yggS</td>
<td>PROSC</td>
<td>In progress in E. coli</td>
<td>Manuscript in prep</td>
</tr>
</tbody>
</table>
Papers and other products delivered

Papers (5)


Reviews (3)


Manuscripts in preparation (5)


Conference presentations (18)
Posters

Invited Oral presentations
13. de Crécy-Lagard V. and Hanson A.D. Phylogenomics-guided validation of function for


Contributed oral presentation


