PseudoCAP

Community-assisted genome annotation: The *Pseudomonas* example

Geoff Winsor, Simon Fraser University

Burnaby (greater Vancouver), Canada

Overview

***Pseudomonas* Community Annotation Project (PseudoCAP)**

- Past and present

Annotation issues

***Pseudomonas* Genome Database Version 2**

- New ability to compare annotations within or between species
- New analyses/updates

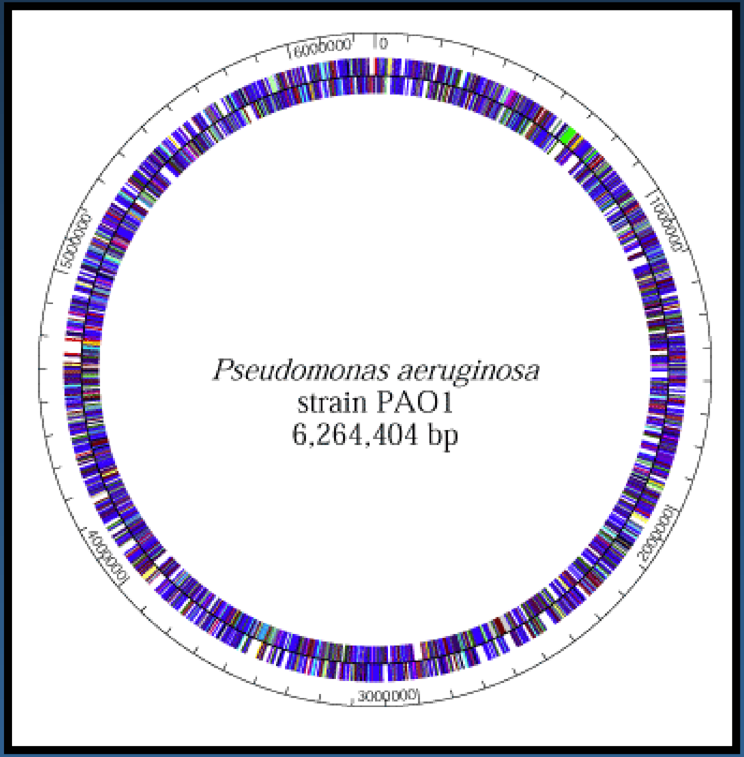
PseudoCAP: *Pseudomonas aeruginosa* Community Annotation Project

Goals

- Critical and conservative genome annotation
- Minimize project costs
- Capitalize on large *Pseudomonas aeruginosa* research community

Solution

- Community-aided and internet-based approach to continually updated, reviewed genome annotation



Winsor *et al.* (2005) *Nucleic Acids Res.* 33:D338-43.
Brinkman *et al.* (2000) *Nature* 406:933
Stover *et al.* (2000) *Nature* 406:959-964

PseudoCAP: Past and present

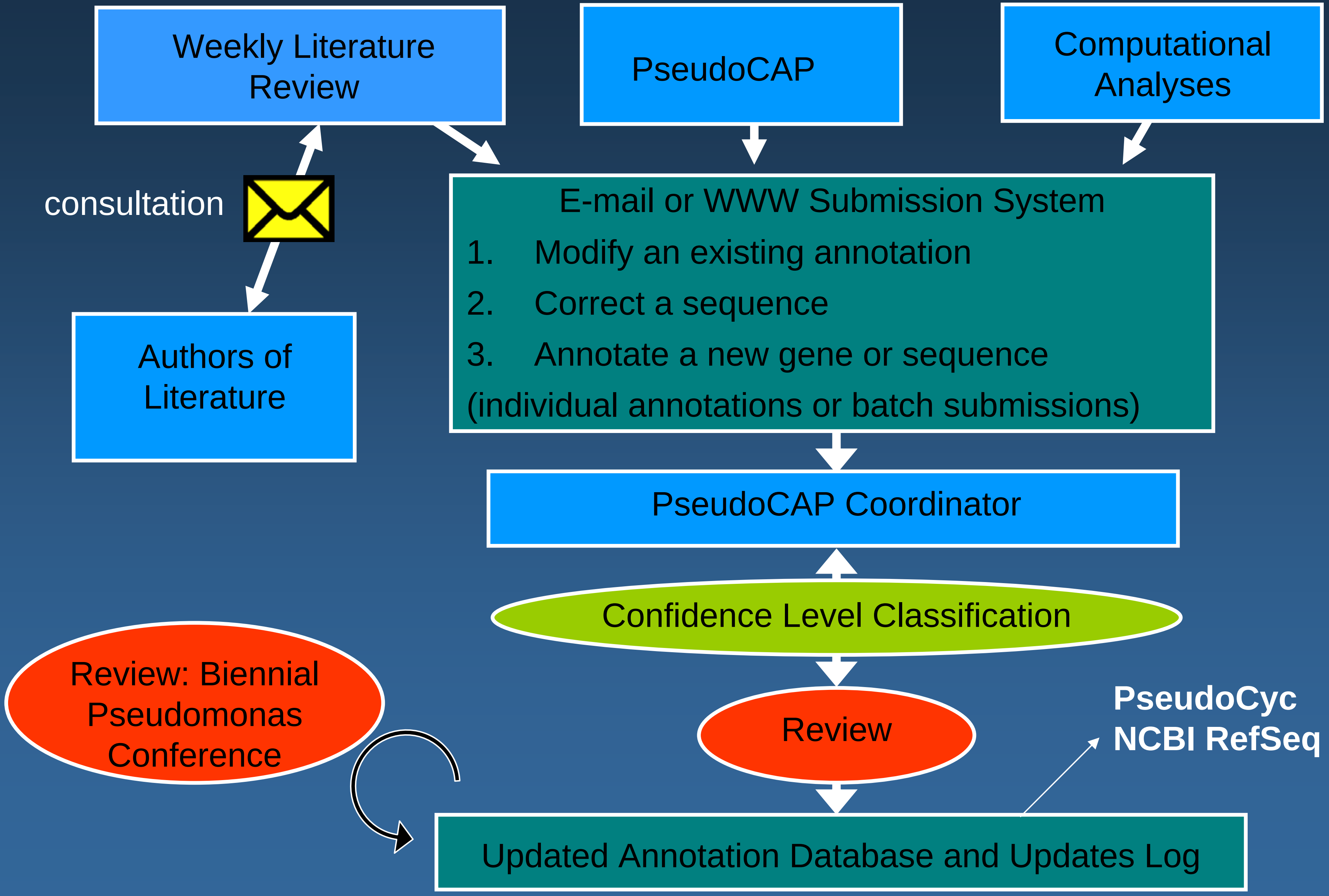
Initial PseudoCAP (1997 – 2000)

