Figure Captions

Joshua J. Hamilton

September 16, 2016

## Figure 1 (figures/fig1-tree/tree-abbrev.pdf)

Phylogenetic placement of the SAGs and MAGs within the acI lineage, relative to other sequenced actinobacterial genomes in the class Actinobacteria (Gao and Gupta 2012). The tree was built using RAxML (Stamatakis 2014) from a concatenated alignment of single-copy marker genes (Darling et al. 2014). The class Acidimicrobiia forms the outgroup.

## Figure 2a(b) (figures/fig2-sampling/completeness-2.svg and figures/fig2-sampling/completeness-1.svg)

Mean estimated completeness of tribe(clade)-level population genomes. For each tribe(clade), genomes were randomly sampled (with replacement) from the set of all genomes belonging to that tribe(clade). Completeness was estimated using single-copy marker genes (Parks et al. 2015). Error bars represent the 95% confidence estimated from 1000 iterations.

## Figure 3 (figures/fig3-workflow/Figure3.pdf)

Overview of reverse ecology pipeline. **(A)** Annotate microbial contigs using [KBase](http://kbase.us/), and build a metabolic network reconstruction from the annotations. For each genome, convert the metabolic network reconstruction to a metabolic network graph using custom Python scripts. In these graphs, metabolites are represented as nodes (circles) and reactions by arcs. **(B)** Create a composite network graph for each clade by joining graphs for all genomes from that clade, and compute seed compounds for the composite graph. In individual genome graphs, grey nodes and edges indicate components of the composite graph missing from that genome graph. In the composite graph, seed compounds are shown in red. **(Inset)** Three seed compounds which indicate an auxotrophy for L-homoserine, a methionine precursor. **(C)** Map metatranscriptomic reads to each individual genome using [BBMap](https://sourceforge.net/projects/bbmap/). For each clade, identify orthologous gene clusters using OrthoMCL (Li, Stoeckert, and Roos 2003). For each cluster, count all unique reads which map to any gene within that cluster using HTSeq (Anders, Pyl, and Huber 2014) and compute the relative gene expression using RPKM (Mortazavi et al. 2008).

## Figure 4 (figures/fig4-seeds/acI-aux-GH.pdf)

Seed compounds of members of the acI lineage. Compounds in the lower panel are degraded by peptidases and glycoside hydrolases. For these compounds, the intensity of the color indicates the percentile average log2 RPKM of the encoding gene cluster. For compounds acted upon by multiple gene clusters, the percentile of the most highly-expressed cluster was chosen.

## Figure 5 (figures/fig5-transporters/acI-transporters.pdf)

Compounds actively transported by members of the acI lineage. The intensity of the color indicates the percentile average log2 RPKM of the encoding gene cluster. For multi-subunit transporters, the percentile of the most highly-expressed subunit was chosen.

## Supplementary Figure 1 (figures/fig1-tree/tree-full.pdf)

Phylogenetic placement of the SAGs and MAGs within the acI lineage, relative to other sequenced actinobacterial genomes in the class Actinobacteria (Gao and Gupta 2012). The tree was built using RAxML (Stamatakis 2014) from a concatenated alignment of single-copy marker genes (Darling et al. 2014). The class Acidimicrobiia forms the outgroup.

## Supplementary Figure 2 (figures/fig3-workflow/FigureS2.pdf)

Pruning a metabolic network graph, using glycolysis as an example. **(A)** Metabolic reconstruction for glycolysis. This panel also shows the KBase reaction IDs corresponding to each step of glycolysis, as well as the gene(s) encoding the enzyme for that step. Genes are shown only for the TE02754 genome. **(B)** Glycolysis represented as a metabolic network graph. Currency metabolites are circumscribed by a dotted rectangle. For clarity, protons are not shown. **(C)** Representation of glycolysis after pruning. The images in (A) and (B) are modified from (Ma and Zeng 2003).

## Supplementary Figure 3 (figures/fig3-workflow/FigureS3.pdf)

Identifying seed compounds in metabolic networks. **(A)** Metabolic network graph for an imaginary network. Grey circles represent metabolites and arcs represent reactions or parts of reactions. **(B)** To identify seed compounds, the metabolic network graph is first decomposed into its strongly connected components (SCCs), sets of nodes such that each node in the set is reachable from every other node. Here, each colored set of nodes corresponds to a unique SCC. **(C)** SCC decomposition enables seed sets to be identified from source components (components with no incoming edges) on the condensation of the original graph. In the condensation of the original graph shown here, each colored node corresponds to a unique SCC (and a potential seed se). Seed sets are enclosed in dotted circles. The pink seed set contains three seed compounds, the orange contains two, and the blue and yellow each contain one seed compound. The images in (A) and (B) are modified from (Borenstein et al. 2008).

## References

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2014. “HTSeq A Python framework to work with high-throughput sequencing data.” *Bioinformatics* 31 (2): 166–69. [doi:10.1101/002824](http://doi.org/10.1101/002824).

Borenstein, Elhanan, Martin Kupiec, Marcus W Feldman, and Eytan Ruppin. 2008. “Large-scale reconstruction and phylogenetic analysis of metabolic environments.” *Proceedings of the National Academy of Sciences* 105 (38): 14482–87. [doi:10.1073/pnas.0806162105](http://doi.org/10.1073/pnas.0806162105).

Darling, Aaron E, Guillaume Jospin, Eric Lowe, Frederick A Matsen, Holly M Bik, and Jonathan A Eisen. 2014. “PhyloSift: phylogenetic analysis of genomes and metagenomes.” *PeerJ* 2 (January): e243. [doi:10.7717/peerj.243](http://doi.org/10.7717/peerj.243).

Gao, Beile, and Radhey S Gupta. 2012. “Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria.” *Microbiology and Molecular Biology Reviews* 76 (1): 66–112. [doi:10.1128/MMBR.05011-11](http://doi.org/10.1128/MMBR.05011-11).

Li, Li, Christian J Stoeckert, and David S Roos. 2003. “OrthoMCL: identification of ortholog groups for eukaryotic genomes.” *Genome Research* 13 (9): 2178–89. [doi:10.1101/gr.1224503](http://doi.org/10.1101/gr.1224503).

Ma, Hongwu, and An-Ping Zeng. 2003. “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.” *Bioinformatics* 19 (2): 270–77. [doi:10.1093/bioinformatics/19.2.270](http://doi.org/10.1093/bioinformatics/19.2.270).

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and quantifying mammalian transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. [doi:10.1038/nmeth.1226](http://doi.org/10.1038/nmeth.1226).

Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Phil Hugenholtz, and Gene W Tyson. 2015. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.” *Genome Research* 25 (7): 1043–55. [doi:10.1101/gr.186072.114.Freely](http://doi.org/10.1101/gr.186072.114.Freely).

Stamatakis, Alexandros. 2014. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.” *Bioinformatics* 30 (9): 1312–13. [doi:10.1093/bioinformatics/btu033](http://doi.org/10.1093/bioinformatics/btu033).