Input and Output Files:

Each tool in the Proteomic-Genomic Nexus uses different input and output files for its purpose, some of which are optional. A summary of the input and output files used in the Proteomic-Genomic Nexus software package is shown in Supplementary Tables 1 and 2.

Supplementary Table 1: Input files used by each of the tools in the Proteomic-Genomic Nexus software package.

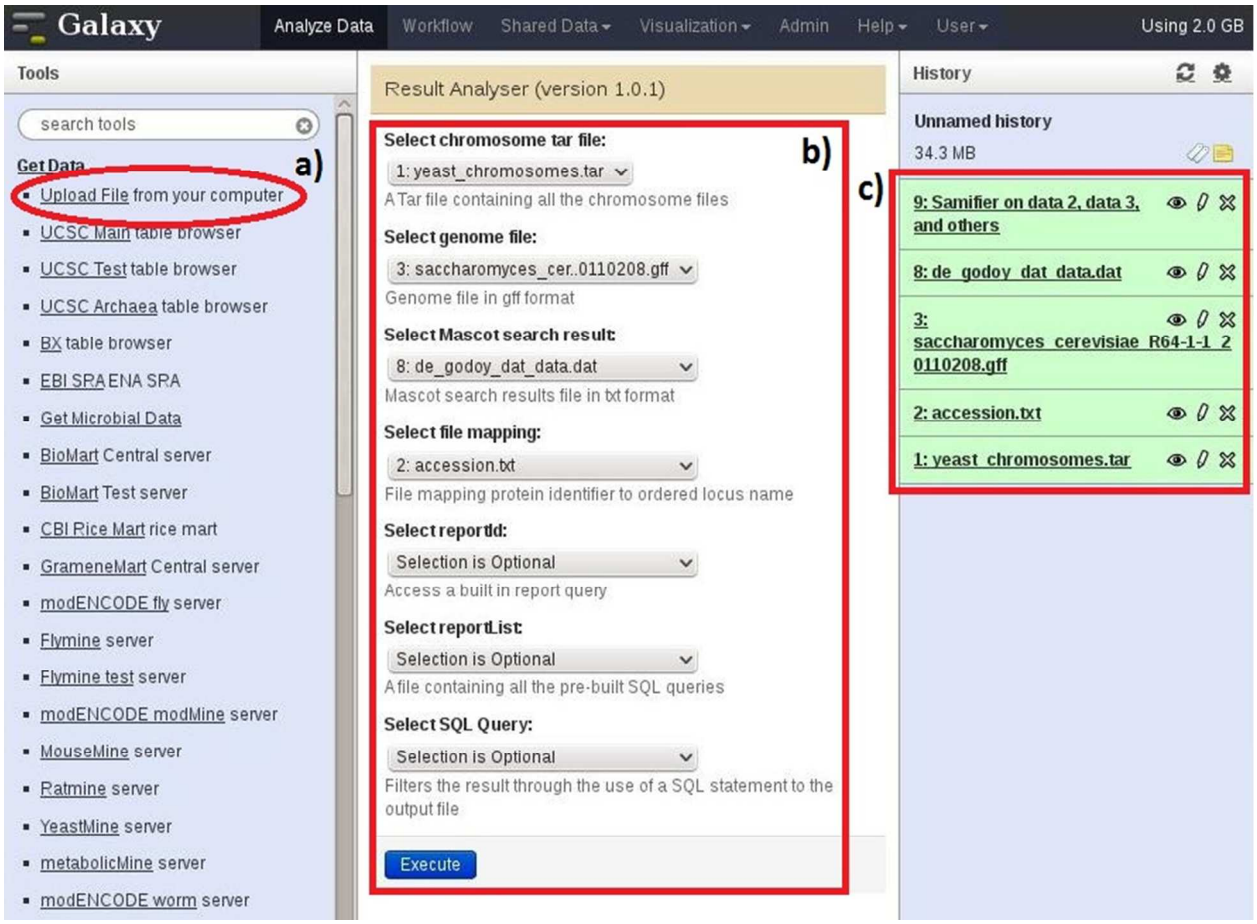
	Protein Generator	Virtual Protein Merger	Samifier	Results Analyzer
Gene prediction file in Glimmer3 format	x			
Translation table in NCBI format	x	x		
Genome sequence in FASTA format	x	x	x	x
Annotation results in GFF3 format		x	x	x
Mascot search results in .dat or .mzIdentML format (versions 1.0 or 1.1 supported)		x	x	x
A text file mapping protein IDs to ordered locus name			x	x
File containing pre-built SQL queries				x

Supplementary Table 2: Output files produced by each of the tools in the Proteomic-Genomic Nexus software package.

