PATRIC Genome Curation

Eric E. Snyder, Ph.D.
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0447
USA

Office: (540) 231-5428
Mobile: (540) 230-5225
Fax: (540) 231-2891
Email: eesnyder@vbi.vt.edu
JDAM: N 37 12'01.6", W 80 24'26.9"

PATRIC Genome Annotation

PATRIC Pathogens

PATRIC Organisms	Named Isolates	RefSeq Genomes	PATRIC Genomes	Associated Diseases
Bacteria				
Brucella sp.	24	2	4	Brucellosis
Coxiella burnetti	1	4	2	Q Fever
Rickettsia sp.	186	6	9	Typhus, RM Spotted Fever
Viruses				
Caliciviruses	1036	13	52	Foodborne Gastroenteritis
Coronaviruses	286	11	170	SARS
Hepatitis A viruses	14	1	17	Hepatitis
Hepatitis E viruses	9	1	37	Hepatitis
Rabies viruses	13	1	9	Rabies

PATRIC Organisms	Genome Properties					
	kb	topology	NA	Sense	PolyA	ORFs
Bacteria						
Brucella sp.	3200	circular	dsDNA			3300
Coxiella burnetti	2000	circular	dsDNA			2134
Rickettsia sp.	1200	circular	dsDNA			834
Viruses						
Caliciviruses	8	linear	ssRNA	+	+	3
Coronaviruses	30	linear	ssRNA	+	-	14
Hepatitis A viruses	7	linear	ssRNA	+	+	1
Hepatitis E viruses	7.5	linear	ssRNA	+	+	3
Rabies viruses	12	linear	ssRNA	-		5

Reference Gene Set

Compile sets of orthologous genes

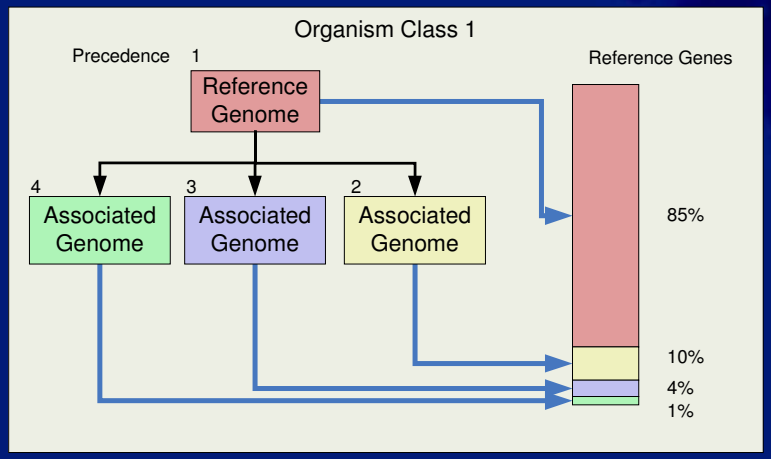
- Within and between organism categories (genera)

