# TOWARDS A CONSENSUS ANNOTATION SYSTEM.

**Sequencing Finishing Analysis and the Future**

**May 2009**

**Owen White**
**Institute for Genome Science**

# Questions

- What if annotation generation can be easily out-sourced?
- How would multiple centers rationally contribute annotation?
- What is the role of the increasing number of closely related species?

# Annotation systems

- IGS: Annotation Engine
- JCVI: Annotation Service
- JCVI: Genome Properties
- Victor Markowitz, JGI: IMG
- Folker Myer, Argonne: RAST
- Swiss-Prot: HAMAP rules
- Genoscope: Microscope
- Ensembl?

# Probing genes with high quality HMMs

- TIGRFam HMM
  - Highly accurate HMM rigorously assigns function.
- Carries assertion datatypes:
  - Functional name
  - E.C. Number
  - Genetic Name
  - GO assignment
  - Literature info.

# Annotation Evaluation

TIGRFam HMM

Functional names
GO assignments
Genetic names
EC numbers

Did they assign them all?
Did they assign them consistently?

# Data Volume

| Source Data | Organisms | Genes | Tested | Yield |
|---|---|---|---|---|
| BHB | 27 | 98,896 | 12,973 | 13.12% |
| ERIC | 96 | 409,242 | 83,216 | 20.33% |
| NMPDR | 120 | 387,407 | 42,963 | 11.09% |
| Pathema | 72 | 366,928 | 48,225 | 13.14% |
| Patric | 25 | 53,664 | 9,534 | 17.77% |
| SwissProt | 586 | 390,696 | 118,443 | 30.32% |
| HAMAP | 585 | 188,779 | 112,636 | 59.67% |
| IMG | 950 | 3,197,329 | 493,385 | 15.43% |
| RefSeq | 718 | 2,398,558 | 143,052 | 5.96% |
| Genbank | 729 | 2,461,596 | 401,323 | 16.30% |
| Subsystems | 659 | 797,078 | 257,766 | 32.34% |
| CMR | 403 | 1,114,859 | 196,659 | 17.64% |
| KEGG | 494 | 1,697,018 | 269,874 | 15.90% |

Did they assign them consistently?

# Consistency

The frequency of genes that have a identical product name.

$$\frac{\sum_{f}\left[n_f * \sum_{s}\frac{\binom{m_s}{2}}{\binom{n_f}{2}}\right]}{\sum_{f} n_f}$$

where

$n_f$ = the number of genes in a TIGRFAM *f*

$m_s$ = the number of genes sharing a string name *s*

# Consistency: Site4/hisA

**Campylobacter**
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)

**Listeria**
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)

**Staphylococcus**
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)

**Vibrio**
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)
5.3.1.16 Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16)

# Site5/hisA

**Bacillus anthracis**
5.3.1.16 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase **x 7**
NO_EC phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase

**Burkholderia mallei**
5.3.1.16 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase **x 4**

**Burkholderia pseudomallei**
5.3.1.16 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase **x 8**
5.3.1.16 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide isomerase

**Clostridium botulinum**
5.3.1.16 phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase

```
ACCESSION   ZP_00234783
VERSION     ZP_00234783.1  GI:47097220
DBSOURCE    REFSEQ: accession NZ_AADQ01000053.1
KEYWORDS    .
SOURCE      Listeria monocytogenes str. 1/2a F6854
  ORGANISM  Listeria monocytogenes str. 1/2a F6854
            Bacteria; Firmicutes; Bacillales; Listeriaceae; Listeria.
REFERENCE   1  (residues 1 to 56)
  AUTHORS   Nelson,K.E., Fouts,D.E., Mongodin,E.F., Ravel,J., DeBoy,R.T.,
            Kolonay,J.F., Rasko,D.A., Angiuoli,S.V., Gill,S.R., Paulsen,I.T.,
            Peterson,J., White,O., Nelson,W.C., Nierman,W., Beanan,M.J.,
            Brinkac,L.M., Daugherty,S.C., Dodson,R.J., Durkin,A.S., Madupu,R.,
            Haft,D.H., Selengut,J., Van Aken,S., Khouri,H., Fedorova,N.,
            Forberger,H., Tran,B., Kathariou,S., Wonderling,L.D., Uhlich,G.A.,
            Bayles,D.O., Luchansky,J.B. and Fraser,C.M.
  TITLE     Whole genome comparisons of serotype 4b and 1/2a strains of the
            food-borne pathogen Listeria monocytogenes reveal new insights into
            the core genome components of this species
```

/product="hypothetical"

```
            Fedorova,N., Forberger,H., Tran,B. and Fraser,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-APR-2004) The Institute for Genomic Research, 9712
            Medical Center Dr., Rockville, MD 20850, USA
COMMENT     PREDICTED REFSEQ: This record has not been reviewed and the
            function is unknown. The reference sequence was derived from
            EAL05376.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..56
                     /organism="Listeria monocytogenes str. 1/2a F6854"
                     /strain="1/2a F6854"
                     /db_xref="taxon:267409"
```

/product="lacitehtopyh"

Did they assign them all?

