GSC 8
towards a richer set of information to describe our complete genome collection

Session I: Setting the Stage: Progress and Prospects within the GSC

Session Chair: Lynette Hirschman, MITRE

9:00 Introduction and Goals of the Workshop

- First issue of SIGS published (July 20th, 2009)
- Launch of the NSF RCN (2009-2013)
- Creation of the GSC's non-for-profit (Spring 2009)
- From MIGS/MIMS to MIENS: The "MIGS family of standards" (MIGS/MIMS/MIENS)
- GCDML, the Genomic Rosetta Stone, the "GEM Catalogue", and the "MIGS/MIMS for 1000 genomes and metagenomes project"
- GSC Finishing Standards Working Group
- M3 – Metagenomics, Metadata and MetaAnalysis – community outreach – to M5
- GSC Governance (GSC Board, RCN Steering Committee, SIGS Editorial Board, Officers of the GSC non-for-profit)
- GSC Roadmap

Dawn Field, NERC Centre for Ecology and Hydrology *Video* *Transcript*

9:15 The GSC's eJournal: "Standards in Genomic Science"
George Garrity, Michigan State University *Video* *Slides(?)* *Transcript*

9:30 The GSC's Research Co-ordination Network – RCN4GSC
John Wooley, UCSD *Video* *Slides* *Transcript*

9:45 Encyclopedia of Systems Biology: Eugene Kolker, Seattle Children's Hospital *Video* *Transcript*
biosharing website: Susanna Sansone, EBI *Video* *Transcript*

10:00 Curation of MIGS compliant data: Towards 1000 genomes and metagenomes
Peter Sterk, NERC Centre for Ecology and Hydrology and the Sanger Institute *Video* *Slides* *Transcript*

10:15 The MIENS (Minimum Information about an ENvironmental Sequence) specification
Frank Oliver Glöckner and Pelin Yilmaz, MPI-Bremen *Video* *Slides(?)* *Transcript*

10:30 Standards and the INSDC: Submission of MIGS/MIMS/MIENS
Ilene Mizrachi, NCBI *Video* *Slides* *Transcript*
Bob Vaughan, EMBL *Video* *Transcript*

11:00 Coffee Break

11:30 The MIGS* database, GCDML and the vision of the future
"Genomes and Metagenomes" (GEM) Catalogue
Renzo Kottmann, MPI-Bremen *Video* *Slides* *Transcript*

11:45 Outcomes of the ISA-GCDML workshop: towards multi-omic data sharing
Susanna Sansone, EBI *Video* *Slides* *Transcript*

12:00 The GSC's Finishing Standards Working Group
Patrick Chain, Los Alamos National Lab *Video* *Slides* *Transcript*

12:15 "A GSC Global Genome Census?"
Nikos Kyrpides, DOE JGI *Video* *Slides(?)* *Transcript*

12:30 Discussion/wrap up of session

- Including Brief Announcements of other work of interest to the GSC (verbal)
- Announcements of working lunches, dinners, and evening drinks

13:00-14:00 Lunch

Working Lunch: M5 Platform (Chair: Folker Meyer, Argonne)

Session II: MIGS/MIMS/MIENS in the real world

Session Chair: Dawn Field, NERC Centre for Ecology and Hydrology

14:00 Flash Updates (Verbal) *Video* *Transcript*

- The MegX.net database (Renzo Kottmann, MPI-Bremen)
- The MG-RAST metagenomic server implementation of MIGS/MIMS (Folker Meyer, Argonne National Laboratory)
- GOLD and MIGS/MIMS compliance (Nikos Kyrpides, DOE Joint Genome Institute)
- The IMG/m implementation of MIGS/MIMS and user-compliance (Victor Markowitz, Lawrence Berkeley National Labs)
- VAMPS and Microbis and compliance with MIENS (Linda Amaral-Zettler, MBL, Woods Hole)
- The RDP and MIENS (James Cole, Michigan State University)
- MIGS/MIMS/MIENS and the Human Microbiome Project DACC (Jennifer Wortmann, University of Maryland)

Mega-sequencing projects – the future

14:15 The Genomic Encyclopedia of Bacteria and Archaea

- Overview and scope
- MIGS compliance
- Genome notes in SIGS

Nikos Kyrpides, DOE JGI *Video(?)* *Slides(?)* *Transcript(?)*

14:30 The Human Microbome Project (HMP)

- Overview and scope
- Data sets to be generated
- MIGS/MIMS/MIENS compliance
- The DACC

George Weinstock, Washington University in St. Louis *Video* *Slides* *Transcript*

14:45 The Terragenome Initiative

- Overview and scope
- Data sets to be generated
- Soil Survey/MIENS
- MIGS compliance

Janet Jansson, LNBL *Video* *Slides(?)* *Transcript*

15:00 The Tara-Oceans Project
Jeroen Raes, University of Brussels *Video* *Slides(?)* *Transcript*

15:15 The DOE KnowledgeBase
Robert W. Cottingham, ORNL *Video* *Slides(?)* *Transcript*

15:30 Coffee

16:00 Discussion: The M5 Platform

- Introduction by Folker Meyer (Argonne)(15 minutes)
- Building the M5 Roadmap
- Towards a Unified set of genomes and metagenomes

Leads: Folker Meyer, Owen White, Eugene Kolker, Dawn Field

Day 2 Thursday 10th of Sept

Session III: Unifying concepts in genomic annotation: from SOPs to standards

Session Chair: Nikos Kyrpides (DOE, JGI)

9:00 Gene Calling Standards
Nikos Kyrpides, DOE Joint Genome Institute *Video* *Slides* *Transcript*

9:20 Discussion

11.00 Annotations in RefSeq
Tatiana Tatusova, NCBI *Video* *Slides(?)* *Transcript*

11:20 From multiple sources to consensus annotations: a vision for the future
Owen White, University of Maryland *Video* *Slides* *Transcript*

11:40 Discussion: Chair Victor Markowitz (LNBL)

Working Lunches:

- MIENS (Chairs: Frank Oliver Glöckner and Pelin Yilmaz)
- Annotations (Chairs: Owen White and Victor Markowitz)

Session IV: Working Group Meetings

14:00-15:30

1. I. MIENS Working Group (Chair: Frank Oliver Glöckner and Pelin Yilmaz)
2. II. GSC Biocomputing Consortium – the M5 platform (Chair: Folker Meyer)

16:00 Presentations from MIENS and Biocomputing Working groups (5 minutes each plus questions)

Session V: The GSC journal: Standards in Genomic Sciences (SIGS)

Chair: Peter Sterk (NERC Centre for Ecology and Hydrology and the Sanger Institute)

16:30 The Evolution of SIGS: Next Steps

- Genome Reports
- Metagenome reports
- SOPS
- Other content
- Advertising
- Special Issue from M3/GSC 8

George Garrity, Editor-In-Chief (Michigan State University) *Video(?) Slides(?) Transcript(?)*

Day 3 Friday 11th of Sept

Session VI: GSC Roadmap – the Horizon

Chairs: George Garrity

9:00 The BioCurator Society
Pascal Gaudet (Northwestern University) *Video Slides(?) Transcript*

9:40 Discussion
George Garrity

11:00 Reviews of morning session and setting actions for the future GSC Board

Discussion

- Summary of sessions and actions
- Setting actions for core projects: a Roadmap
- Contributions to SIGS: specific submissions
- GSC 8 meeting report: authors and content
- Planning for M3 @ PSB in Jan 2010
- Future meetings: GSC9 – and beyond

Facilitators: George Garrity Michigan State and Frank Oliver Glöckner, MPI Bremen

14:00 Wrap up: Final Review of Actions and Strategy development Workshop Co-organizers

This project has received funding from NIEeS and a NERC International Opportunities Award (NE/3521773/1) 2005-2008)

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Defining genome project standards in a new era of sequencing

Patrick Chain

GSC 8 Meeting
JGI, Walnut Creek, CA
Sept. 2009

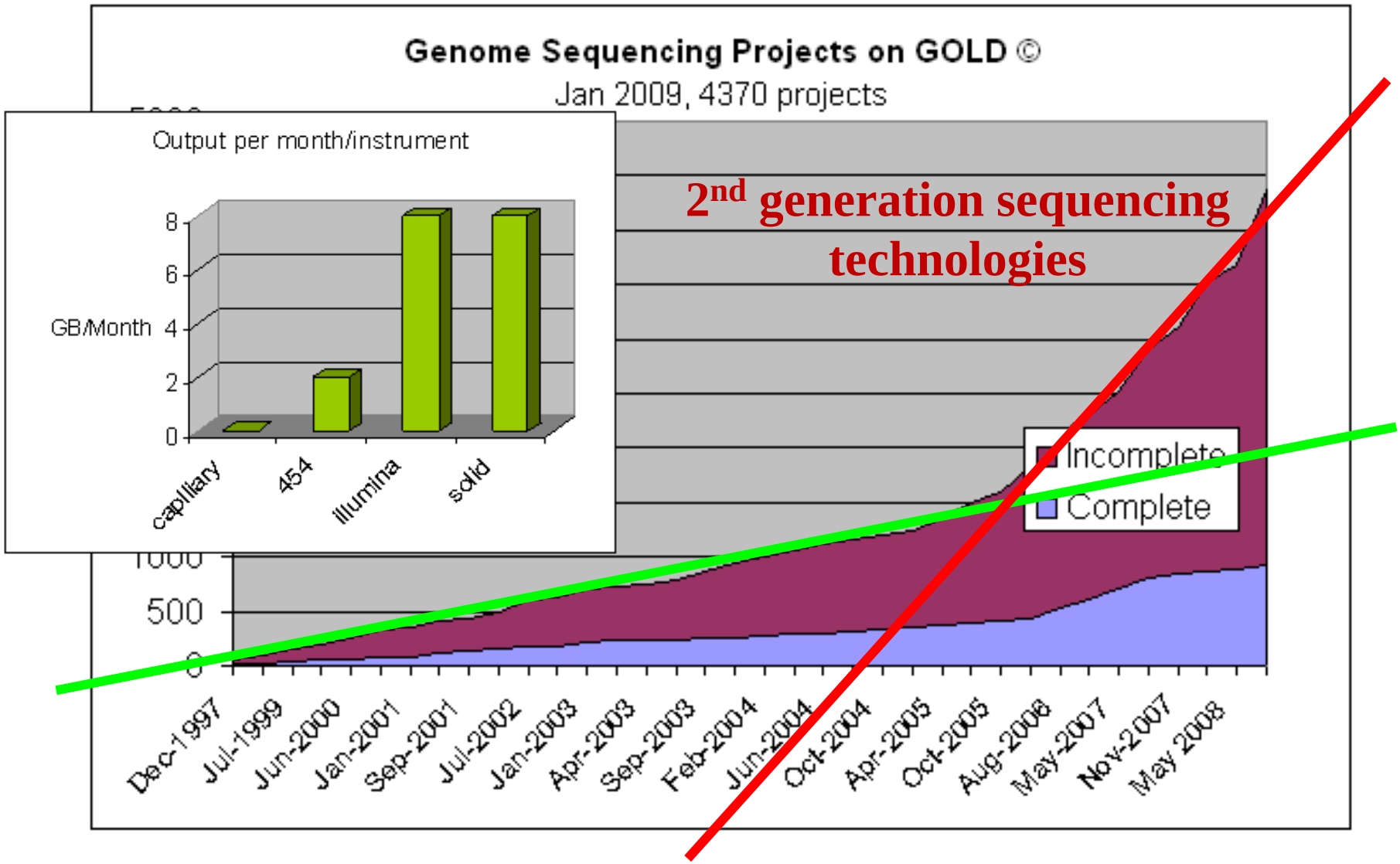
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Past genome sequencing standards

- **1995: First complete genome of a free-living organism**
- **1997: 2nd International Strategy Meeting on Human Genome Sequencing in Bermuda**
  - **Defined "Finished" sequence – aka "Bermuda" standard**
- **Until very recently: genomes fell into 1 of 2 categories - draft or finished**
  - **Finished in bacteria/archaea is defined as base pair perfect sequence in a single contig**
  - **Finished in eukaryotes typically conforms to the bermuda standards, though most genomes are only regionally finished**
  - **With only 2-categories, it is apparent that this model requires significant re-evaluation…**

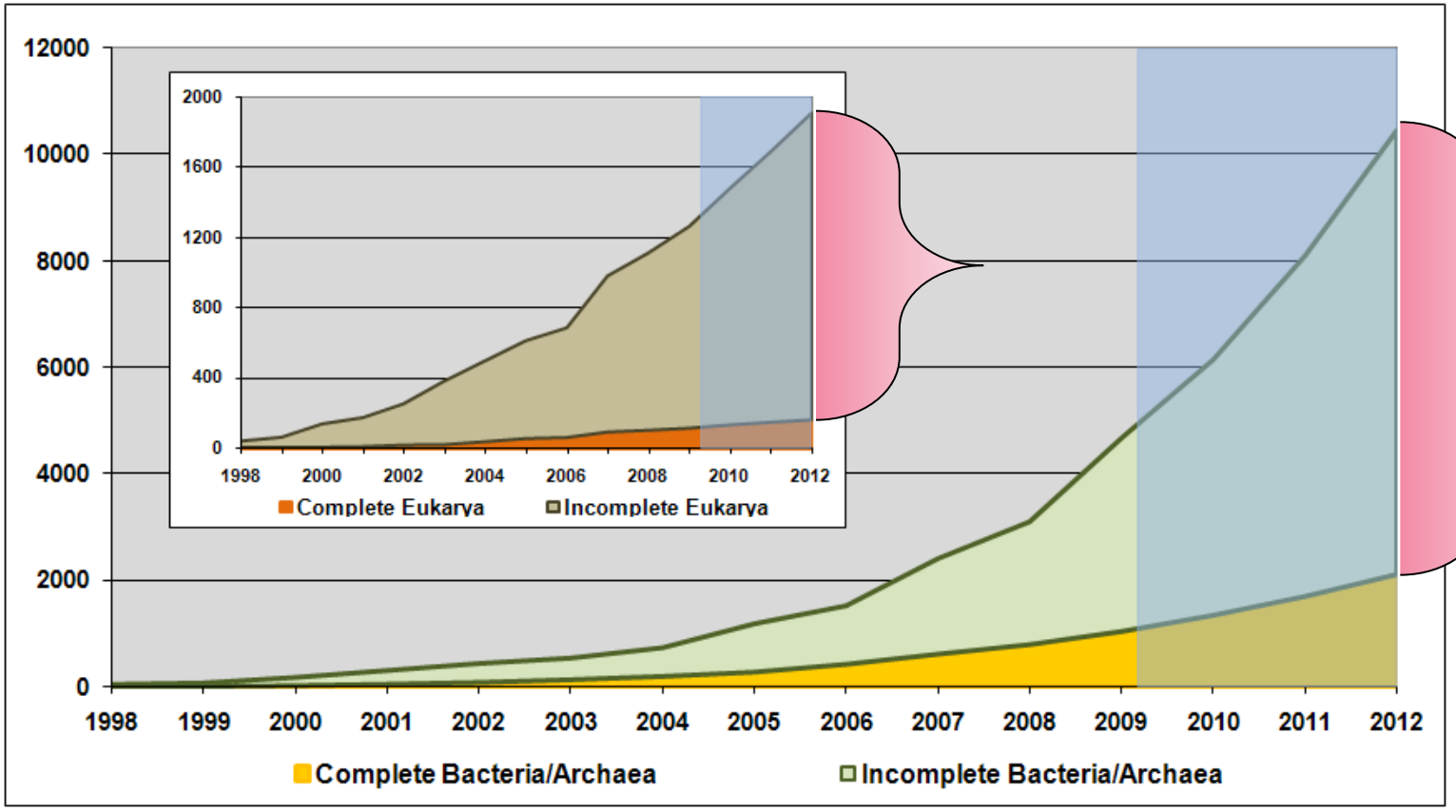
JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

2 categories: A growing problem



JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

A conservative projection: 12000 by 2012



What kind of draft?

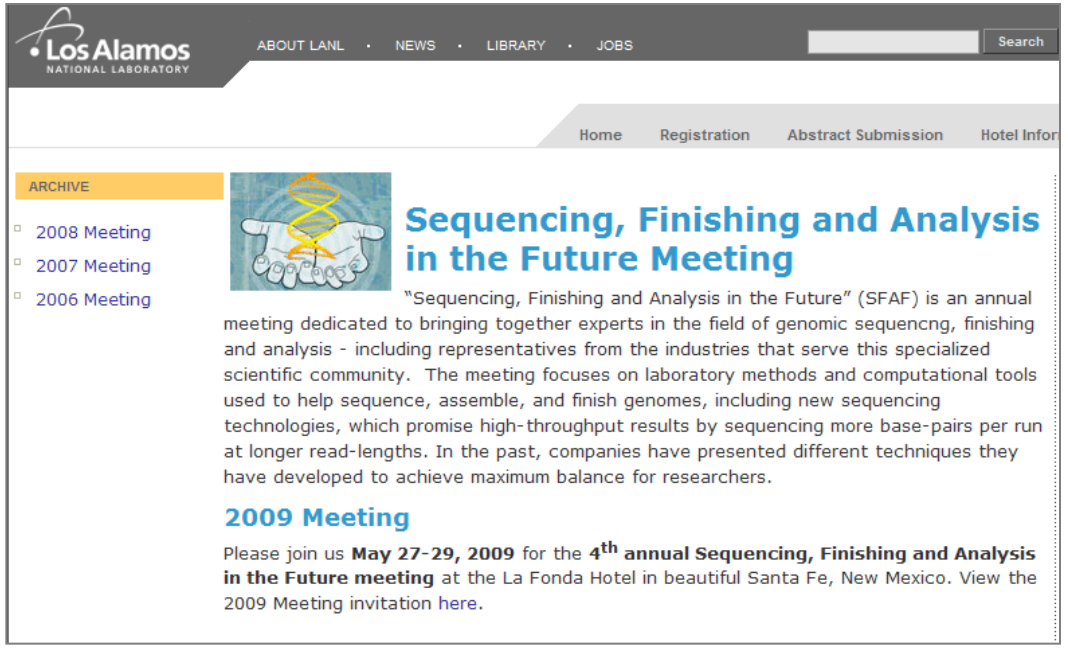
Blakesley, R.W. et al. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res* 14, 2235-2244 (2004)

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

"Dawn" of New Finishing Standards

- Finishing effort needs to be applied effectively and some gradations between draft and finished are required.
- Community needs to be better informed of the quality of product they are receiving.
- Granting bodies need to understand what will be delivered.
- MIGS paper

Formed the international genome sequencing standards working group

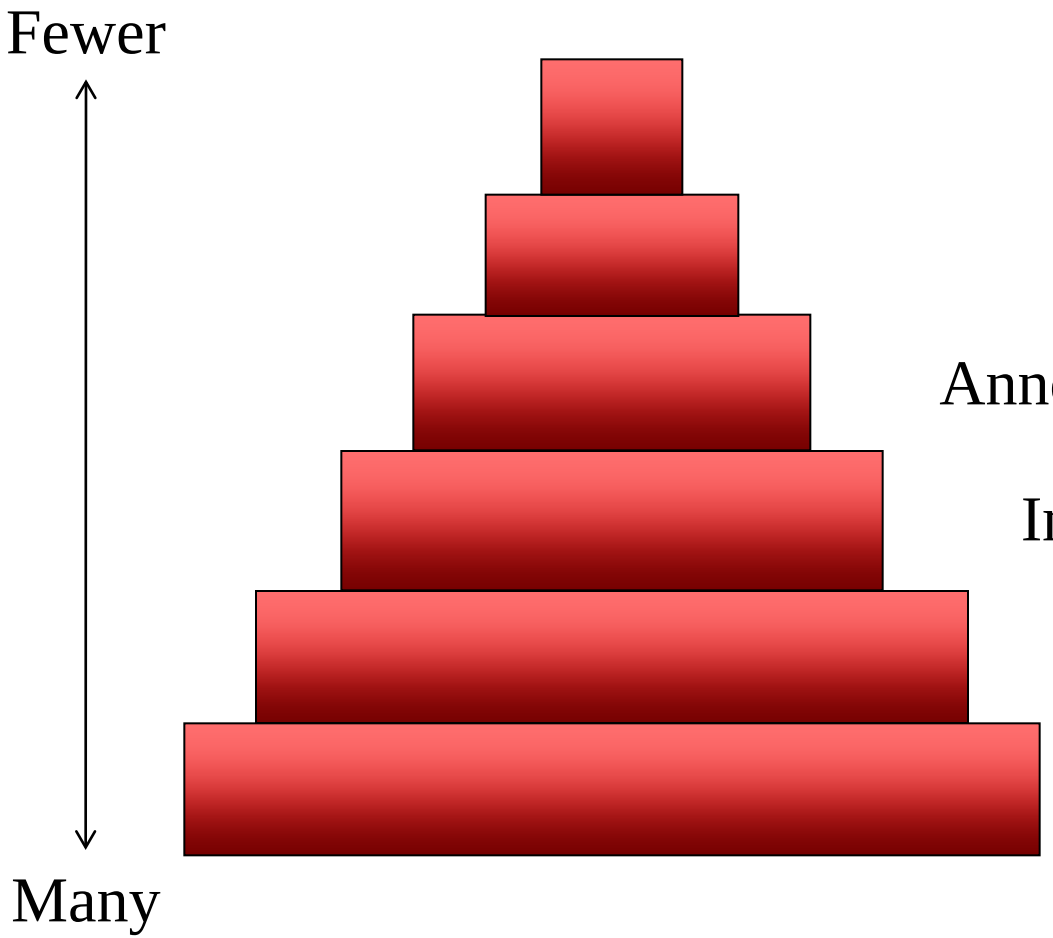


Joint Announcement on Genome Sequencing Standards

- **The International Genome Sequencing Standards Working Group:**

  DOE JGI (LANL, LBNL, LLNL), Sanger Institute, Human Microbiome Project (WashU Genome Center, The Broad Institute, JCVI, Baylor College of Medicine Human GSC, NHGRI, NIAID, U. of Maryland Institute for Genome Sciences), Emory U., HudsonAlpha GSC, Michigan State U., Natural Environmental Research Council Centre for Ecology and Hydrology, Ontario Institute for Cancer Research, NCBI, Naval Medical Research Center

An agreement on gradation of finishing



Also, regionally improved…

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

An agreement on gradation of finishing

- **Six genome sequencing standards:**

- **Standard Draft: minimally or unfiltered data from any number of different sequencing platforms, that are assembled into contigs. This is the minimum standard for a submission to the public databases. Sequence of this quality will likely harbor many regions of poor quality and can be relatively incomplete. It may not always be possible to remove contaminating sequence data. Despite its shortcomings, Standard Draft is least expensive to produce and still possesses useful information.**

- **High-quality-draft: overall coverage representing at least 90% of the genome or target region. Efforts should be made to include only sequence of the target organism and exclude contaminating sequences. This is still a draft assembly with little or no manual review of the product. Sequence errors and misassemblies are possible, with no implied order and orientation to contigs. This level is appropriate for general assessment of gene content.**

- **Improved-high-quality-draft: additional work has been performed beyond the initial shotgun sequencing and High-quality-draft assembly, by using either manual or automated methods. This standard should contain no discernable misassemblies, and should have undergone some form of gap resolution to reduce the number of contigs and supercontigs (or scaffolds). Undetectable misassemblies are still possible, particularly in repetitive regions. Low quality regions and potential base errors may also be present. This product is normally adequate for comparison to other genomes.**

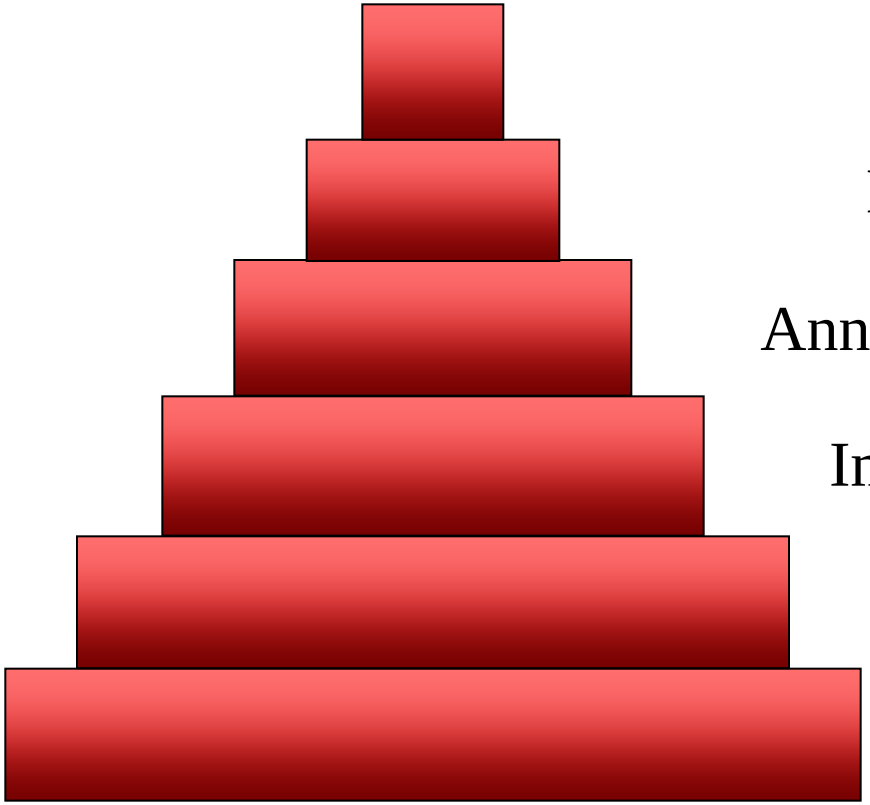
- **Annotation-directed-improvement: may overlap with the previous standards, but the term emphasizes the verification and correction of anomalies within coding regions such as frameshifts, and stop codons. This standard will most often be used in cases involving complex genomes where improvement beyond this category fails to outweigh the associated costs. Gene models (gene calls) and annotation of the genomic content should fully support the biology of the organism and the scientific questions being investigated. Exceptions to this gene-specific finishing standard should be noted with comments in the submission. Repeat regions at this level are not resolved, so errors in those regions are much more likely. This standard is useful for gene and pathway comparisons.**

- **Non-contiguous finished: describes high quality assemblies that have been subject to automated and manual improvement, and where closure approaches have been successful for almost all gaps, as well as misassembled and low quality regions, however some exceptions exist. All gaps and sequence uncertainties have been attempted to be resolved, and only those recalcitrant to resolution remain, but are specifically noted in the genome submission as to the nature of the uncertainty. This product is thus of finished quality with the only exception being repetitive or recalcitrant gaps, thus making it appropriate for most analyses.**

- **Finished: refers to the current gold standard; genome sequences with less than 1 error per 100,000 bp and where each replicon is assembled into a single contiguous sequence, with a minimal number of possible exceptions commented in the submission record. All sequences are complete and have been reviewed and edited, all known misassemblies have been resolved, and repetitive sequences have been ordered and correctly assembled. Any remaining exceptions to highly accurate sequence are commented in the submission. This product is appropriate for all types of detailed analyses to other sequences and acts as a high quality reference genome for comparative purposes.**

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

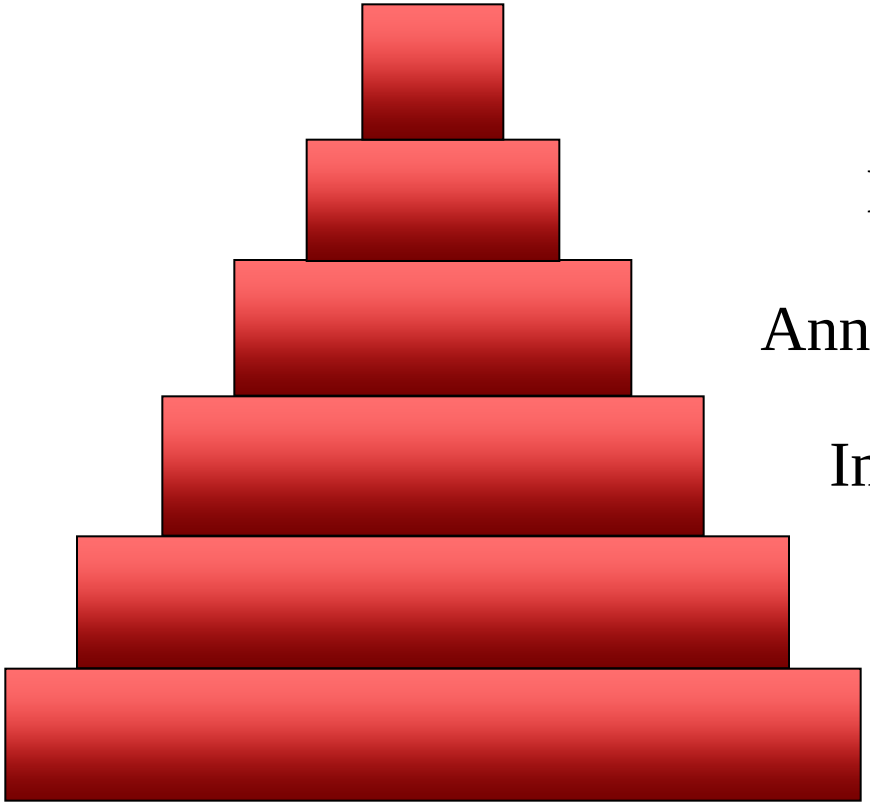
Improved high quality draft

High quality draft

Standard draft

may not always be possible to remove contaminating sequence data from this incomplete dataset