- 61 researchers, 1741 annotations
- Experts provide clarification on gene names
- Tables of data submitted by e-mail.
- Annotations incorporated by 3 annotators
- Increased initial annotation quality/consistency

Current PseudoCAP (2000 – present)

- 82 researchers, 1231 additional annotations (not including 24174 computationally-derived annotations)
- Submissions made using web-based forms or e-mail
- Annotations subject to review process
- Continually updated genome annotation

PseudoCAP approach to continually updated annotations



Weekly Literature Review

PseudoCAP

Computational Analyses

consultation

Authors of Literature

E-mail or WWW Submission System

1. Modify an existing annotation
2. Correct a sequence
3. Annotate a new gene or sequence

(individual annotations or batch submissions)

PseudoCAP Coordinator

Confidence Level Classification

Review

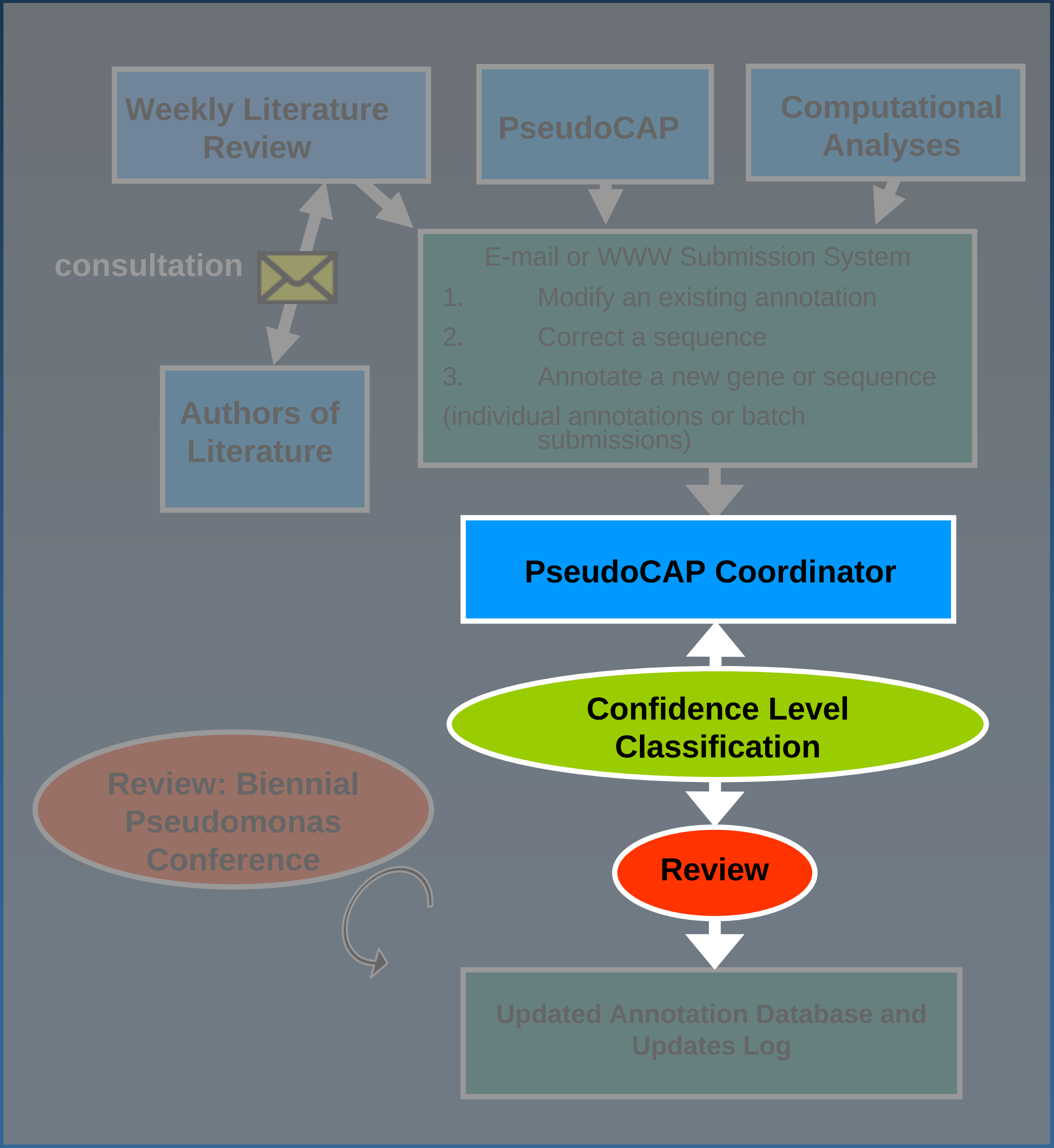
Review: Biennial Pseudomonas Conference

PseudoCyc
NCBI RefSeq

Updated Annotation Database and Updates Log

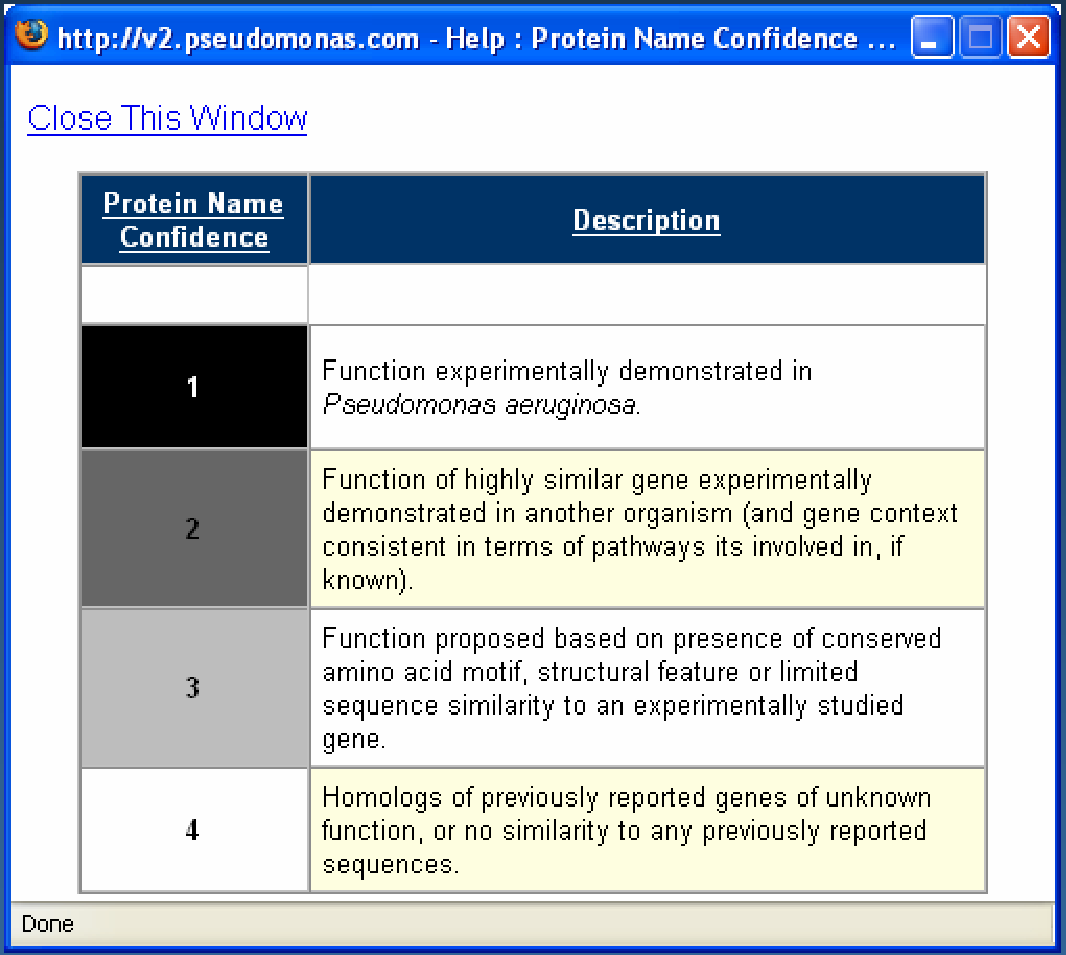
Annotation submission review process