# Completeness

For all genes that *could* receive an assertion:

the percent of genes that *did* get an assertion.

# 1,061 TIGRFams-GO, Counts

| | Possible | Assigned |
|---|---|---|
| Site1 | 3,147 | 896 |
| Site2 | 26,293 | 761 |
| Site3 | 11,964 | 0 |
| Site4 | 5,064 | 0 |
| Site5 | 19,844 | 15,330 |
| Total | 66,312 | 16,987 |

# 736 TIGRFams-ECs, Counts

| | Possible | Assigned |
|---|---|---|
| Site1 | 2,064 | 0 |
| Site2 | 17,674 | 3,494 |
| Site3 | 7,415 | 6,938 |
| Site4 | 2,856 | 716 |
| Site5 | 11,174 | 9,787 |
| total | 41,183 | 20,935 |

# hisA

| | # | Gene Name | GO | EC# |
|---|---|---|---|---|
| Site2 | 36 | 19 | 1 | 7 |
| Site3 | 14 | | | 14 |
| Site4 | 4 | 1 | | 1 |
| Site5 | 21 | 1 | 18 | 20 |

All genetic names were identical.
All EC numbers were identical.

# Are spotty results bad?

## hisA

| | # | Gene Name | GO | EC# |
|---|---|---|---|---|
| Site2 | 36 | 19 | 1 | 7 |
| Site3 | 14 | | | 14 |
| Site4 | 4 | 1 | | 1 |
| Site5 | 21 | 1 | 18 | 20 |

All genetic names were identical.
All EC numbers were identical.

# GO Assignments

| | Common name | | GO assignments | | | | Gene symbol | |
|---|---|---|---|---|---|---|---|---|
| | | | Source | | Completeness-Consistency | | | |
| Source Data | Completeness | Consistency | Completeness | Consistency | Completeness | Consistency | Completeness | Consistency |
| BHB | [illegible] | 45.14% | 86.89% | 51.74% | [illegible] | [illegible] | 81.59% | 75.54% |
| ERIC | [illegible] | 38.51% | 81.10% | 86.89% | [illegible] | 50.11% | 32.37% | 54.03% |
| NMPDR | [illegible] | 93.26% | 51.74% | 81.39% | [illegible] | [illegible] | 33.06% | 47.68% |
| Pathema | [illegible] | [illegible] | 90.10% | 97.90% | [illegible] | [illegible] | 84.88% | 79.29% |
| Patric | [illegible] | 90.79% | 81.39% | 90.10% | [illegible] | [illegible] | 90.06% | 91.92% |
| SwissProt | 100% | 90.80% | 97.90% | 90.10% | [illegible] | [illegible] | 98.83% | 91.66% |
| HAMAP | 100% | 92 | | | 59.08% | 75.03% | 99.11% | 92.73% |
| IMG | 92 | | | | 38.72% | 53.03% | 0.00% | NA |
| RefSeq | 97 | 51 | | | 40.2 | 50.11% | 54.21% | 50.25% |
| Genbank | 95 | 35. | | | 34.21% | 48.48% | 48.84% | 43.61% |
| Subsystems | 99 | 93.15% | 0.00% | NA | 64.71% | NA | 0.00% | NA |
| CMR | 99 | 75.34% | 77.99% | 0.00% | 56.23% | NA | 84.82% | 78.50% |
| KEGG | 97 | 37.83% | 0.00% | 0.00% | 42.95% | NA | 58.82% | 51.95% |

Sort

| Source Data |
|---|
| BHB |
| ERIC |
| NMPDR |
| Pathema |
| Patric |
| SwissProt |
| HAMAP |
| IMG |
| RefSeq |
| Genbank |
| Subsystems |
| CMR |
| KEGG |

# Choosing best of breed annotations

| | Quality | G1 | G2 | G3 | G4 | G5 | G6 | ....$G_N$ |
|---|---|---|---|---|---|---|---|---|
| Source 1 | Best | | | | | | | |
| Source 2 | Good | | | | | | | |
| Source 3 | Okay | | | | | | | |
| Source 4 | Bites | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Final | | | | | | | | |