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

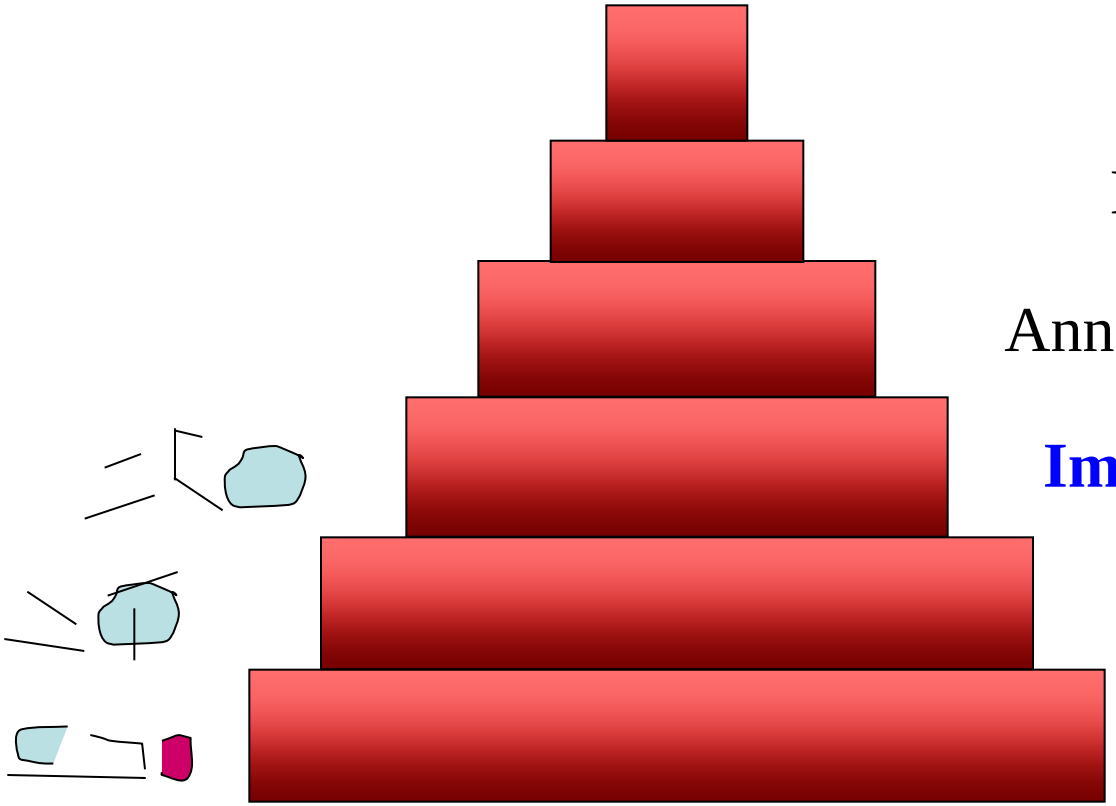
Improved high quality draft

High quality draft

Standard draft

still a draft assembly (90% coverage) with little or no manual review of the product, thus sequence errors and misassemblies are possible, with no implied order and orientation to contigs

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

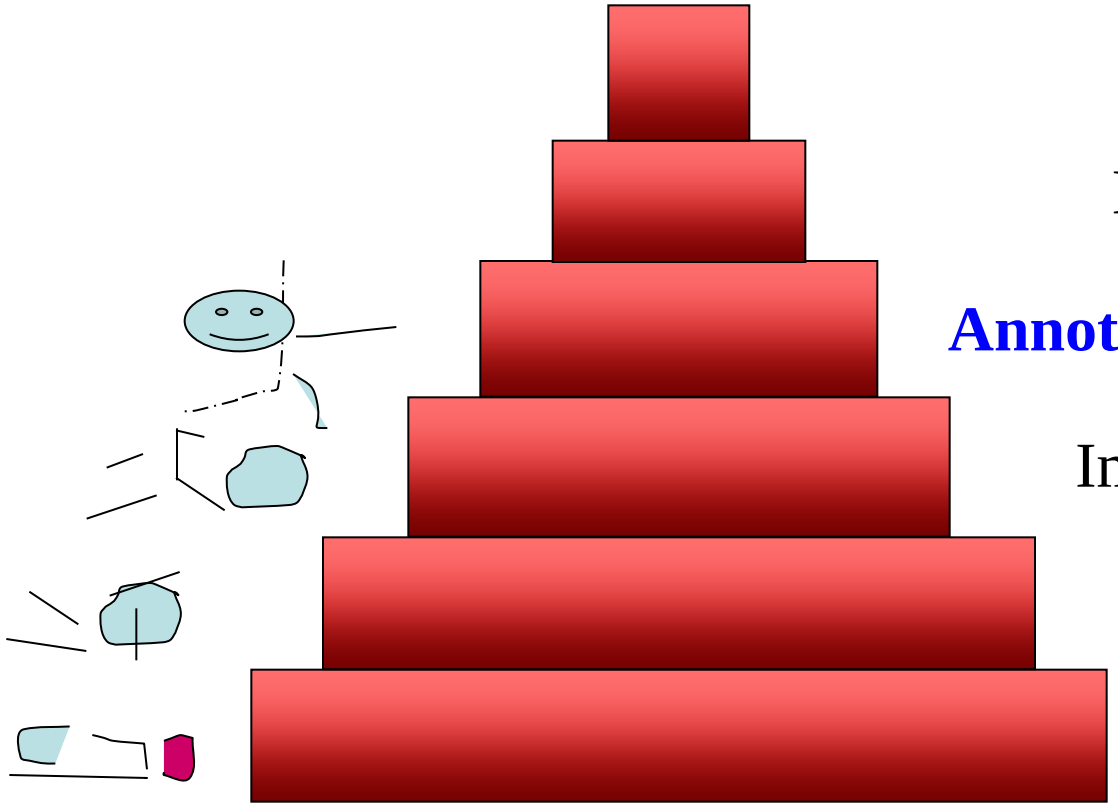
Improved high quality draft

High quality draft

Standard draft

Automated or manual improvement – with all but undetectable misassemblies addressed…low quality regions and potential bp errors may also be present, but the sequence is of high quality

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

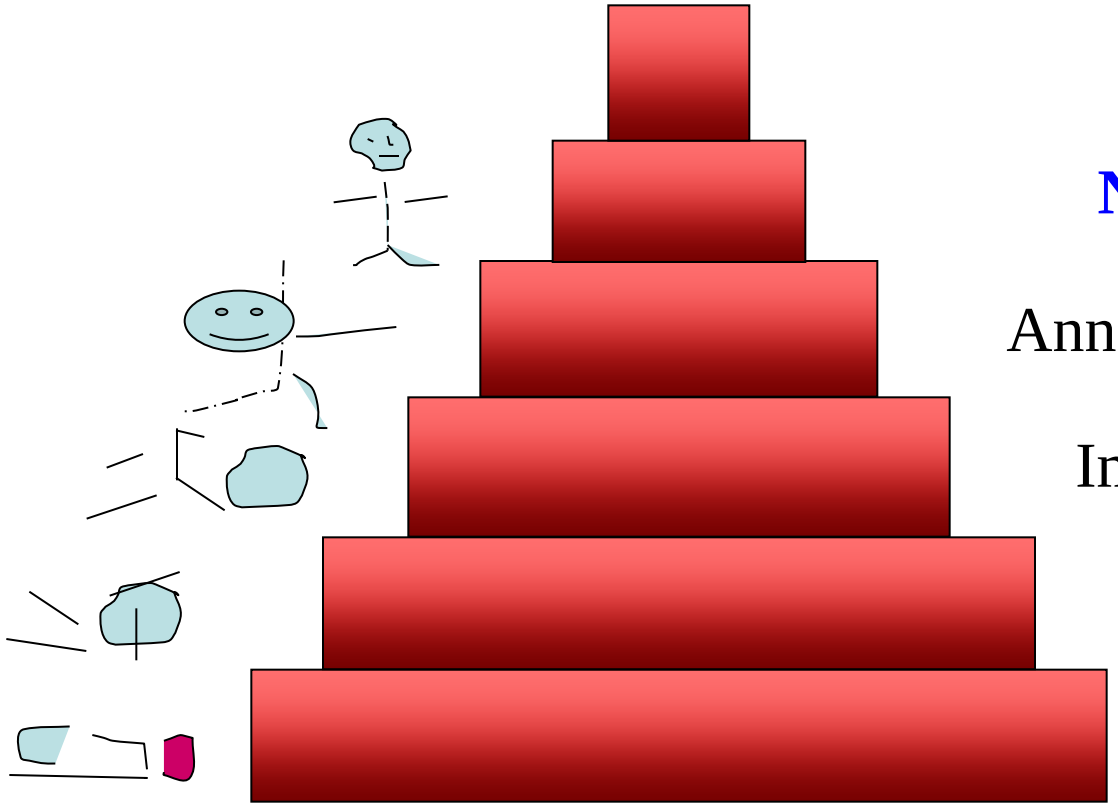
Improved high quality draft

High quality draft

Standard draft

Efforts have been made to resolve all errors in gene regions (exceptions to this gene-specific finishing standard should be noted with comments)…repeat regions are not necessarily resolved, so errors in those regions are much more likely

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

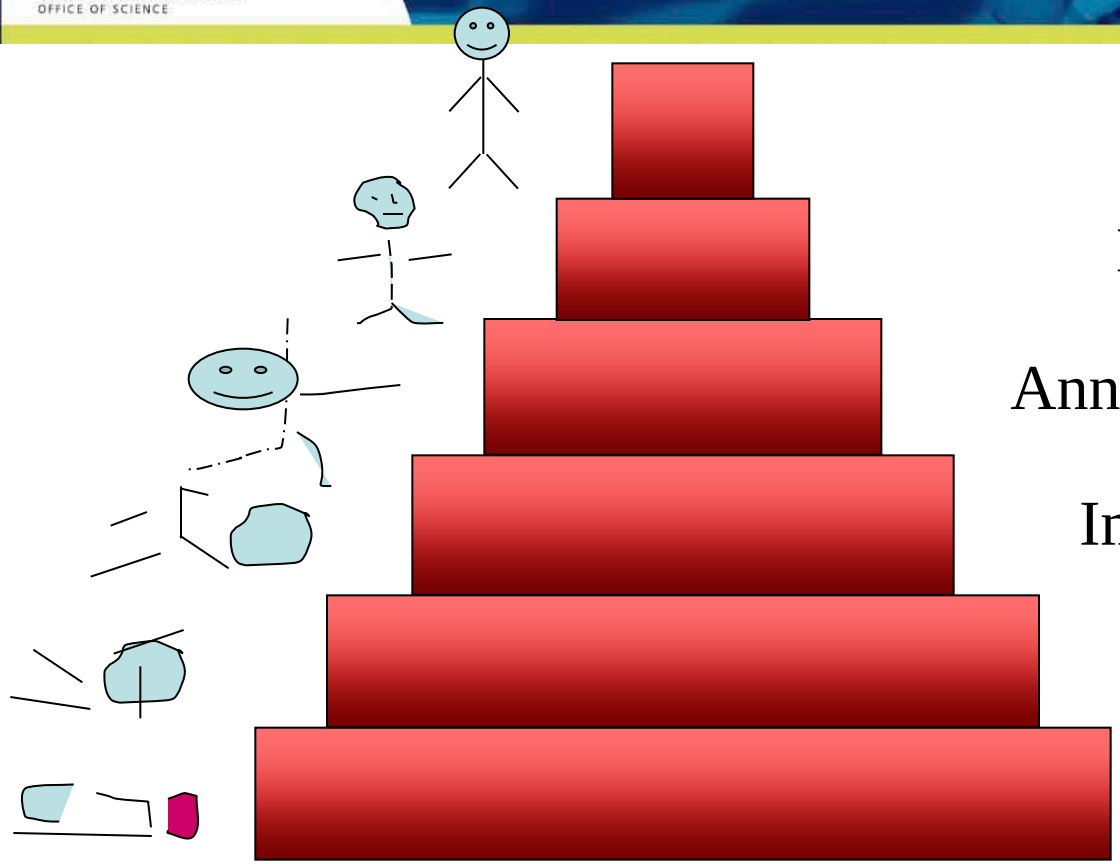
Improved high quality draft

High quality draft

Standard draft

All gaps and sequence uncertainties have been attempted to be resolved, and only those recalcitrant to resolution remain, but are specifically noted as to the nature of the uncertainty

An agreement on gradation of finishing



Finished

Non-contiguous finished

Annotation-directed improvement

Improved high quality draft

High quality draft

Standard draft

Gold Standard (1 error per 100,000 bp, no misassemblies). Any remaining small exceptions to highly accurate sequence are commented in the submission.

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

One system for all

- Agreed upon by all members of our consortium
- Will vastly improve user understanding
- Publicise: publication and on institute webpages
- Discussions with databases underway
- Technology-agnostic so can adapt as 3rd (next) generation of sequencers arrives
- Project/centres can add more detail as required such as sequence coverage and platform type
- Fits very well with the GSC
  - Used by SIGS (www.*standardsingenomics.org*)
- Already implemented by GEBA, HMP

The MIGS* database, GCDML and the vision of the future “Genomes and Metagenomes” (GEM) Catalogue

- There is now an official database for the MIGS/MIMS/MIENS checklists
- GCDML is mature enough for implementation
- The vision for the “Genomes and Metagenomes” Catalogue is being re-invented

Outcomes of ISA-GCDML workshop

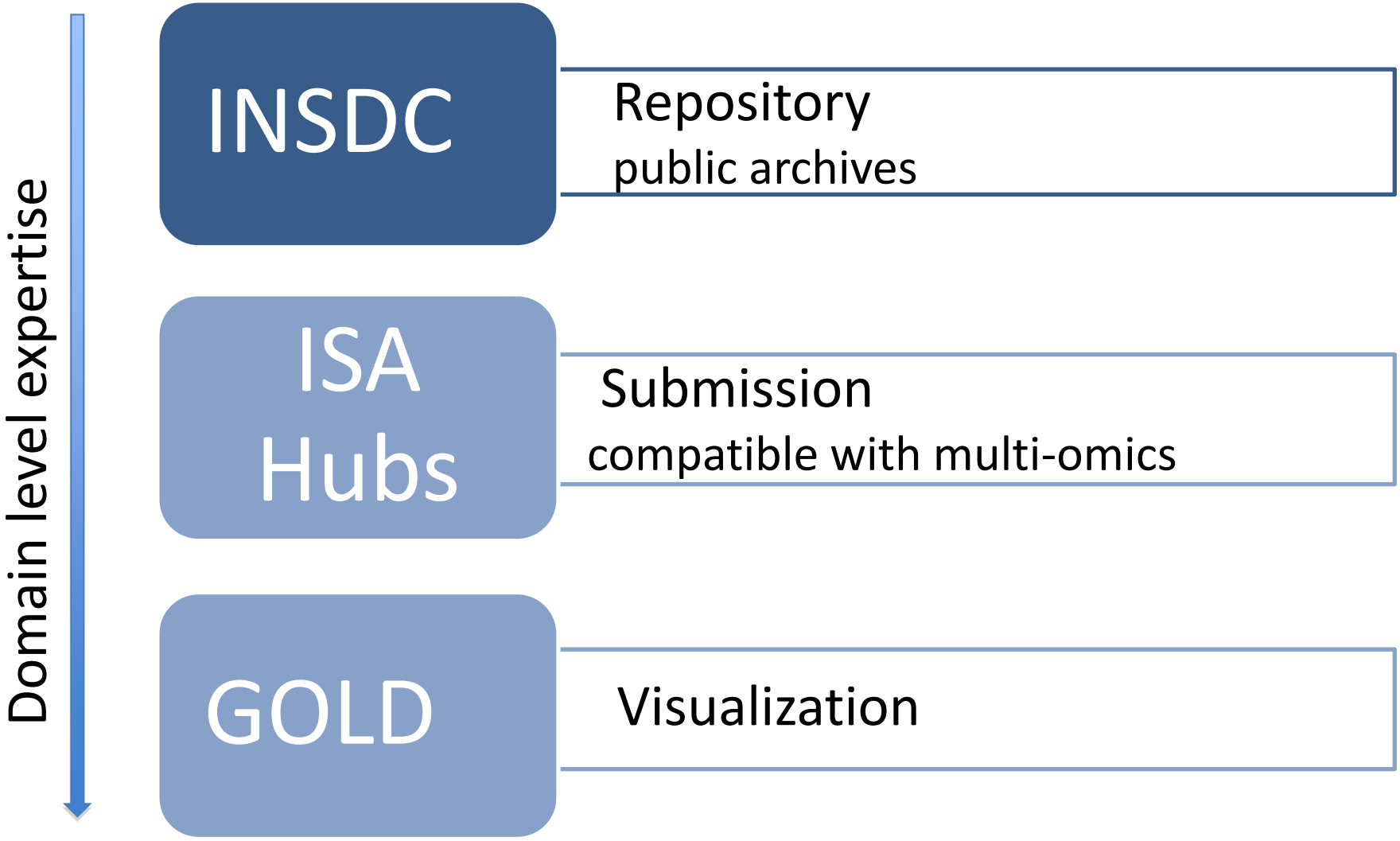
Sept 7th: discussed alignment of GCDML and ISATAB - 1 slide summary of discussions: Overview

Sept 8th: discussed the GSC's GEM Catalogue and significantly evolved requirements document

Conclusions:

- the efforts of this workshop should be continued in the form of a working group for the GEM Catalogue and this group should work on a multi-author roadmap paper that could be submitted to SIGS
- the proposal for the GEM Catalogue was advanced significantly - currently exploring creating the collaborative GEM Catalogue from a combination of the GOLD database (visualization), an ISA Hub (submission/mulit-omic) and the INSDC (authoritative databases)

GEM Catalogue Proposal



Next Steps

- Form GEM Catalogue working group
- Working Lunch: Friday

- See wiki page on GEM Catalogue

Genomic Standards Consortium

Gene Calling Standards

Nikos Kyrpides
Genome Biology Program
DOE Joint Genome Institute

Finding unique genes

Obligate parasite of horses

Causes human disease in tropical areas (melioidosis)

☐	Buchnera aphidicola (subsp. Baizongia pistaciae)			
☐	Buchnera aphidicola (subsp. Schizaphis graminum)			
☐	Burholderia cepacia strain 383 ATCC 17660 (R-18194)	B	Draft	JGI
☐	Burkholderia cenocepacia HI2424	B	Draft	JGI
☑	Burkholderia mallei (strain ATCC 23344)	B	Finished	TIGR
☑	Burkholderia pseudomallei (strain K96243)	B	Finished	Sanger Institute; Porton Down
☐	Burkholderia vietnamiensis strain G4 (R1808)	B	Draft	JGI
☐	Burkholderia xenovorans LB400			
☐	Campylobacter jejuni (strain NCTC 11168)			
☐	Candida glabrata CBS138 (Torulopsis glabrata)			
☐	Candidatus Blochmannia floridanus			
☐	Caulobacter crescentus (strain CB15 / ATCC 19089)	B	Finished	TIGR

Statistics For User-selected Organisms

Taxonomic Domains(D): B = Bacteria, A = Archaea, E = Eukarya.

Organism Name	D	Total Bases	GC Perc	# Scaffolds	CDSs	w/ Func	COG Genes	Pfam Genes	Pfam Clusters
Burkholderia mallei (strain ATCC 23344)	B	5835527	68%	2	4764	3561	2640	2204	1299
Burkholderia pseudomallei (strain K96243)	B	7247547	68%	2	5855	4423	3292	2673	1439

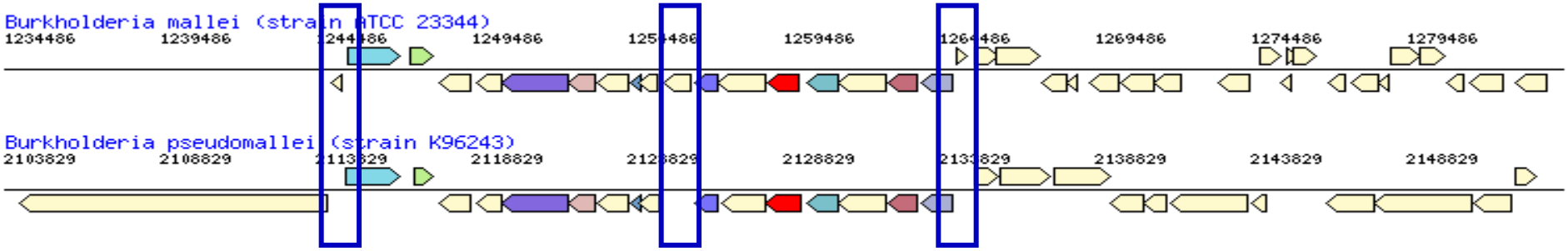
Genomic Standards Consortium

Gene Ortholog Neighborhoods

Loaded

Neighborhoods of orthologs in user-selected organisms.
Genes of the same color (except light yellow) are from the same orthologous group (COG).
Light yellow = no COG assignment.

hint: *Mouse over a gene to see details (once page has loaded).*



- Phylogenetic profiler finds 548 unique genes in *B. mallei*

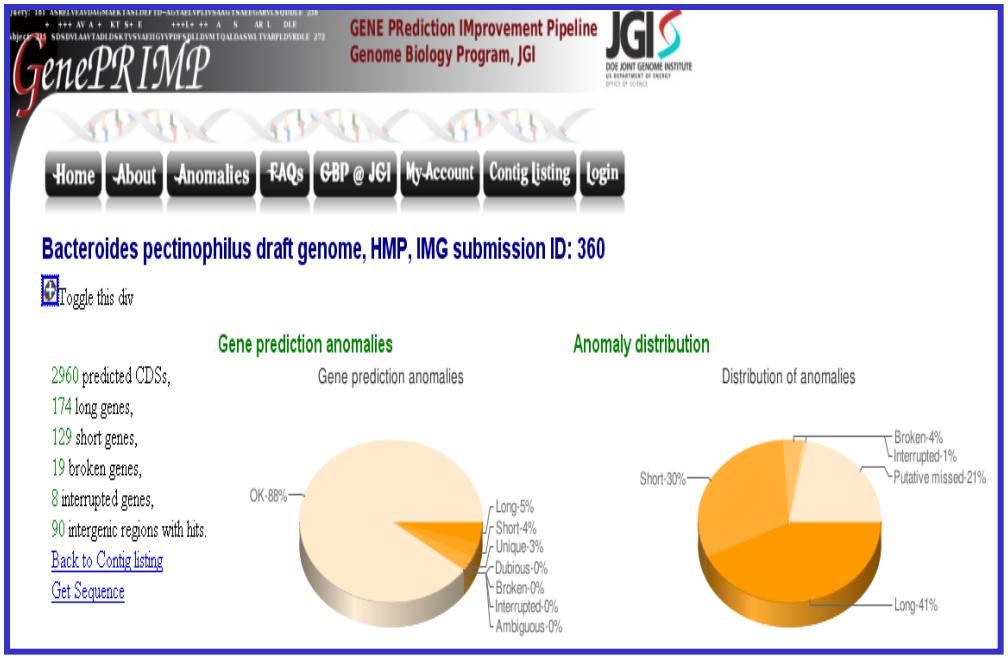
- However, 497 of them in fact exist in B. pseudomallei, but they have not been called as real genes.
- The difference in gene models reveals 89.2% error rate in unique genes

Genomic Standards Consortium

	Mycobacterium sp. Spyr1 GC% = 67.9, Size=6 Mb					Methanosphaerula palustris E1-9c GC% = 55.35, Size=2.9 Mb				
	1	2	3	4	5	1	2	3	4	5
CDSs	5553	5395	5296	5304	4888	2974	3334	2819	2940	3177
Short genes	482	398	267	672	79	235	230	158	315	115
Long genes	83	53	62	34	992	46	59	55	44	294
% CDSs w/ anomalous starts (short + long)	10.17%	8.36%	6.21%	13.31%	21.74%	9.49%	8.69%	7.55%	12.2%	12.87%
Missed genes	607 (10.93%)	569 (10.54%)	451 (8.51%)	735 (13.9%)	658 (13.46%)	196 (6.59%)	206 (6.18%)	165 (5.85%)	264 (8.97%)	106 (3.33%)
Unique genes	67	118	23	206	99	190	522	121	554	277
Dubious genes	11	0	2	0	10	25	0	3	0	8
Broken genes	30	33	27	22	34	41	50	38	25	71
Interrupted genes	51	62	48	60	53	23	36	23	22	60
Frameshift anomalies (broken + interrupted)	81	95	75	82	87	64	86	61	47	131

Gene Prediction Quality Assurance

GenePRIMP



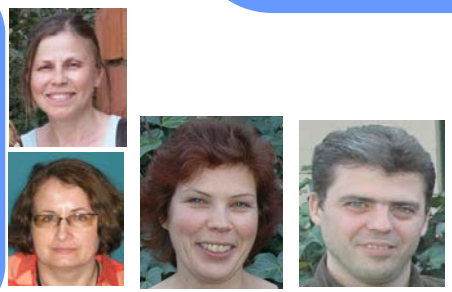
Gene Prediction Improvement Pipeline

GenePRIMP is a pipeline that consists of a series of computational units that identify erroneous gene calls and missed genes and correct a subset of the identified defective features.

APPLICATIONS

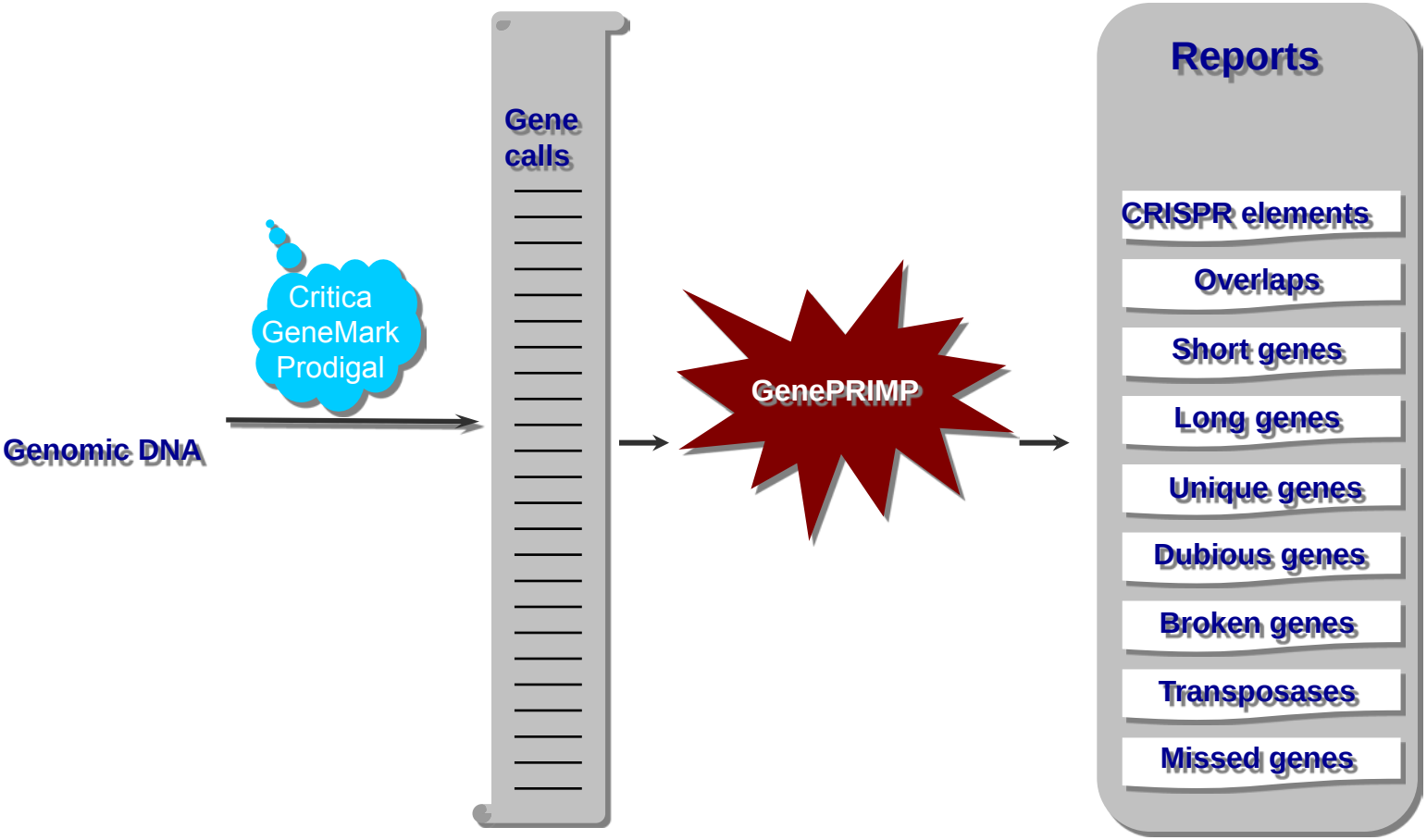
- Identify gene prediction anomalies
- Benchmark the quality of gene prediction algorithms
- Benchmark the quality of combination / coverage of sequencing platforms
- Improve the sequence quality

- **Manual QA**
- **Complete Genomes curated: >300**
- **Manually examined Genes: ~1,000,000**
  - Modified Genes: ~100,000
- **Average modifications: 10.4% / genome**