Purpose

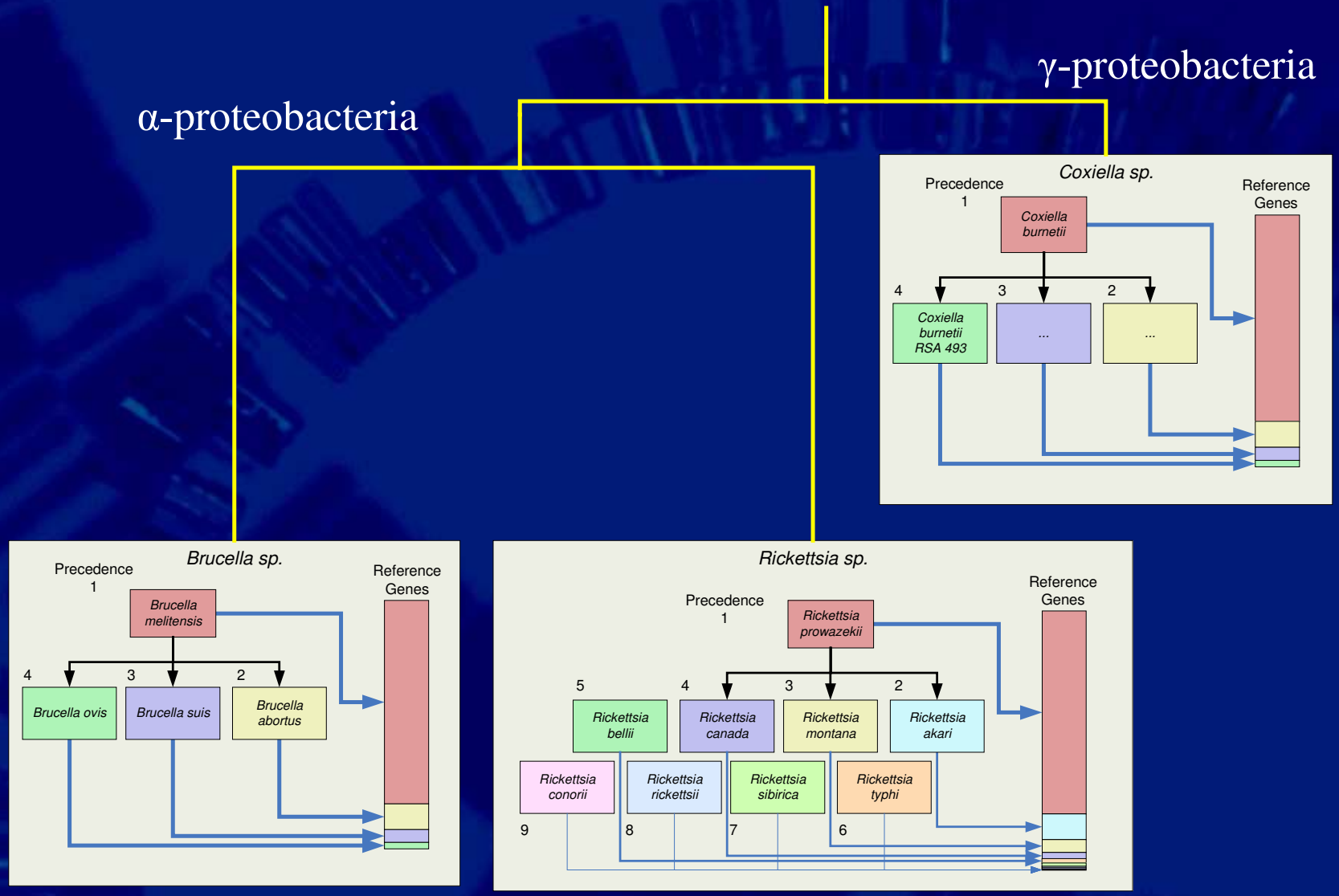
- Identify conserved genes characteristic of categories
- Identify organism-specific genes
- Facilitate functional characterization