# Assertion types

- Common name
- EC number
- Genetic name
- GO
  - Function
  - Process
  - Cellular component

# Assertion types

- Common name
- EC number
- Genetic name
- GO
  - Function
  - Process
  - Cellular component
- Mutant phenotype
- Molecular interaction
- Regulation

# Choosing best assertion

| Description | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Best | |
| Source 2 | Good | |
| Source 3 | Okay | |
| Source 4 | Bites | |

| EC Number | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Bites | |
| Source 2 | Good | |
| Source 3 | Best | |
| Source 4 | Okay | |

| GO Assignment | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Best | |
| Source 2 | Okay | |
| Source 3 | Good | |
| Source 4 | Bites | |

| Gene 1 - Final | | | |
|---|---|---|---|
| | Data | Quality | |
| Source 1 | Desc | Best | |
| Source 3 | EC# | Best | |
| Source 1 | GO | Best | |

# Refinement of annotation data

- Bioinformatics resource centers
  - Sponsor: NIAID
  - 8 Sites
  - Annotation split across many centers
  - Requirement: tight interoperation
- Approach:
  - Assign assertions
  - Describe assertions with evidence codes

# EV codes

- ISS – Curated from sequence similarity
- EXP - Inferred from experiment
- LIT - Literature
- IEA - Electronic annotation
- ICE - Inferred from genomic context
- ICL - Inf. from presence in cluster
- ISR - Inf. from system reconstruction

Thank you: GO consortium

# Volume of Assertions

| BRC | Function | Process | Cell. component |
|---|---|---|---|
| Site 1 | 17,249 | 12,924 | 8,709 |
| Site 2 | 32,608 | 27,760 | 7,176 |
| Site 3 | 220,056 | 172,690 | 206,286 |
| Site 4 | 240,470 | | |
| Site 5 | 314,304 | 310,530 | 124,157 |
| Site 6 | 81,387 | 13,733 | 5,380 |
| Site 7 | 29,220 | | 29,689 |
| **total** | 935,294 | 537,637 | 381,397 |

Sites: ApiDB, BHB, ERIC, NMPDR, PATRIC, Pathema, VBRC

# Volume of evidence

| BRC | Genes* | Rows |
|---|---|---|
| Site 1 | 19,470 | 79,458 |
| Site 2 | 35,584 | 86,696 |
| Site 3 | 419,682 | 1,476,691 |
| Site 4 | 240,470 | 528,515 |
| Site 5 | 363,979 | 997,537 |
| Site 6 | 67,654 | 119,110 |
| Site 7 | 59,387 | 924,473 |
| **total** | 1,206,226 | 4,212,480 |

* Genes – of those supplied in ev-code files

# Evidence Abundance

| | Site1 | Site2 | Site3 | Site4 | Site5 | Site6 | Site7 |
|---|---|---|---|---|---|---|---|
| **EXP** | 236 | | 4,404 | 334 | 211 | | |
| **LIT** | 1,316 | | 12,162 | 49,472 | | | 180 |
| **ICL** | | | | 57,896 | | | |
| **IEA** | 19,272 | 2,205 | 372,197 | 224,424 | 288,854 | 21,979 | |
| **ISR** | | | | 114,316 | 513 | | |
| **ISS** | 1,296 | 29,527 | | 35,285 | 31,732 | 41,902 | 57,409 |

EXP –Inferred from experiment
LIT – Literature
ICL – Inf. from presence in cluster
IEA – Inferred from electronic annotation
ISR – Inf. from system reconstruction
ISS – Inferred from sequence similarity

# Previous: choose best assertion

| Description | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Best | |
| Source 2 | Good | |
| Source 3 | Okay | |
| Source 4 | Bites | |

| EC Number | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Bites | |
| Source 2 | Good | |
| Source 3 | Best | |
| Source 4 | Okay | |

| GO Assignment | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Best | |
| Source 2 | Okay | |
| Source 3 | Good | |
| Source 4 | Bites | |

| Gene 1 - Final | | | |
|---|---|---|---|
| | Data | Quality | |
| Source 1 | Desc | Best | |
| Source 3 | EC# | Best | |
| Source 1 | GO | Best | |