Genomic Standards Consortium

GenePRIMP



Amrita Pati, Natalia Ivanova, GBP

Genomic Standards Consortium

PROCESS FLOW

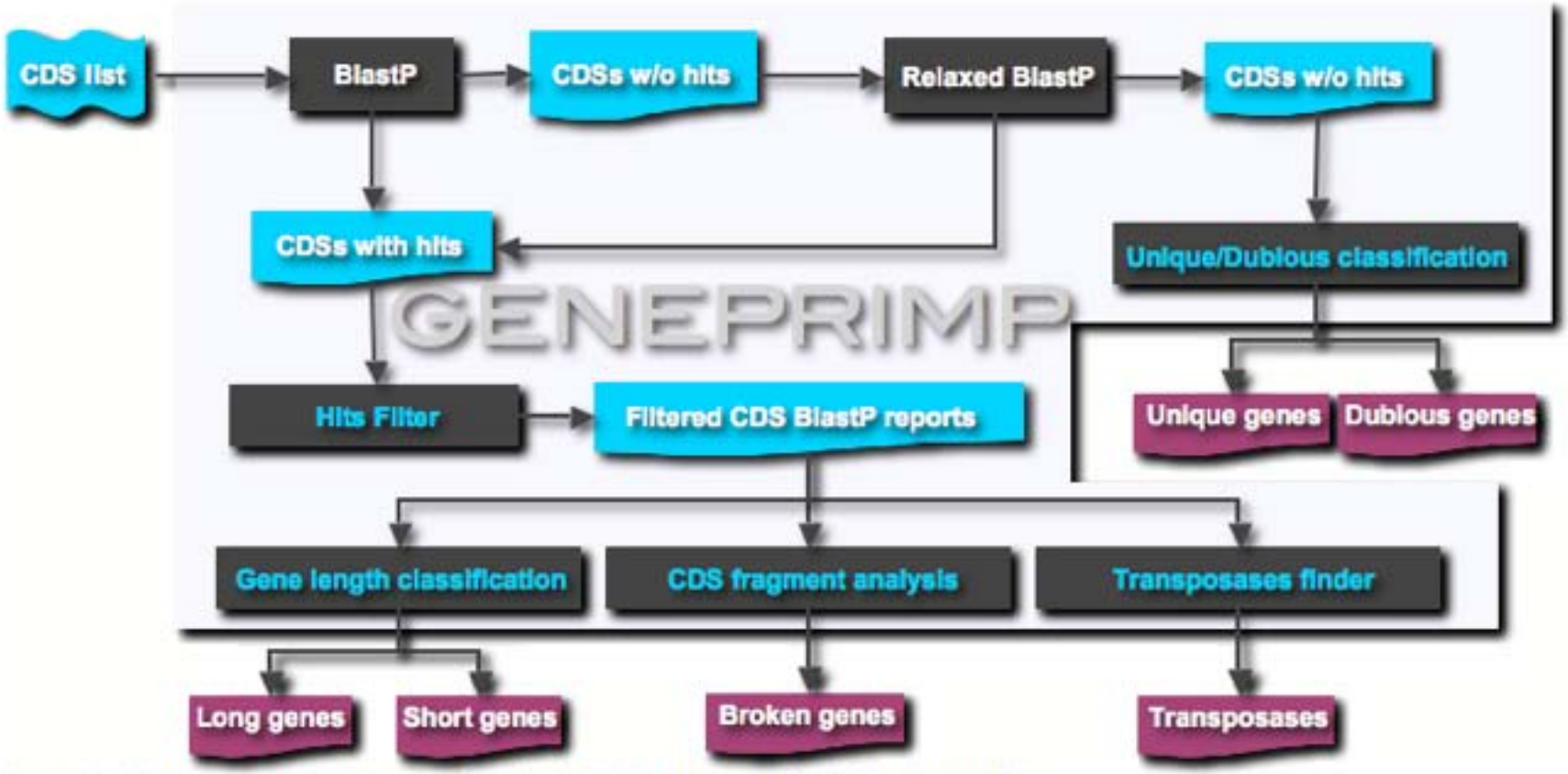


Figure 2. GenePRIMP analysis of called CDSs to identify defective calls

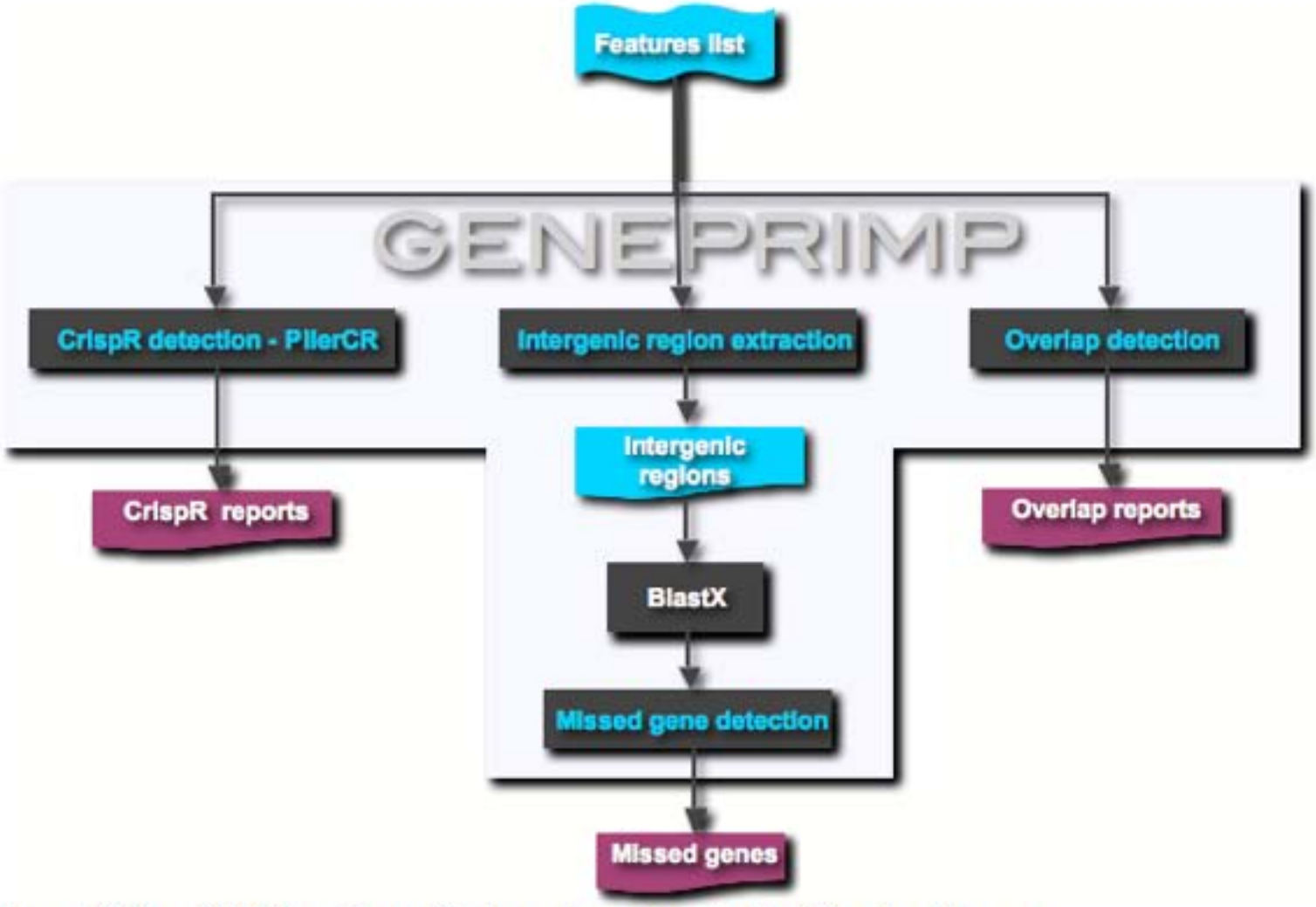


Figure 3. GenePRIMP analysis of intergenic regions to identify missed genes

Genomic Standards Consortium

Short genes

A gene is called 'short' when it has been truncated significantly at the 5'-end. In such a case, the gene is significantly shorter than its homologs in other species as illustrated in Figure 1. Commonly, as a result of a gene being called short, important functional domains get truncated (Figure 2), resulting in theoretical loss of function of the gene.

```
QUERY      3   QPASRTGKLSSPVANPVTNSPFGYRTNPLSGAAGELHTGLDFAAGCGTPVFSAGKGTVTE 62
116669360 121    .As..S.GA.LeQm..a....F.VS.IT.G...F.r.q.Y..q...s..A.AS....f 178
119962355 144 .r..sV.tMA..LdTlT.a......iS..T.G..DF.r.q..V.q...a.hA.At.k..f 203
119952752 164                      ....i.SS.IT.S-........LV.A.qVa..A..S...V. 203
163839258 208        .m....LdSlLmT......V..VT.........q.Y..S..Sk.YA.Ag....f 261
119963507 123        .S.y..LevlnkS.sY...yS..T.l...F.w.q.Y..A...r.YA.da.V.ra 176
119952436 148          .NA.L.QmTVS....F....MT..G.DF.N.t.LSS..e...Mxxxxxxxxx 199
116662150 126          .GA.L.amsVT....l..S.IT.Gp........Lq.A.r.xxxxxxxxxxx 177
116669124 121        .S.mA.LevlTeT.g..l.VS..T.T...F.w.q....A...r.YA.da.V.ra 174
166032791 263                    S.t....Na..A..StN-.k.V....St...IYA.Aa....S 303
167761642 255                    S......eQ..A..StN-.k.I.Y...i...IYA.Aa...VS 295
62425839  266        .E.AN.AkGyp.T......iH.IT..k-K..S.t..GVp....IrA..d.i.Vi 318
158338472 285       ..QMvr..VG.I.SN-....SH.If.T-.RF.A.t..G.pt.A.I.ASdS...iv 337
158340041 260       ..Q.vr..VG...SN-....iH.Vl.Gr-RF.A.t..G.pt.A.I.A.dS...iv 312
22299392  254  .lp.S.R.Ay.Iqa.L.-....W.iH.Il.rq-RF.A.V..G.df...I.A.ea...if 310
83943198  302     .iAaqKA.f...LkS......rD.kT.Gr-RM.S.t.....m...Ih.TAE.V..H 361
```

Figure 1. Alignment of a 'short' gene with its homologs in other species

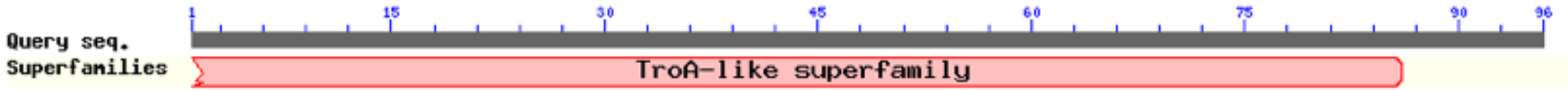


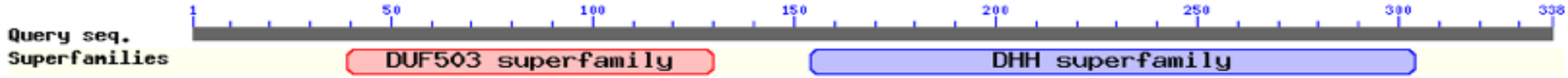
Figure 2. Truncation of a conserved functional domain as a result of a a gene being called short

Long genes

A gene is called 'long' when it has been called much longer than it should be at the 5'-end. In such a case, the gene is significantly longer than its homologs in other species as illustrated in Figure 3. A long gene can result in overlaps with neighboring features resulting in neighboring genes being called short or features being missed in the flanking intergenic regions.

```
QUERY      43  APTEVHERQAVLLQPRMAALSPHPGPAADLLQGCTMIDYAQALNEAQYEAATSGDGPVLV 102
120603850  1                                          ..EF.Ke..a.....V.Tl...... 25
46578573   1                                          ..EF.Ke..a.....V.Tl...... 25
78358610   1                                          ....kKe..a.......TlE..M.. 25
94987114   7                                              ..S..l..V.Ap...... 24
51246731   21                                             ..p..Hl.V.TtE..... 38
94265519   32                                           Eq..S..r..VcApp..I.. 51
94265519   510 ..A.Ll.QvlIyy..lLrr.y.Y.rrqr..e.lrpLLagyESMDDfla.l.lDpSS.eVd 571
94264521   32                                           Eq..S..r..VcApp..I.. 51
163726334  5                                         .Q.eKd...S....V.TlE..M.. 28
83816462   23                                        .E..E....q..a...A.k..L.I 46
116328254  1                                        MSWkEe..a..l..VlTqE..... 24
166833290  20                                        ...rt....e..a.VS.ep..A.. 43
74317324   20                                          .lSr..pe.rR.VVA....L.. 47
148652066  66                              .sA.hnSV.SS-SeV...ge..pS.ll...TtE.k... 102
24214893   2                                      .i.I.EWkEe..p..m..VlTl...... 31
```

Figure 3. Alignment of a 'long' gene with its homologs in other species



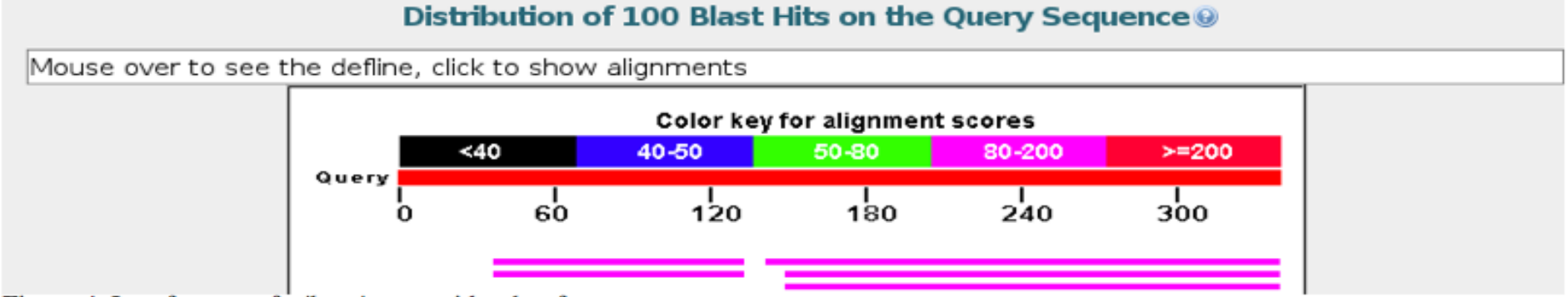


Figure 4. Interference of a 'long' gene with other features

Unique genes

A gene is called 'unique' when it has no known homologs in other species. Such a gene results in no hits when its amino acid sequence is compared using Blast to genes in other organisms. Often, such a gene has been called in error and results in other errors such as neighboring genes being called short.

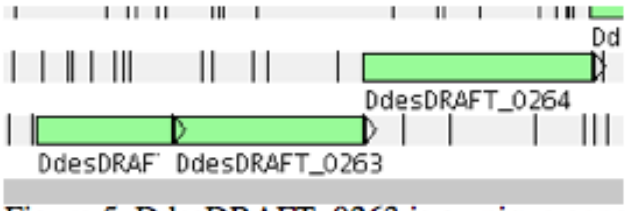


Figure 5. DdesDRAFT_0263 is a unique gene. If DdesDRAFT_0264 were detected as a short gene, DdesDRAFT_0263 would actually be responsible for this short call

Go to top

Dubious genes

A unique gene that is too short to actually result in a functional gene is classified as 'dubious'. In reality, very few (1-10) dubious genes are actually found among gene calls, but when found, unique and dubious genes can be included within intergenic regions to discover missing genes.

Broken genes

A gene is 'broken' into two or more parts when it is not called as a single gene, but as a series of smaller genes. Broken genes could result from gene prediction errors or frameshifts. Typically, the blast results of two parts of a broken gene have many hits in common. DdesDRAFT_1032 and DdesDRAFT_1033 were reported as parts of a broken gene by GenePRIMP (Figure 6). Figure 7. illustrates these genes in terms of their alignments.

Figure 6. Broken genes DdesDRAFT_1032 and DdesDRAFT_1033

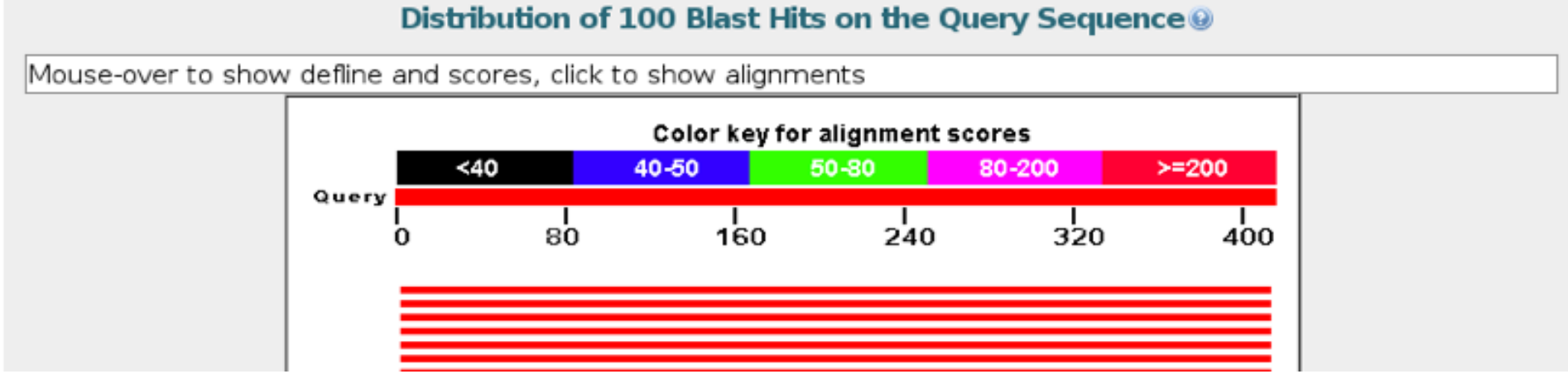


Figure 9. Predicted intergenic region

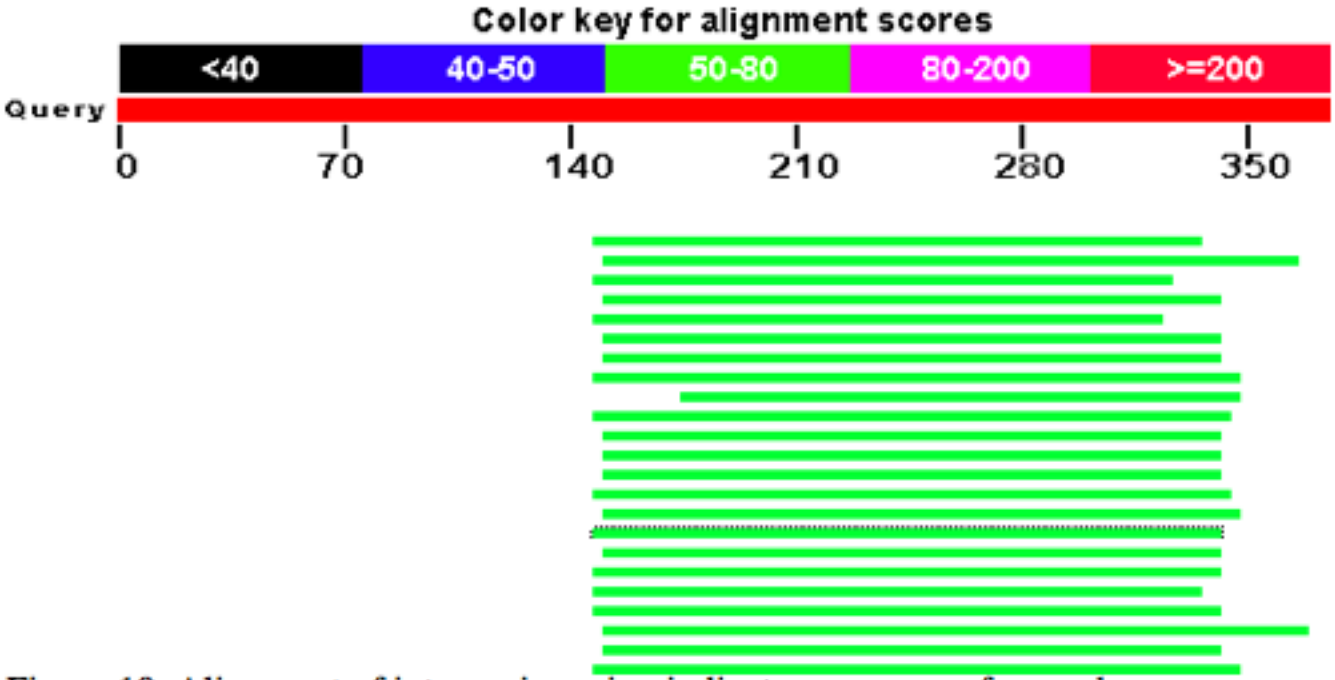


Figure 10. Alignment of intergenic region indicates presence of a good gene

Genomic Standards Consortium

4 Discussion

1. Minimal CDSs length (not sure if this is relevant) – JGI's standard is 15 aa. The shortest experimentally characterized protein in bacteria is 13 aa PatS – a diffusible peptide regulating heterocyst formation in Anabaena.

2. CDSs on top of rRNAs (excluding intron-containing RNAs) – JGI is deleting them, because there is no experimental evidence that these genes are real.

3. Pseudogene annotation:
   a) identification of all fragments
   b) merging the fragments to identify the correct coordinates of the gene feature
   c) CDSs composed of joined fragments and translations

JGI standard is to identify all fragments based on sequence similarity, merge them to generate translation and gene coordinates. Reasoning: evolutionary studies including protein family abundance, translation is useful because we're not always certain whether the gene is a pseudogene or not (e. g. frameshift fragments that produce a full-length translations when joined, gene fragments separated by repeat sequences that can be joined as a result of recombination; example is Anabaena, where nitrogenase expression is regulated by nitrogen fixation gene rearrangement).

Genomic
Standards
Consortium

Standards and the INSDC: Submission of MIGS/MIMS/MIENS

GSC8, September 9, 2009

Ilene Mizrachi
NCBI, NLM, NIH

- Culture and DNA sample availability
- Structured comment
- BioSamples Database
- Assembly User Object
- New Projects Redesign

Capturing Standards

- Two column metadata table in the COMMENT section of the FlatFile with customizable and expandable list of tags
- BioSamples Database – database which describes samples used in an experiment or series of experiments
- Fields within SC and/or BioSamples can be validated for MIGS/MIMS/MIENS compliance

The Structured Comment

LOCUS CP001688 3110487 bp DNA circular BCT 03-SEP-2009
DEFINITION Halomicrobium mukohataei DSM 12286, complete genome.
ACCESSION CP001688 ABTY01000001 ABTY01000002
VERSION CP001688.1 GI:257168392
DBLINK Project:27945

##Metadata-START##
Organism Display Name Halomicrobium mukohataei arg-2, DSM 12286
Culture Collection ID DSM 12286, ATCC 700874, JCM 9738, NCIMB 13541
GOLD Stamp ID Gi02248
Greengenes ID 245441
Funding Program DOE-GEBA 2007
Gene Calling Method Prodigal
Isolation Site Salinas grandes from Andes highlands in Jujuy Argentina
Collection Date 1991
Isolation Country Argentina
Latitude -22.66332
Longitude -66.23672
Depth Sea level
Oxygen Requirement Facultative
Cell Shape Rod-shaped
Motility Motile
Sporulation Nonsporulating
Temperature Range Mesophile
Temperature Optimum 45C
Salinity Halophile
Gram Staining gram-
Biotic Relationship Free living
Diseases None
Habitat Salt marsh, Soil
##Metadata-END##

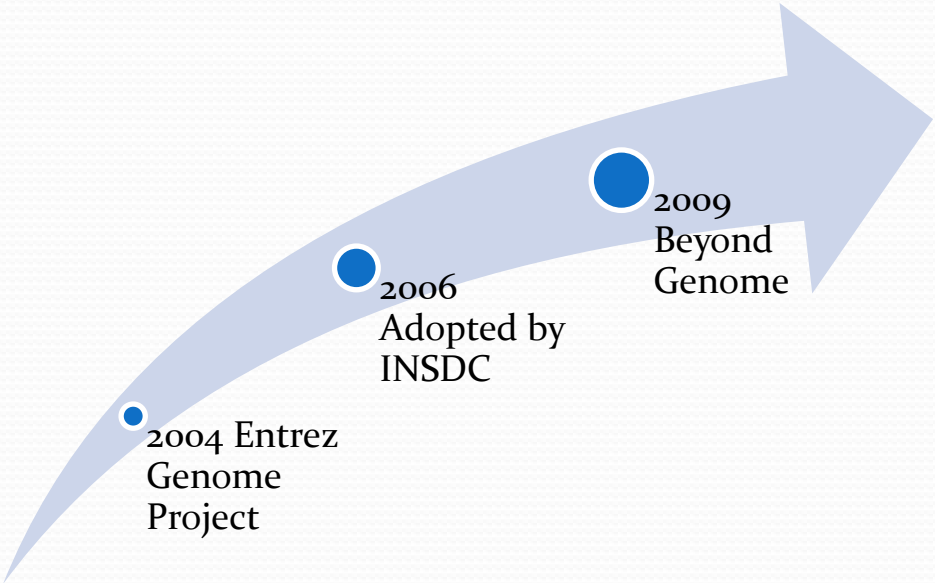
Assembly Details

- Structured Comment to capture details of the assembly

  ##Genome-Assembly-Data-START##
  Finishing Goal High Quality Draft
  Current Finishing Status Standard Draft
  Assembly Method Celera Assembler v. 5.1.1
  Assembly Name: EscCol_1.0
  Genome Coverage 15x
  Sequencing Technology 454
  ##Genome-Assembly-Data-END##

Genome Project to Project

Project



Genome Project

Project redesign

Problems:

- Biased to genome sequencing project
- Fixed Project types
- Parent-child relationships
- Batch submissions
- Database implementation
- Data pointers (accessions) stored in project database

Solutions:

- General project types with extendable attributes
- Flexible vertical and horizontal grouping (links)
- Generic submission spec for primary data archive
- New database XML schema
- Primary data resources point to Project

Genome Project

Definition: A collection of complete and incomplete (in-progress) large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms.

Required fields: Project type, Organism/Project name, Taxonomy id, Sequencing center and/or Submitting organization, Contact info

Project types: genome sequencing, assembly, annotation , map

Locus-tag prefix is required for every project

Project

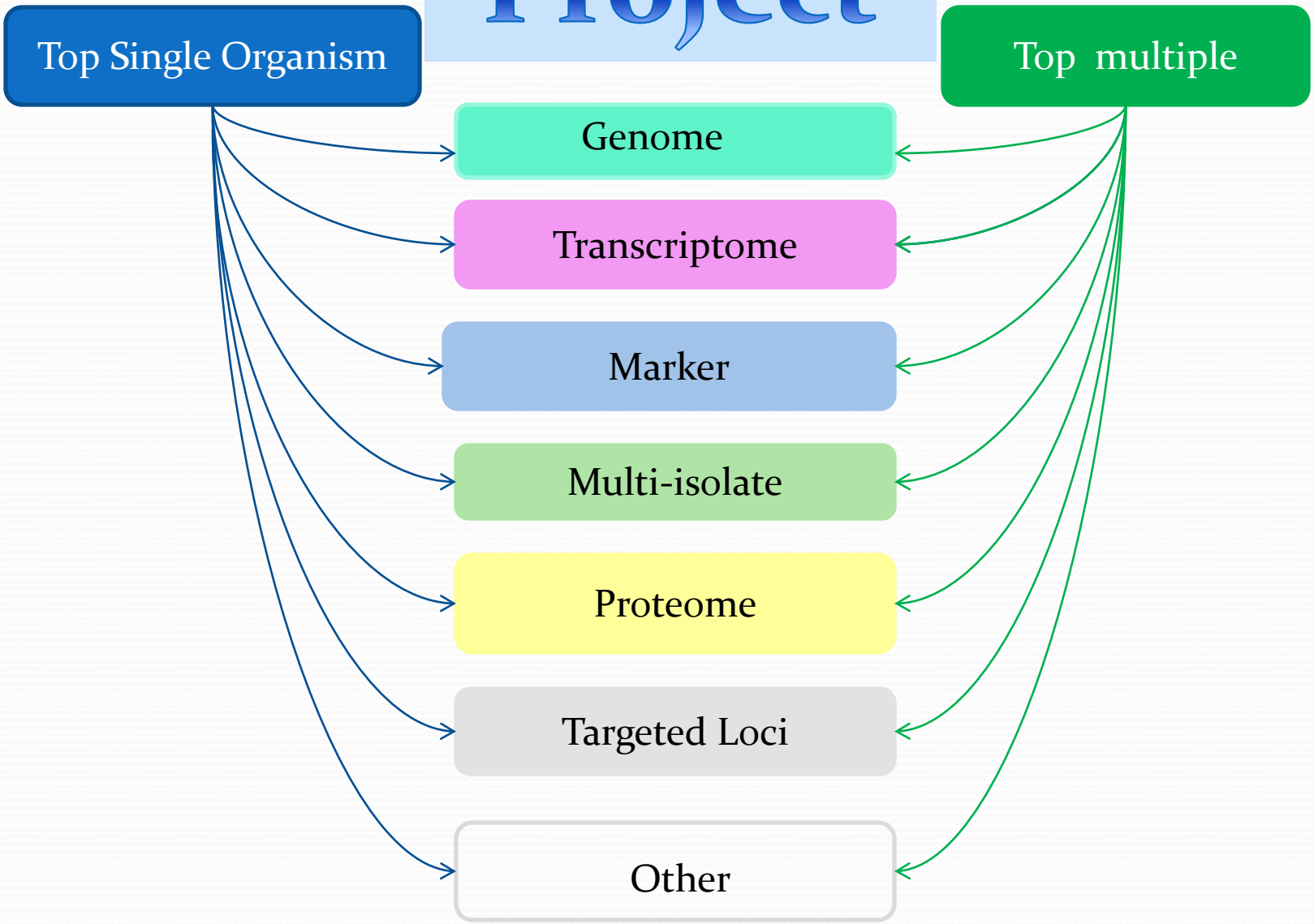
Definition: A collection of data or pointers to that data. (for example, a group of sequences, short reads or microarrays or projects).