Methodology

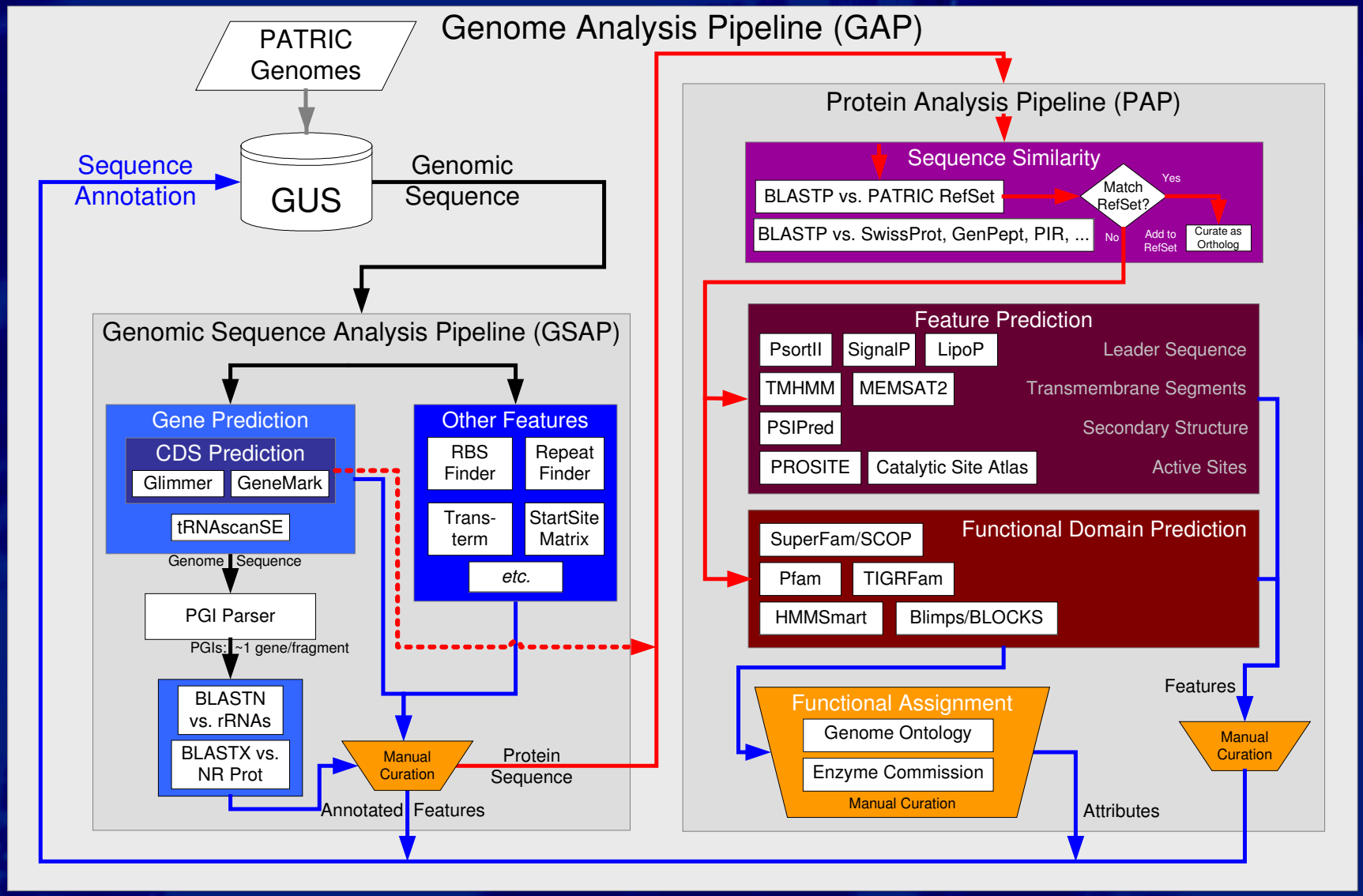
- Construct reference gene set by iterative comparison with individual genomes
- Orthology defined by:
  - bidirectional best (BLAST) hit
  - Conserved synteny



Reference Gene Set



Curation Infrastructure



Putative Gene Intervals

Goal

Divide genome into segments, each containing a single gene (plus flanking sequence)

Genome Sequence (Linear)