# Now: Choose best assertion.evcode

| GO Assignment.ISS | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Bites | |
| Source 2 | Good | |
| Source 3 | Best | |
| Source 4 | Okay | |

| GO Assignment.ICL | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | NA | |
| Source 2 | Best | |
| Source 3 | NA | |
| Source 4 | NA | |

| GO Assignment.IEA | | |
|---|---|---|
| | Quality | G1 |
| Source 1 | Best | |
| Source 2 | Good | |
| Source 3 | Okay | |
| Source 4 | Bites | |

| Gene 1 - Final | | | |
|---|---|---|---|
| | Data | Quality | |
| Source 3 | ISS | Best | |
| Source 2 | ICL | Best | |
| Source 1 | IEA | Best | |

# Datatype Saturation:Summary

- Rich data types can be combined, to:
  - Improve annotation
  - Present an audit trail for users
  - Create cooperative (v. competitive) model of ann.
  - Aid addition of old annotation on top of new
  - Make exchange of data possible.
- Future: More complex methods could be used to combine data.

# Rational Approach

- Minimal data types
- Assertion types, Evidence codes
- Perform on-going evaluation of sources
- SOPs
- Annotation Services and clouds

# Minimal data types

- Defines what we guarantee
  - Is passed on to GenBank
  - Supported by all validation tools
  - Often evaluated by QC tools
  - The central web site displays
- Simplify conversion to ontologies
- Incremental improvement is based on
  - Usability studies
  - Committees, e.g. I/HMP Research Network
- For example: EC numbers, GO assignments, genetic names, pathways

# Increase in Assertion types

- Common name
- EC number
- Genetic name
- GO
  - Function
  - Process
  - Cellular component
- Mutant phenotype
- Molecular interaction
- Regulation
- Expression
- Phylogenetic assertion
- Pathway reconstruction
- Network analysis

# On-going evaluation of:


- IGS: Annotation Engine
- JCVI: Annotation Service
- JCVI: Genome Properties
- Victor Markowitz, JGI: IMG
- Folker Myer, Argonne: RAST
- Swiss-Prot: HAMAP rules
- Genoscope: Microscope
- Ensembl

OMICS A Journal of Integrative Biology
Volume 12, Number 2, 2008

DOI: 10.1089/omi.2008.0017

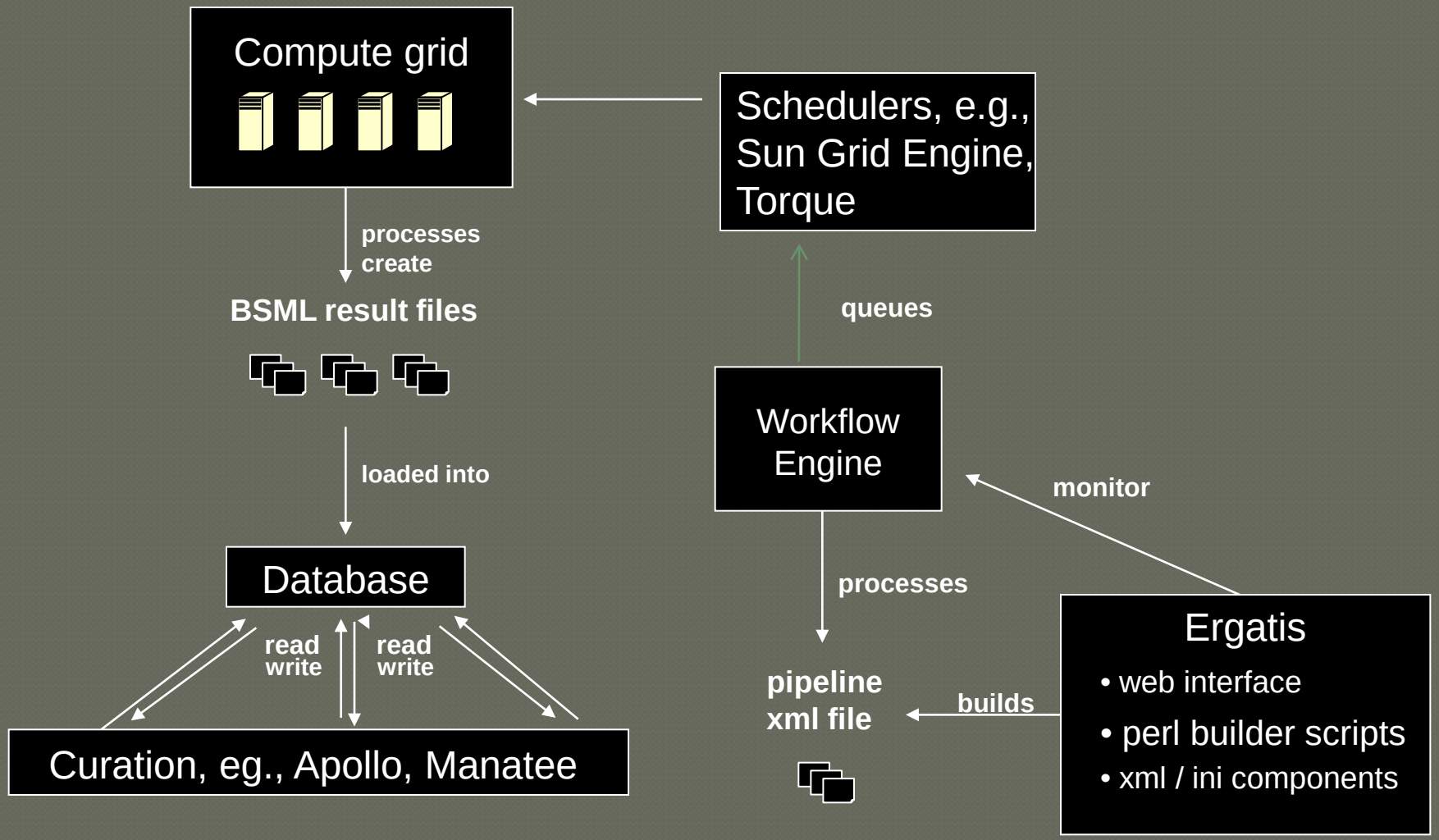
# Toward an Online Repository of Standard Operating Procedures (SOPs) for (Meta)genomic Annotation