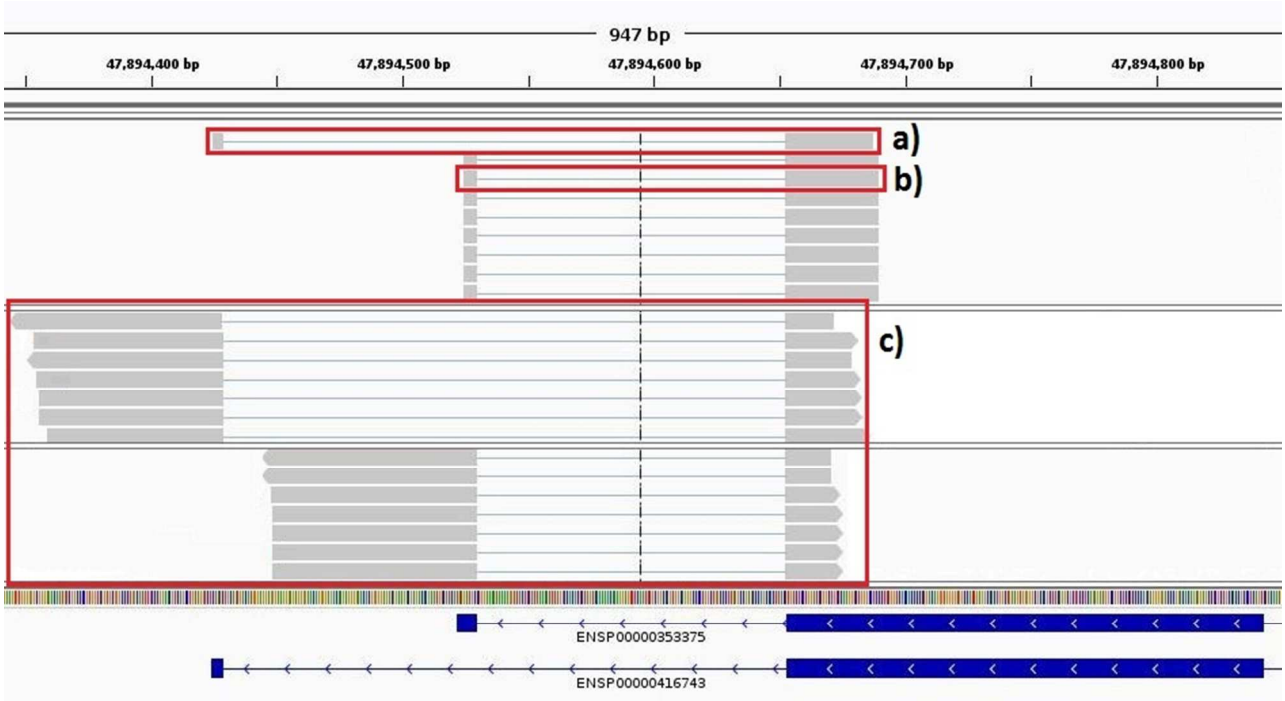
	Protein Generator	Virtual Protein Merger	Samifier	Results Analyzer
Protein database in FASTA format	x			
A text file mapping protein IDs to ordered locus name	x			
Annotation results in GFF3 format	x	x		
Sequence alignment file in SAM format			x	
Regions of interest in BED format			x	
A text file which describes reported errors			x	
A table file of the sequence alignment SAM file				x

Supplementary Table 3. Mascot identity threshold scores for analysis of two strains of *Campylobacter concisus*