$\mathbf{S} = (s_1, s_2, s_3, \ldots, s_{n-1}, s_n)$ sequence length = n

$\mathbf{P} = (p_1, p_2, p_3, \ldots, p_{z-1}, p_z)$ **S** divided into z PGIs

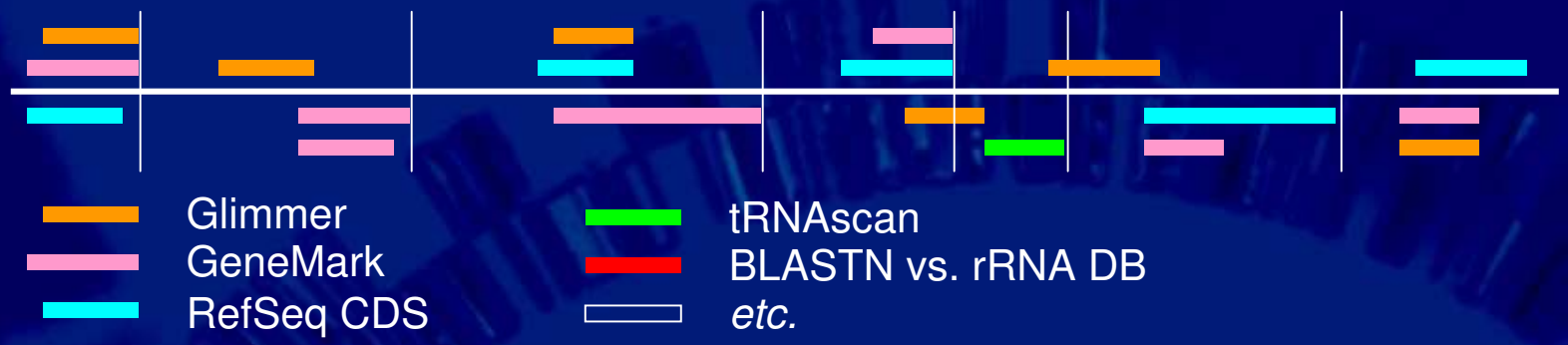
$p_1 = (s_1, s_2, \ldots, s_a)$

$p_2 = (s_{a+1}, s_{a+2}, \ldots, s_b)$

$\ldots$

$p_z = (s_{y+1}, s_{y+2}, \ldots, s_z)$

Putative Gene Intervals



Problem

How to choose the optimal set of gene predictions from heterogeneous and potentially conflicting sources?

Solution

Dynamic programming

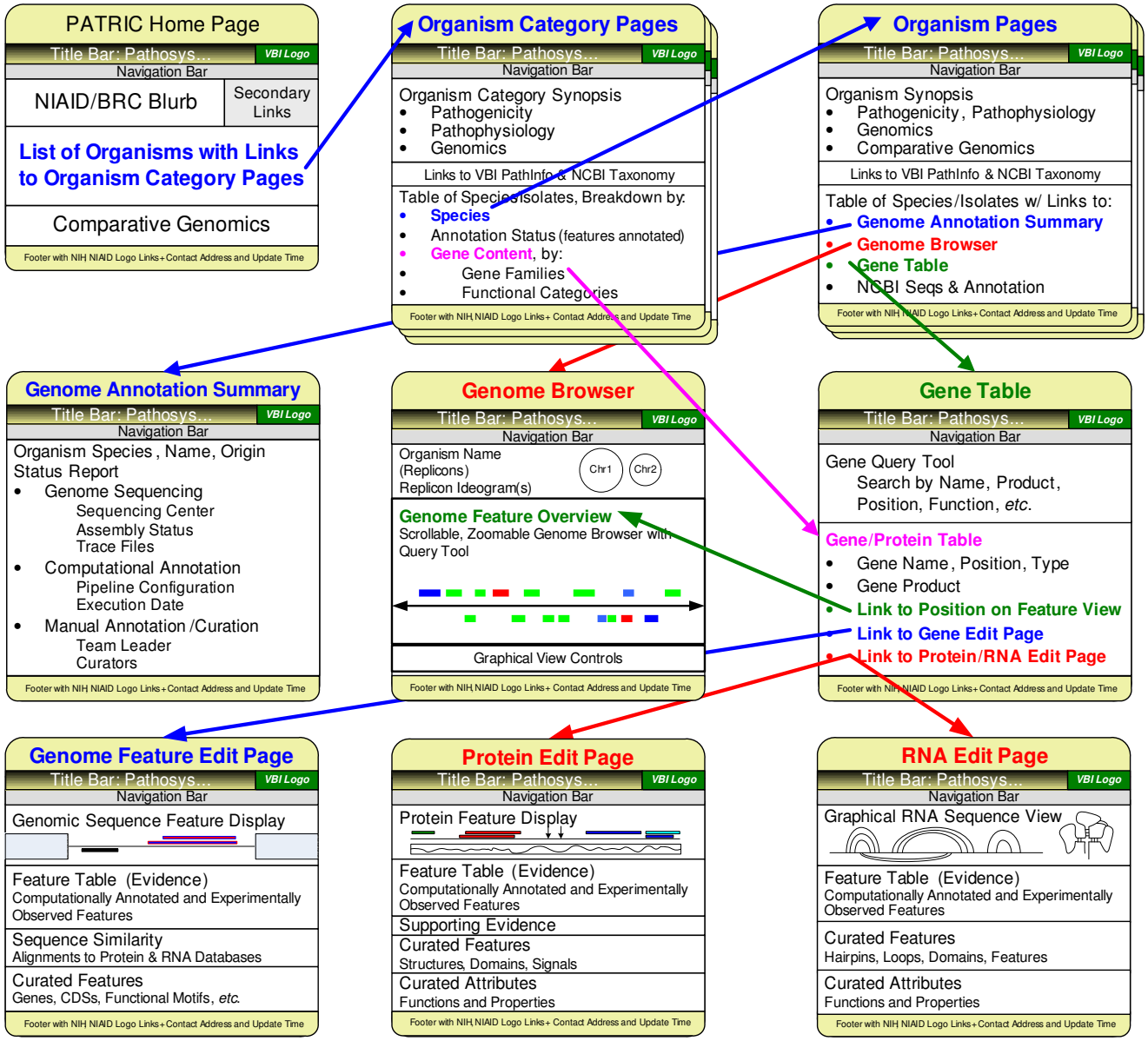
$$D_j = \max_{z \in [pgms]} \begin{cases} L^z_{1,j} \\ \max\limits_{k:1 \to j-1} \left[L^z_{k,j} + D_{k-1} \right] \\ D_{j-1} \end{cases}$$