Samuel V. Angiuoli,[1,2] Aaron Gussman,[1] William Klimke,[3] Guy Cochrane,[4] Dawn Field,[5] George M. Garrity,[6] Chinnappa D. Kodira,[7] Nikos Kyrpides,[8] Ramana Madupu,[9] Victor Markowitz,[10] Tatiana Tatusova,[3] Nick Thomson,[11] and Owen White[1]

***An E-journal to serve as an SOP repository***

“Standards in Genome Sciences"

# Ergatis infrastructure

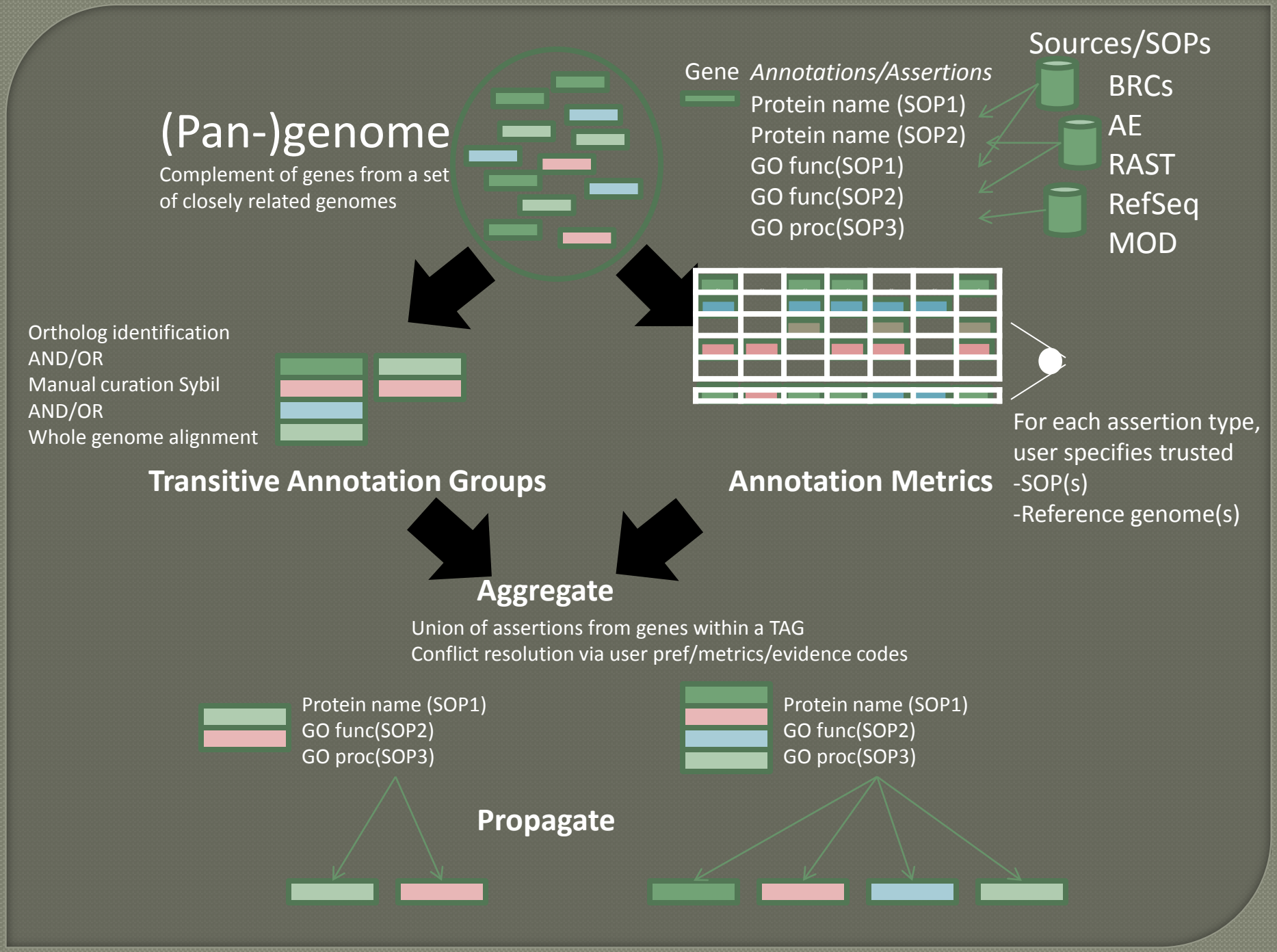


# Clouds

- Huge, thousands of slots available on demand
  - Amazon EC2
  - Google
- Free: Science Clouds
  - Terragrid
  - University of Chicago
  - University of Florida

# Microsoft Cloud

- 2,500 servers per shipping container.
- Truck hauls container into a 50,000sqft warehouse
- Connect electric, network and water cooling
- Upload of software
- 200 containers on line.
- 198MW electrical capacity: (aluminum smelter)


(Pan-)genome
Complement of genes from a set of closely related genomes
Gene
Annotations/Assertions
Protein name (SOP1)
Protein name (SOP2)
GO func(SOP1)
GO func(SOP2)
GO proc(SOP3)
Sources/SOPs
BRCs
AE
RAST
RefSeq
MOD