- Coordinator examines initial submission
  - Looks for reference and contact info for corresponding author
- Responds with requests for any additional information and clarification, if required.
- Entry reviewed by additional reviewer from the research community, if required.
- Coordinator assigns product name confidence classification


Weekly Literature Review
PseudoCAP
Computational Analyses
consultation
Authors of Literature
E-mail or WWW Submission System
1. Modify an existing annotation
2. Correct a sequence
3. Annotate a new gene or sequence
(individual annotations or batch submissions)
PseudoCAP Coordinator
Confidence Level Classification
Review: Biennial Pseudomonas Conference
Review
Updated Annotation Database and Updates Log

Product confidence level classification system

- Reflects the type of evidence upon which the gene/protein name was based.
- Similar confidence classification used for other data
- TIGR has requested our Class 1 annotations



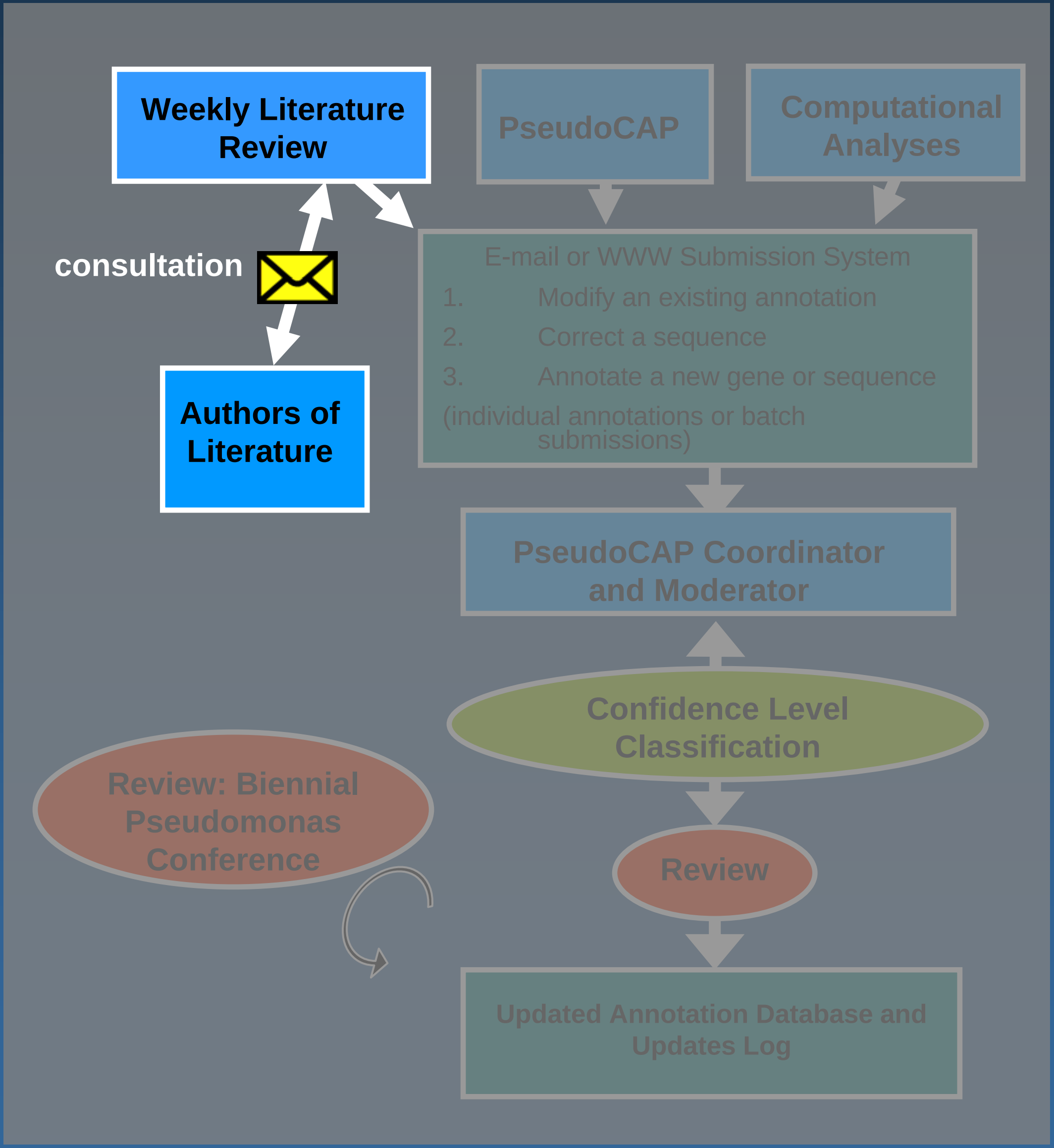
http://v2.pseudomonas.com - Help : Protein Name Confidence ...