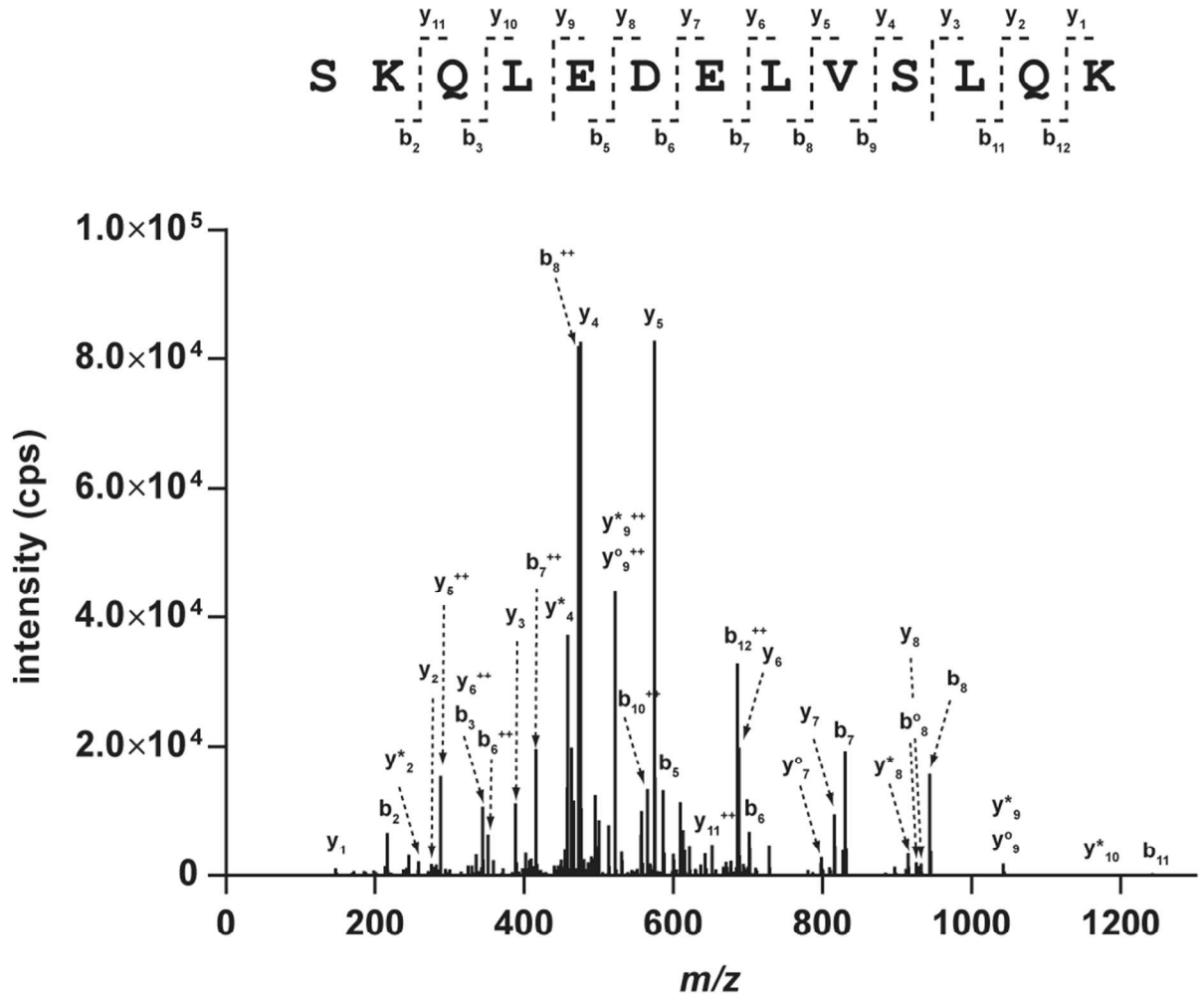
Mascot protein sequence database	Mascot threshold score for each strain	
	***C. consisus* strain 13826 (RefSeq: NC_009802.1)**	***C. consisus* strain UNSWCD (RefSeq: AENQ01000001 - AENQ01000086)**
NCBI RefSeq [1]	6	7
Genes predicted by Glimmer [2]	7	6
Six-frame translation of genome by Virtual Protein Generator	25	25



Supplementary Figure 1. Graphical user interface of Results Analyser in Galaxy. a) A hyperlink to an interface to upload the input files. **b)** Users can then run the Results Analyser tool by selecting their files from the drop-down menus and applying filters through customized SQL queries. **c)** Output files that are ready for download are listed in the right-most panel.



Supplementary Figure 2. Alternative splicing of human microtubule 4, confirmed by RNA-seq and GeLC-MS/MS. **a)** A peptide of sequence EAQTLDSQIQETSI was found to span the splice junction of one isoform of this protein (ENSP00000353375). This peptide had a Mascot peptide score of 88. **b)** A different peptide, of sequence EAQTLDSQIQETN, was found to span across an alternative splice junction for another isoform (ENSP00000416743). This peptide had a score of 63. The gene structure and splicing pattern for each isoform from the Ensembl database is shown in the blue tracks below. **c)** RNA-seq reads confirm the two alternatively spliced junctions. RNA reads that do not span across splice junctions were manually removed to facilitate this visualization in IGV. Although high-confidence peptides and RNA-seq reads validated the two alternatively spliced isoforms of the human microtubule-associated protein 4, the Cufflinks sequence assembly program did not detect both isoforms.



