Comparative Proteogenomics

Eli Venter, Samuel H Payne

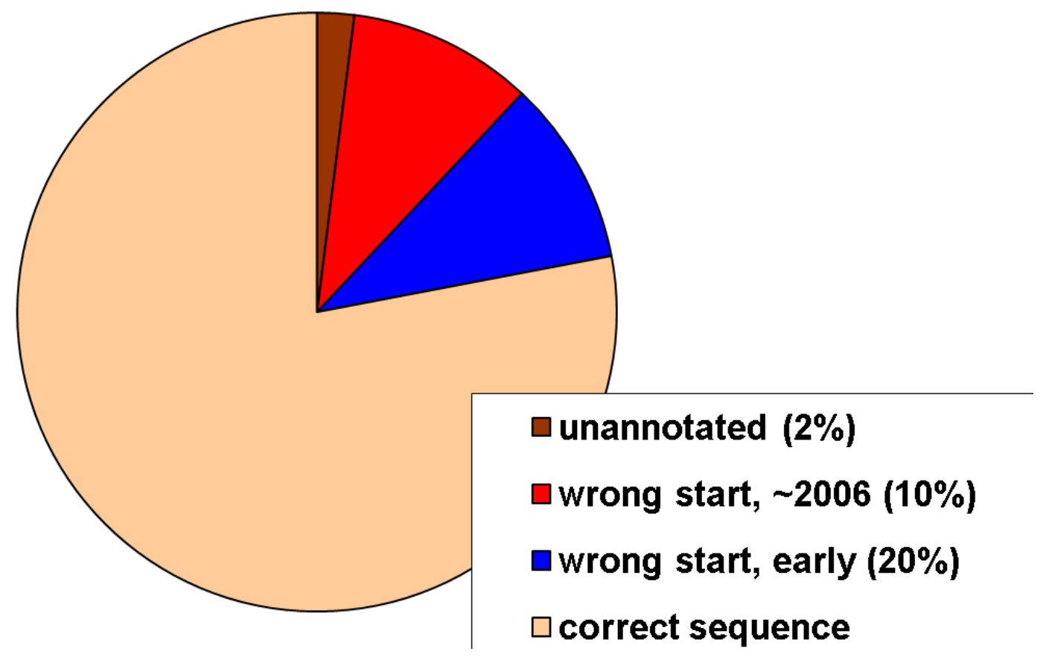
J Craig Venter Institute, Rockville MD

Abstract

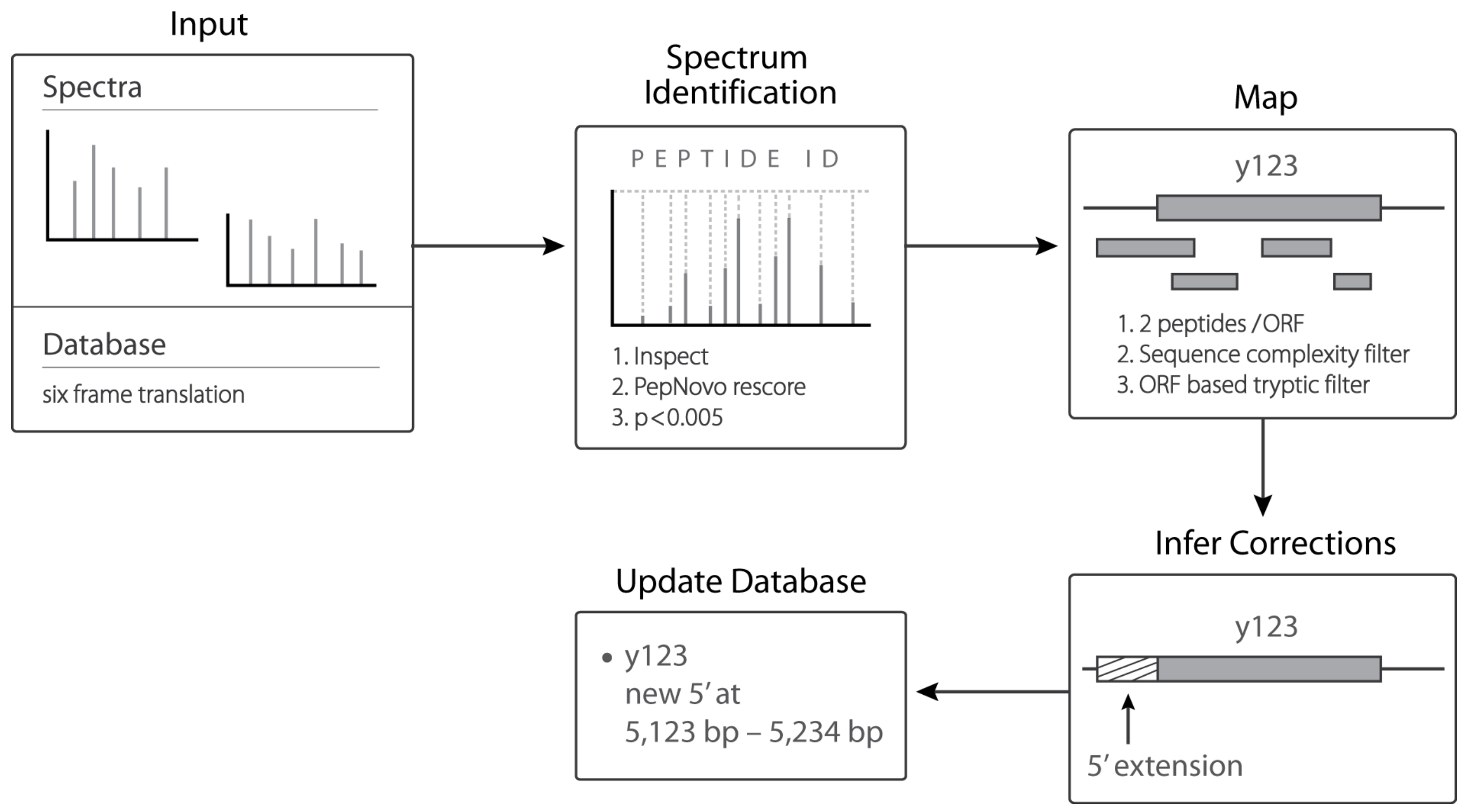
Almost all prokaryotic genomes receive only a single round of automated annotation. Thus gene sets contain numerous errors in even the most basic form of annotation: protein primary structure. Proteogenomics can quickly and efficiently discover misannotations. We analyze seven datasets from five bacterial 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.