Close This Window

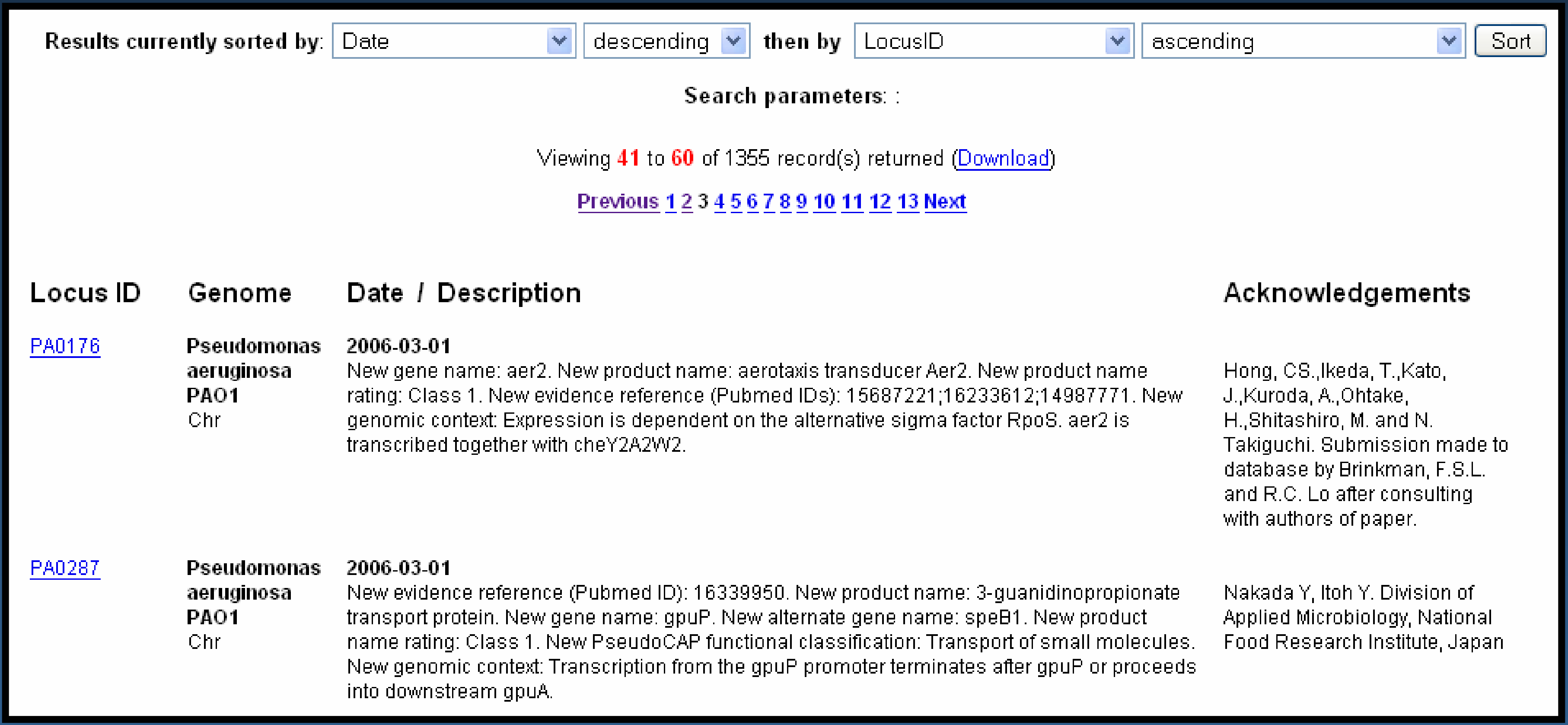
Protein Name Confidence	Description
1	Function experimentally demonstrated in *Pseudomonas aeruginosa*.
2	Function of highly similar gene experimentally demonstrated in another organism (and gene context consistent in terms of pathways its involved in, if known).
3	Function proposed based on presence of conserved amino acid motif, structural feature or limited sequence similarity to an experimentally studied gene.
4	Homologs of previously reported genes of unknown function, or no similarity to any previously reported sequences.

