Comparative Proteogenomics
Eli Venter, Samuel H Payne
J Craig Venter Institute, Rockville MD

Abstract
Almost all prokaryotic genomes receive only a single round of automated annotation. Thus gene sets contain numerous errors in even the most basic form of annotation: protein primary structure. Proteogenomics can quickly and efficiently discover misannotations. We analyze seven datasets from five bacterial phyla, and correct hundreds of genes. We also speculate on reasons for errors in gene prediction software.

Introduction
Accurate gene models are a prerequisite for meaningful use of a genome. Annotations consistently miss genes, and start sites may be wrong for an additional 20%. Gene prediction software is often trained on too narrow a set of proteins, and thus has difficulty with novel, or irregular proteins. Unfortunately, as software improves, dubious predictions remain in public databases, confusing comparative analysis.

Methods
Our automated pipeline reports the observed proteome, including novel genes, dubious genes, translational start sites, signal peptides, and evidence of frame shift. We begin to analyze the conservation of protein start sites across taxa.

Results
Our automated pipeline reports the observed proteome, including novel genes, dubious genes, translational start sites, signal peptides, and evidence of frame shift. We begin to analyze the conservation of protein start sites across taxa.

Patterns of Error

Conclusions
Annotation accuracy correlates well with GC content, distance to model organisms, date of annotation. Subsets of proteins can be isolated to further train gene prediction algorithms.

Acknowledgements
We thank Richard Smith, PNNL, for data; Bill Klimke, NCBI, for help incorporating results into RefSeq; Ari Frank, UCSD, for help with PepNovo. This work was funded by NSF grant EF-0949047 to SHP. Contact at spayne@jcv.org