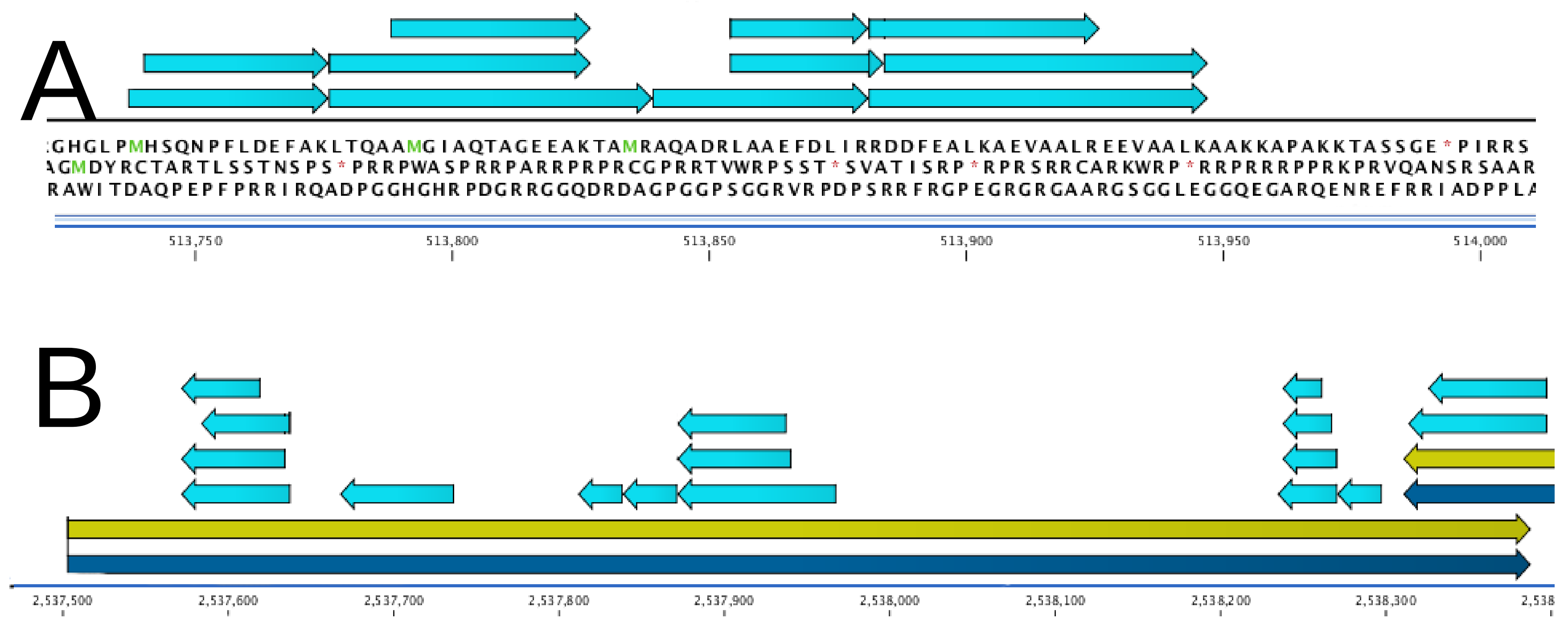
Methods



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.

	Novel Genes	Wrong Start
Caulobacter	65	105
Synechocystis	7	18
Arthrobacter	12	65
Desulfovibrio	40	75
Leptospira	18	26
B. anthracis	4	5
Y. pestis	4	6

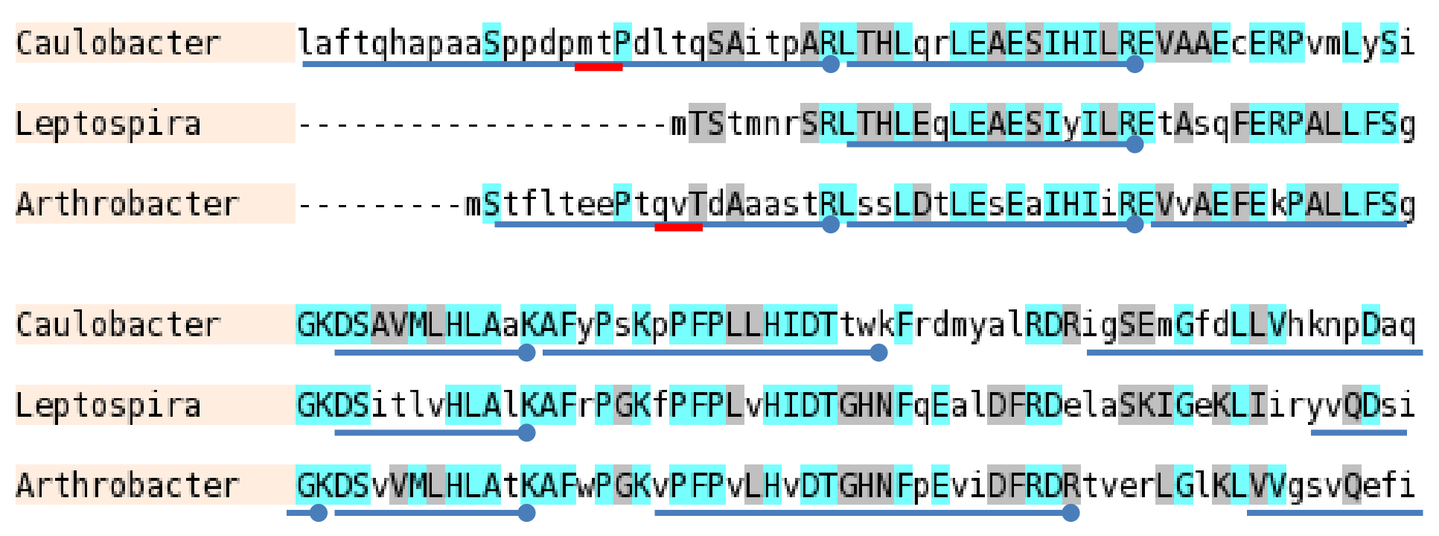


Types of Novel Genes:

A. Peptide mapping to a region of the genome lacking any annotation.

B. Peptides mapping to the same locus as a currently predicted gene, but in a different frame (or strand), implying a dubious gene. Purging dubious gene models is a necessary, yet laborious, task.