Literature review process

- Review literature once per week in order to keep number of submissions from growing too large
- Consult with author of paper to make sure they agree with submission
    - Not frequently subject to additional review because it has already been peer-reviewed


Weekly Literature Review
PseudoCAP
Computational Analyses
consultation
E-mail or WWW Submission System
1. Modify an existing annotation
2. Correct a sequence
3. Annotate a new gene or sequence
(individual annotations or batch submissions)
Authors of Literature
PseudoCAP Coordinator and Moderator
Confidence Level Classification
Review: Biennial Pseudomonas Conference
Review
Updated Annotation Database and Updates Log

Updated annotations are recorded in an updates log



Results currently sorted by: Date descending then by LocusID ascending

Search parameters: :

Viewing 41 to 60 of 1355 record(s) returned (Download)

Previous 1 2 3 4 5 6 7 8 9 10 11 12 13 Next

Locus ID	Genome	Date / Description	Acknowledgements
PA0176	**Pseudomonas aeruginosa PAO1** Chr	**2006-03-01** New gene name: aer2. New product name: aerotaxis transducer Aer2. New product name rating: Class 1. New evidence reference (Pubmed IDs): 15687221;16233612;14987771. New genomic context: Expression is dependent on the alternative sigma factor RpoS. aer2 is transcribed together with cheY2A2W2.	Hong, CS.,Ikeda, T.,Kato, J.,Kuroda, A.,Ohtake, H.,Shitashiro, M. and N. Takiguchi. Submission made to database by Brinkman, F.S.L. and R.C. Lo after consulting with authors of paper.
PA0287	**Pseudomonas aeruginosa PAO1** Chr	**2006-03-01** New evidence reference (Pubmed ID): 16339950. New product name: 3-guanidinopropionate transport protein. New gene name: gpuP. New alternate gene name: speB1. New product name rating: Class 1. New PseudoCAP functional classification: Transport of small molecules. New genomic context: Transcription from the gpuP promoter terminates after gpuP or proceeds into downstream gpuA.	Nakada Y, Itoh Y. Division of Applied Microbiology, National Food Research Institute, Japan

- Individual gene pages link to all entries in log for that gene and protein
- Boolean search page allows search by locus ID, participant, update description and date updated
- Sort results and download to text file
- Updates log made available for review at *Pseudomonas* conference

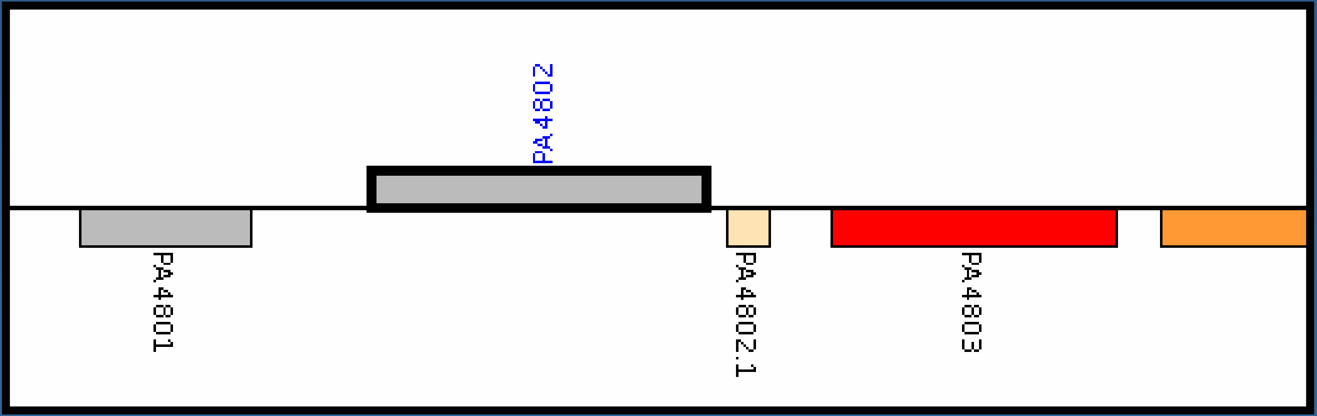
Annotation issues

Annotation issues encountered

- **Unpublished data submitted**
  - Review by at least one additional member of the research community
  - If gene and protein names are offered, recorded as alternate names until published or other equivalent consensus is reached.

- **Addition of new genes** (e.g. recently added rRNA, tRNA genes)
  - Decimal numbering system

    Example, PA4802.1 would be used to indicate the gene PA number for a new gene identified between PA4802 and PA4803. PA4802.01 would be used to identify a new gene between PA4802 and PA4802.1.