D_j = optimal score at *j*

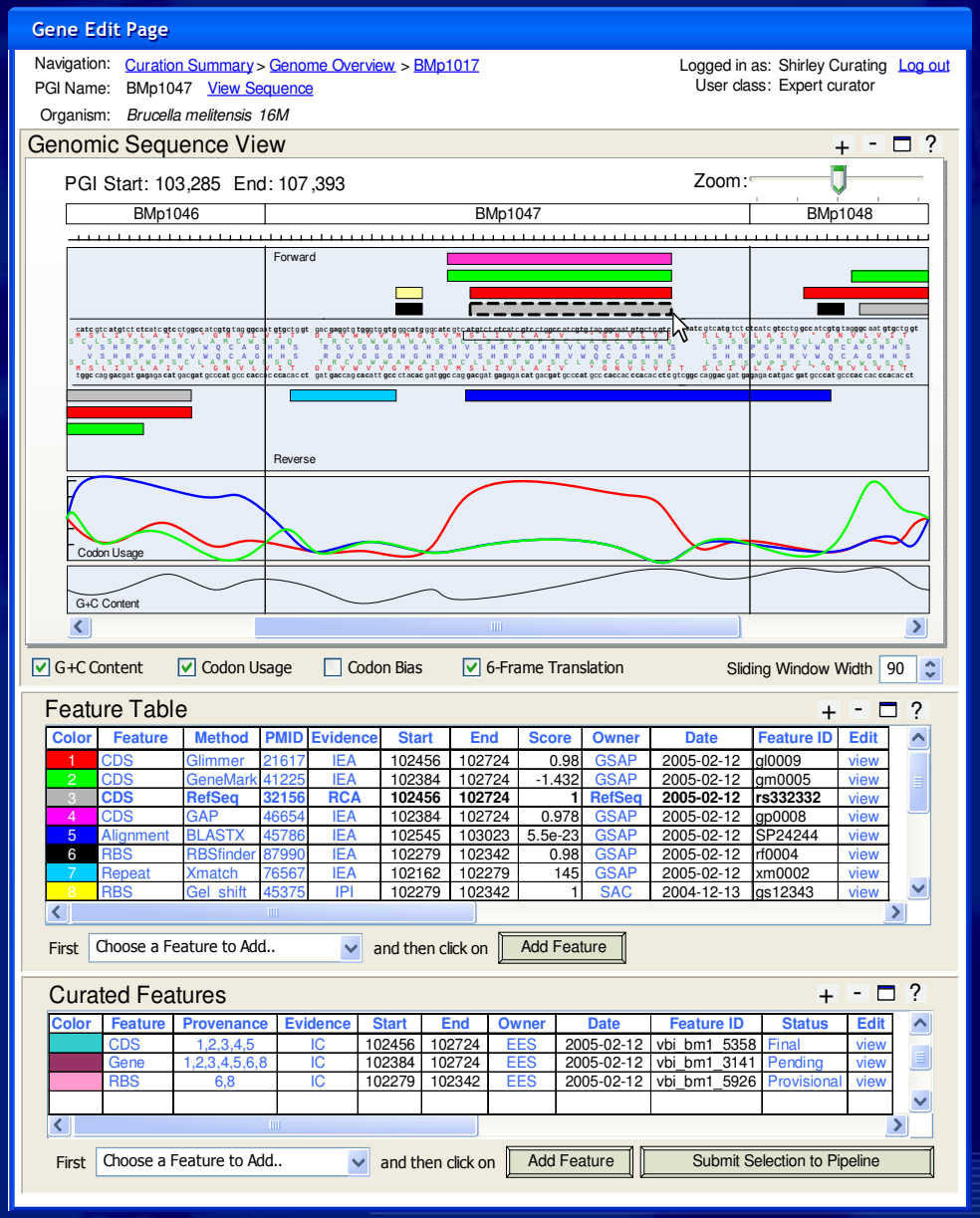
$L^z_{i,j}$ = score of interval predicted by method *z*