Ortholog identification
AND/OR
Manual curation Sybil
AND/OR
Whole genome alignment
Transitive Annotation Groups
Annotation Metrics
For each assertion type, user specifies trusted
-SOP(s)
-Reference genome(s)
Aggregate
Union of assertions from genes within a TAG
Conflict resolution via user pref/metrics/evidence codes
Protein name (SOP1)
GO func(SOP2)
GO proc(SOP3)
Protein name (SOP1)
GO func(SOP2)
GO proc(SOP3)
Propagate

sybil: strepneumo: home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://strepneumo-sybil.igs.umaryland.edu/cgi-bin/strepneumo-v1r9b1/shared/index.cgi?site=strepneumo

Google Gmail SOM Mail Dictionary Rathema W Wikipedia HMPDACC HMP-base Blogs1

Number

1

2

Sybil

Sybil t

2250

rRNA(2) rRNA rRNA rRNA rRNA rRNA

Tettelin et al. Figure 1

1.2603VR

2.NEM316

6.COH1

7.H36B

8.18RS21

0 Mb 0.25 Mb 0.5 Mb 0.75 Mb 1 Mb 1.25 Mb 1.5 Mb 1.75 Mb 2 Mb 2.25 Mb

Google™

sybil streptococcus

# Deliverables

- Ergatis scheduling systems
- Annotation services
- Installable annotation workflows
- Objective annotation measures