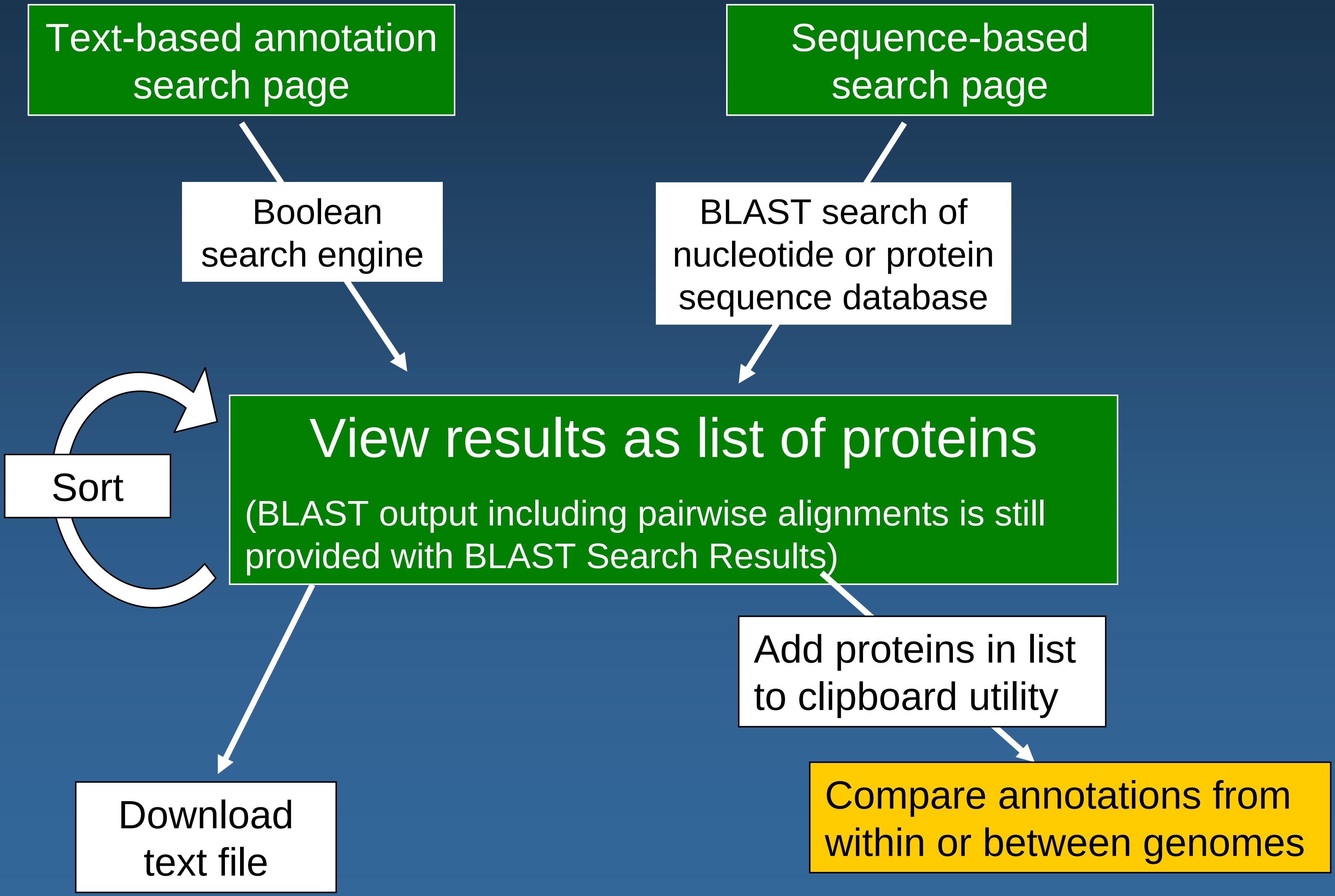
Annotation issues encountered

- **More than one gene with same name OR one gene with multiple proposed primary names**
  - Conflicts should be resolved within reasonable time frame to avoid confusion in the literature
  - All researchers involved in the conflict should be involved in the resolution
  - In absence of researcher consensus, the more predominant name used in the literature and by research groups is favored
  - In the absence of literature and researcher consensus, the first published name in the literature will be given priority
  - Alternate names are still recorded under the “alt gene” name field
  - Biennial *Pseudomonas* conference review

Pseudomonas Genome Database Version 2

New ability to compare annotations within or between species

Search annotations associated with any *Pseudomonas* genome made publicly available



Text-based annotation search page

Sequence-based search page

Boolean search engine

BLAST search of nucleotide or protein sequence database

Sort

View results as list of proteins

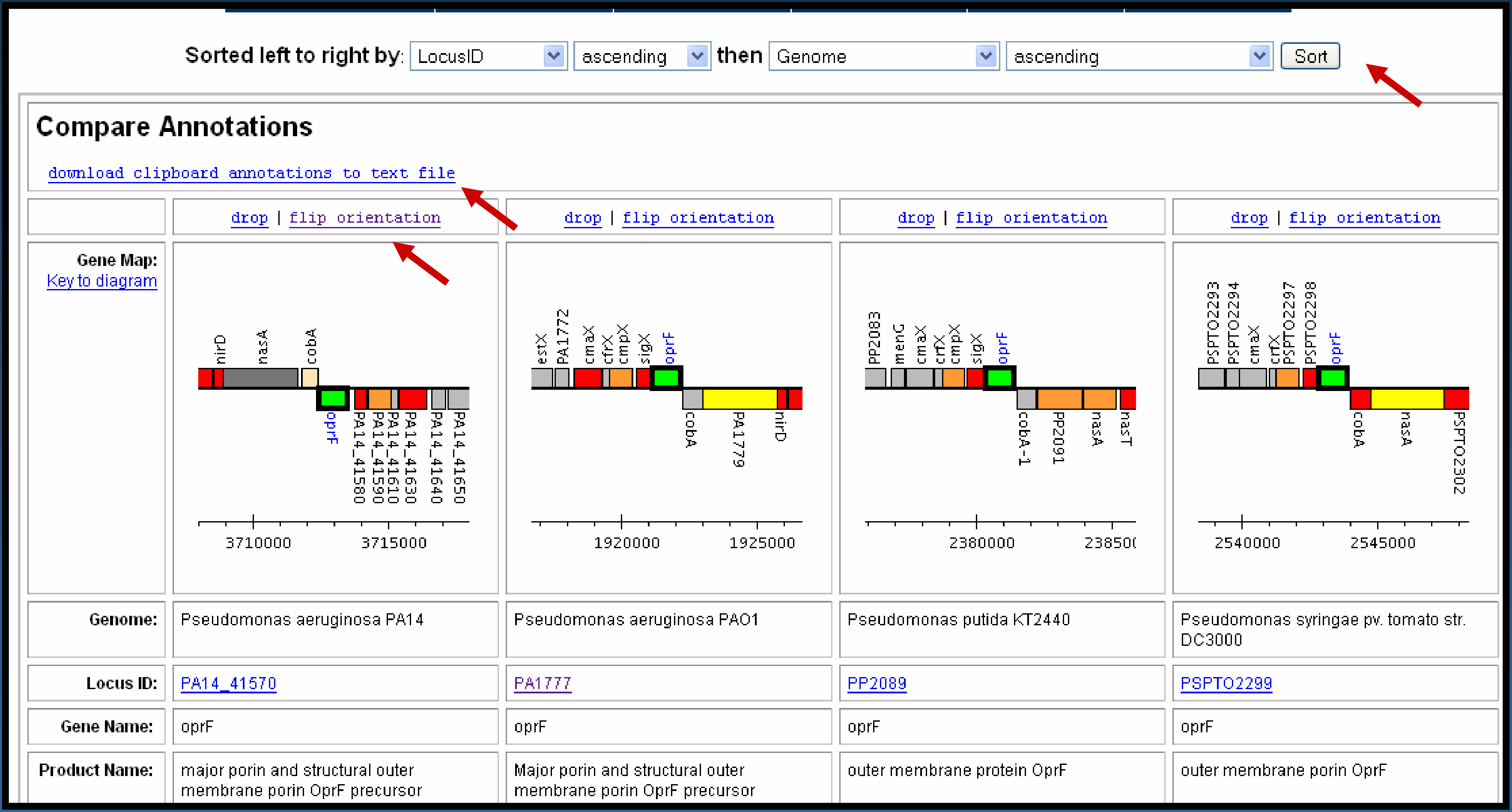
(BLAST output including pairwise alignments is still provided with BLAST Search Results)