Required fields: Project type, Organism/Project name, ~~Taxonomy id,~~ Sequencing center and/or Submitting organization, Contact Info, **Project Title, Project Description.**

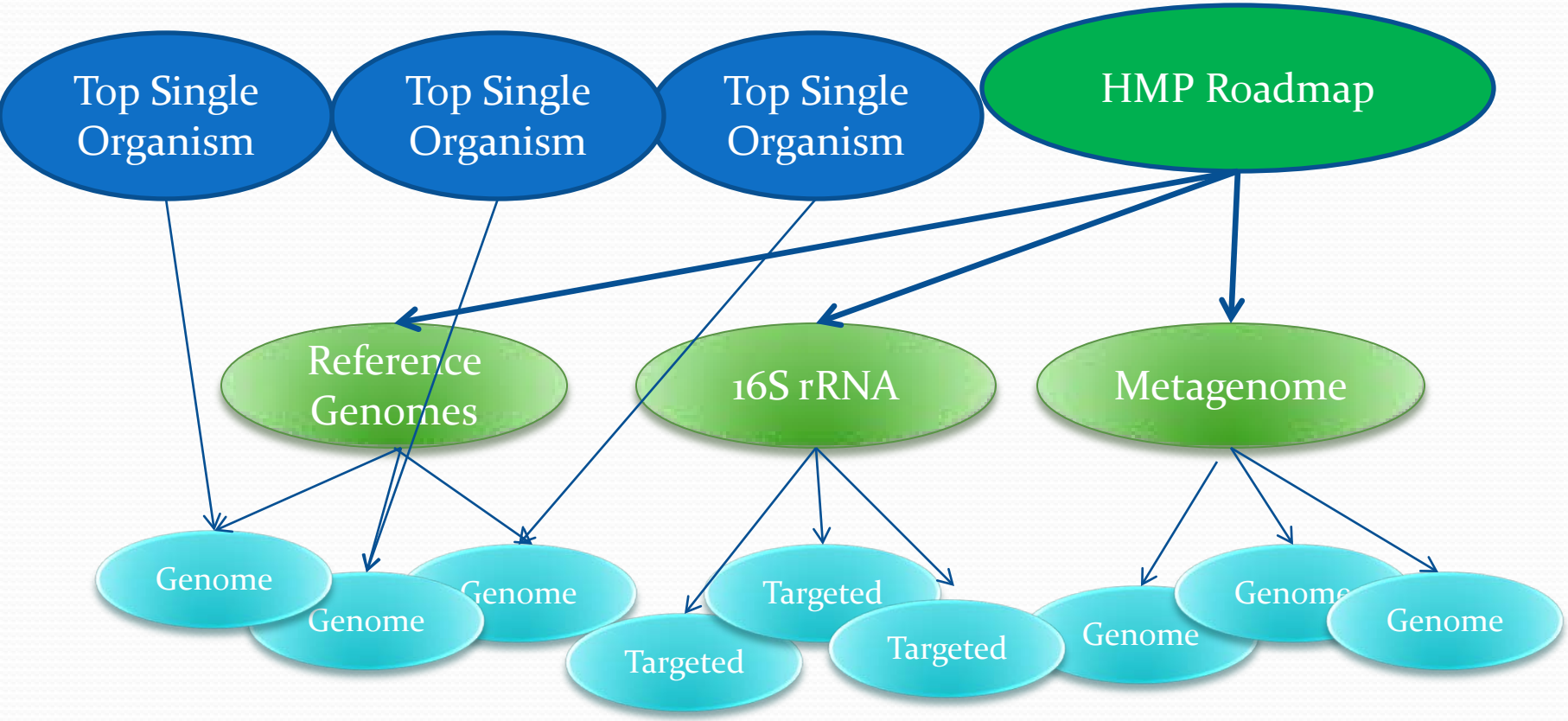
Project types: Genome, Transcriptome, Marker, Multi-isolate, Proteome, Targeted Loci, Other

Locus-tag prefix: is required for **Genome Project only.**

Project


Top Single Organism
Top multiple
Genome
Transcriptome
Marker
Multi-isolate
Proteome
Targeted Loci
Other

Human Microbiome Project



New Entrez Project database

PROJECT

Search Clear

Title: **Human Microbiome Project**
Project ID: 28331

Contributor: **NIH Human Microbiome Consortium**

Relevance: **Medical: Human health**

Description:
The human microbiome refers to the community of microorganism that live in or on the human body. The Human Microbiome Project will sequence the genomes of microoganisms that have been previously isolated from the human body. In addition, samples from the digestive tract, mouth, skin, nose, and female urogenital tract will be sequenced for 16S ribosomal RNA analysis and considered for metagenomic analysis.

Keywords: NIH Roadmap

Project Type: General Overview

+ Related Projects: 4331 sequencing projects

Publications:
1. **Human Microbiome Project. Science, 2009** More...

Related projects:
- Reference genomes
- 16S RNA sequencing
- Metagenomic WGS

Related resources:
- NIH HMP Roadmap
- DACC HMP

Links:
- Genome
- Nucleotide
- Protein
- Pubmed
- Taxonomy
- SRA
- dbGaP

TOWARDS A CONSENSUS ANNOTATION SYSTEM.

Genome Standards Consortium

Sept 2009

Owen White
Institute for Genome Science

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?

No food fights.

How can we evaluate annotation data?

…and why would we want to.

Consensus annotation pipeline for the HMP project

Who: the DACC and the sequencing centers (Broad, JCVI, WashU, Baylor)
What: take the best-of-the-best components and build a consensus pipeline
How: evaluate existing pipelines, develop metrics, evaluate components

Existing pipelines:
IGS
RAST
JCVI
JCVI with BioName
IMG
Broad
RefSeq

Metrics:
- Consistency within a source
  - Compare set of genes that belong to defined family across all test organisms
- Consistency between sources
  - Compare same genes across sources
  - Counts of annotation elements
- Comparison with "gold standard" reference annotation
  - Computationally with distance scores
  - Manual review of subset

Pipeline components:

- Pairwise search databases
  - UniProt
  - FigFams
  - eggNOG
  - CDD
  - Priam
  - TCDB
  - Char
  - PRK
- HMMs
  - Pfam
  - TIGRFAM
  - Panther
- Motifs
  - PROSITE
  - TMHMM
  - SignalP
  - LipoP

Establish a ranking hierarchy and decision tree:
Which datasource ranks highest? What cut-off criteria should be used? Can annotation elements from different sources be combined for one protein?

Test organisms:
- *Bacteroides intestinalis* 341, DSM 17393
- *Clostridium leptum* DSM 753
- *Rhodobacter sphaeroides* 2.4.1
- *Staphylococcus aureus* N315
- *Bacillus anthracis* Ames
- *Burkholderia mallei* ATCC 23344
- *Bacillus subtillis* 168
- *Escherichia coli* K12 substr MG1655

Mixture of Gram +\-, high/low GC, some draft, some with "gold

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

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

Source Data	Common name		Gene symbol	
	Completeness	Consistency	Completeness	Consistency
BHB	99.99%	45.14%	81.59%	75.54%
ERIC	99.99%	38.51%	32.37%	54.03%
NMPDR	99.99%	93.26%	33.06%	47.68%
Pathema	99.99%	90.84%	84.88%	79.29%
Patric	99.99%	90.79%	90.06%	91.92%
SwissProt	100.00%	90.80%	98.83%	91.66%
HAMAP	100.00%	92.[illegible]	99.11%	92.73%
IMG	92.[illegible]	[illegible]	0.00%	NA
RefSeq	97.[illegible]	51.[illegible]	54.21%	50.25%
Genbank	95.[illegible]	35.[illegible]	48.84%	43.61%
Subsystems	99.99%	93.15%	0.00%	NA
CMR	99.[illegible]	75.34%	84.82%	78.50%
KEGG	97.[illegible]	97.83%	58.82%	51.95%

GO assignments

Source Data	Completeness	Consistency
BHB	86.89%	50.11%
HAMAP	98.25%	94.40%
ERIC	81.10%	36.05%
SwissProt	97.90%	64.63%
NMPDR	51.74%	98.60%
Patric	81.39%	55.03%
Pathema	90.10%	64.63%
CMR	77.99%	53.05%
Genbank	91.96%	NA
IMG	0.00%	NA
Subsystems	0.00%	NA
RefSeq	77.99%	53.05%
KEGG	91.96%	NA

Sort

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	

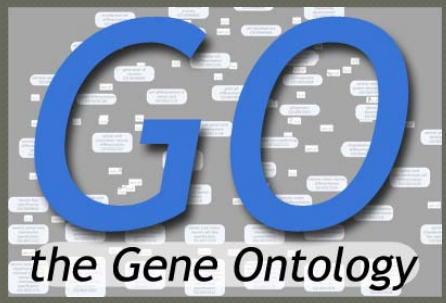
Gene 1 - Final			
	Data	Quality	
Source 3	ISS	Best	
Source 2	ICL	Best	
Source 1	IEA	Best	

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.

Special thanks

Heather Huot
Michelle Gwinn
and HMP Annotation working group



Sam Angiuoli
Aaron Gussman
and Bioinformatics Resource Centers IOWG

NIAID
Bioinformatics
Resource Center

Subject: RE: CAFAE (name optional) is a go

From: Gregurick, Susan [mailto:Susan.Gregurick@science.doe.gov]
Sent: Friday, August 21, 2009 1:57 PM
To: White, Owen; Drell, Daniel
Cc: Gregurick, Susan
Subject: CAFAE (name optional) is a go

Hi Owen,

It was just fabulous to see you last week at the exascale workshop and I appreciate your contribution to this effort. I am sure that the ASCR office will get the materials they need to move forward with their exascale initiative and I hope that our programs get enough to also move forward in our own computational biology/bioinformatics initiative.

I had a nice chat with my colleagues at NIH and I believe we now have an interagency funded initiative to move forward on the CAFAE focus/scoping group. We can't call this a workshop due to the way we are now doing business at DOE. I'm sending to you the last written document on this Critical Assessment of Functional Annotation Experiment (CAFAE) focus group for you to consider. BTW, CAFAE name is my idea and you are most welcome to alter this in any way.

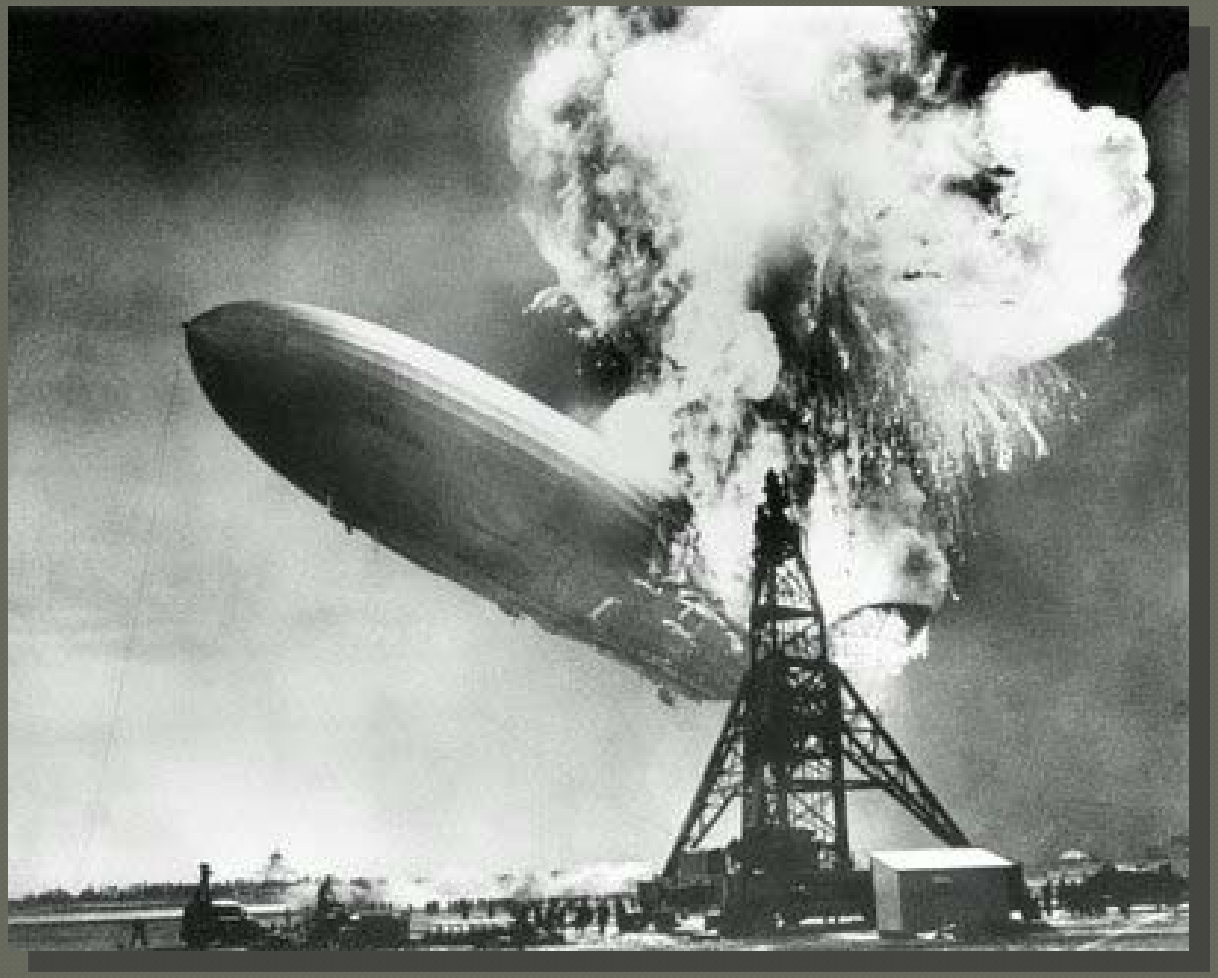
I believe that we will want to formulate the following:

*Timeframe—we believe December or January is best for the meeting, but you will know your schedule and the actual time a community of people can convene.
*Location---feds have limited budgets, so DC area is better for us, but you may have a different idea, its up to you.
*Invitees---when you visited we drafted an initial list, I have more names from NIH, but you will know the people you wish to have there.
*Charges—we drafted these in this enclose document. Are these the questions we really need addressed. Let's bat this around.

In early September perhaps a conference call between yourself, Nikos (?), DOE and NIH can be made to go over these bulleted points.

Future Direction

- Need general rules of engagement
- Need quality metrics and gold standards
- Ev code and assertion types buy-in?



!?

NERC Environmental Bioinformatics Centre

NATURAL ENVIRONMENT RESEARCH COUNCIL

The National Center for Toxicological Research (NCTR)

Investigation / Study / Assay

ISA infrastructure

standards and tools for managing experimental metadata

Susanna-Assunta Sansone, Philippe Rocca-Serra, Eamonn Maguire, Marco Brandizi, Natalyia Sklyar, Chris Taylor

and ISA contributors/collaborators

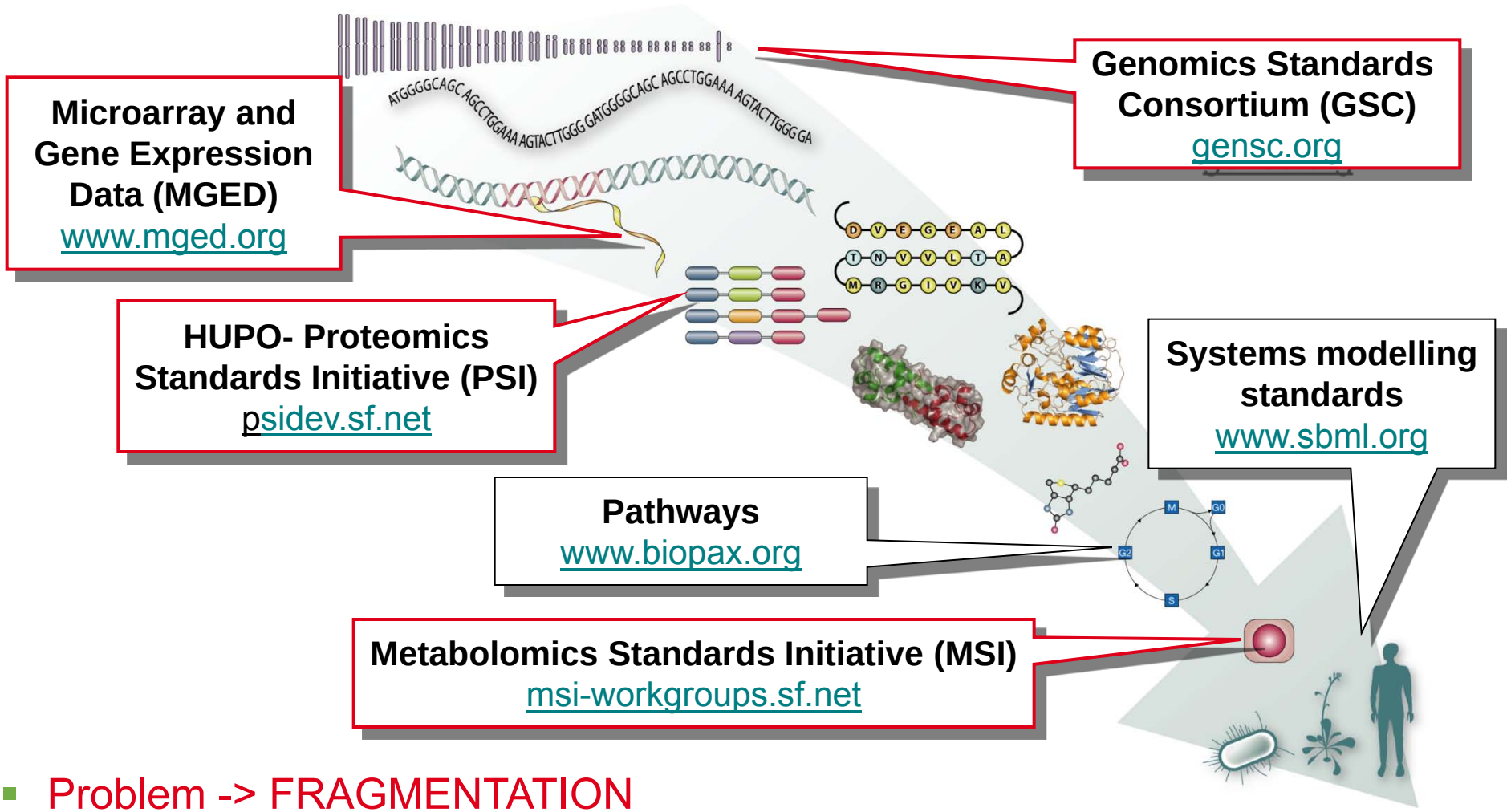
http://isatab.sf.net

GSC meeting, Walnut Creek, JGI, 9th - 11th September 2009

Experiments growing in size and complexity

- Metagenomics and metatranscriptomics study looking at the effect of ocean acidification on phytoplankton and bacterioplankton by characterizing/measuring
  - Genomic sequence by pyrosequencing technology
  - Gene expression by pyrosequencing technology
    - Gilbert et al *PloS ONE*, 2008

Grass root omics initiatives (*de facto* standards), e.g.:

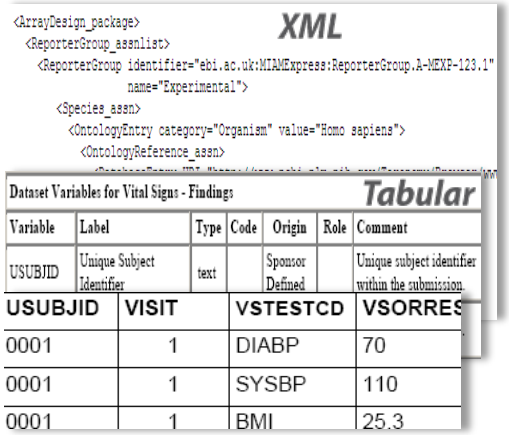


- Problem -> FRAGMENTATION
  - Being focused on particular technology/domain leads to **duplication of effort** and the development of **different standards**, severely **hindering data integration**

Synergistic efforts to overcome fragmentation

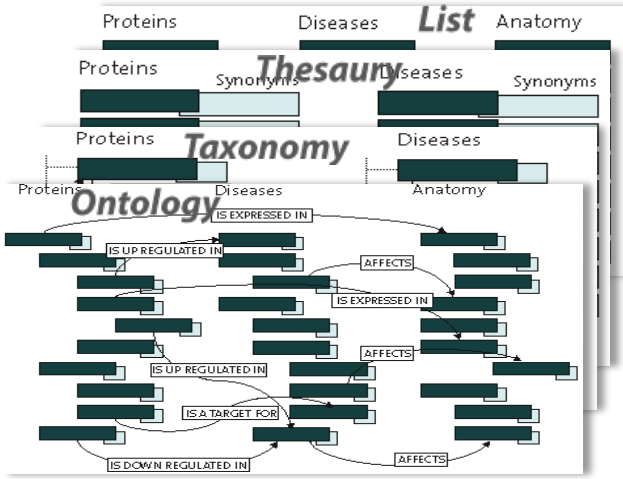
Scope

minimal information to be reported



Syntax

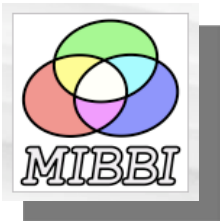
format(s) for the communication



Semantics

terminology(s) for the description

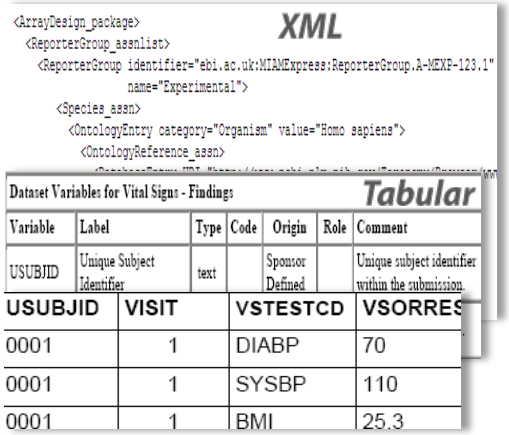
Synergistic efforts to overcome fragmentation

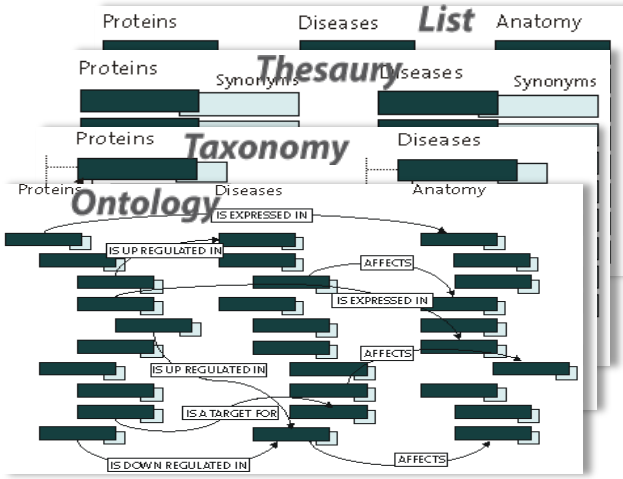


Scope

minimal information to be reported

MIBBI: http://mibbi.org





Syntax

format(s) for the communication

Semantics

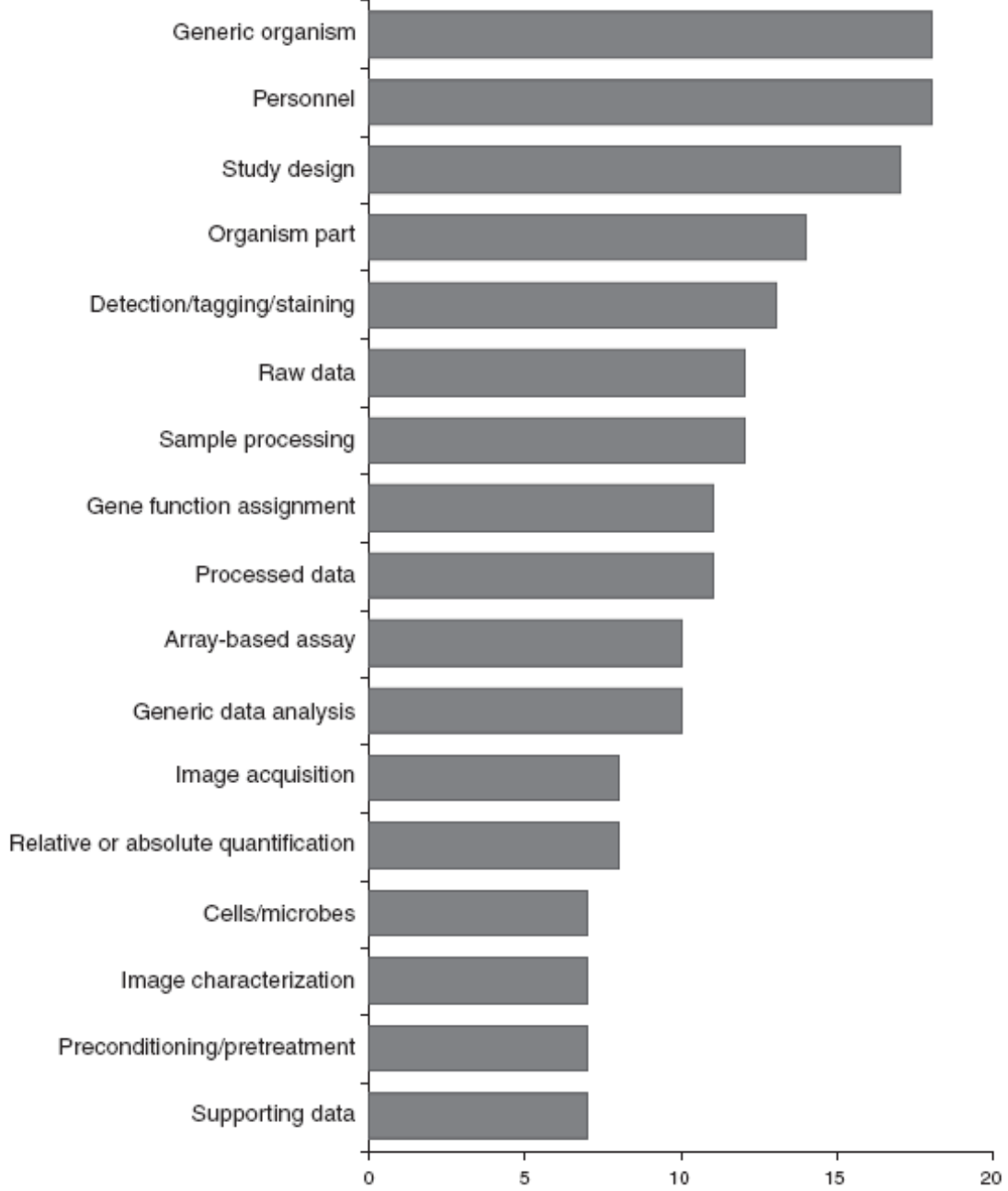
terminology(s) for the description

CIMR	Core Information for Metabolomics Reporting
MIABE	Minimal Information About a Bioactive Entity
MIACA	Minimal Information About a Cellular Assay
MIAME	Minimum Information About a Microarray Experiment
MIAME/Env	MIAME / Environmental transcriptomic experiment
MIAME/Nutr	MIAME / Nutrigenomics
MIAME/Plant	MIAME / Plant transcriptomics
MIAME/Tox	MIAME / Toxicogenomics
MIAPA	Minimum Information About a Phylogenetic Analysis
MIAPAR	Minimum Information About a Protein Affinity Reager
MIAPE	Minimum Information About a Proteomics Experimen
MIARE	Minimum Information About a RNAi Experiment
MIASE	Minimum Information About a Simulation Experimen
MIENS	
MIFlowCyt	
MIGen	
MIGS	
MIMIx	
MIMPP	
MINI	
MINIMESS	
MINSEQE	
MIPFE	
MIQAS	
MIqPCR	
MIRIAM	
MISFISHIE	
STRENDA	
TBC	

Projects/MIGS

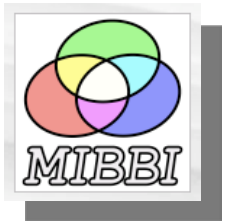
Minimum Information about a

1	
1.1	Domain
1.2	Document Type
1.3	Group
1.4	Main Website
1.5	MI Checklist's Name
1.6	MI Checklist's Acronym
1.7	Current Version Designation
1.8	Release Date for Current Versio
1.9	General Comments



Highest ranked common concepts across the checklists

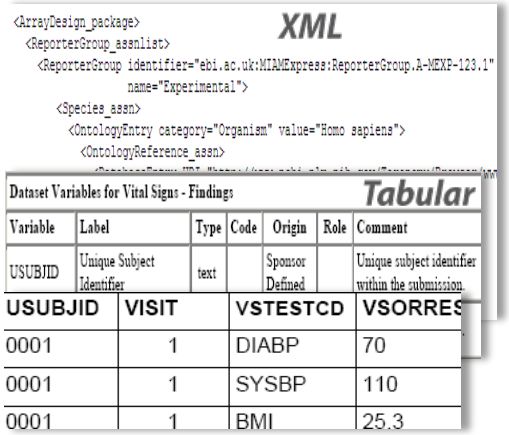
Synergistic efforts to overcome fragmentation