Traceback gives optimal PGI parsing and best prediction

PATRIC Curation System Schematic



PATRIC Curation System



Gene Edit Page

- DNA-level curation interface
- Display of annotation from pipeline (GSAP)
  - Graphical feature view
  - Feature table
- Feature Table

  Consists of primary data from:
  - Pipeline (gene prediction, *etc.*)
  - Programs run off-line
  - Literature
- Curated Features

  Based on the curator's interpretation of electronic annotation and experimental data from the literature
- Derivative sequences for downstream analysis
  - CDS translations (Proteins)
  - Ribosomal RNAs
  - tRNAs

PATRIC Curation System

Protein Edit Page

Protein Edit Page

Navigation: Curation Summary > Brucella melitensis Genome Overview > GeneID > ProteinID

LOG OUT

Protein ID: **BM0345**
Mol. Weight (kDa): **45.3**
pI: **5.6**
Start: **123** End: **345**
Length (aa): **345**

Organism: ***B. melitensis 16M***

Login: **Shurley Curating**
Class: **Supervisor**

Amino acid sequence
Nucleotide sequence

Protein Sequence View

Global View Display top 25 alignments Minimum E-value, 10e - 9 Zoom:

MKPLVDFSGVALLAKPDMSYRYGGAMALKELVEELGYAMPSWYPGKLVAMPQRSTVHGTRREVCYSTEEMPKLVEVVVKVDDMSYRYGGAMALKELVEEAARF

Kyte-Doolittle Hydrophobicity Plot Sliding Window Width 25 Codon Usage Table

Evidence

Features

Color	Method	ID	Feature	Evidence	Start	End	Owner	Date	Details
1	Pfam	204.9	domain: PF00204.9: DNA topoisomerase	IEA	343	450	PAP	5/2/2005	view
2		986.8	domain: PF00986.8: DNA gyrase	IEA	500	600	PAP	5/2/2005	view
3	SMART	433	signature: SM00433: DNA topoisomerase II	IEA	290	650	PAP	5/2/2005	view
4	SignalP	134789	Signal peptide: XYZ signal	IEA	1	12	PAP	5/2/2005	view
5	LipoP	765881	Signal peptide: ABC signal	IEA	1	34	PAP	5/2/2005	view
6	PSIPRED	463567	helix	IEA	1	23	PAP	5/2/2005	view
7			coil	IEA	24	55	PAP	5/2/2005	view
8			sheet	IEA	56	100	PAP	5/2/2005	view
9	TMHMM	346789	inner domain	IEA	1	23	PAP	5/2/2005	view
10			TM domain	IEA	24	55	PAP	5/2/2005	view
11			outer domain	IEA	56	199	PAP	5/2/2005	view
12			TM domain	IEA	200	250	PAP	5/2/2005	view
13			inner domain	IEA	251	279	PAP	5/2/2005	view
14	BLASTP	432115	DNA gyrase-like protein [Streptomyces coelic	IEA	343	450	PAP	5/2/2005	view
15			putative a subunit of topoisomerase IV [Strep	IEA	395	522	PAP	5/2/2005	view
16			DNA gyrase subunit A [Leifsonia xyli subs p. xy	IEA	405	545	PAP	5/2/2005	view
17			DNA gyrase subunit A [Propionibacterium aci	IEA	401	543	PAP	5/2/2005	view

Attributes

First Choose a Feature to Add.. then click Add

Ontology	ID	Attribute	Evidence	Method	Owner	Date	Details
COG	1329	transcriptional regulators, similar to M. xanthu	IEA	Pfam	PAP	5/2/2005	view
GO	3918	DNA topoisomerase activity	IEA	Pfam	PAP	5/2/2005	view
EC	5.99.1.3	DNA topoisomerase (ATP-hydrolyzing)	IEA	SMART	PAP	5/2/2005	view