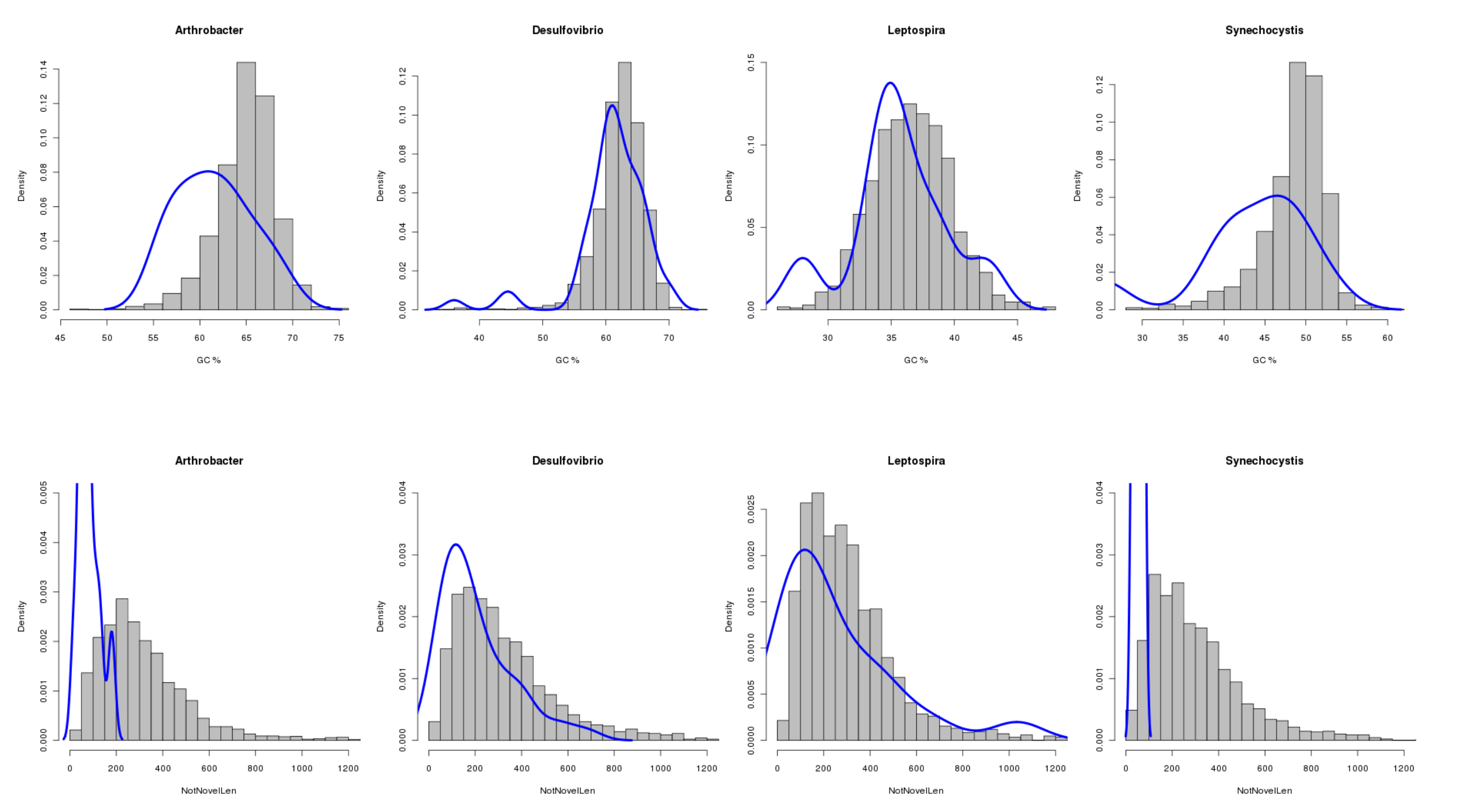
Start Site Conservation


Caulobacter laftqhapaaSppdpmtPdltqSAitpARLTHLqrLEAESIHILREVAAEcERPvmLySi
Leptospira -------------------mTStmnrSRLTHLEqLEAESIyILREtAsqFERPALLFSg
Arthrobacter ---------mStflteePtqvTdAaastRLssLDtLEsEaIHIiREVvAEFEkPALLFSg
Caulobacter GKDSAVMLHLAaKAFyPsKpPFPLLHIDTtwkFrdmyalRDRigSEmGfdLLVhknpDaq
Leptospira GKDSitlvHLAlKAFrPGKfPFPLvHIDTGHNFqEalDFRDelaSKIGeKLIiryvQDsi
Arthrobacter GKDSvVMLHLAtKAFwPGKvPFPvLHvDTGHNFpEviDFRDRtverLGlKLVVgsvQefi

Start site conservation is often assumed, but rarely tested. With our diverse dataset, we can observe how often proteins restrict their start site. In the figure CysD is shown, with two sequences corrected by proteogenomics (original start in red). Each protein utilized the longest sequence possible in its ORF. The n-terminus shows relatively poor conservation.

