Figure Captions

Joshua J. Hamilton

September 16, 2016

## Figure 1 (figures/fig1-tree/Figure1.pdf)

Phylogenetic placement of the genomes used in this study within the acI lineage. The tree was built using RAxML (Stamatakis, 2014) from a concatenated alignment of protein sequences from 37 single-copy marker genes (Darling *et al.*, 2014). The order Actinomycetales forms the outgroup. Vertical black bars indicate groups of genomes belonging to defined clades/tribe within the acI lineage, as determined using 16S rRNA gene sequences (for SAGs and bin FNEF8-2 bin\_7 acI-B only) and a defined taxonomy (Newton *et al.*, 2011). SAGs are indicated with italic text. Supplemental Figure S5 shows the position of the acI lineage relative to other orders within the class Actinobacteria.

## Figure 2 (figures/fig2-sampling/Figure2.pdf)

Mean estimated completeness of tribe-level (clade-level) population genomes as a function of the number of sampled 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 204 single-copy marker genes from the phylum Actinobacteria (Parks *et al.*, 2015). Error bars represent the 95% confidence interval estimated from 1000 iterations.

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

Overview of the seed set framework and metatranscriptomic mapping, using three genomes from the acI-C clade as an example. **(A)** Microbial contigs are annotated using [KBase](http://kbase.us/), and a metabolic network reconstruction is built from the annotations. For each genome, the metabolic network reconstruction is converted to a metabolic network graph using custom Python scripts. In these graphs, metabolites are represented as nodes (circles) and reactions by arcs. Grey nodes and edges indicate components of the composite graph missing from that genome graph. Additional information on this step of the workflow is available in Figure S1. **(B)** A composite network graph is created for each clade by joining graphs for all genomes from that clade, and seed compounds are computed for the composite graph. Seed compounds are shown in red. Additional information on this step of the workflow is available in Figures S2, S3, and S4. **(Inset)** Three seed compounds which indicate an auxotrophy for L-homoserine, a methionine precursor. **(C)** Metatranscriptomic reads are mapped to each individual genome using [BBMap](https://sourceforge.net/projects/bbmap/). Orthologous gene clusters are identified using OrthoMCL (Li *et al.*, 2003). For each cluster, unique reads which map to any gene within that cluster are counted using HTSeq (Anders *et al.*, 2014) the relative gene expression is computed using RPKM (Mortazavi *et al.*, 2008).

## Figure 4 (figures/fig4-seeds/Figure4.pdf)

Seed compounds of members of the acI lineage. **(A)** Auxotrophies and nutrient sources, not including peptides and glycosides. **(B)** Peptides and glycosides. These compounds represent those inferred from genome annotations, rather than the seed compounds themselves. In panel (B), 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/Figure5.pdf)

Transporters that are actively expressed by members of the acI lineage, as inferred from consensus annotations of genes associated with transport reactions present in metabolic network reconstructions. The intensity of the color indicates the average log2 RPKM of the encoding gene cluster. For multi-subunit transporters, the RPKM of the substrate-binding subunit was chosen.

## Supplementary Figure 1 (figures/fig3-workflow/FigureS1.pdf)

Converting an unannotated genome to a metabolic network graph, for a simplified genome containing only glycolysis. **(A)** Microbial contigs are annotated using KBase, and a metabolic network reconstruction is built from the annotations. The reconstruction provides links between protein-encoding genes in the genome and the enzymatic reactions catalyzed by those proteins. **(B)** The metabolic network reconstruction represents metabolism as a hypergraph, in which metabolites are represented as nodes and reactions as hyperedges. In this representation, an edge can connect more than two nodes. For example, a single hyperedge (denoted by a heavy black line) connects the metabolites glucose and ATP to glucose-6P, ADP, and Pi. For clarity, protons are not shown. **(C)** However, the algorithm used by the seed set framework requires metabolism to be represented as a metabolic network graph, in which an edge can connect only two nodes. In this representation, a reaction is represented by a set of edges connecting all substrates to all products. For example, the heavy hyperedge in (B) is now denoted by six separate edges connecting glucose to ADP, glucose to Pi, glucose to glucose-6P, ATP to ADP, ATP to Pi, and ATP to glucose-6P (again denoted by heavy black lines). Of these, only one (glucose to glucose-6P) is biologically meaningful. The dotted line surrounds the currency metabolites. **(D)** The metabolic network graph is then pruned, a process which removes all currency metabolites and any edges in which those metabolites participate. Of the six heavy edges in (C), only the biologically meaningful one is retained, connecting glucose to glucose-6P (again denoted by a heavy black line). The images in (B) and (C) are modified from (Ma and Zeng, 2003). Note: The visual representations shown here are intended to illustrate the metabolic network reconstruction process, and are not indicative of the data structures used by our pipeline.

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

Construction of composite metabolic network graph for clade acI-C. Beginning with metabolic network graphs for genomes Actinobacterium\_10 and ME00885, nodes and edges unique to ME00885 are identified (in red). These nodes and edges are added to the Actinobacterium\_10 graph, giving the composite metabolic network graph for these two genomes (Actinobacterium\_10 + ME00885). Then, this graph is compared to the graph for ME03864, and nodes and edges unique to ME03864 are identified (in red). These nodes and edges are added to the Actinobacterium\_10 + ME00885 metabolic network graph, giving the composite metabolic network graph for clade acI-C. Note: The visual representations shown here are intended to illustrate the metabolic network reconstruction process, and are not indicative of the data structures used by our pipeline.

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

Identifying seed compounds in metabolic networks, using the same metabolic network as in Supplemental Figure S1. **(A)** 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 set of circled nodes corresponds to a unique SCC. **(B)** 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 node corresponds to a unique SCC. This network has a single seed set, SCC\_1, enclosed in a dotted circle. **(C)** Seed compounds can be found from the mapping between SCCs and their constituent metabolites. In this example, glucose is the sole seed compound. While this particular result is probably intuitive, real metabolic networks are considerably more complex. Note: The visual representations shown here are intended to illustrate the metabolic network reconstruction process, and are not indicative of the data structures used by our pipeline.

## Supplementary Figure 4 (figures/fig3-workflow/FigureS4.pdf)

Complete composite metabolic network graph for clade acI-C, showing disconnected components and the giant strongly connected components. Gray nodes and edges represent disconnected components which are dropped prior to computing the network’s seed sets. Red nodes represent those present in the giant strongly connected component which contains the majority of the metabolites in the network.

## Supplementary Figure 5 (figures/fig1-tree/FigureS5.pdf)

Phylogenetic placement of the genomes used in this study within the acI lineage, relative to other sequenced actinobacterial genomes in the class Actinobacteria (Gao and Gupta, 2012) (Table S17). The tree was built using RAxML (Stamatakis, 2014) from a concatenated alignment of protein sequences from 37 single-copy marker genes (Darling *et al.*, 2014). The class Acidimicrobiia forms the outgroup. Vertical black bars indicate groups of genomes belonging to defined clades/tribe within the acI lineage, as determined using 16S rRNA gene sequences (for SAGs and bin FNEF8-2 bin\_7 acI-B only) and a defined taxonomy (Newton *et al.*, 2011). SAGs are indicated with italic text.

## References

Anders S, Pyl PT, Huber W. (2014). HTSeq A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**: e243.

Gao B, Gupta RS. (2012). Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. *Microbiology and Molecular Biology Reviews* **76**: 66–112.

Li L, Stoeckert CJ, Roos DS. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**: 2178–89.

Ma H, Zeng A-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**: 270–277.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621–628.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews* **75**: 14–49.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**: 1043–1055.

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.