Add proteins in list to clipboard utility

Download text file

Compare annotations from within or between genomes

View and compare annotations within a *Pseudomonas* species or between species.



Sorted left to right by: LocusID ascending then Genome ascending Sort

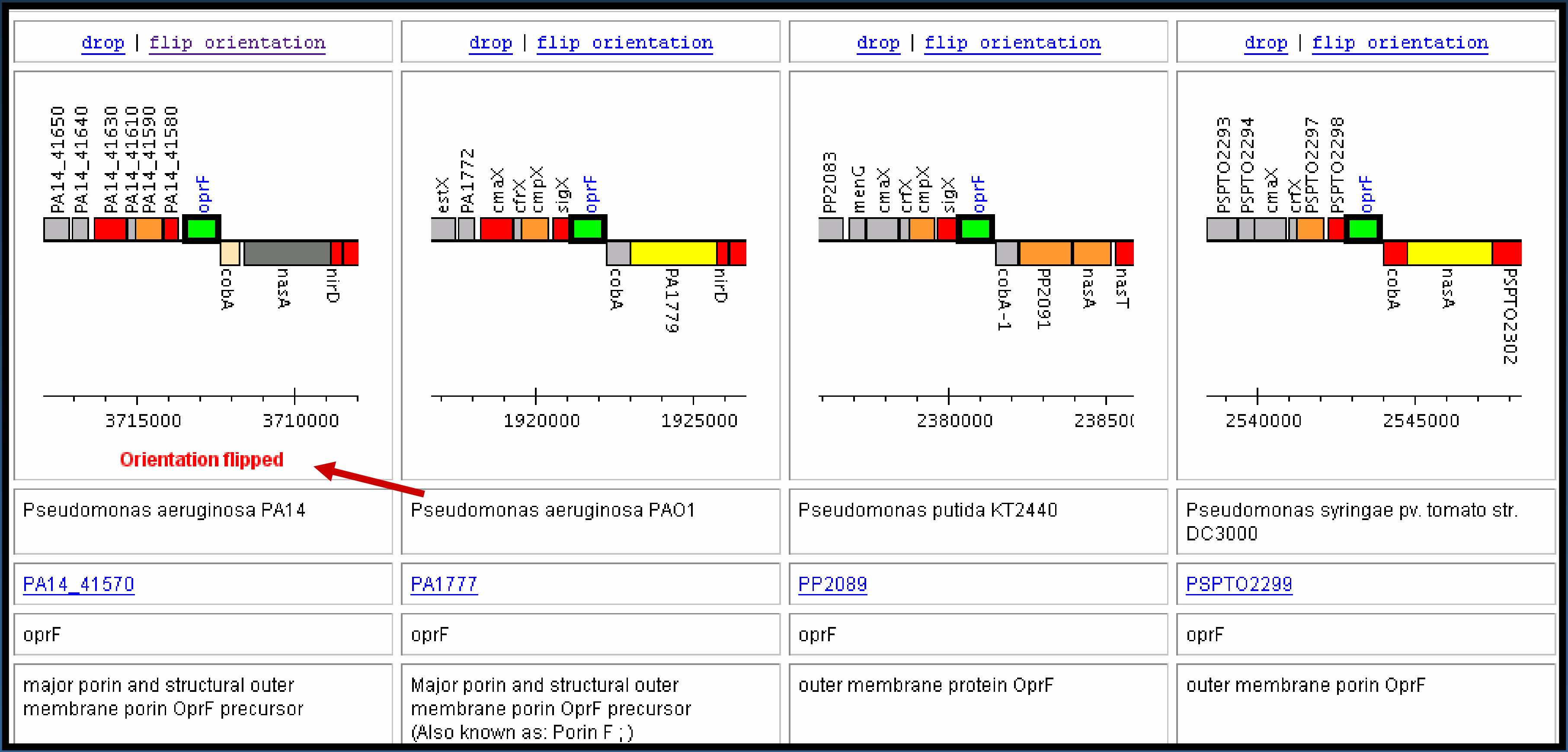
Compare Annotations

download clipboard annotations to text file

	drop | flip orientation	drop | flip orientation	drop | flip orientation	drop | flip orientation
Gene Map: Key to diagram				
Genome:	Pseudomonas aeruginosa PA14	Pseudomonas aeruginosa PAO1	Pseudomonas putida KT2440	Pseudomonas syringae pv. tomato str. DC3000
Locus ID:	PA14_41570	PA1777	PP2089	PSPTO2299
Gene Name:	oprF	oprF	oprF	oprF
Product Name:	major porin and structural outer membrane porin OprF precursor	Major porin and structural outer membrane porin OprF precursor	outer membrane protein OprF	outer membrane porin OprF

- Get to this point using text- or sequence-based search
- Sort results, download clipboard annotations to text file, flip genome orientation

View and compare annotations within a *Pseudomonas* species or between species.