Patterns of Error

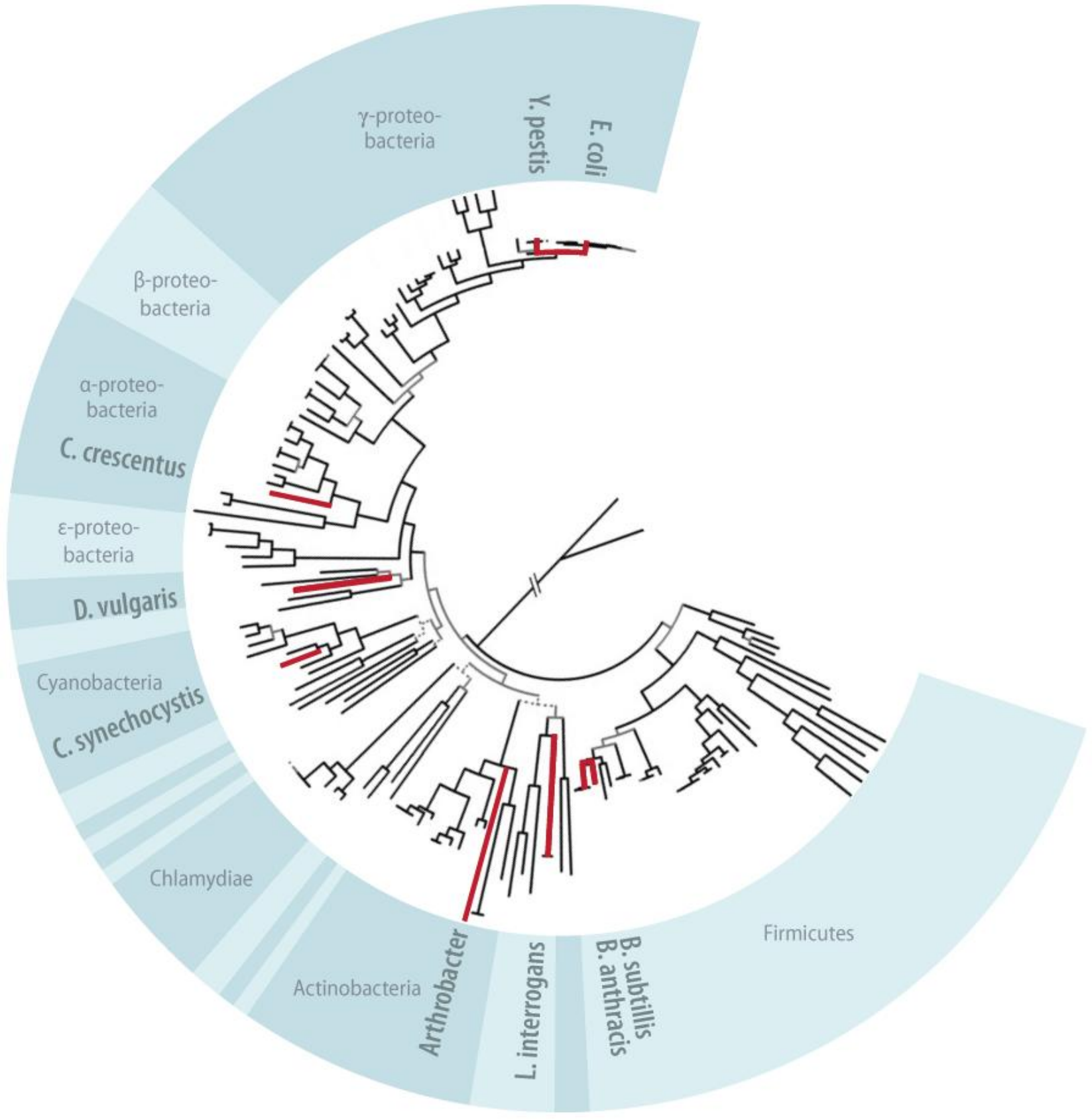
What distinguishes mispredicted genes? Abnormal GC content? Codon usage? Length? Commonness ?



The background distribution of %GC and gene length (grey bars) is compared to the novel genes (blue line). Novel genes are more often short, but not unreasonably so. GC content does not appear drastically different.

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@jcvi.org