Scope

minimal information to be reported

MIBBI: http://mibbi.org



Syntax

format(s) for the communication

OBO

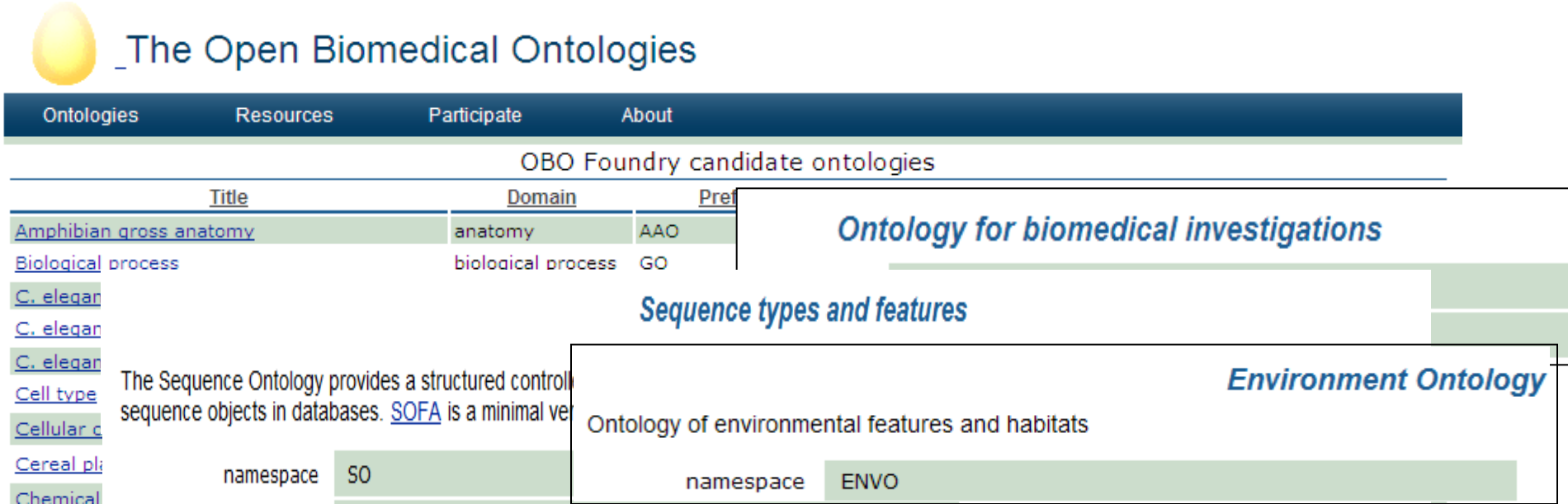
Semantics

terminology(s) for the description

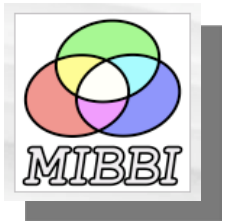
OBO foundry: http://obofoundry.org

OBO Foundry: *Nat Biotech* 2007

- Create a suite of orthogonal and interoperable ontologies
  - establish a set of principles for ontology development
  - overcome the different degree of completeness and quality
  - give attribution at all levels- due credits- and establish metrics
- The *Portal* includes ~80 different freely available ontologies
  - the candidate members of the *Foundry* will ultimately provide with interoperable, orthogonal, well structured ontologies



Synergistic efforts to overcome fragmentation



Scope

minimal information to be reported

MIBBI: http://mibbi.org

Syntax

format(s) for the communication

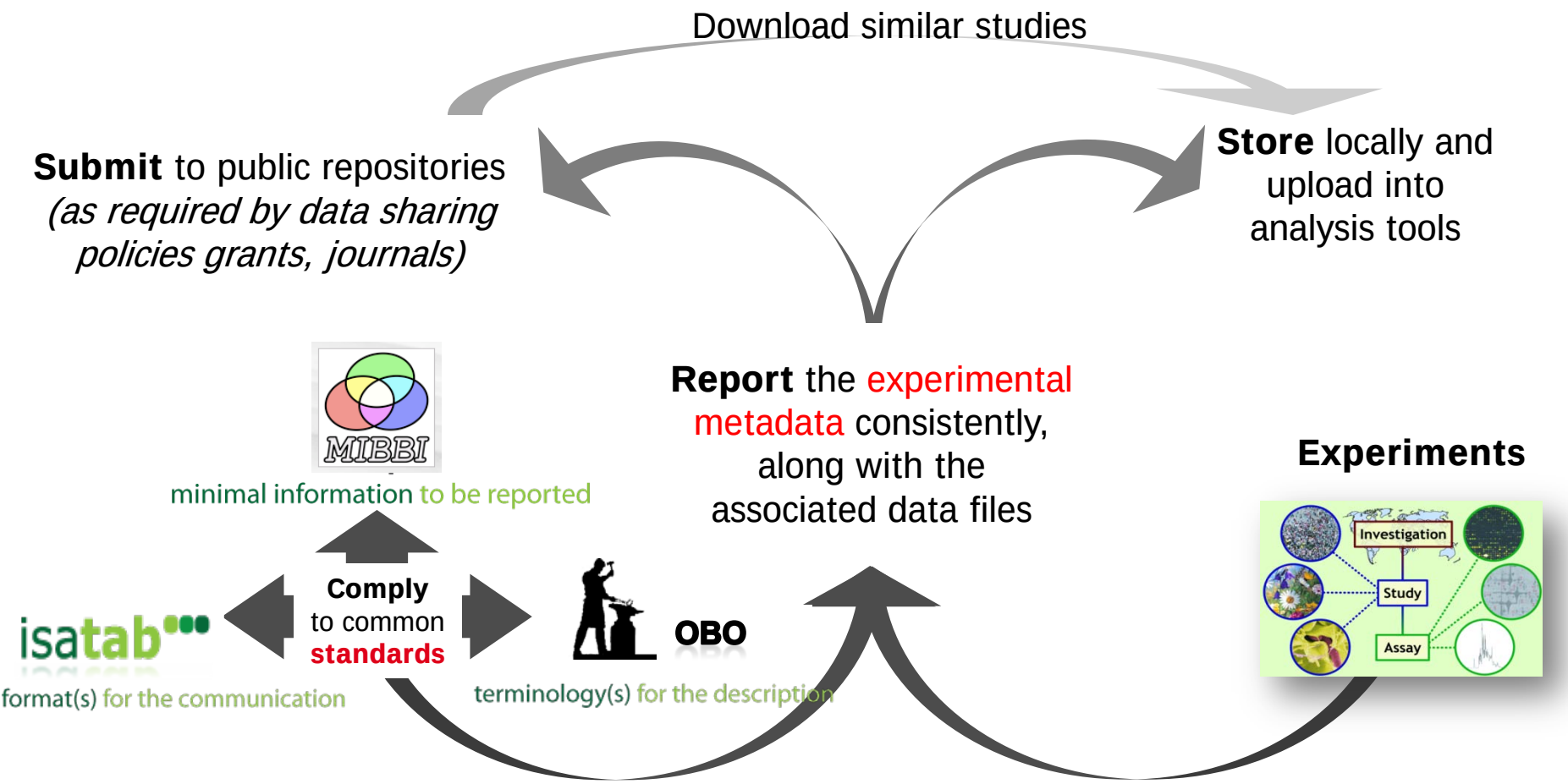
ISA-tab: http://isatab.sf.net

Semantics

terminology(s) for the description

OBO foundry: http://obofoundry.org

From theory to practice: tools for the community



Java standalone components, for **local installation** that can work **independently**, *or* as **unified system**

From theory to practice: tools for the community

Download similar studies

Submit to public repositories *(as required by data sharing policies grants, journals)*

Store locally and upload into analysis tools

Report the experimental metadata consistently, along with the associated data files

POWERUSER

isatab

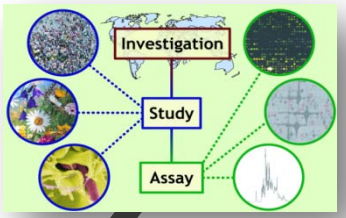
ontologies

BioPortal & OLS

mibbi

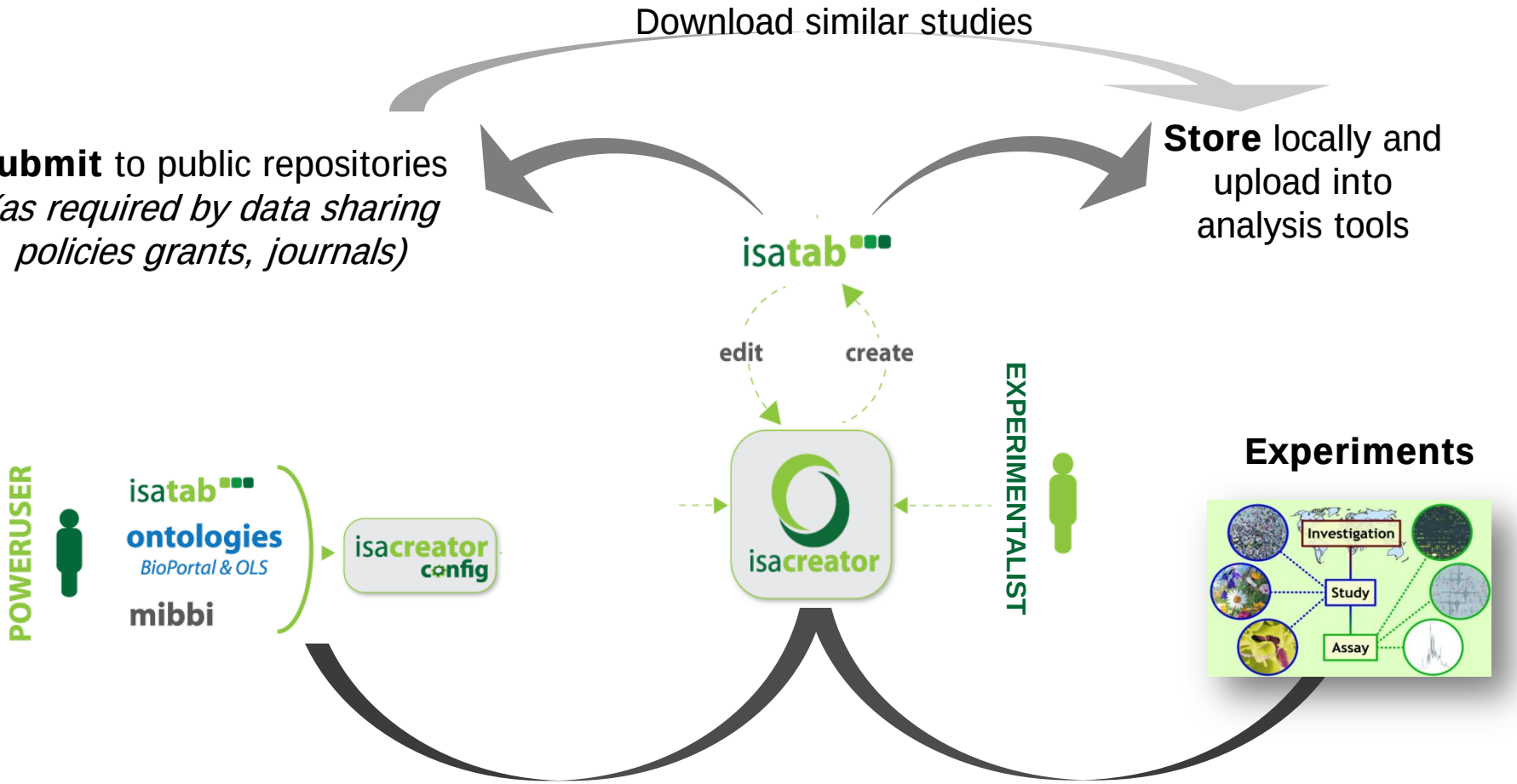
isacreator config

Experiments



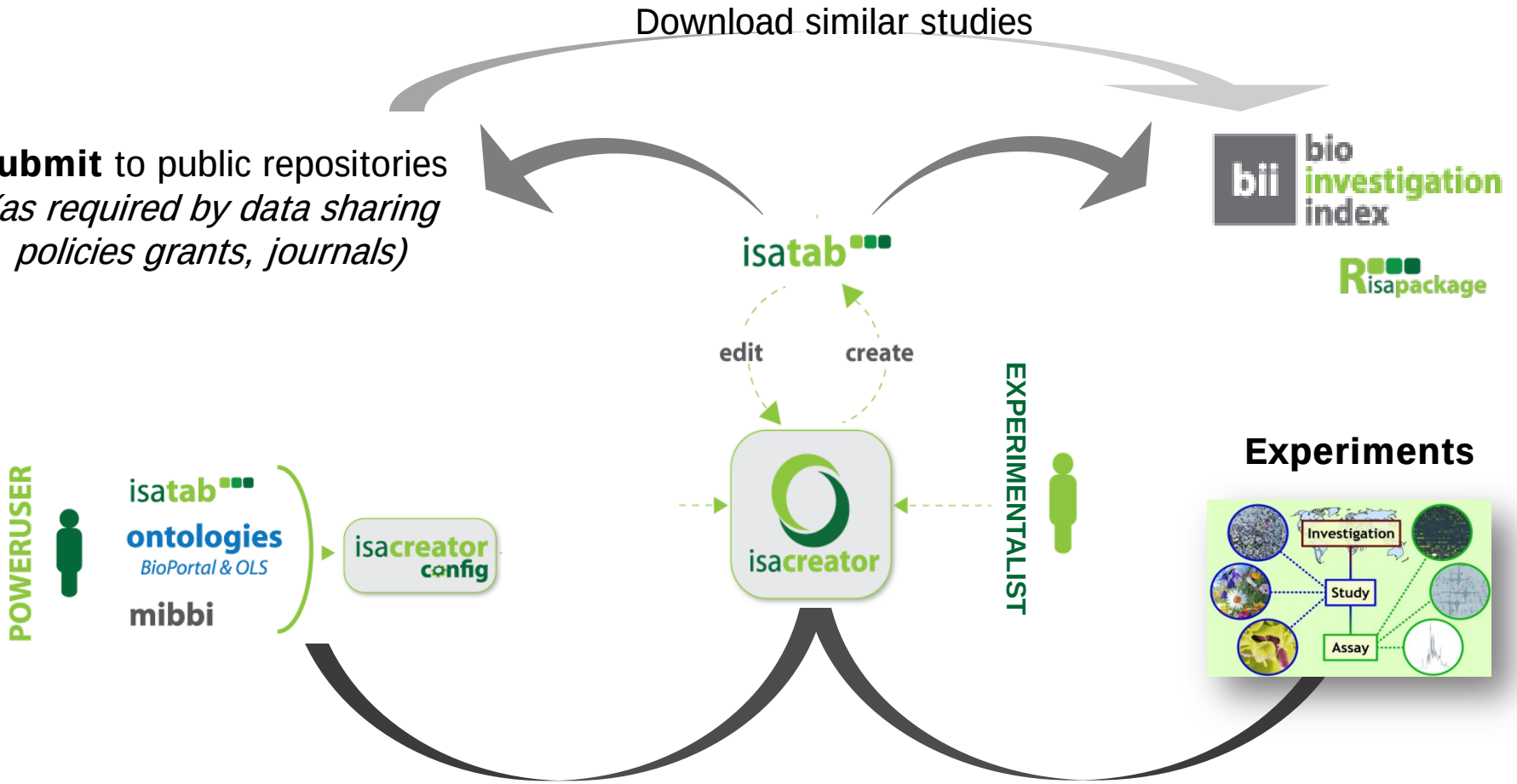
Java standalone components, for **local installation** that can work **independently**, *or* as **unified system**

From theory to practice: tools for the community



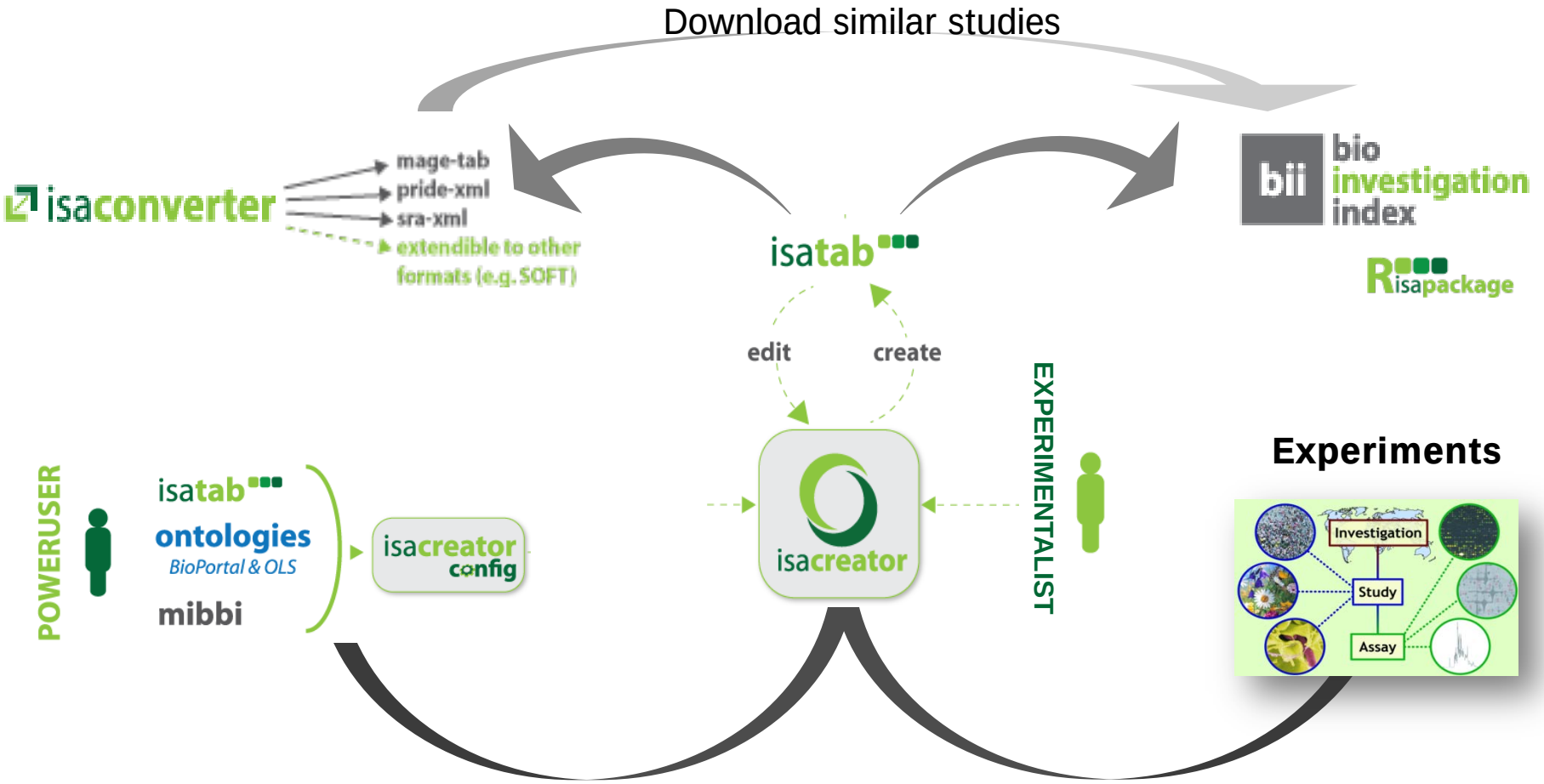
Java standalone components, for **local installation** that can work **independently**, *or* as **unified system**

From theory to practice: tools for the community



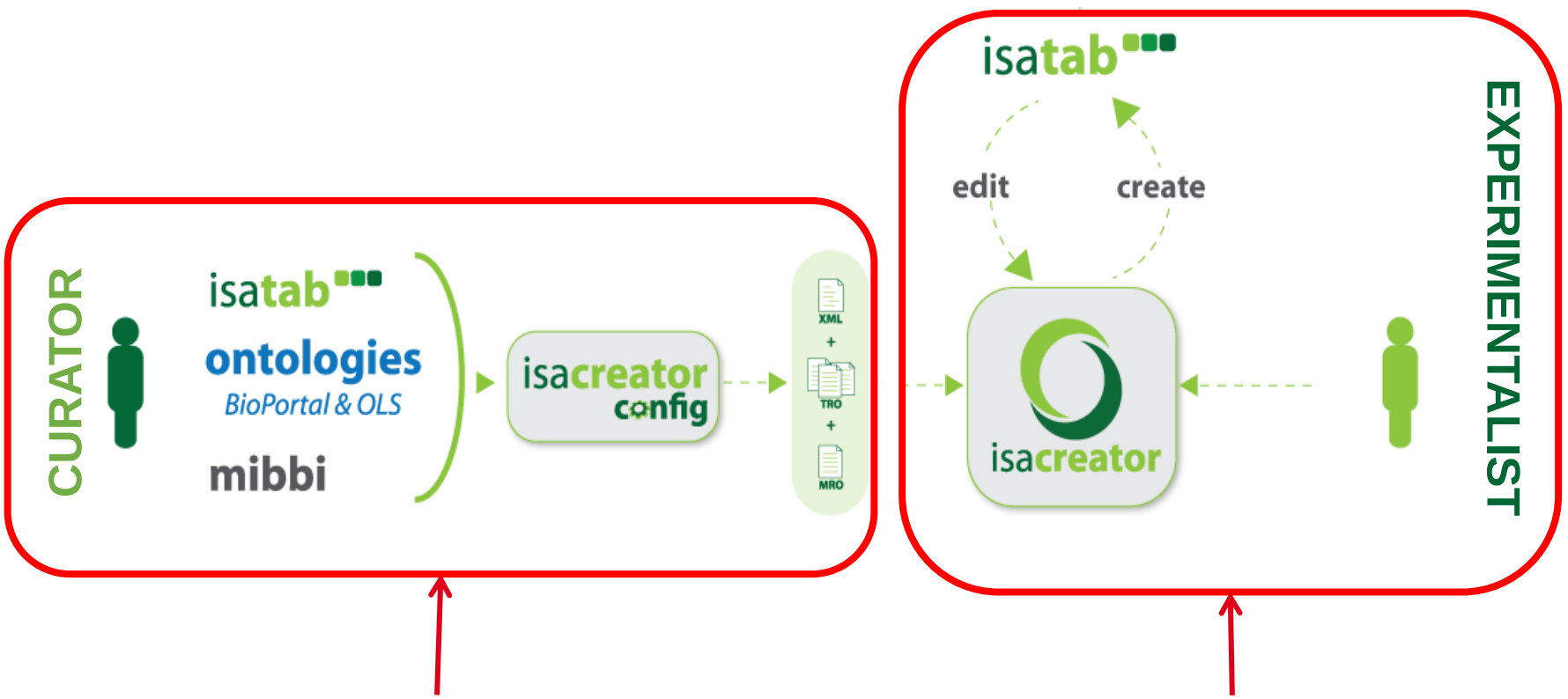
Java standalone components, for **local installation** that can work **independently**, *or* as **unified system**

From theory to practice: tools for the community



Java standalone components, for **local installation** that can work **independently**, *or* as **unified system**

isacreator config to configure the editor tool



Target users for this tool:

curator (power user) assisting experimentalists in the reporting process

Target users for this tool:

Experimentalists, often unaware of what requirements to meet and terms to use

Projects/MIGS

Minimum Information about a Genome Sequence

- **Study**
  - **Environment**
    - Geographic location (latitude and longitude (float (point, transect and region)), depth and altitude of sample (integer)
    - Time of sample collection (UTC)
    - Habitat
    - **MIMS extension: select to report a set of uniform measurements for a given habitat:**
      - Water body: (temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production) (integer, unit)
  - **Nucleic acid sequence source**
- **Assay**
  - **Sequencing**
    - Nucleic acid preparation (extraction method(CV); amplification(CV)
    - Library construction (library size (integer), number of reads sequenced (integer), vector(CV)
    - Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony) (OBI)
    - Assembly (assembly method (CV), estimated error rate (unit) and method of calculation (CV))
    - Finishing strategy (status—e.g., complete or draft (CV), coverage (integer), contigs (integer))
    - Relevant Standard Operating Procedures (SOPs)
    - Relevant electronic resources

Oisacreator config

Environment

Environment description[1]

- Environment type.
- Local feature of interest.
- Specific entity of interest.

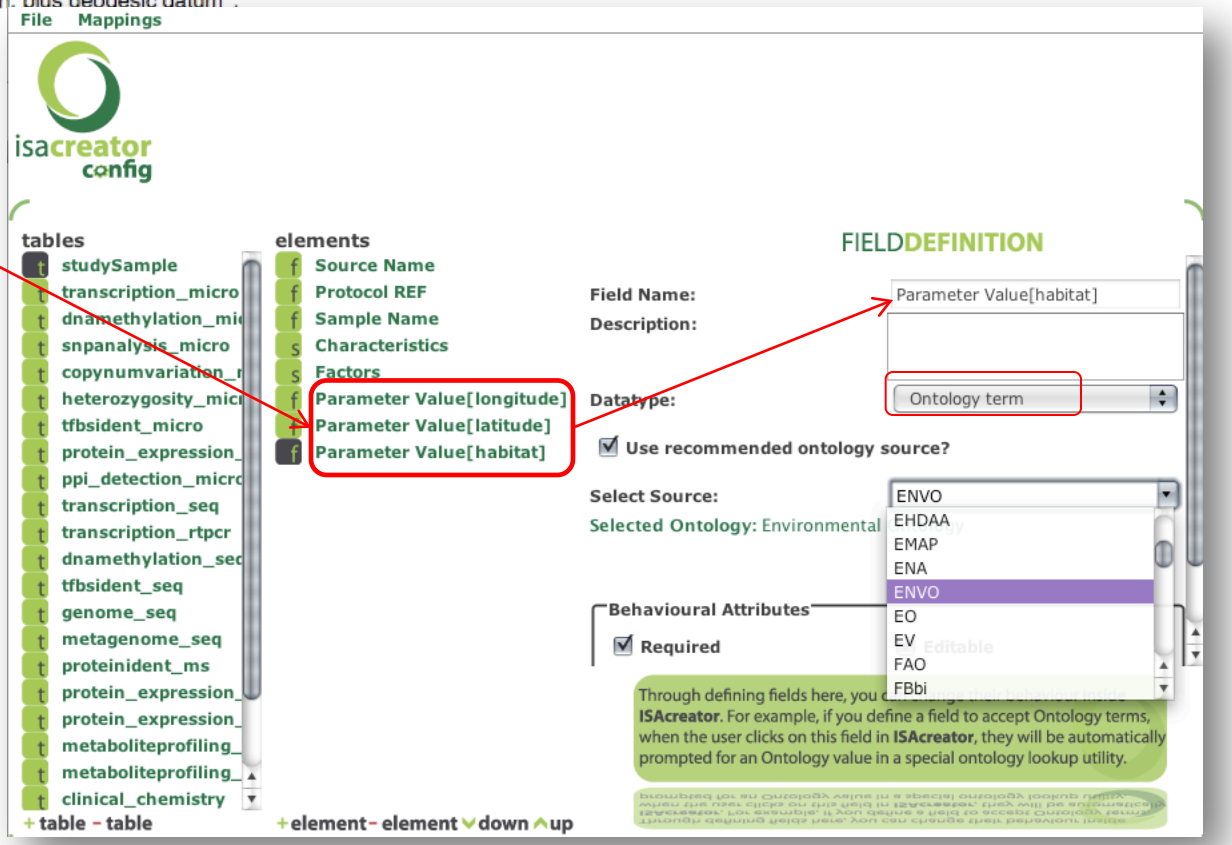
Environment location

- Whether the location is a point, transect or area.
- Latitude(s) and longitude(s) appropriate to identify the location, plus geodesic datum[2].
- Altitude(s) or depth(s) appropriate to identify the location.