Orientation flipped

drop | flip orientation	drop | flip orientation	drop | flip orientation	drop | flip orientation
Pseudomonas aeruginosa PA14	Pseudomonas aeruginosa PAO1	Pseudomonas putida KT2440	Pseudomonas syringae pv. tomato str. DC3000
PA14_41570	PA1777	PP2089	PSPTO2299
oprF	oprF	oprF	oprF
major porin and structural outer membrane porin OprF precursor	Major porin and structural outer membrane porin OprF precursor (Also known as: Porin F ;)	outer membrane protein OprF	outer membrane porin OprF

- Perform pre-formatted BLAST searches, multiple sequence alignment (ClustalW)
- Go to individual gene pages for links to:

GBrowse, PseudoCyc, KEGG, TIGR, updates log

New analyses and updates

New and updated analyses from Brinkman Lab

PSORTb…very accurate protein location prediction

- Multi-component (or module) approach to localization prediction

Modules included with PSORTb:

Signal peptides: Non-cytoplasmic
- HMM

Amino acid composition/patterns: All localizations
- Support Vector Machine’s trained with frequent subsequences

Transmembrane helices: Cytoplasmic membrane
- HMMTOP

PROSITE motifs with 100% precision: All localizations

Outer membrane motifs: Outer membrane
- Identified by association-rule mining

Homology to proteins of experimentally known localization: All localizations
- “SCL-BLAST” against database of proteins of known localization

Integration with a Baysian Network

96% precision
82% recall

New and updated analyses from Brinkman Lab

PSORTb…very accurate protein location prediction

96% precision
82% recall

•Most precise subcellular localization prediction method available

•First computer-based method that exceeded the accuracy of high-throughput laboratory methods (and more than 500x faster)

• New version in development (improved recall)

http://www.psort.org/psortb

Gardy *et al.* (2005) *Bioinformatics 21(5):617-623*

Rey et al. (2005). BMC Genomics 6:162.

New and updated analyses from Brinkman Lab

Ortholuge

Improving the specificity of high-throughput ortholog prediction

Improved (more precise) ortholog predictions using Ortholuge

- http://www.pathogenomics.ca/ortholuge
- Fulton *et al.* (2006) *BMC Bioinformatics* 7:270
- High throughput method for evaluating ortholog predictions
- Examines phylogenetic distance ratios between two comparison species and an outgroup species
- Identifies predicted orthologs undergoing unusual rates of divergence
- Assigns ortholog predictions as ‘probable orthologs’, ‘uncertain’ or ‘probable paralogs’
- Notable number of orthologs predicted by reciprocal best-BLAST-hit analysis are likely false positives (i.e. are paralogs)
  - Bacterial genome dataset (~5%), eukaryotic genome dataset (~10%)

New and updated analyses

PseudoCyc

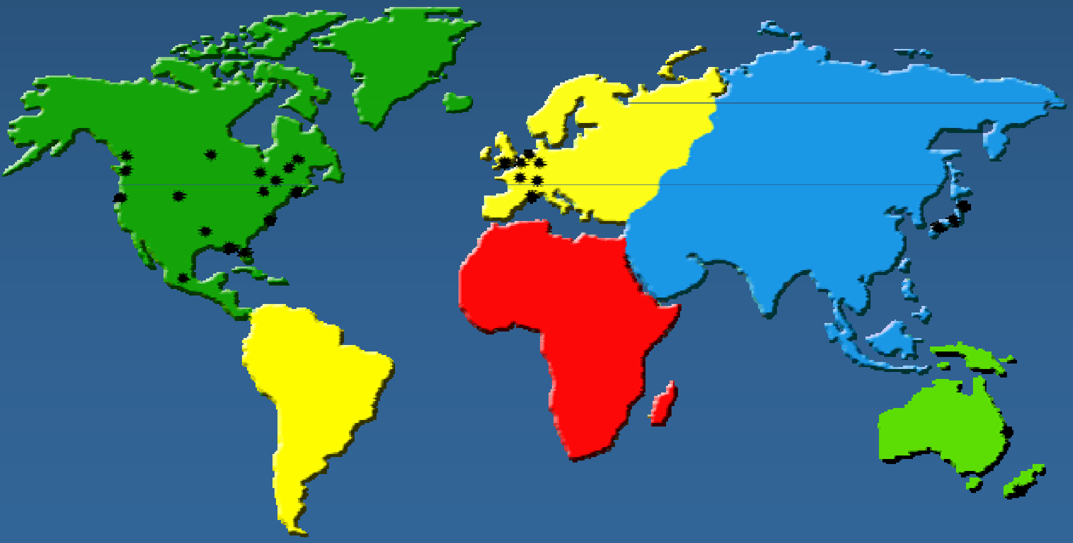
- Romero and Karp (2003) *J Mol Microbiol Biotechnol*. 5(4):230-239.
- Pathway/Genome database for *Pseudomonas aeruginosa* PAO1 developed by SRI International
- Now maintained by PseudoCAP
- Using Pathway Tools version 9.5

Updates

- Gene, product names updated to include latest annotation submissions
- Genomic DNA sequence and base pair coordinates updated to accommodate nucleotide insertion at position 2669175
- Primary keys in database updated

Acknowledgments

- Fiona SL Brinkman
- Robert EW Hancock
- Raymond Lo, Brinkman Lab
- SRI International – Pathway Tools Support
- Cystic Fibrosis Foundation Therapeutics Inc.
- PseudoCAP participants



- v2.pseudomonas.com
- www.pseudomonas.com
- All web site source code freely available under GNU public license