Curated Features and Attributes

Color		Prov	Feature	Evidence	Start	End	Owner	Date	Edit
		1,4	DNA topoisomerase domain	IC	123	223	AP	6/20/2005	view
		3	DNA gyrase domain	IC	284	363	AP	6/20/2005	view
		5	signal peptide	IC	1	34	AP	6/20/2005	view

Attributes

Provenance	Ontology	ID	Attribute Name	Evidence	Status	Owner	Date	Edit
8	COG	1329	transcriptional regulators, similar to M. xanthu	IC	Final	AP	6/20/2005	view
9	GO	3918	DNA topoisomerase activity	IC	Pending	AP	6/20/2005	view
10	EC	5.99.1.3	DNA topoisomerase (ATP-hydrolyzing)	IC	Provisional	AP	6/20/2005	view

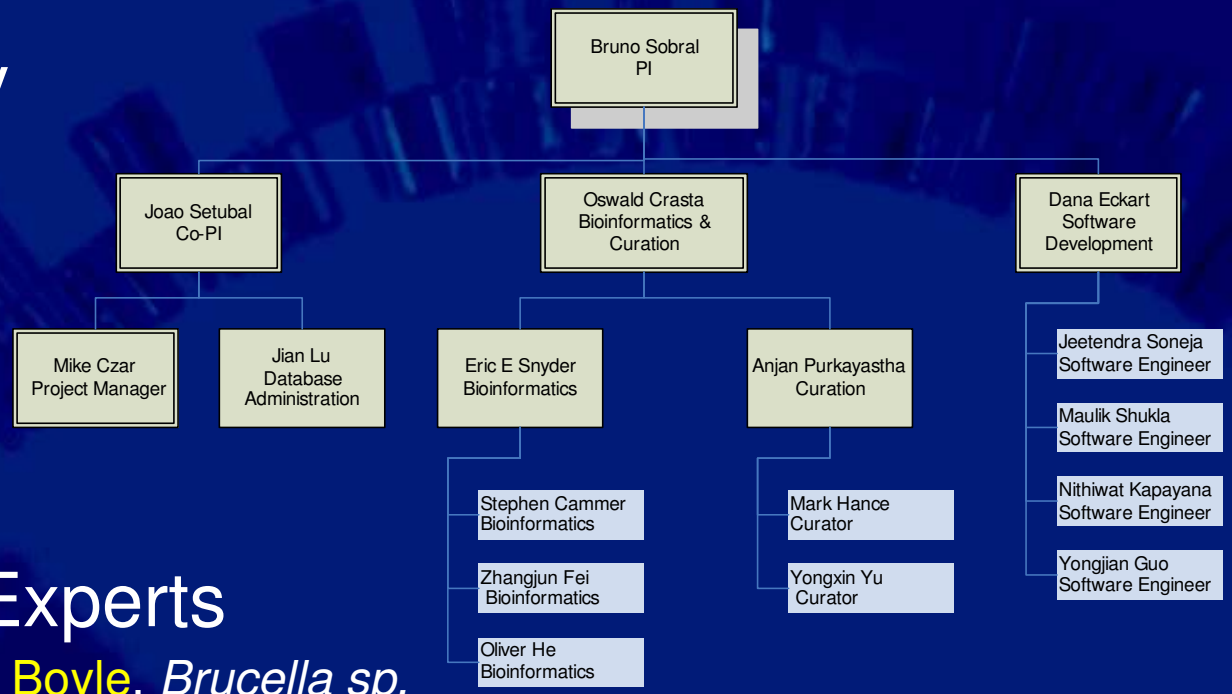
First Choose a Feature to Add.. then click Add Delete Submit to Database

- Protein-level curation interface
- Display of data from pipeline (PAP)
  - Graphical feature view
  - Feature & Attribute Tables
- Feature Table
  - Functional, structural domains
  - Alignments
- Attribute Table
  - Function, process, location
  - As suggested by sequence analysis or experimental evidence
- Curated Features & Attributes

  Based on the curator's interpretation of electronic annotation and experimental data from the literature
- Curated Attributes for downstream analysis:
  - Metabolic and signaling pathways
  - Comparative genomics