Projects/MIGS

MIBBI

Minimum Information about a Genome Sequence



❶ Identify mandatory fields (e.g. following minimal requirements set by the community)

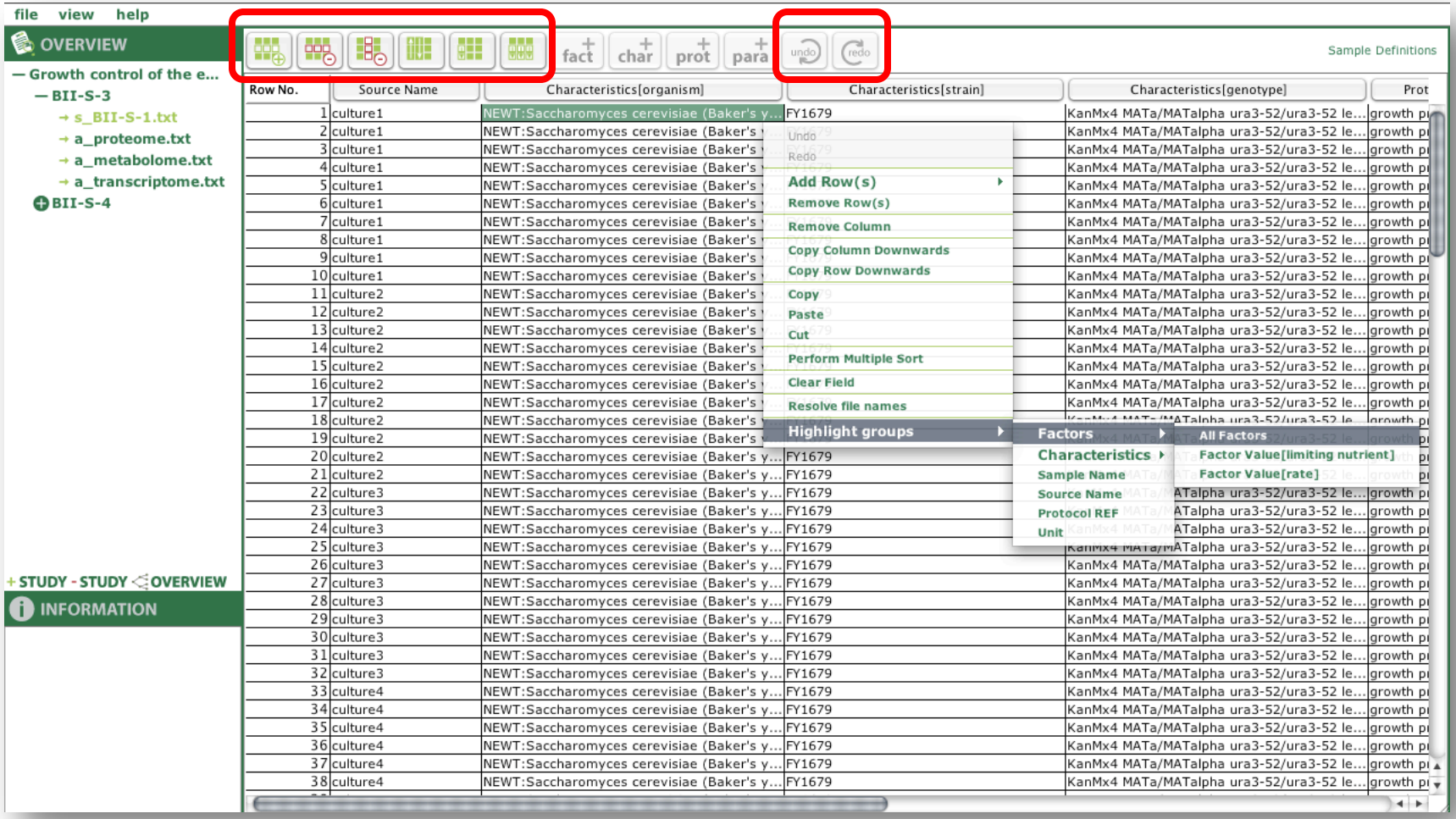
❷ Define type of value allowed (e.g. ontology term)

Example of restriction of the search to the EnvO ontology,

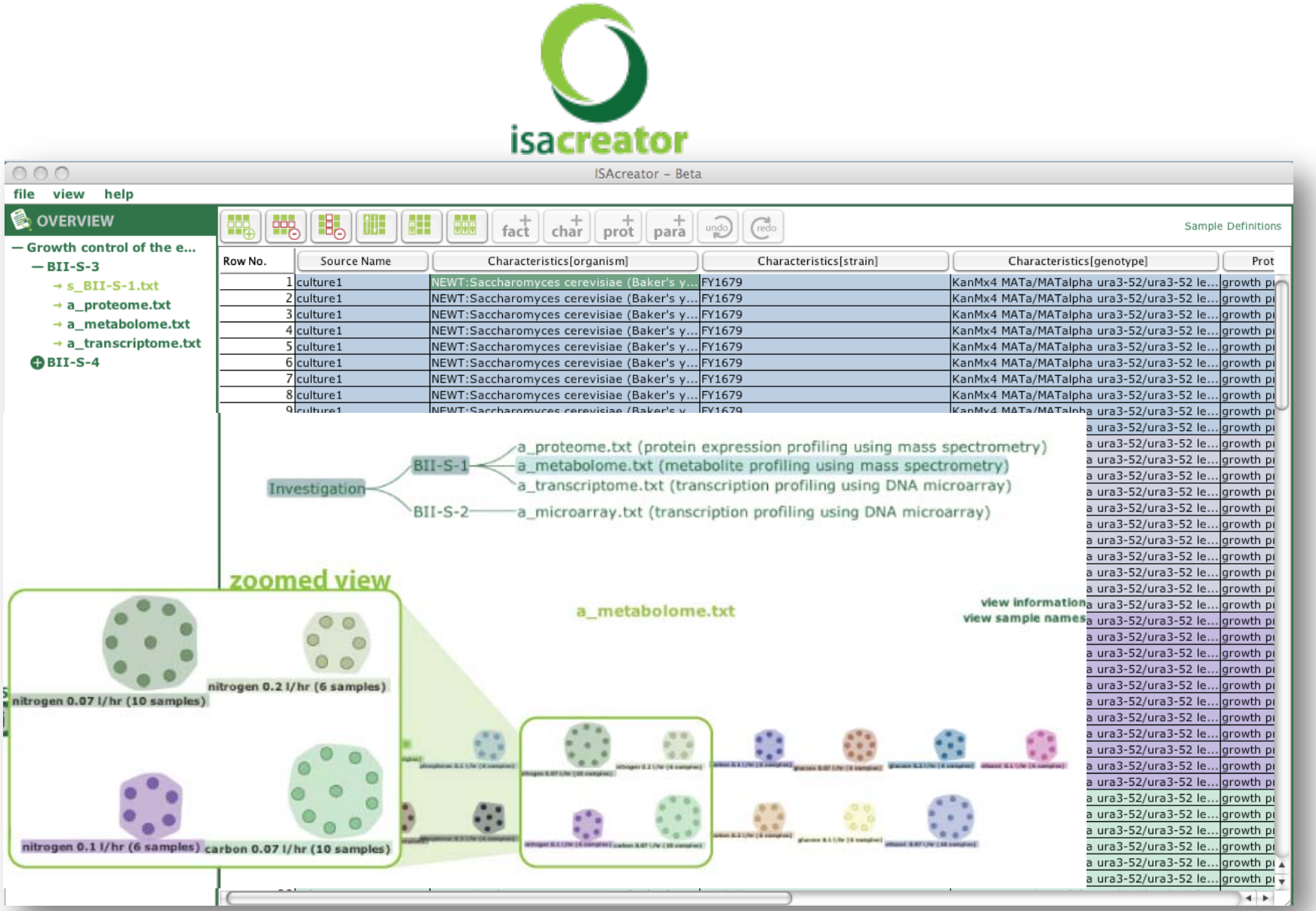
for the describing the ‘habitat’ of a sample

Ontologies, accessed in real time via the Ontology Lookup Service and BioPortal

isacreator



Spreadsheet functionalities, including: move, add, copy, paste, undo, redo and right click options



Groups of samples are colour coded

bio investigation index public instance deployed @ EBI

search

_freetext

_organism: Filter on organisms

_measurement: Filter on measuremen

_technology: Filter on technolog

_platform: Filter on Platform

clearfields searchindex

browsestudies

6 public studies containing 282 assays

Investigation	Study				Assay		
Acc	Acc	Title	Organism	Factor	Measurement	Technology	#
BII-I-1	BII-S-2	A time course analysis of transcription response in yeast treated with rapamycin, a specific inhibitor of the TORC1 complex: impact on yeast growth	Saccharomyces cerevisiae (Baker's yeast)	compound, exposure time, dose	transcription profiling	DNA microarray	14
BII-I-1	BII-S-1	Study of the impact of changes in flux on the transcriptome, proteome, endometabolome and exometabolome of the yeast Saccharomyces cerevisiae under different nutrient limitations	Saccharomyces cerevisiae (Baker's yeast)	limiting nutrient, rate	protein expression profiling transcription profiling metabolite profiling	mass spectrometry DNA microarray mass spectrometry	3 48 111
	BII-S-6	The Influence of Pharmacogenetics on Fatty Liver Disease in the Wistar and Kyoto Rats: A Combined Transcriptomic and Metabonomic	Rattus norvegicus (Rat)	time, compound, strain	transcription profiling metabolite profiling	DNA microarray NMR spectroscopy	17 79
	BII-S-3	Metagenomes and Metatranscriptomes of phytoplankton blooms from an ocean acidification mesocosm experiment	marine metagenome	compound, dose, collection time	metagenome sequencing transcription profiling	nucleotide sequencing nucleotide sequencing	4 4
	BII-S-4	An initial characterisation of the Fasciola hepatica transcriptome using 454-FLX sequencing	Fasciola hepatica (Liver fluke)		transcription profiling	nucleotide sequencing	1
	BII-S-5	Determination of the complete genome sequence of Salmonella paratyphi A str. AKU_12601	Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601		genome sequencing	nucleotide sequencing	1

Acknowledgements

Groups and individuals participating in:

MIBBI http://mibbi.org

OBO Foundry http://obofoundry.org

ISA-Tab format http://isatab.sf.net

isainfrastructure team:

Marco Brandizi

Eamonn Maguire

Nataliya Sklyar

Chris Taylor

Manon Delahaye

Richard Evans

Philippe Rocca-Serra

Susanna-Assunta Sansone

http://isatab.sf.net

carcinoGENOMICS a Project of the European Union

NuGO The European Nutrigenomics Organisation

FDA Food and Drug Administration

The National Center for Toxicological Research (NCTR)

NEBC

NERC Environmental Bioinformatics Centre

Natural Environment Research Council

University of Cambridge,
Jules Griffin

CNRS,
Magali Roux

BC Cancer Research Center,
Ryan Brikman

University of Bordeaux,
Antoine Daruvar

Leibniz Institute Plant Biology,
Steffen Neumann

Southwestern Medical Center,
Richard Scheuermann

Harvard Medical School,
Win Hide, Oliver Hofmann

GSC community, in particular:

Dawn Field

Peter Sterk

Curation of MIGS compliant data: Towards 1000(s) of genomes and metagenomes

Peter Sterk

NERC Centre for Ecology and Hydrology/
Visitor of the Wellcome Trust Sanger Institute

This project has received funding from NIEeS and a NERC International Opportunities Award (NE/3521773/1) 2005-2008)

Outline

- GSC, MIGS/MIMS, gcdml
- Curation of Sanger pathogens and other projects
  - Approach, tools
- Evaluation of this curation effort
  - Ontologies and controlled vocabularies
- Incorporation of MIGS/MIMS data into INSDC records
- Evaluation of ISA-creator as a curation tool

The Genomic Standards Consortium

- The goal of this international community is to promote mechanisms that standardize the description of genomes and the exchange and integration of genomic data
- MIGS/MIMS Standard and MIGS Checklist
  - descriptions of different taxa (metagenomes, eukaryotes, prokaryotes, viruses, plasmids, and organelles) and concepts (Organism, Environment, Phenotype, Sample Processing, and Data Processing).
- …

http://gensc.org/

Checklist version 2.0

Nat. Biotechnol.26(5):541-547(2008)

Investigation	Report type EU	BA	PL	VI	OR	ME
• Submit to trace archives and INSDC	M	M	M	M	M	M
• Investigation type (i.e., report type)	M	M	M	M	M	M
• Project name[2]	M	M	M	M	M	M
• Study						
• Environment						
• Geographic location (latitude and longitude[float (point, transect and region)], depth and altitude of sample)[(integer)]	M	M	M	M	M	M
• Time of sample collection[(UCT)]	M	M	M	M	M	M
• Habitat[EnvO]	M	M	M	M	M	M
MIMS extension: select to report a set of uniform measurements for a given habitat:						M
• Water body: (temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production)[(integer, unit)]						
• Nucleic acid sequence source						
• Subspecific genetic lineage (below lowest rank of NCBI taxonomy, which is subspecies) (e.g., serovar, biotype, ecotype)[(CABRI)]	M	M	M	M	M	–
• Ploidy (e.g., allopolyploid, polyploid)[(PATO)]	M					
• Number of replicons (EU, BA: chromosomes (haploid count); VI: segments)[(integer)]	M	M	–	M	–	–
• Extrachromosomal elements[(integer)]	X	M				
• Estimated size (before sequencing; to apply to all draft genomes)[(integer; base pairs)]	M	X	X	X	X	–
• Reference for biomaterial (primary publication if isolated before genome publication; otherwise, primary genome report)[(PMID or DOI)]	X	M	X	X	X	X
• Source material identifiers: (cultures of microorganisms: identifiers[(alphanumeric)] for two culture collections[(OBI)]; specimens (e.g., organelles and Eukarya): voucher condition and location[(CV)])	M	M	M	M	M	M
• Known pathogenicity		M		M		
• Biotic relationship (e.g., free-living, parasite, commensal, symbiont)[(OBI)]	X	M		X		
• Specific host (e.g., host taxid, unknown, environmental)[EnvO]	X	M	M	M		
• Host specificity or range[(taxid)]	X	X	X	M		
• Health or disease status of specific host at time of collection (e.g., alive, asymptomatic)[PATO]		M		M		
• Trophic level (e.g., autotroph, heterotroph)[PATO]	M	M	–	–	–	–
• Propagation (phage: lytic or lysogenic; plasmid: incompatibility group)[(CV)]	M		M	M	–	–
• Encoded traits (e.g., plasmid: antibiotic resistance; phage: converting genes)[(CV; see caption)]		X	M	M		X
• Relationship to oxygen (e.g., aerobic, anaerobic)[PATO]		M	–	–	–	–
• Isolation and growth conditions[(PMID or DOI)]	M	M	M	M	M	M
• Biomaterial treatment (e.g., filtering of sea water)[(OBI)]						M
• Volume of sample[(integer)]						M
• Sampling strategy (enriched, screened, normalized)[(CV)]						M
• Assay						
• Sequencing						
• Nucleic acid preparation (extraction method[(CV)]; amplification[(CV)]	M	M	M	M	M	M
• Library construction (library size[(integer)], number of reads sequenced[(integer)], vector[(CV)]						M
• Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)[(OBI)]	M	M	M	M	M	M
• Assembly (assembly method[(CV)], estimated error rate[(unit)] and method of calculation[(CV)])	M	M	M	M	M	M
• Finishing strategy (status—e.g., complete or draft[(CV)], coverage[(integer)], contigs[(integer)])	M	M	X	X	X	X
• Relevant Standard Operating Procedures (SOPs)	M	M	M	M	M	M
• Relevant electronic resources	M	M	M	M	M	M

MIGS: Investigation

- ⊙ **Submit to trace archives and INSDC**
- ⊙ **Investigation type (i.e., report type)**
- ⊖ **Project name**
  - ⊖ **Study**
    - ⊕ **Environment**
    - ⊕ **Nucleic acid sequence source**
  - ⊖ **Assay**
    - ⊕ **Sequencing**

http://www.mibbi.org/index.php/Projects/MIGS:Summary_list
Thanks to Chris Taylor @ EBI

http://gensc.org/

What is gcdml?

- GCDML is implemented using XML Schema.
- GCDML aims to take full advantage of the benefits of an XML representation of genomic contextual data.
- XML provides a machine readable representation of metadata that facilitates the capture, exchange and comparison of large amount of data.
- XML is widely used to build data capture and exchange formats.
- The MIGS/MIMS checklist has been implemented in gcdml by Renzo Kottmann.
- The schema is extensible and allows for the capture of a richer dataset if required.

Checklist and xml schema defined; next step: curation of genome metadata

- Start with finished and published bacterial and archaeal genomes
  - 800 curated genomes from GOLD (thanks Nikos!)
- Initial focus on 42 Sanger Institute bacterial pathogens
  - Further curation from literature
- Creation of MIGS/MIMS reports in gcdml
- Feedback to GOLD
- Curation of more (including eukaryotic) pathogens with help from Sanger curators

Step 1: batch conversion of GOLD data into gcdml

22	23	24	25	26	27	28	29	30	
RELEVANCE	DISEASE	HABITAT	PH	OXYGEN REQUIREMENTS	CELL SHAPE	CELL ARRANGEMENT	MOTILITY	SPORULATION	ENERGY S
Biotechnological	None	Marine		Aerobe	Sphere-shaped				Photosynt
Biotechnological	None	Host, Soil, Wastewater		Facultative	Coccus-shaped		Nonmotile		
Environmental, Bioremediation, Biotechnological	None	Fresh water, Acid mine drainage, Hy		Aerobe	Rod-shaped		Motile		Heterotro
Bioleaching, Bioremediation, Environmental	None	Acid mine drainage	2	Obligate anaerobe	Rod-shaped		Motile	Nonsporulating	Obligate c
Bioleaching, Bioremediation, Environmental	None	Acid mine drainage	1.3 - 4.0	Obligate anaerobe	Spiral-shaped		Motile	Nonsporulating	Obligate c
Biofuels, Energy production, Ethanol production,	None	Fresh water, Hot spring		Aerobe				Sporulating	
Agricultural, Plant Pathogen	Bacterial fruit blotch	Host		Aerobe	Rod-shaped		Motile	Nonsporulating	
Biotechnological	None	Soil		Aerobe	Rod-shaped		Motile	Nonsporulating	
Human Pathogen, Medical	Nocosomial infection, Pneumonia	Fresh water, Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Human Pathogen, Medical	Meningitis, Pneumonia, Septicemia	Fresh water, Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Medical, Human Pathogen	Nosocomial infection, Pneumonia	Fresh water, Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Human Pathogen, Medical	Nosocomial infection, Pneumonia	Fresh water, Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Human Pathogen, Medical	Nosocomial infection, Pneumonia	Fresh water, Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Medical	Nosocomial infection	Host		Aerobe	Rod-shaped	Singles	Nonmotile	Nonsporulating	Chemohet
Medical, Human Pathogen	Nosocomial infection	Soil, Fresh water, Food, Human skin		Aerobe	Rod-shaped		Nonmotile	Nonsporulating	Heterotro
Animal Pathogen, Medical, Swine Pathogen	Porcine pleuropneumonia	Host, Respiratory tract		Facultative	Rod-shaped	Chains, Pairs, Singles	Nonmotile		
Animal Pathogen, Medical, Swine Pathogen	Necrotizing pleuropneumonia	Host, Respiratory tract		Facultative	Rod-shaped	Pairs, Singles	Nonmotile		
Animal Pathogen, Medical, Swine Pathogen	Necrotizing pleuropneumonia	Host, Respiratory tract		Facultative	Rod-shaped	Chains, Pairs, Singles	Nonmotile		
Biotechnological, Succinic-acid production	None	Bovine rumen, Host		Facultative	Rod-shaped				

Gold data in spreadsheet

This was the easy part!

```
<?xml version="1.0" encoding="UTF-8"?>
<NasReports xmlns="http://gensc.org/gcdml" xmlns:xsi="http://www.w3.org/
    xsi:schemaLocation="http://gensc.org/gcdml http://gensc.sf.net/ns/gc

    <MIGSReports>
        <prokaryote gcatID="migs41" sourceName="migs38" sourceVersion="1
            <ncbiOrganismName>migs08 migs09</ncbiOrganismName>
            <ncbiTaxId>migs42</ncbiTaxId>
            <genomeProjectID>migs43</genomeProjectID>
            <studyData>
                <projectName></projectName>
            </studyData>
            <originalSample>
                <organismalMaterial>
                    <samplingTime>
                        <instance></instance>
                    </samplingTime>
                    <location>
                        <name>migs34</name>
                        <lat></lat>
                        <lon></lon>
                    </location>
                    <habitat>
migs23
                    </habitat>
                    <name>
                        <genus>migs05</genus>
                        <species>migs06</species>
                        <strain>migs09</strain>
```

gcdml template

```
}elsif ($data[30] =~ /chemoheterotroph/) {
    $data[30] = "heterotroph, chemoorganotroph";
}elsif ($data[30] =~ /chemoautotroph, chemolithotroph/ || $data[30] =~ /chemoautotroph/) {
    $data[30] = "autotroph, chemolithotroph";
}elsif ($data[30] =~ /chemoorganotroph/) {
    $data[30] = "heterotroph, chemoorganotroph";
}
$data[6] =~ s/^\w+ //;
$data[12] =~ s/x//gi;
$data[11] =~ s/sanger/dideoxysequencing/i;
$data[11] =~ s/454/pyrosequencing/;
$data[11] =~ s/,/ and/;

my $spacer = "                    ";
my $indent = "    ";
my $indent2 = "        ";
if ($data[23] =~ /marine/ || $data[23] =~ /water/ || $data[23] =~ /ponds/) {
    $data[23] = $spacer . $indent . "<aquatic>\n$spacer$indent2" . "<waterBody>\n$spacer$indent2$i
    $data[23] .= $spacer . $indent2 . "</waterBody>\n$spacer$indent" . "</aquatic>";
} elsif ($data[23] =~ /food/ || $data[23] =~ /milk/ || $data[23] =~ /dairy/ || $data[23] =~ /wine/
        || $data[23] =~ /beer/ || $data[23] =~ /meat/) {
    $data[23] = $spacer . $indent . "<food></food>";
} elsif ($data[23] =~ /host/ || $data[23] =~ /intestinal/) {
    $data[23] = $spacer . $indent . "<hostAssociatedHabitat></hostAssociatedHabitat>";
} elsif ($data[23] =~ /plants/) {
    $data[23] = $spacer . $indent . "<plantAssociatedHabitat></plantAssociatedHabitat>";
} elsif ($data[23] =~ /soil/) {
    $data[23] = $spacer . $indent . "<soil></soil>";
} elsif ($data[23] =~ /oil fields/ || $data[23] =~ /solar saltern/ || $data[23] =~ /solfataric fie
                || $data[23] =~ /hydrothermal vent/ ) {
    $data[23] = $spacer . $indent . "<extremeHabitat></extremeHabitat>";
} elsif ($data[23] =~ /rhizosphere/) {
    $data[23] = $spacer . $indent . "<soil></soil>";
} elsif ($data[23] =~ /hot spring/ || $data[23] =~ /sediment/ || $data[23] =~ /sludge/) {
    $data[23] = $spacer . $indent . "<data[23]></$data[23]>";
    $data[23] = $spacer . $indent . "<soil></soil>";
```

Perl script gold2gcdml

800 gcdml files

http://gensc.org/

Step 2: editing of gcdml files

- Curation from literature, culture collections and other resources
- Editing gcdml files in oXygen XML editor
- This is time-consuming!
- Assigning latitute-longitude in Google Earth
- Curation of 42 Sanger pathogens finished to high standard
- About 750 genomes have had a first-pass

```
<?xml version="1.0" encoding="UTF-8"?>
<NasReports xmlns="http://gensc.org/gcdml" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://gensc.org/gcdml http://gensc.sf.net/ns/gcdml/1.7.0/base/gcdml.xsd">
    <MIGSReports>
        <prokaryote gcatID="001223_GCAT" sourceName="not_assigned" sourceVersion="1">
            <ncbiOrganismName>Mycobacterium marinum M</ncbiOrganismName>
            <ncbiTaxId>216594</ncbiTaxId>
            <genomeProjectID>16725</genomeProjectID>
            <studyData>
                <projectName>Mycobacterium marinum M project at Wellcome Trust Sanger Institute</projectName>
            </studyData>
            <originalSample>
                <organismalMaterial>
                    <samplingTime>
                        <instance>1992</instance>
                    </samplingTime>
                    <location>
                        <name>Moffet Hospital, University of California, San Francisco, USA</name>
                        <lat>37.77</lat>
                        <lon>-122.46</lon>
                    </location>
                    <habitat>
                        <animalAssociatedHabitat></animalAssociatedHabitat>
                    </habitat>
                    <name>
                        <genus>Mycobacterium</genus>
                        <species>marinum</species>
                        <strain>M</strain>
                        <typeStrain>false</typeStrain>
                    </name>
                </organismalMaterial>
            </originalSample>
            <isolate>
                <isolationConditions>
                    <doi>10.1101/gr.075069.107</doi>
                </isolationConditions>
                <cultureCollection>
                    <identifier>ATCC BAA-535</identifier>
```

http://gensc.org/

Evaluation of this effort

- Not always easy or possible to find all data items in literature
  - Data not reported or ambiguous
  - Referenced papers hard to obtain (mostly 'old' publications)
- Missing terms in ontologies/MIGS controlled vocabularies
- Evaluation of the need for certain MIGS fields (proposals to add or delete certain fields)
- Not always obvious how to represent data. Problem cases can be documented and discussed on the GSC wiki: http://gensc.org/gc_wiki/index.php/MIGS_Problem_Cases

Addressing checklist, ontologies, CVs is crucial

- Accurate definitions of all terms and incorporation of terms in relevant ontologies promotes consistent curation among different groups as well as interoperability between studies
- Adhere as much as possible to approved terms

A series of wiki pages for MIGS terms

(http://gensc.org/gc_wiki/index.php/GSC_Ontology_Terms)

- I have made a start listing and defining terms suggested in the MIGS/MIMS checklist
- I have proposed new terms where needed
- everyone can add or change terms and definitions or comment

Environment

* Ontology-Habitat: completed subject to verification

Nucleic acid sequence source

* Ontology-Subspecific Genetic Lineage: completed subject to verification
* Ontology-Ploidy: completed subject to verification
* Ontology-Trophic Level: completed subject to verification
* CV-Propagation
* CV-Encoded Traits
* Ontology-Biotic Relationship: completed subject to verification
* Ontology-Specific Host: completed subject to verification
* Ontology-Health or Disease status: completed subject to verification
* Ontology-Oxygen Relation: completed subject to verification
* Ontology-Biomaterial Treatment: input from metagenomics community desired
* CV-Sampling Strategy

Sequencing

* Ontology-Sequencing Method
* CV-Assembly
* CV-Finishing Strategy

http://gensc.org/

Trophic level ontology wiki page

Term (1st level)	Term (2nd level)	Term (3rd level)	Term (4th level)	Term (5th level)	Definition	(List of) ontologies and IDs (or propose ontology)	Status	Comments	Proposer if applicable
trophic level	autotroph				organism that produces complex organic compounds from simple inorganic molecules using energy from light or inorganic chemical reactions (Wikipedia).	PATO	proposed		sterk
trophic level	heterotroph				organism that requires organic substrates to get its chemical energy for its life cycle (Wikipedia)	PATO	proposed		sterk
trophic level	facultative autotroph				An organism that can make organics can from inorganic carbon or obtain organics produced by other organisms.	PATO	proposed		sterk
trophic level	facultative heterotroph				An organism that can obtain organics produced by other organisms or, can make it from inorganic carbon.	PATO	proposed		sterk
trophic level	oligotroph				organism that can live in a very low carbon concentration, less than one part per million (Wikipedia)	PATO	proposed	consider also the opposite, copiotroph, an organism that prefers environments rich in carbon	sterk
trophic level	methylotroph				microorganisms that can utilize reduced one-carbon compounds, such as methanol or methane, as the carbon source for their growth (Wikipedia)	PATO	proposed		sterk
trophic level	methanotroph				methylotroph that can degrade methane.	PATO	proposed		sterk
trophic level	lithotroph				organism that uses an inorganic substrate (usually of mineral origin) to obtain reducing equivalents for use in biosynthesis (e.g., carbon dioxide fixation) or energy conservation via aerobic or anaerobic respiration (Wikipedia)	PATO	proposed	perhaps redundant for MIGS as chemolithotrophs and photolithotrophs cover the lithotrophs	sterk
trophic level	lithotroph	chemolithotroph			organism that uses inorganic compounds for aerobic or anaerobic respiration (Wikipedia)	PATO	proposed		sterk
trophic level	lithotroph	photolithotroph			organism that uses light as energy source (Wikipedia)	PATO	proposed		sterk
trophic level	organotroph				organism that uses organic molecules as an energy source	PATO	proposed	perhaps redundant for MIGS as this is covered by chemoorganotroph	sterk

Incorporating MIGS/MIMS data into INSDC records

- As agreed during the GSC 6 meeting in October 2008, the collaborating databases DDBJ/EMBL/GenBank (INSDC) will incorporate MIGS/MIMS data into their genome records
- EMBL now accepts (my) MIGS data in XML (XSLT transformed gcdml files)

ISAcreator as a curation tool

- Better user-friendly curation tools are needed
- ISAcreator developed at the EBI is a promising tool
- The 'ISA-team' and I have worked together to generate MIGS data in ISA-TAB format
  - a MIGS-compliant Sanger pathogen is publicly available (http://www.ebi.ac.uk/bioinvindex)
  - Some issues need to be addressed, but we expect many more genomes to follow

studyinformation

Study ID: Bll-S-5

Title: Determination of the complete genome sequence of Salmonella paratyphi A str. AKU_12601

Organism(s): Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601

Description: Determination of the complete genome sequence of Salmonella paratyphi A, a cause of Typhoid fever in many parts of the world. The genome is 4,581,797 bp in size and has a GC content of approximately 52.2 %. There is also a plasmid of 212,711 bp.

Design(s): dideoxy sequencing

Publication(s): Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, Quail MA, Norbertczak H, Walker D, Simmonds M, White B, Bason N, Mungall K, Dougan G, Parkhill J,Pseudogene accumulation in the evolutionary histories of Salmonella enterica serovars Paratyphi A and Typhi. CiteXplore:19159446

Sample attribute(s):

Attribute name	Attribute value(s)
pathogenicity	human
number of chromosomes	1
energy source	chemoorganotroph
trophic level	heterotroph
organism	Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601
culture collection id	not applicable
host health disease status	alive
number of extrachromosomal elements	1
habitat	animal-associated habitat
finishing strategy	finished
project id	30943
longitude	67 E
oxygen relation	facultative aerobic
subspecific genetic lineage	serovar
host	Homo sapiens
sampling time	2002year
number of contigs	2count
latitude	25 N
location	Karachi, Pakistan
biotic relationship	pathogen
fold coverage	21.9 X
host class	natural host

Guideline(s) followed: MIGS

Download:

STUDYMETADATA in isatab

more information about the study including protocols applied and sample processing steps...

open isatab in spreadsheet software or download, import and view it in the isacreator

ASSAYDATAFILES&RECORDS

the assays associated with this study are listed below with links to their raw and processed data files (if available) as well as links to submission records in other repositories (where applicable)...

assaytype

Measurement: **genome sequencing**
Technology: **nucleotide sequencing**
Platform: **ABI PRISM 3730**

View EMBL-Bank Entry For AM412236

Contact(s): Kathryn E Holt, Peter Sterk

http://gensc.org/

Acknowlegdments

Thanks to:

- The GSC, in particular Dawn
- The pathogen sequencing group at the Wellcome Trust Sanger Institute for having me as a visitor
- The GOLD team for providing a lot of data
- The 'ISA team' @EBI
- The EMBL database team @EBI

This project has received funding from NIEeS and a NERC International Opportunities Award (NE/3521773/1) 2005-2008)

The Human Microbiome Project (HMP)

George Weinstock

Washington University in St. Louis
School of Medicine

The Genome Center at Washington University

Players

- http://nihroadmap.nih.gov/hmp/
  - See “Funded Research”
- HMP Genome Centers
  - Baylor, Broad, JCVI, Wash U
- The DACC (many members here)
- Demonstration Projects (15)
  - 3 Genome Centers
  - U Md, NYU, VCU, U Penn, U Mich, Indiana U, UCLA, NHGRI intramural, Woods Hole, …
- Development Projects

Overview and scope

Jumpstart Program + White papers (Year 1)
- 500 Ref genomes
- Human sampling
- 15/18 sites
- 375 subjects
- 12,000 specimens
- Metagenomics
  - SSU, Shotgun
  - Apply to specimens

HMP Centers RFA
- 400 Ref genomes
- Metagenomics
- Jumpstart specimens
- Virome
- Euk. Microbes
- Other (e.g.Txome)

Years 2-4

Well coordinated

Demonstration Projects RFA
- 15 awards – 1 year
- 5 awards – 4 years
- 10 involve large genome centers
- Mainly metagenomics
- Some whole genomes

Years 2-4

Technology Development
e.g. not yet cultured organisms

Informatics Development
e.g. metagenomic assembly

Ethical, Legal, Social Issues

Data Analysis & Coordination Center

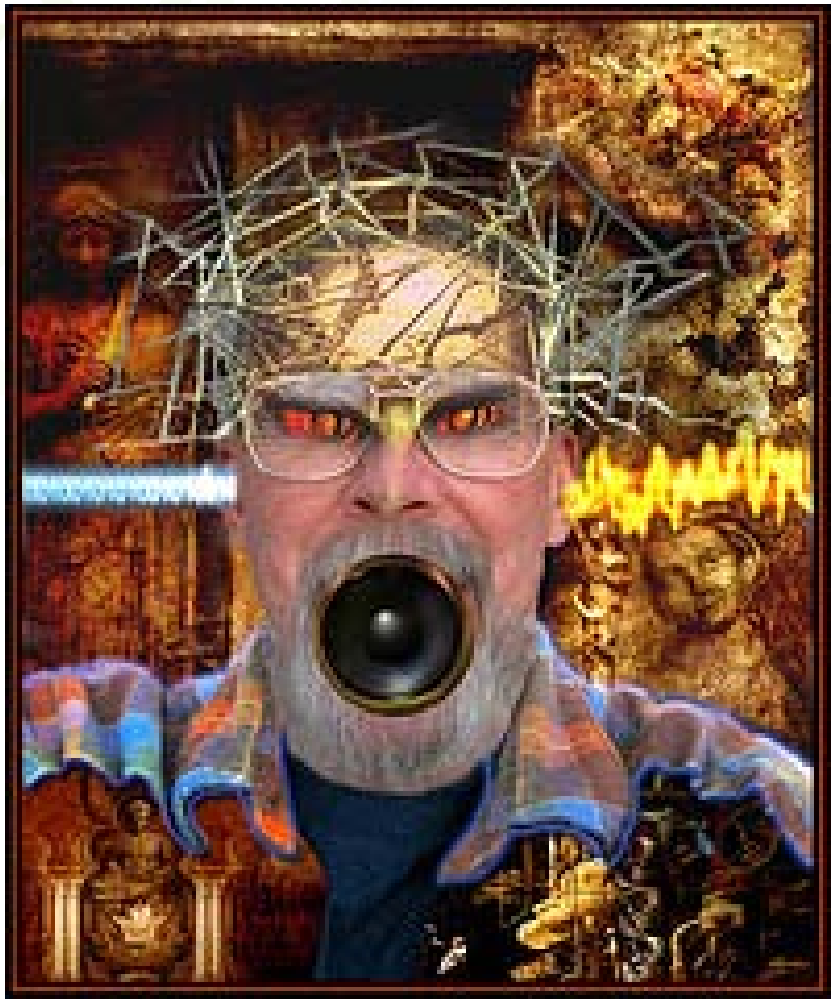
Strategy for the different parts

- Selection of 900 reference genomes
  - NIH Working Groups, Community recommendations
  - Center-initiated or Legacy Collaborations
- Metagenomics of 12,000 specimens (provisional plan)
  - 454-based SSU sequencing to bin samples
  - Deeper SSU analysis of representatives
  - Shotgun sequencing of representatives
  - Analysis?
- Virome? Euks?
- Demonstration Projects
  - Greater variability in methods and analysis
  - Varying degrees of experience

Data sets to be generated

- All Data Rapid Submission
  - Some controlled access
  - 12 month embargo (?)
- Whole genomes
  - 900 assemblies and annotations
  - > 900 submissions to SRA/TA (15% to be upgraded)
- SSU metagenomic data (submission not discussed yet)
  - 454: 12,000 + Demon Proj data sets
  - Sanger: ~500,000 full length sequences
- Shotgun metagenomic data (not counting cDNA)
  - 454: 200 runs (?) ~ 10^8 reads
  - Illumina: 200 runs (?) ~ 10^{10} reads
  - How will metagenomic reads be annotated? Pathways, gene families, BLAST hits, taxa, viruses, etc.