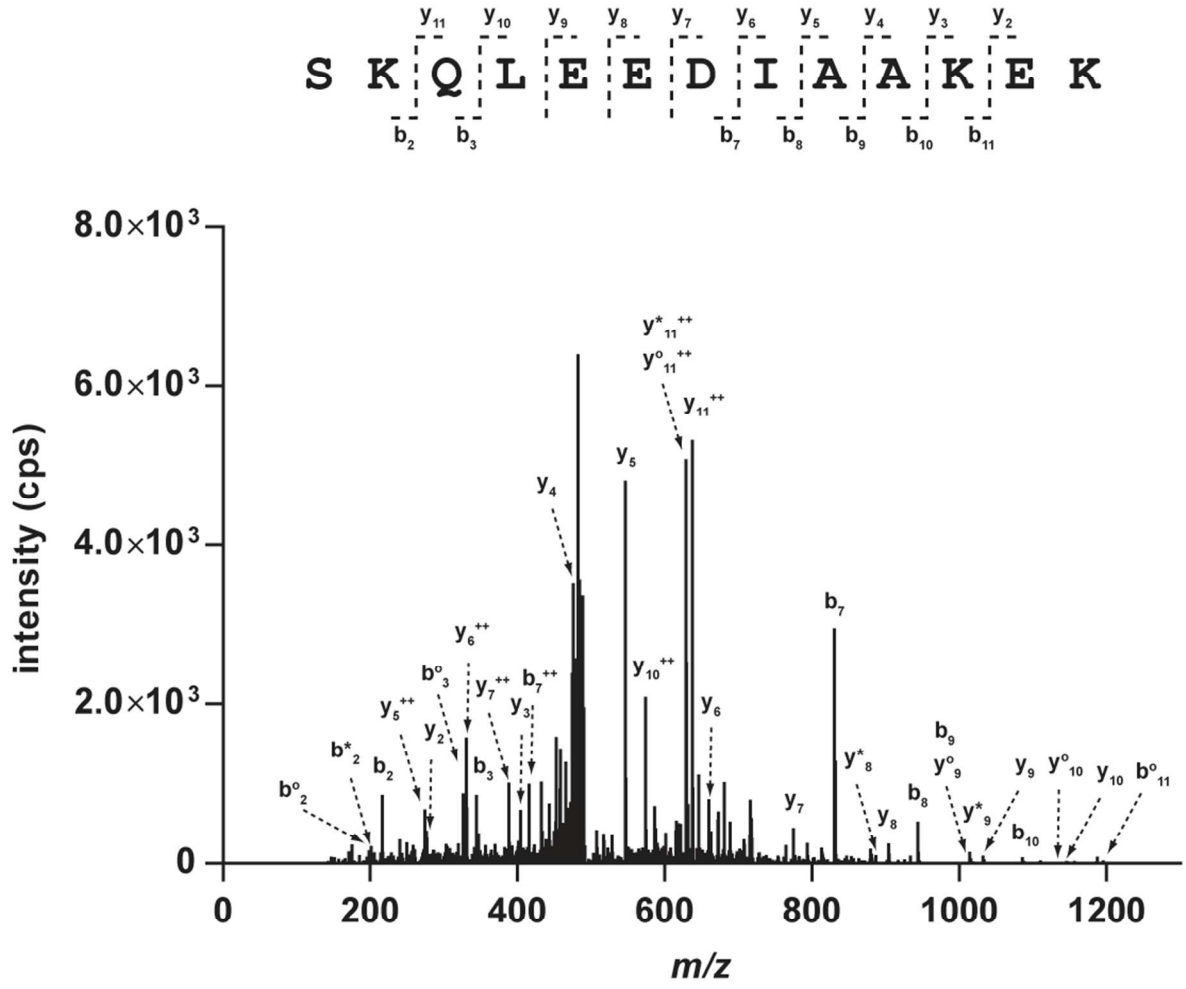
Notes:

[1]**Doubly charged ions are denoted by (++)**

[2]**Ions formed after a neutral loss of NH_3 are denoted by (*)**

[3]**Ions formed after a neutral loss of H_2O are denoted by (°)**

Supplementary Figure 3. CID-MS/MS spectrum of a peptide of sequence SKQLEDELVSLQK. This peptide was observed as a triply-charged ion at 506.2787 m/z, and identified with a Mascot ion score of 61 (E-value = 1.7×10-5). Observed b and y ions and their derivatives are labeled in the spectrum. The associated fragment ion coverage is summarised in the illustrated peptide sequence.



Supplementary Figure 4. CID-MS/MS spectrum of a peptide of sequence SKQLEEDIAAKEK. This peptide was observed as a triply-charged ion at 496.9347 m/z, and identified with a Mascot ion score of 47 (E-value = 3.6×10-5). Observed b and y ions and their derivatives are labeled in the spectrum. The associated fragment ion coverage is summarised in the illustrated peptide sequence.

References

1. Pruitt KD, Tatusova T, Brown GR, Maglott DR: NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research* 2012, 40(Database issue):D130-135.

2. Delcher AL, Bratke KA, Powers EC, Salzberg SL: Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)* 2007, 23(6):673-679.