Acknowledgements

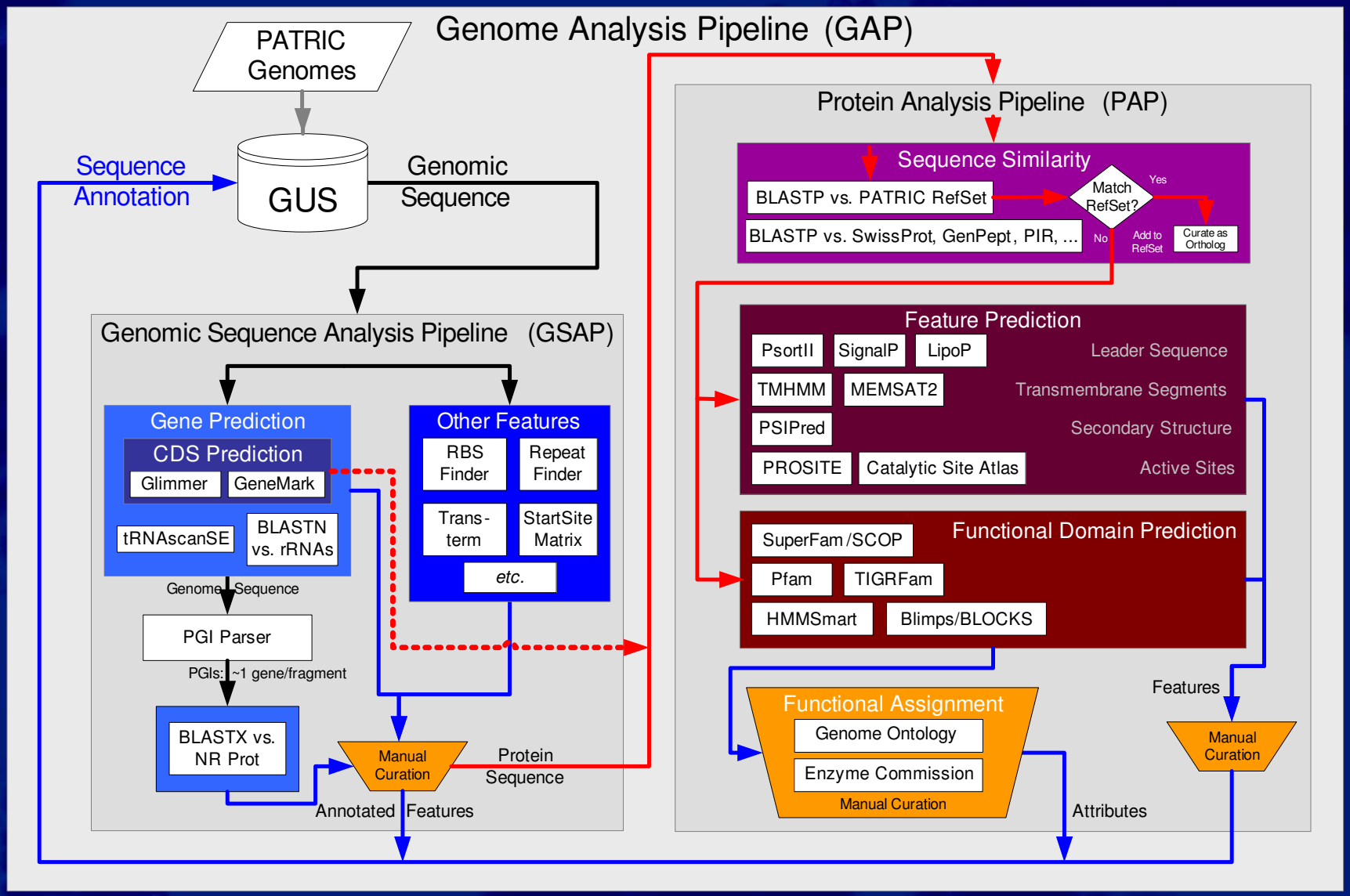
VBI Faculty



Organism Experts

Stephen Boyle, *Brucella sp.*
Abdu Azad, *Coxiella burnetii, Rickettsia sp.*
Susan Baker, Corona viruses
Xiang-Jin Meng, Hepatitis E viruses

NIH/NIAID Funds

Questions, comments, suggestions:
eesnyder@vbi.vt.edu

Curation Infrastructure