The DACC



Related projects

- NHLBI RFA: HIV and respiratory microbiome
- NIDCR: Oral Metagenome
- NCRR: Primate microbiome
- Same species sequencing: MRSA etc.
- International
  - EU: Multiple projects with MetaHIT the major one
  - Asia: Japan, China, Korea all have HMPs
  - Smaller programs in Africa, Australia, etc.

Washington University in St. Louis
School of Medicine

The Genome Center at Washington University

A New Arm of the GSC:
the "RCN4GSC"

A Research Coordination Network is an NSF mechanism for funding projects that connect / network researchers toward community goals, such as those of the GSC

Research Coordination Network for the Genomic Standards Consortium

- Sustain GSC leadership in establishing and integrating genomic standards through community based efforts.
- Extends GSC activities at 100K (total costs) per year for five years. (Renewal of any RCN is - per NSF - unlikely.)
- Funding is to support outcome-focused working meetings, the exchange of early-career scientists between GSC groups to advance key standards, selectively encourage informative outreach / networking at major bioscience meetings.

RCN4GSC (see SiGS (2009) 1: 87-90)

- Includes a focus on all GSC core projects.
  - GCDML, Genomic Rosetta Stone, Habitat-Lite, SIGS, Genome and Metagenome Catalogue
- Promises NSF to reach out to new communities beyond traditional genomics:
  - Enhancing effort to bring molecular standards to environmental research, via GSC's GCDML to inter- connect with ecological data standards, e.g. EML;
  - Exploring options with the Biodiversity Community, notably including standards for sequence data and metadata of museum specimens for phylogenetic considerations.

RCN Governance

- PI: John Wooley, UCSD
- **Steering Committee**: John C. Wooley, UCSD (PI); Dawn Field, CEH Oxford; Frank Oliver Glöckner, MPI-Bremen; George Garrity, MSU; Nikos Kyrpides, JGI; Karen Nelson, JCVI; Owen White, University of Maryland, as drawn from the GSC Board. Other RCN members responsible for leading core projects and activities include James Cole, MSU; Peter Dawyndt, University of Ghent; Renzo Kottmann, MPI-Bremen; Lynette Hirschman, MITRE; Victor Markowitz, LBNL; Inigo San Gil, UNM; and Lynn Schriml, University of Maryland. Other steering committee members and core leaders will be selected to cover the expanded ef-forts in ecological, environmental and biodiver-sity data.

Patrick Chain GSC Finishing Standards

Announcer: [0:04] Our next speaker is Patrick Chain, who is going to talk about a new standards project for sequencing.

Patrick Chain: [0:14] Thank you. Thanks Don for having me here. Same with Nikos, although I'm part of the JGI so I come here anyway. [0:27] I'll just give you a very brief overview on discussions that have lasted well over a year now, on trying to define new standards for genome projects. Like many others, I'll give you just a very brief history.

[0:46] The first GSC meeting in '97, the first complete genome came out in '95. These two things set the standards for what genome projects are. At this second international strategy meeting on human genome sequencing, in Bermuda, they set the Bermuda standards.

[1:06] They should have just called it the GSC back then, the original. They defined exactly what a finished genome is. And that was perfect, complete coverage, down to one error in 10,000 bases.

[1:23] Until recently, you either had this finished category, or an amorphous draft category. For finished Bacteria and Archaea it meant base pair perfect pretty much, with a possible couple of errors.

[1:42] For eukaryotes, generally they conform to the Bermuda standards, which also includes not tackling centromere or telomere tough repetitive regions. So it was clear relatively recently that with the flood of new genomes coming through we needed to reevaluate these standards.

[2:04] We've seen this picture from NCBI, which shows the growth of GenBank. Similarly from GOLD you see this nice growth of genome projects. You have in purple, the top band, the incomplete piles, and then the lower complete.

[2:22] And you can see incomplete is starting to really diverge away from the number of complete genomes. If you take a look at this curve, you've got this one nice slope, maybe another nice slope.

[2:32] This change of slopes actually coincides very well with the implementation of some of the second generation sequencing technologies, to do the drafting as well as starting to help out in some of the finishing processes.

[2:47] Just to illustrate exactly how fast this field is moving we have this little diagram I got from a friend at Sanger. This is about a year old. This was their output per month, almost eight gigabases per month, per solid or aluminum machine.

[3:06] Now, you can get maybe 30 or greater gigabases per run, with the run taking three to five days or so. We're certainly improving the throughput of these machines, and this will only make our problems a bit worse.

[3:26] So we've tried to extrapolate what this will look like in the future. We did it for the five years, but the numbers were too high. So, I'm just showing you to 2012. We're estimating approximately 12,000 genomes. I suspect this is a very low estimate, since it only takes into account the last few years.

[3:52] In terms of draft, finished is still pretty much finished, base pair perfect. What kind of draft is there? There was a proposal in 2004, which was way down here. They already realized we needed an additional category of sequencing project flavor. They proposed an intermediate grade, finishing.

[4:19] Both Los Alamos National Lab and JGI have hosted for the past four years this finishing meeting, now called the "Sequencing, Finishing and Analysis in the Future" meeting.

[4:34] This has primarily involved large genome sequencing centers to discuss some of the issues with newer technologies, and how we can progress forward some of these sequencing projects.

[4:48] We've certainly come to the conclusion that finishing efforts need to continue, but they need to be applied very effectively. Perhaps some gradations between these two categories of draft and finished are required, and we should standardize these. This certainly helps the scientific community to try and interpret these genomes.

[5:14] I chatted with Dave Ussery about a year and a half-ago who mentioned that he did this large study, these pan-genome studies. He found these really odd Burkholderia genomes from a pile of other Burkholderias that had a large number of genes in them, despite the genome size was the same and also a lot of frame shifts and other things.

[5:38] It was really because you have to drill down in GenBank to find out whether of not it's just a draft genome with 454 only data that was assembled using the first generational assembler with the first generation chemistry.

[5:53] So it's very hard for people to interpret the data that's actually out there, and granting bodies also need to understand exactly what sequencing centers provide.

[6:03] The MIGS paper certainly provided a very good boost to this initiative. We formed this international genome sequencing standards working group, and this comprised of many of the larger sequencing centers. So, including JGI, Sanger, the members of the HMP Jumpstart sequencing group, which consists of JCVI, Broad, Wash U, Baylor, et cetera.

[6:36] So, we began working on these standards. We've got very good feedback from many large and small centers. I should mention that just coming to this agreement amongst sequencing centers has been this massive undertaking, which I probably will never try doing again.

[7:00] Anyway, so we started with this picture of a lot of standard genomes, standard draft genomes, and far fewer finished sequences. We've added interim qualities, which are pretty much technology agnostic, so we're not dependent on the technologies since these are going to be changing all the time.

[7:22] So we've kept very loose definitions on purpose. This bottom bar should actually be very small because we'd be taking out of that box and filling out the others. And then of course there

are also many genomes, particularly eukaryotic genomes that are better regionally improved where certain parts are really targeted for finishing.

[7:49] I was jealous by all the very small text I've seen and some of the other talks. So, I've decided to put the definitions here, which I will give you one second to read.

[laughter]

Patrick: [8:00] And I'll go through this in just a bit more detail, and I'll fly through this. [8:06] So, standard draft really targeted more for the non-large centers that are just pumping out some genome sequencing, and some group that decided, "Well we can buy a four-five-four machine and pump out ten genomes within the next couple of years, then we'll just release this in Genbank."

[8:24] So this is standard draft, where we anticipate they're still giving some contending sequence, probably highly incomplete. It doesn't necessarily have good coverage and probably perhaps not very well assembled.

[8:38] Then there's higher quality draft, which would be more or less the JGI standard draft, so large sequencing centers do produce very good coverage, generally of the genomes.

[8:52] So when you release a draft it's generally more than 90% complete. But, this particular standard has no manual review or very little, so sequencing errors are quite possible, misassemblies, and there's no order by the context.

[9:10] There's improved high quality draft, which has been manually or automatically curated in some fashion to try and resolve misassemblies, and so most of these should be addressed with the data in hand, without necessarily generating additional data. So it's a very high quality and can be trusted for many analyses.

[9:38] There's annotation directed improvement, so particularly slated towards eukaryotic genomes where you may want to characterize specifically genes or non-coding RNAs, so there have been extensive efforts to improve or resolve these regions. However, the rest of the genome may have errors and there may be misassemblies, et cetera.

[10:06] And there's non-contiguous finished, where basically this captures all the genomes we just could not get to the perfect level of finishing.

[10:16] So, these are genomes with recalcitrant regions that just are, do not want to get sequenced through, regardless of the technology used. But all other gaps and sequence uncertainties, like homopolymers and low quality regions that have all been tried at least a couple of times.

[10:36] And then we have perfectly finished, which is our gold standard. We've bumped it up a notch to only one error in 100,000 base pairs. But, that's pretty much what we aimed for anyway, and this one is what you're probably most familiar with or used to.

[10:52] So basically we've come to this set of six standards that hopefully fit just about all projects, including isolate genomes or genomes that come out of metagenomic projects. These are the, like I said it's agreed upon by all members of our consortium which was a feat in itself.

[11:17] Hopefully, this will be very useful for users and we do plan on publicizing this. We have a paper coming out shortly, at least we hope. And then we plan on posting some of the standards on our institute web pages as well as the GSC page since we could possibly be adopted by them, so that I don't have to do this myself.

[11:43] We had discussions with databases and it looks like everything is going to be implemented, which is perfect. Like I mentioned, this is technology independent so we can hopefully deal with, next even, next or third-gen technologies.

[12:02] And it also allows a lot of flexibility in what every sequencing center does or whatever pipeline they have in place to do their protomer type of genome improvements. And it is already in the individual EV and the HMP projects. So, I don't have a list of all the contributors, but they're, I'm simply a spokesperson so there are many people from all of those groups. Thanks.

[clapping]

Announcer: [12:35] Peter.

Participant 1: [12:36] Somebody is going to reassign keywords of something that's not INSCP, just to let you know.

Patrick: [12:50] As part of SIGS, but when you're going over the previously published genomes, I just rely on Nikos for a bunch of this. [laughter]

[loud ringing] [13:02]

Patrick: [13:04] He's only smiling so that I don't think he's going, I don't, there are no plans, there are no specific plans to address that.

Participant 2: [13:12] What kind of sequencing of genomes for the purpose of assaying of polymorphism?

Patrick: [13:16] Right. This is something that, I don't think we've even reached a consensus on the format to describe these, directed evolution studies or sequencing multiple time series, time points, as strains evolve. Perhaps that we go along with a slightly separate format saying this is a resequencing project. [13:39] These are the regions that it could fall under, the regionally improved where we specifically annotate what regions have coverage, which ones do not and then also list which regions of the genome have changed.

[13:54] But, at least it fits within some of these categories, and then that would depend on how much coverage you have, what type of platform, et cetera. We haven't tackled that. It's a difficult question.

Participant 3: [14:09] Patrick, I know I've asked you this before in this situation. This is pretty much an improvement to the description of the genome? What are you going to get?

Patrick: [14:21] Are they typical?

Participant 3: [14:22] Single, composite, all the files that go along with genome, single pieces of the [inaudible 14:23]

Patrick: [14:31] That would be ideal if we could compare quality scores from the different platforms. Right now, it's not entirely standardized even though it should be. [14:43] But, it's not clear how to convert or combine all the quality metrics together, particularly when you have some bases covered with one platform type, other bases covered with multiple or different platform type.

[15:02] And I'm not aware of that specific area. I'm sure some of the assembly groups need to take this into account to do their consensus column, so I assume that would probably come out when they endorse it.

Robert Cottingham on the DOE KnowledgeBase

Robert Cottingham: [0:02] OK, I guess first I wanted to say I'm really sorry that John Wooley's not here. For those of you who may not know, John was a program officer at DOE more than 20 years ago and was one of the first people there overseeing biological research that actually was starting and promoting the idea of the role of computing in doing biological research. [0:30] I wish he was here to be able to see the outgrowth of that, because this is the first time I've come to this meeting, but my sense is that what you all are doing is absolutely key to the future success of a scientific endeavor in biology. It's hard to imagine how we're going to go forward in the future and advance substantially beyond where we are today without the overall standardization effort.

[1:01] I mean that. When I talk about standards, I know the name of the group is Genomic Standards, but I think standards has to extend well beyond the genome in order to really ultimately be successful. And I mean it not just even the biological sense but also very much in the technological sense as well.

[1:24] So you may know that the DOE has for years actually had workshops on the topic of where computing was going in the future and what could be done to make it more effective in scientific research. The more recent instantiations of these meetings have focused on this concept of a knowledge base.

[1:50] I'm not sure that I totally know exactly what a knowledge base is, but I think I'll know it when I see it.

[1:59] I got involved in this about a year ago and have a very, I think, different perspective on how this needs to be moving forward in order to be successful. So I'm going to tell you a little bit about that.

[2:17] To give you a little bit of history, the most recent of these workshops was a little over a year ago. It was in May of 2008. I know some of you were there; I remember Don was there. The outgrowth of that workshop was a report that was published earlier this year, and it's online on DOE's website.

[2:38] Mostly this is focused at a very high level on a vision of what computing could be and what it needs to really accomplish. The main focus, I think, is really on the notion that there's a lot of research the DOE and others are funding -- which, to a large extent, these research efforts are isolated.

[2:56] So it's very difficult to be able to integrate data that hasn't already been integrated and that there are many groups that are working in isolation. It's very difficult to take the results of their work and somehow relate it to the larger community effort. I know if you've ever been involved in reviewing some of these papers, you'll have a sense of how difficult that is.

[3:17] I can say for myself, I find it very frustrating to get a paper submission that's based on a set of data that was generated in analysis without being able to take that in an effective way and

be able to reanalyze it my own way, put it in tools that I'm familiar with which might be slightly different from the ones that the authors have put out.

[3:40] So in many ways, I feel like what we're doing is really not science. I have a brother who's a physicist, and he's very critical of this, and I understand that criticism. So part of what I think biology overall needs to do is begin to become more rigorous. The way that will happen is through the standards effort to a large extent.

[4:03] I know that I'm at some level preaching to the choir here, but I thought I'd show you something and talk to you a little bit about some examples and also tell you a little bit about the funding situation.

[4:14] One of the things that I think is a big change over the last year or so that I see in the context of DOE, is that they're actually getting serious about funding some of these efforts. I think that Folker talked about this a little bit also on the computing side of DOE. They're very interested in funding computing support associated with biological science.

[4:37] In the other sciences they're funding orders of magnitude more. They're providing orders of magnitude more money than they are to biological efforts. So I think there are easy opportunities if there are good credible arguments for utilizing those resources.

[4:55] The Office of Biological and Environmental Research at DOE is funding both national lab efforts and non-national lab efforts, so that should cover everybody in this room. [laughter]

Robert: [5:07] And this year they're funding on the order of about eight million, something like that, which is on the order of about double what it was a year ago. I'm reasonably confident that that level of funding or more will happen over the next few years. So I think we're at the right time and at the right place to really advance this substantially further. [5:36] I'll cite a local example from Oak Ridge. This is an example from the bioenergy center there. You see in the photograph in the upper left, there are some folks that are out on an expedition, sampling microbes that are in a hot spring in Yellowstone and then bringing those back.

[5:59] There's a whole variety of analyses that are being done, ranging from genomic sequencing, transcriptomic analysis, culturing, looking at metabolites, doing proteomics.

[6:13] And as well, there are efforts and projects that we're working on around modeling data, looking at how we can use those models as a basis to automatically generate data structures, data representations, websites, databases. Then combining that all together with our computing resources, trying to provide a useful result that helps advance scientific research. So at a high level, that's kind of our goal.

[6:42] The analytical methods that are being largely used today are shown in green. I won't go over all the exponential growth in data; you've heard that already today. But what I want to focus on are the things that are in red.

[6:59] As an example, an area that I think is really important for the future and does relate to standards, is the ability to do an assessment of the quality of the data. An assessment, ultimately,

of the quality of the analytic results as a basis to be able to compare scientifically how certain methods are working and which ones are better than others.

[7:22] I think one of the things that probably concerns me the most is the realization that most researchers who are using the computational tools that people like me produce, is that they're using them like black boxes. So they're trusting that the results they get out are, let's say, of uniform quality, when anybody who works in this area knows that they're not.

[7:43] But partly the reason why that doesn't happen with this is because A) there's no expectation in the field of that happening, and B) no standards on which to base it. So there's another important role for standards.

[7:57] I think, in addition, because many of the projects have been working in isolation, it's difficult to do assessment, as well. If there was more of an open standards/open software kind of approach to things, then it would be easier for anybody to come in and look at how certain things were being done, and ultimately to compare and contrast them and assess.

[8:22] Finally, I think things related to this, which are understandable in the past not having been done, are things like versioning. So the notion being that if you have different analytical systems gathering or reassessing the same data over time, you want the ability to be able to compare and contrast that.

[8:42] Even doing that on essentially the same samples and the same equipment isn't even very standardized. So that's a serious problem. I think the standards effort helps to provide a basis on which we can build platforms that do away with that kind of situation.

[9:00] So ultimately I think the goal is that we want to be able to present to anybody who's looking at the results of these systems, what was the basis on which the results they're seeing displayed on the screen, what was the basis for that? Where did that come from? Being able to drill down to where was the data, being able to drill down to the analytical methods and understanding where those things came from. That's the goal at a high level of how I think of the knowledge base.

[9:25] So in many ways, this is a huge challenge. I'm not trying to suggest that we're going to build some systems over the next few years that are going to totally address this. But I am trying to challenge everybody here to think about what do we need to do to accomplish that over a long period of time.

[9:44] I think, in many ways, the tools that we have accessible to ourselves today are tools that were riding on the coattails of a lot of technologies, especially computational technologies that none of us were involved in developing. But it's not clear to me that we can go forward into the future that way.

[10:02] Partly I think that's because biology is extremely complicated. And if we start developing the standards that you're talking about, the computational systems are inherently going to be a lot more complicated than the ones that we've built over the last decade or two.

[10:16] But that's OK, because I think the underlying computation infrastructure that is being built up in the word today is up to that challenge. Now the challenge to us is to figure out how to take advantage of that situation and figure out how to build such systems.

[10:35] I think sometimes when people have talked about the idea of the knowledge base, at least at some of the meetings I've been at in the last year, it's been very much -- I'm exaggerating a bit here. But it's very much been a discussion of, well, let's gather the data from this place and this place and this place, and we'll just push it all together and we'll try a couple of routines that we all agree on, and that will provide us a step ahead.

[10:58] I don't accept that. I think that's a very low bar, and a very low standard.

[11:05] One of the things that we're going to have to do is figure out how to start working together collaboratively. That needs to become part of the overall research community's effort. So this is, in many ways, kind of a social engineering exercise, as much as it is a technical one.

[11:19] I guess I have many more slides I can talk about here, but the one things I want to leave you with is that the overall effort, as I see it right now, is one of a software engineering project, in many ways. Developing standards is essentially providing a platform and a foundation to do that software engineering exercise.

[11:38] So much of the work that we're probably going to do, at least over the next year, is going to be very much establishing requirements and giving a software engineering community -- in the computer science community and elsewhere -- an opportunity to actually build stuff for us that's actually going to help us out.

[11:53] OK? So I think I'll leave it at that.

Field on GEM Catalogue

Dawn Field: [0:03] I'm going to give a really short discussion, and this is sort of to segue between "here's the GSC and we've organized the coming up of the standards in the community." The point is, none of us really care about the standards, the formats, anything. We want to get on with them and doing something interesting with it.

[0:17] So now we're going to shit a bit more into how we actually implement it and how we end up with a single point of entry where you can go and click and get all this rich meta-data that's been aggregated across all these fantastic databases that are doing different things. We're now heading down that road.

[0:33] Just a few high-level things to point out. The MIGS checklist was published in "Nature Biotech", but as a PDF. It's not in the publication. So when people come to comply with it, it is still evolving to a certain extent. We had to overcome this issue of giving people the latest up-to-date version.

[0:52] So I just wanted to point out that there's a new development, that MPI-Bremen, which is Renzo, Palin, and all, is actually holding a special sort of last MIGS/MIMS/MIENS relational database. And what this means to people is that it's held in a secure state, with all the descriptors and examples, all the little tweaks, the controlled vocabulary, the whole thing up to date.

[1:15] It's accessible to everybody, and it came out of doing the MIENS checklist. Because Palin can do a dump into a spreadsheet and it can then circulate for discussion. Then any changes that are proposed go back in and it is maintained.

[1:29] So that's a huge advance in keeping one record. Because the finalist fields have to go the INSDC, so we can really start pushing the benefits.

[1:40] GCDML, as we said, is the XML exchange language. We're really not going to go through the details. As you heard, CAMERA is implemented on it. I think Access has a backend bases on it. Several different databases are using it now.

[1:51] It's just to say again, because Renzo has presented on this a couple times, that it is mature enough for implementation. I think it also now holds MIENS descriptors as of a few days ago. So you can look on the Wiki if you want more about that.

[2:04] What I'm going to talk about is what we would want to do with the checklist and the XML, this exchange format. And talk a little about the genomes and metagenomes catalogue.

[2:13] This has been around from the very beginning when we were prototyping the checklist, because you wanted to have something that looks sort of like a webform, so that biologists could come in and sort of look the descriptors that you might be looking for and say yay or nay.

[2:25] So we just built a system that was a prototype. We now would like to completely relaunch it. Now it's just listed as a requirements document. How would you want to interact with this system?

[2:36] It's a genome and metagenomes catalogue, so the GEM Catalogue, and it would be very much a distributed activity that can link directly to these key databases MG-RAST, CAMERA, GOLD, et cetera.

[2:50] We redeveloped this idea. Renzo came to visit me and we talked with Peter Sterk and Alan Tepperton, who's now interested contributing. We had some ideas, but most of the development of this new vision came from the workshop we had yesterday about ISA and GCDemo, which are two formats.

[3:06] But again, out of it the main thing to figure out is if we start generating content, where does it go and how do we manage it? Susanna is going to talk next about the outcomes and mostly talk about the ISA concept.

[3:17] But the two conclusions were that we would form a GEM Catalogue working group. So that's quite a large group, but we invite anybody in here to participate. And we'll be talking to give you different from the IMG about how to participate in this. And that one way forward would be to try to write a multi-author roadmap paper.

[3:37] So before we try and code anything at this point - this has gone in fits and starts - we get all the requirements from the users and the providers of the meta-data.

[3:47] We really do have a quite developed proposal, based on what we had before, plus new things. It is up in the Wiki. After much, much discussion yesterday, we came to a view that we might not have to build it, we might be able to assemble it with modifications from key things that are there.

[4:06] And when we were discussing about scope, there's sort of three aspects of this. A repository, a way to visualize the meta-data and in the future do analysis of it, and some way to deal with submissions.

[4:19] We realize that we have these three parts, with the GOLD database being the front end. It's already got such a high profile and a heavy user base, and is very well-known. Backed with what's called an ISA Hub, that Susanna will talk about. This brings it into a multi-omic domain, with the backend being the INSDC database.

[4:40] We've seen absolutely milestone talk by Bob, where there was a submission interface that has a GSC standard in the mock-up. We might be far enough along that the INSDC can be a backup.

[4:51] This is just a model that we're proposing, and hopefully it will develop over the next three days. This is sort of what it would look like, with the INSDC being the public archive. We fully acknowledge that that is the place where the data should go. The ISA Hub gives you another way to submit in with ISA creator. And GOLD would be the front end, how you would present this to the world.

[5:12] Of course this would need modification into the hub and they would work together. GCDML would be the exchange language. But we've got this nice level long the outside, where

domain level expertise drills down right into the genome/metagenome community, but back up to align with what the broader public repositories are doing with standards.

[5:34] So the next step is to form the GEM Catalogue working group formally. Again, I said we will be talking to people. Hopefully people will step up. But most of the people in the workshop - I think we had about 15 people - would be involved. I think we'll have a working lunch on Friday to sort of get a few more things up in the Wiki. Have a face-to-face before we all disperse again, and we can continue on.

[5:54] There is quite an extensive Wiki page now with specific pages with what we have decided to implement, or a sort of a wish list set of things. We can use that as a basis of discussion.

Dawn Field Introduction

Moderator: [0:00] So, the next person up is Dawn Field and she will be doing the general introduction and another kickoff.

Dawn Field: [0:17] OK. This was just a quick slot to, again, say thank you to people and orientate you about what's going on. This is very much a community driven thing and I just wanted to, instead of putting slides together, just remind people who have been here from the beginning. [0:30] GSC started in 2005 after the meeting in Cambridge. Many of you have been to GSC meetings since. There are still new faces here, so we thank everybody who's here by showing the GSC Wiki. So, hopefully this machine is like Internet leakers.

[0:52] So, huge thanks to the JGI for hosting this. Before, everything had been done at the EBI. And, the real milestone for us, we had one meeting in San Diego which was GSC 7. That's part of the metagenomics meeting with John Wooley. We now have a large Wiki that we need to spend a lot of time working on because it's grown organically, so if people do it's just to remind you that it's completely open.

[1:24] If you have projects that are in here, we're not even going to be able to cover everybody's project in this session, but the agenda is out. But, it is here and you've got and you've got a news item [inaudible] so you can quickly get to a GSC eight meeting with the agenda, the logistics, etc. Even now, we're adding a new Wiki pages. And, I'll just quickly show the GSC board as a way of giving thanks and also to remind people of how many projects are now going.

[1:49] This is a very brief page but we actually got it in there. So, the GSC had no formal structure when it started. It was started with [inaudible] people brought together to have exploratory workshops. It then formalized the working groups. There have been [inaudible] coming up.

[2:03] I just wanted to run down the list of who's on the board and what they've all ready contributed because there are so many people who are chipping in after Energy '08, which was in San Diego in 2007 at [inaudible] .

[2:16] We did formalized a board and hopefully this board will then get to a point where it changes in time. So, if people are willing to step up to the plate over time, we will then take nominations. But, as it stands, right now, we have people like Guy Cochrane who really pushed forward and will in INSDC. Acceptance of nukes, MIMS [inaudible] , we've got some [inaudible] here, and, Ilene, he'll speak about that.

[2:42] I've been lucky to start at the beginning, so I'm still chairing. But, George Garrity managed to take away an idea from GSC five and actually launch a journal which is now here, so he'll talk about that, the Standards in [inaudible] Scientist Journal, and he's contributed to many other projects.

[2:59] Frank Oliver Glöckner came to the sirA meeting and said we need to extend MIGS, the Minimum Information about a Genome Sequence, further into the environment with habitat, so we extended it to metagenomics, which is MIMS and he's now here with Paylor leading MEEMS, so applying mixed MIMS to 16-S, we'll hear about that.

[3:21] Lynette Hirschman joined us from the very beginning and led the Habitat-Lite initiative, so controlled vocabulary is how we describe Habitat in detail. Eugene Kolker came to the second meeting and then let us do two special issues of the journal, Onix, which outlayed several GSC road map papers, which are there online now.