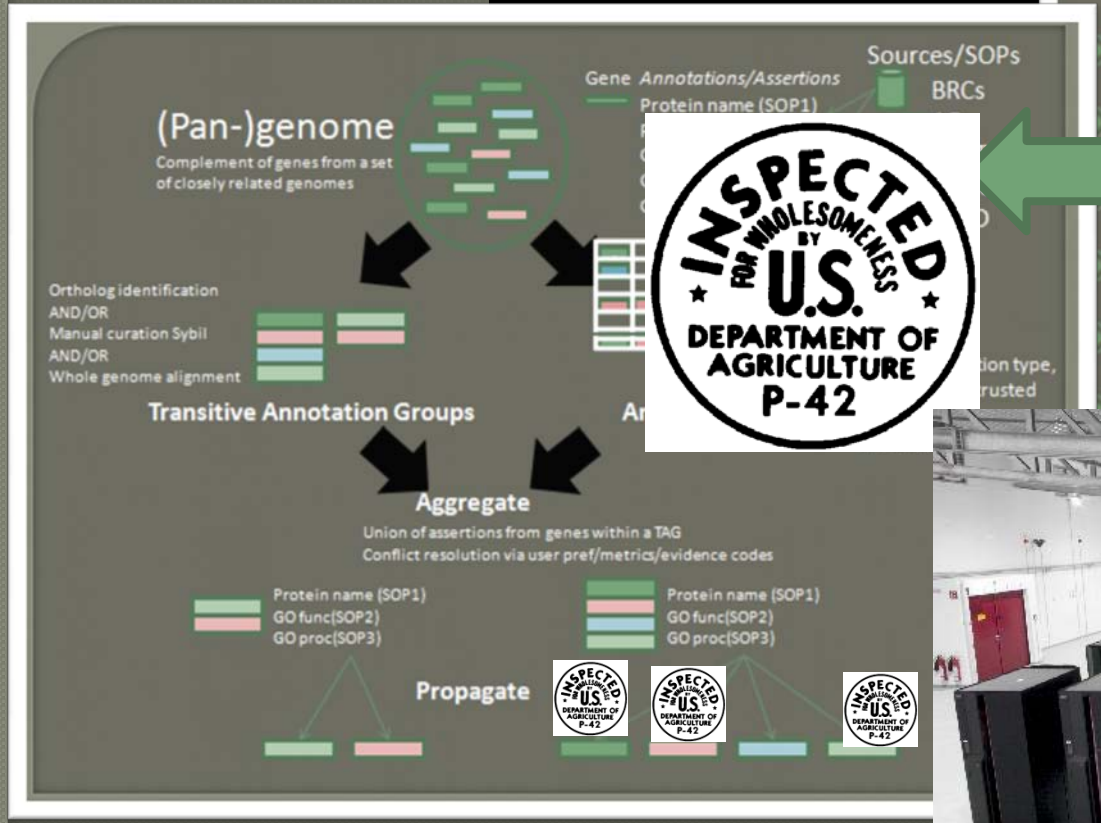
# Deliverables

- Sybil:
  - Workflow
  - Chado
  - Comparative analysis
  - Pan-genome analysis system
  - Web interface

# Near term delivery

- Full-blown data type saturation
- Objective annotation measures repository
- Resequencing annotation system
- Microbiome analysis tools

IGS: Annotation Engine
JCVI: Service
JCVI: Properties
Victor GI: IMG
Swiss ne: RAST
Geno rules
Ensembl cope

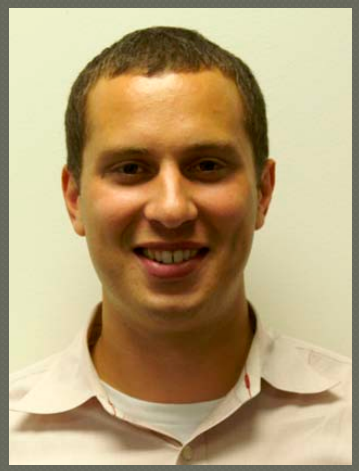







# Special thanks

Sam Angiuoli
Data modeling, algorithms



David Riley
Pan-genome

Aaron Gussman
Data management

Joshua Orvis
Ergatis

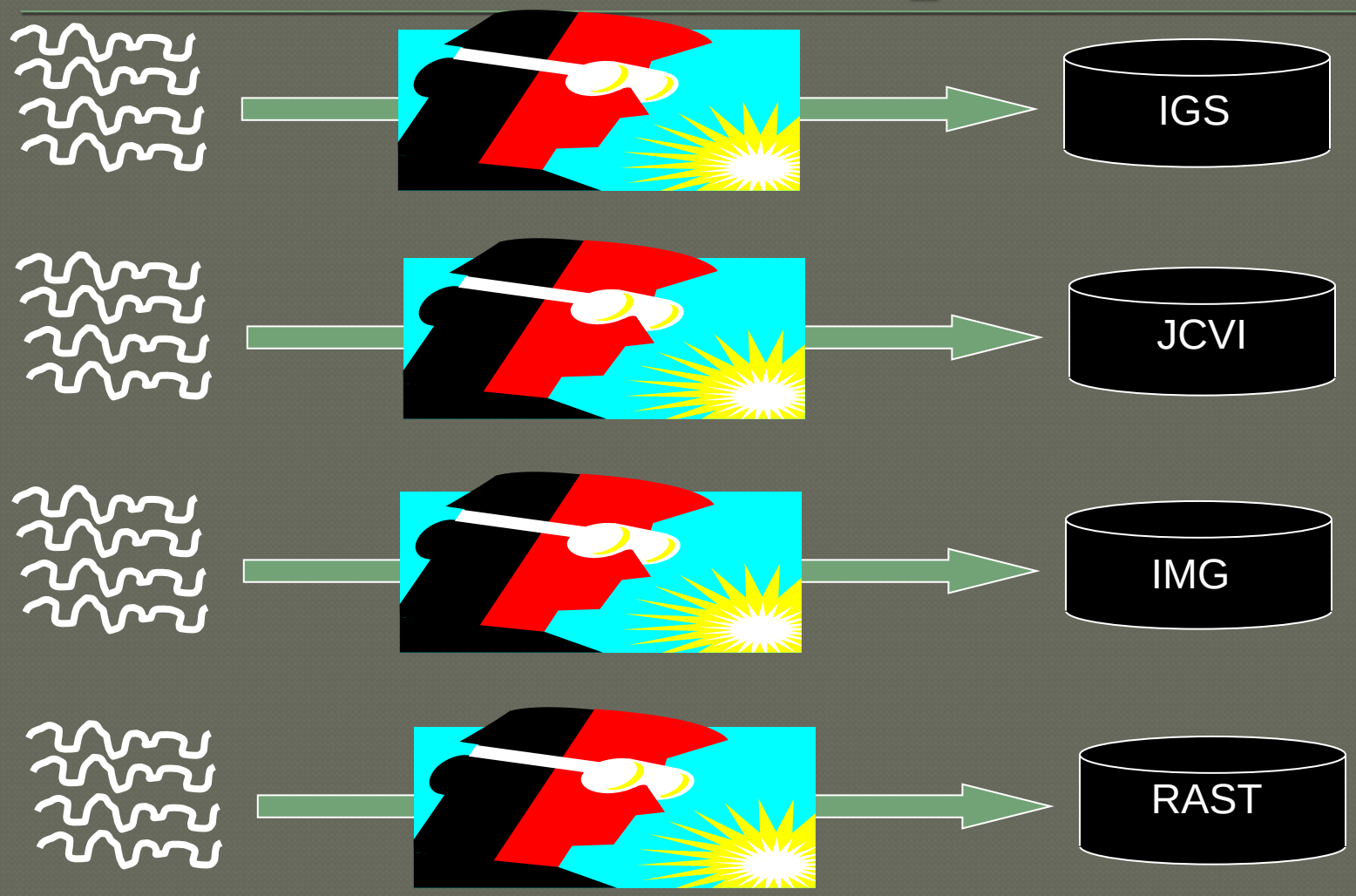
NIAID
Bioinformatics
Resource Center

# HMP Reference Genomes: Assembly

- Multiple centers generating draft genomes
- SOPs published by centers.
- Draft genome criteria
- Role of Centralized repository…

Attribution: HMP Jumpstart annotation working group

# Multiple Centers

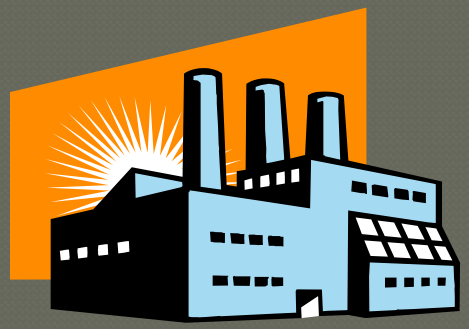


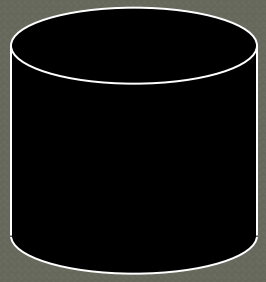
# Potential Systems

- Assembly
- Gene identification (GenePRIMP)
- Name resolution
- GO assignments
- Systems
- Orthologs (Functionally related families)
- Community comparison

# Community annotation pipeline.

- Friendly participation
- Based on strict datatypes
  - Assertions: Function, process, virulence factor, pathway
  - Include evidence codes
- Stamp with an SOP
- Develop object measures for above
- Select best of breed
- Move into one common pipeline





# HMP Reference Genomes: Annotation

- Multiple centers generating annotation
- SOPs published by centers.
- Role of centralized repository…
  - Attribution: HMP Jumpstart annotation working group

Thank you

# A Rational Proposal

- Could we use objective measures to select best of breed methods?