[inaudible] . We have an exchange language, an XML based exchange language, GCDML, which we talked about briefly, which is now mature enough to be implemented but he led the development of that. Nicos has been out playing poodle since the beginning and certainly at GSC five was the geva genomes [inaudible]

[3:45] the fact that it went MIMS/MIGS compliant and is now in gold was really a turning point for us to help us finish the specification. It was a real milestone.

[4:11] Victor pushed with the IMG to have one of the first MIGS/MIMS compliant interfaces and work with digionics to start prototyping the descriptors. Folker Meyer is new to the board but he's leading something called M5, so we had a SIG, a Special Interest Group, at ISMV, we had a power session. The day was called M3, metagenomics, metadata, and metaanalysis. We got talking about cloud computing and how the metagenomic community would deal with this glut of data, and he's now pulling the community together to kind of find one voice about how he overcomes some of these challenges and he's doing that with Ellen White and Eugene Kolker. We'll have sessions about that here.

[4:57] Suzanne is somebody who has always been there for multi OMICs, keeping us on track for what's going on at the international standards level, and we'll talk about how we hope now to merge efforts with [inaudible] , with the ISDC repository and with her work on [inaudible] to make something called the genomes and metadome catalog, the federated set of databases sharing the spread of the rich metadome that we're trying to capture.

[5:22] Peter Sterk came to the first meeting and then took the initiative to host the second meeting at BBI. He's been at all the meetings now and he's doing actual curation, so he really has forged ahead in getting [inaudible] genomes online. He's doing a lot of stuff with SIGS, and it really is to underscore the curation and contact that really pushes to that [?] topic.

[5:43] Kevin Wright came in at GSC five and pushed the idea of SOPs and it's his first contribution and since then he's done other things including M5. And John Willing, a very special thanks, he's never actually been to a GSC meeting, but he's shown such good support. When we were looking for funding he agreed to lead a RCN application, a Research Coordination Network, [inaudible] and it gives us five more years of funding for the GSC for meetings. It's how we've managed to meet today, so many, many thanks to John Willing. Unfortunately, I have to give apologies. So, he's not here but he's actively watching the list and the email and he'll be in touch.

[6:27] All right. So, many, many thanks to everybody. To stick on schedule, the agenda's in there. It's absolutely packed. It's going to be a mix of updates. The first day is usually updates on current activities, there will be a few proposals near the end of the day, and then the first session.

[6:43] I'll be moving into these mega-sequencing projects and you'll get an idea of how much data is going to start coming down the pipeline. Then the next two days are packed with

activities. Two of the main things will be this M5, this infrastructure, how we all work together through new technologies to share data and to compute, and the means checklist which has a very strong working group right now, and we'll hopefully start to work towards drafting a paper after this meeting [inaudible] . So, I welcome everybody. Enjoy the meetings. And I am certainly looking forward to it.

George Garrity on SIGS

Mr. GEORGE GARRITY: [0:02] Nico sat down and mentioned that those looming problem and that - and when all that they wound up doing is getting us involved in that, you know, project called SIGS. And this is the -- it's taken us about a year now to essentially get moving forward on SIGS. But, as Nicholas was mentioning, the role problem at that point in time was the explosion in data as far as their coming out in the -- new sequencing methodologies and the changing technologies. [0:49] At that point in time, I had - been just finished the stint of my teaching and my teacher, of course, in biotechnology in the way which we teach bio-technologies through publication here with you. It's an interesting way of actually doing we've been using some open access tools for publishing for about five years now. And that with a lot of success I have mentioned it and little brainstorming section and I wound up doing something that isn't quite expected and that is getting involved in doing a journal.

[1:15] And so the problem that Nicholas was mentioning was this concept about a loss of genome papers and scientific literature. The problem was that not only the papers disappear but there's also a loss of data, annotational and actual data. And this was the figure that Nicholas had shown us during GSC5 in December of 2007. I think the condition has gotten progressively worse overtime. And so the idea was what can we actually do to solve this problem.

[1:43] And that was to create a more open access journal, the idea was, that if we were - if we'll do this we could actually enforce on all standards amount of force, but actually encourage to use those standards. And then also see whether or not we might be able to publish the content very rapidly. And that was the concept of SIGS. We actually wrote a paper that appeared Nomex that's framed and that paper then set forth of what was used to say small grant proposal to something that was funded in-house through the Mission of State University Foundation.

[2:11] And a lot of this actually move forward with a publication that would allow us to publish an article so it's actually, we were thinking where we use forms, including short genome reports for the lack of other standardized and easy read articles that would allow to maybe we can continue to see what was going on with on-going genome projects. We want to extend that to many genomes as well as in a robotic.

[2:34] But also, we've got other things like detailed standard operating procedures, reports of meetings such as the ones we have here and others review some commentaries on a number of other types of reports that are typically not part of the index scientific literature called Gray Literature. So, as we've been going through this process we now have open access publication that went live in July of this past -- past summer.

[2:59] And it allows us now to see what another concept in links compliance in art and serving written report will become acceptable to the community. We also think that this is going to provide a very reliable way of communicating to the community, provide people for instance, a good glimpse of the GiBV genomes, certain sequence here in collaboration with the post of the DSM cells and begin time together also the text on differentiation as well. And time getting that literature.

[3:27] So, what we've been doing in the past year is building the infrastructure where stain is kind of an effort on whole as the publishing the 600 puree new articles between now and June 2011. We've established an editorial office. We have an editorial board. We have an advisory board and plus both of them are scientific, the community as well as from the publishing community of guys purely selection of the appropriate kinds of tools we want to use. And we're in a process of trying to attract as many credible authors and reviewers if possible to talk about the issue of reviewing it momentarily.

[3:58] The other thing that we are looking at doing is establishing a cost effective way of actually putting out these papers as quickly as possible. The business walls for open access journals still remains somewhat uncastids. A few of the journals have been successful but other journals have met with a little bit success, trying to keep these kinds of operations and our life for a long period of time. It's difficult. There is a cause for article against media logic.

[4:21] And one of the other things we discussed that one time when the article is being created was to establish a more provocative entity for the GFC. So, we would be able to receive funds and then be able - just correspond under the US TATCA. The location of the editorial office is in Michigan State University. I was - became the noble editor in chief. But a managing editor that worked with us for about five months, Scott Harrison six months. He just left the position. It's open right now. I think that provides with an opportunity and look at whether or not we want to restructure the office.

[4:54] Peter Stork has been doing Genome Editing Force along with others on the project, and we have a masters pre-student by the name if Remy Anne Nelson who is working on the project with the system production in moving the articles along. So, if you see Remy Ann's name, that means that she is working on your article and she's looking for a response. So, please don't ignore. We have driven down editorial board. Its involved a number of old established members in the community. We're also going out as -- we've decided to do it last year and attracting four to five additional associate editors.

[5:31] And so, what we're trying to do now is to move forward with the secure review process to make sure that the short genome articles that we have in fact actually go through a good script name, but also wanted to - is expedite this as possible. All the standard compliance we're looking not only for the science but we're also looking for the name verification in his check list and at some point in time, I think we will go to the extended links and GCDML checklist as well.

[5:57] On the SOPs, we have one SOP that's published. I think that we're still looking with Sam and Julie to develop some additional SOPs and to develop more over to check lists to make sure the address of piece to published we're in standardized format. And that people understand what the target of the article is what should actually be included in the SOP as well. Other articles that are being published undergo a normal -- that are peer review process, submitted to the editor, go out for the reviewers, make sure that the quality of work is as good as possible.

[6:29] Now one of things with candles, we can accelerate the review cycle. Those items that are bluer than items that are under the controls of the authors, the names, scripts and peer review process. It's something that's a little bit difficult but what we're trying to do on our own short cycle is by moving up on things like popular in the early stages since we know most of the lead

articles, which one to be published sooner or later, we can actually pull up a clean number quickly.

[6:54] Likewise, we can also short circuit some of the things by using slow tools down the string. We've been doing this now for five years in a classroom environment and be able to do peer review articles of large number of articles in a 28 day cycle. And one of the things that's very nice is we can control the peer review process. This is our current bottleneck and if we could get peer reviewers to return the material within seven to ten days, it would be wonderful, but our experience so far has been much longer than seven to ten days.

[7:23] It was actually in the month of August which tends to be particularly difficult if we're just getting started. That's probably not the worst time actually, one kind story journal. The other thing that we've actually been dealing on the production side as we move over to production of the content and SML -- it had some very nice tools as well as paper word document and converted to the one compliant SML, and one has. It's very, very fast that are also allows us is down, produced HT mode directly from the SML.

[7:54] Now, we're looking at the necessary modifications through corporate SML FL tool that will allow some -- produce pdfs of what as well as directly from word document is supposed to happen to do this directly from printed documents and how - what handle it can do. So editorial 112 was pretty straightforward right now. It's taking us about, I will say about six to seven hours long. Each one of the articles as it goes through process at the editorial level, we're using mailing words, OJS which is the editorial environment, the open ground system, so we, Malcus and distribute the content through.

[8:35] And the other tool that we're actually using is something called X-analysis. A plug in that works very nicely out of org. At this point in time, before publishing 200 articles a year, we estimate the cost per articles, going to be at $930. This is below the current rate for most publications. We think we can actually drive this thing on for about five to $600 per article. We have published 11 articles so far. We have 11 articles that are queued for the next issue.

[9:03] These are all the copy editing or in the late stages. We've already produced the HTML. And so that authors can actually see both the word documents and the HTML. We have 15 articles on the queue and we totally have a number of additional articles that are in play. The website has been up and running now for a number of months and since July -- this is the website -- the articles that complete the accessible online that being indexed in Google is following right now. They're also available directly from the website -- once again, available in both pdf and HTML form.

[10:06] The other is that all the content is now being essentially set up. So we'll go directly into Automated Central as soon we get all the -- more content, then we can pass the correct area code of Automated Central. We're doing all the standard things directly out of following on what's expected for Automated Central articles. All power articles and are now priced, and the articles are completely arranged right now using digital object right down a far and the rest of the literature databases and everything allows us to do both forward linking with using both Mc dawns and identifiers.

[10:41] And so that's where the project stands. If you can give me a couple more minutes -- one form actually, maybe one other thing I think was the showing. This is where we stand right now. We're actually getting a lot of activity on the site. We're having any advertise or anything is pretty much words now, or in seven days we've actually have done close to $2,000 moves in the articles and so I think we are making some headlines, and -- thank the others...

Pascal Gaudet on the Biocurator Society

[background noise and inaudible speech in the background]

Pascal: [0:02] OK. Good morning, I'm Pascal Gaudet. I'm interim executive board member of International Society for Biocuration. For those that don't know me, what I usually do is I'm a curator for the [dictodays inaudible 0:27] database, which is a model organism database for dictodays phylum, also, very involved in the book and social activities, like we discussed yesterday. [0:29] OK. So, you guys are all, I think, deeply aware of the issues here, but just to state why we're doing this. As you know, there's a lot of databases that exist that cover all kinds of areas in biology, like genomes, genes, protein interactions, protein DNA, metabolic and signaling pathways and many other things.

[0:55] There are an exploding number of databases everyday. The nucleic acid research is to, I think...So, there's a thousand or how many databases are out there now? 1500? It's very big.

[1:09] At the same time, all the resources are essential to the researchers. If you call researchers, they will all say to you they use...I think 95% of the researchers use some database on a daily basis.

Some people spend half their day just getting data from the databases. So, some of these resources have been around for a long time. [inaudible 1: [1:23] 34] and people are a little more familiar with what they do.

[1:39] But, there's newer techniques. So, there's newer ways to present data. So, there's also this explosion of new resources that you want to know how to find them and what they do.

Then, as you guys are deeply aware, the new technology means more data: [1:50] means you need have to have a way for researchers to access that data. So, this is just a small glimpse of the curated databases.

[2:06] So, there's a lot of them. So, the biocuration activity involved transforming data into a form that, basically, researchers doing wetland research can take and use for their own work.

So, it aims to provide a way to present the data, so that people can understand its whole meaning. So, you require biologist biocurators that enter this data. You also require software developers that provide [to develop inaudible 2: [2:22] 44] the tools and researchers in my informatics.

[2:45] So, this is really an area, where it's peripheral to the main research, but it's essential for the research to be happening. The curators are usually Ph.D trained and have a lot of published papers and experience in the area they're innovating.

So, the issues are: [3:08] all these things are sort of left on their own little independent way. A lot of the, even systromes, a lot of the moths were developed by some researchers on the weekend, but he needed a way to organize his own data.

[3:29] Then, eventually, it got funded. So, it's been kind of a side project from the beginning. It's been a little bit difficult to put this out there as one main need for research.

[3:42] I think you guys are all...Everyone has funding problems. We're not the only ones, but it's difficult to explain how much we need this funding. So, that's why we created the society.

[3:58] I'm going to explain the broad goals of the society. One is we would really like to integrate researchers, journals and the funding agencies, into the biocurator work. A lot of us that have resources have our own resources.

[4:17] Usually, depends, but sometimes can be a big group or small group. But it's bi-directional. It would be nice to really have a web, a lot of connections, where it would be easy to find out what else is available out there other than the resources you go to on a daily basis.

[4:45] Yeah. We want to make sure the funding is insured. In order to do this, I think we really need to get together and explain what our needs are and explain that we're being efficient about how we do things.

[4:58] So, if someone starts a database on Saturday mornings, maybe their funding agencies will say, "Is this really a seriously thought through project?" So, having been more organized will help us establish what makes a resource essential or not.

[5:20] So, this is just a little example that we like to present as a goal. There's a lot of these papers out there, where they use the goal to get analyzed date, and they don't even cite the goal.

[5:35] I think this is an example for the goal. I think every database has this issue. Of course, it's out there. Of course, I use the goal. Why do I need to say this? So, it's difficult for us to estimate how many times people have used the goal if we're only cited 50% or 15%.

[5:53] I don't know how many people don't cite. So, it would really make an impact if people really acknowledged where they got knowledge from. So, the mission statement is divided into three broad areas.

[6:12] First is to improve the structure of our work. Then, to improve interactions with the researchers and the journal publishers, because we are really...This is really our way inside the community.

[People inaudible 6:29]

[6:27] , what researchers want to do is they want to have an interesting project, and they want to publish this in a high impact journal. Then, we want to annotate this information.

[6:39] So, at every level, we need help, from the researchers. The researchers are here, who can help us do the innovations, would be inclined to help us. The journals do have the ability to ask the researchers for more stringent ways to submit their data such that the databases can incorporate them more easily.

[7:02] Like we've had so many examples of you're annotating a gene, and you think you're annotating a human gene and you realize it is a mouse gene because there is a small losing difference and I mean a loss is 147.

[7:17] So people don't always even describe to you what gene or protein or entity they are working on. So we would like to work with their journals to force people to describe their work in such a way that it can be incorporated in a computer-readable way.

[7:37] And then promote biocuration as a career path, because there is a lot of work to be done. And it's a new area. I think we need people that have the creativity to make this useful. You know, you can just go and read and do kind of a little bit of monkey work, or you can do it in a smart way and I think it's important that we get really good people that will help us to the next level.

[8:05] So this is the more specific detail. So for example, we organize regular conferences. And I must say that we've had three international conferences so far and they've been really great. People didn't really know each other until then. It's been really helpful to know that people are facing the same issues. Some people have solutions to your problems, or ideas.. Some people don't. And then you can identify areas where it needs a lot of work. It's been very, very good to talk to the people doing the same thing.

[8:43] The second thing is something I'm working really heavily on lately. We would like the curators to publish their work in peer-reviewed journals as scientific work. So it's been kind of the assumption that this is not primary data, this is not new data, you're just organizing data in a certain way.

[9:07] But this organization is essential to understand the data. And so the fact that you're organizing this data needs to be presented out there and this consortium of genome standards is one way to sit together and decide what's the best way to present this data or what are the different ways to present this data. And if this is not available out there, I think this is what will keep on happening is people will continue to develop their own system. And we need, really, ways to put this information out to the whole community.

[9:43] So we have lots of ideas for new journals. So provide the forum for people to discuss together on topics they're interested in. Some people are interested in standards, some people are interested in phenotypes, things like that. So we would like to be a hub where people can go and say "OK where do I find someone who cares about educating undergraduate students for biocuration and databases?" for instance.

[10:13] And another of our goals, which is why I'm connected with Almon is, it's important for us that there is documentation and standards for the annotation tools. And we don't want to impose anything, but at least if the documentation is shared, somewhere. And then we can discuss, again, take the good ideas from everyone and try to see if we can come up to standards for annotation.

[10:39] And of course we would like to make the case that we need stable funding for several resources that are available right now.

[10:51] OK, so this slide describes a little bit. We really want to link with the journal publishers and there's ideas in that respect. Like there's people working on semantic tagging. So that would help if the papers were somehow XML tagged and we could more easily just at least identify which genes a paper is describing. And we really need the support from the journals.

[11:16] And this has been a little challenging because the journals don't know which databases are going to keep on being funded. So they've only accepted a limited number of identifiers. So of course they don't want to link and cite resources that look like they might not exist in the next couple of years. So we need to also establish ways to keep those entitles traceable. And yes the journal publishers are willing to talk with us if we can provide a system that will be reliable.

[11:56] And yes promote biocuration as a career path. So the ideas we have right now is it would be nice if we could talk with people undergrad or graduate students and have like modules or courses that cover certain aspects of bioinformatics and biocuration outside the bioinformatics.

[12:16] Like I'm really thinking we need to target the biologists themselves, the ones doing gene knockouts and things like that. And then coming back to the research to try to understand the gene they identified. So we need to understand how we organize the information. So it would be great if we could educate even biologists at early level. And also train people that want to do biocuration by doing workshops and so on. And we've done this at the biocurator meeting.

[12:51] Yes so just the short history of biocuration is we have had three international conferences so far. The society, the International Society for Biocuration was incorporated this January. And since then we've been working on the technical aspects of getting a bank account and getting a system where we can take members and so on. And I'm part of the interim executive board along with Lydie Bougueleret and Lorna Richardson.

[13:23] Next week we're going to have the first election for the executive board. So in our constitution we're supposed to have nine members. So we have six open positions because we have two years' mandate. So Lydie, Lorna and myself are staying on for another year.

We have so far 220 members the last time I checked in the society Interestingly, there are 36 people running for those six positions. I think this is great. People feel really involved. And I don't know, I think there's going to [this can inaudible 14: [13:46] 03] be a lot of fun.

Janet Jansson on the Terragenome Initiative

Janet: [0:03] Jansson: Now I'll shift gears a little bit and I'll explain something about the Terragenome soil sequencing consortium. First, a little bit about the history. This project was actually started by a group of French investigators in Lyon in 2007, Pascal Simonet and Tim Vogel, and they had a workshop to try to bring researchers together to discuss metagenomic sequencing of soil. [0:29] Since then, there was another, follow-up workshop in Lyon in December of 2008. And just this past June, we've had the third workshop, which was in Uppsala, Sweden. There will be an additional workshop in Seattle, in conjunction with the ISME meeting, next year.

[0:48] This has continuously evolved over these last few years. It began with a small, focal group, but it has now grown. The meeting in Uppsala, we had about 120 people, over 20 countries represented. So, it's really growing interest in metagenomic sequencing of soil.

[1:08] The website has been established. It was Folker Meyer that developed that website. You can see the web address there. Right here.

[1:19] We've also, through this consortium, initiated several grant applications, mainly in the US and in Europe. An additional action of Terragenome is to develop working groups. One of those includes development of standards; coordination with the GSC, which is why I'm here today; and to also communicate that this consortium exists. It's an open consortium, so everybody who is interested in soil metagenomics is free to join the consortium.

[2:01] The organization of Terragenome, we did formalize a steering committee, just for the purpose of getting things organized. We had an election and voted Jim Tiedje to be the coordinator of Terragenome.

Then you can see the other members of the steering committee: [2:17] myself, Pascal Simonet, Folker, Jeroen Raes, Eric Triplett, Joe Zhou did the analysis, Pauline Mele from Australia, and George Kowalchuk, and also Luis Wall. We're currently looking for steering-committee-persons from Japan and China.

[2:42] What the steering committee is in charge of is the basic coordination and oversight. To provide some structure to the consortium--it's all virtual--drafting the planning mission, and also conforming to standards, incorporating standards.

[3:00] Another goal, in particular for the funding agency, is to delineate why soils are so key to focus on. So, we've had a couple of editorials in "Nature" about technology, for example. Then we also provide a mechanism for integration and for data-accumulation skills and material exchange, et cetera.

[3:30] Also, at the last meeting in Sweden, we tried to say, OK, what are the main focus areas? What are the main things that we can do within Terragenome? So, one thing that we have are the standards and tools, so development of standards and tools. Also, as I mentioned, to apply for funding.

[3:49] We have two different focus areas. One focus area is to concentrate on the Rothamsted soil. This is a long-term field station in the UK, and so this is the initial focus point. Everybody focuses on one reference soil.

[4:05] But we also decided to include other reference soils from other countries. This is mainly a political decision because, in some cases, it was easier for individual countries to get funding for their own country's soil than to get funding for Rothamsted.

[4:21] However, there are several Rothamsted sequencing projects that are ongoing already. I think that this is impressive considering it's just been initiated for a couple of years. These are all samples of metagenomic sequencing projects that are focused on the reference site, which is at Rothamsted. As you can see, I'm not going to go through all of these, but there are a large number of different projects.

[4:50] In addition to the steering committee, we have several subcommittees in Terragenome. So, the first one is really focusing on the Rothamsted site and sequencing of the reference-soil metagenome.

[5:01] We have a second subcommittee that is working with the metadata standards and a third on isolates. If Nikos is here, he'd be happy to hear that. So there will be a specific subcommittee working on isolates. Funding opportunities is the fourth. The web portal, I already mentioned.

[5:22] When you're working with soil, it seems like almost every individual laboratory that's working with soil has their own sampling method. We're trying to standardize that. There's the committee that's working on that, and, also, with methods for extraction of DNA, RNA, and proteins.

[5:43] The second methods subcommittee is looking at getting the data about the soil biochemistry, physics, et cetera. So, again, to standardize that. The 9th subcommittee is on bioinformatics, and then we have a 10th, which is a catalog of partner activities and skills. Yeah?

Nikos: [6:04] Is anybody from the metadata standards here? Or meta application?

Janet: [6:08] Yes, yeah, I'm going to talk a little bit about that, Nikos. I'll go into that a little bit more. So, this is the metadata subcommittee, and so we have; Dawn is here, and Jim Cole are here. In addition, there's George Kowalchuk, David Myrold, Cindy Nakatsu, Gustav Teddyson and Jim Tiedje are on the subcommittee. [6:35] You have already had some meetings, as I understand, so these are slides provided from Jim. And the goal of the metadata subcommittee was to establish a set of suggested attributes for the sequence data and this is a collaboration between the GFC and Terragenome.

[6:56] This was a survey that was done through the Terragenome website. I asked Jim yesterday, it's about a 106 participants in that survey. They were asked to evaluate different chemical, physical, biological and a number of additional parameters on two criteria. Whether they were easy or hard to obtain, so that we see on this axis, on the y-axis.

[7:22] Also, whether they were considered to be not very important or very important, and you have that on the x-axis. The ones that are the most interesting, at the end of the survey, are really

the ones that are very important and easy to obtain. It shows where the survey participants originated from.

[7:47] Just briefly, these were at the end of that survey, these were the criteria that were considered easy to obtain and very important. For example, I'm not going to go through all these. For chemical parameters, pH was thought to be easy to obtain and very important. Climate is another example.

[8:07] We're starting to narrow down these vast number of parameters into a smaller grouping. So the different sections that are the foci for the soil are the site descriptions, sampling description, climate, and soil classification of soil analysis.

[8:27] I was supposed to point out these criteria may look different from what you're normally used to from an ocean survey, because you have to take into account the soil parameters. So, it's a different type of data.

[8:40] This is just an example of the types of attributes that you would want to get information about for a site. For example, the latitude and longitude is obvious, but also the current vegetation, the history of the land management, et cetera. The same for sampling descriptions. So there are specific parameters for soil that are important and relevant.

[9:06] I'm going to shift gears and talk just a couple minutes about one of the JGI's programs that's ongoing for soil. This is a pilot project, and it's called the Great Prairie. The rationale for doing a metagenomic sequencing of the Great Prairie. This is the defined rationale that...considering we're also at the DOV laboratories that has to be central to the DOV mission.

[9:35] I think also it's a very important ecosystem to look at because it's one of the largest expanses of the world's most fertile soils. It provides a reference site for understanding biological bases and ecosystem of the community.

It does parallel the ocean gyre as an important ecosystem. If you think about the amount of primary productivity and bio-G chemical sites in [inaudible 10: [9:54] 02] in the Midwest prairie, it's equivalent to the ocean. And it does sequester the most carbon in living soil system which is also important for carbon sequestration, the only objective. It produces a large amount of biomass.

[10:21] For the pilot study, these are the three sampling sites. We have four specific field sites from Wisconsin, two from Iowa, and two from Kansas. So, these are all chosen to be sampled from native prairie adjacent to managed fields.

[10:43] This is a picture of Jim Tiedje in his farm gear in Iowa. So, one of our sampling sites was a over one hundred year tilled corn site. He actually took some of his samples inside.

[11:02] This is the last slide, and I got this from Jill, because I think it very nicely illustrates where we're going in terms of the amounts of sequencing. The field's termite hindgut was 62 Mbp of Sanger sequencing. We get to the cow rumen, so 17 Gbp by illumina sequencing.

[11:24] Now, the flagship pilot projects that are ongoing and the prairie project is one of the flagship projects. We are approaching a hundred gigabase pairs of sequencing by the time those are completed.

The next large project that is being initiated are the real flagship projects and these are going to be terabase pair projects. And so for [inaudible 11: [11:39] 50] we're discussing a Tera Tera project for a soilgram challenge. I think that this diagram nicely illustrates where we're going and also the deeper standards and ways to deal with all of this data.

Woman 1: [12:10] Are there any questions?

Kolker on Encyclopedia of Systems Biology

Eugene Kolker: OK, so it's going to be short. Shorter than 10 minutes. I think we want to let you know about a new project, it's a encyclopedia systems biology. It's an online application, it's like encyclopedia, done by Stringer with help... Actually Stringer, if you Google encyclopedia systems biology you can find it there. So I can give you Website, it's VFworks.stringer.com slash MRW then slash and then we can see different encyclopedias are doing. So they did several before now they're trying to do this encyclopedia of different types of integrated analysis. We have several of the executive editors, I think they're called, with us. One of them is Bernard Dubitsky, he is member of Performance Editorial Board and that's why I know well him. So what we decided to do is... it's not standard book, it's not going to be actually published, it's online book. It's going to be updated for at least several years after it's going to be published so references are going to be updated. They have interesting mechanisms how to do it, how to simplify it more or less on Stringer's site.

I'm going to be one of the editor, editors of a specific chapter. So there's several, I mean I don't know 20 plus and the chapter I'm going to be responsible for is called statistics and analysis. And part of that chapter it's going to be something what we discussed in this meeting important, it's going to be on data standards. So whatever anybody wants to do along those lines or if you have any other ideas about systems calling, you know it's not exactly what we're talking about here, but you have other interests, everybody have different interests, right? So let me know, send email.

What I also wanted to tell you is a little bit about performance standards. We will maybe have a little bit of time to discuss what we're going to do like action items on different parameters of GSC. I can tell you that again our journal is more along the lines of systems biology, integrated biology but when we have appropriate applications to do a couple of cases that people send us and we suggested six.

I think one actually hopefully happening now but we'll see. So again, if you are interested in any type of data analysis of multi-organism type integration, end integration or you are doing competitive studies, I know that some of us mentioned here in our presentations that we're interested in integrated approaches, so send it to us. We are completely online, we have six issues a year we are now in 22nd percentile of all publications in biotechnology and I can tell it's like over 140, we were not in this space a few years ago and our expectation is about three in a couple of years so, and what's really cool, what's really driving us and I would say like semi concealable is special issues.

And we had several special issues for GSC, five to six, but now we don't get it. But five to six, and these special issues actually raise readership and promoted the journal. And in addition the citationing which is variable measure of how good or bad this journal is. Our rejection rate now is grew like 200 percent in last three years. First years was struggle but now it's really grown. We have a social database from Dupont Hills, in Chicago now from different countries or towns so it's, it's growing. It's reaching, I would say, 700 miles now to put in the right perspective.

So anyway, so if you're interested in some topics in general way, just be specific and you'll want to get the voice out through a special issue, let me know. We always welcome cool ideas from people I can tell you there is some special issues, we had one on personalized medicine and one on inter-genomics and next year, early next year we're going to have one on life formings so I think I'm done, thank you very much.

Kolker on GSC Organization

Eugene Kolker: [12:01] Kolker: So we started on the project officially in May. In June on ISMB we kind of first discuss it in public. Before it was always conference calls between board members. So today, five minutes about our proposed organizational structure. [12:21] So basically it's more or less standard structure by many societies and consortia. I was trying to get Pelin to understand the difference, at some point I thought I had but now I don't so. Primarily societies are made of individuals primarily, and consortia of institutions, but that's not always the case. Like we are part of some consortium by NIH, which it's made up of individual PIs.

[12:59] In any case, a structure - an average standard structure - looks like this. We have a society-scale board of directors. Sometimes they're called a council. And we have a board of directors. Another piece is a executive committee, usually it's made of five or so people. Usually it's made of a president. And as everybody knows, or if you don't, recently - it's like months ago - we had the conference for GSC and we voted, obviously, that the president was going to be Dawn Field. So for those that don't know, that's her.

[13:47] And then we also voted I can be vice-president. And Peter Sterk, who you heard a few minutes ago, is going to be secretary. I suggested for treasurer a person I know for years, as someone to understand about finances, because we better know that about ourselves. They with their projects.

[14:12] Usually the executive committee adds one to two people, most of their as part of it. They usually add members at large, one, two, or three. And an executive director to basically run the show. We don't have that person, although some of us have ideas to suggest.

[14:34] But any ideas are more than welcome, so if you want to participate in any of this, let Dawn, Peter, or me know. Do not overlook Dawn. But you may want to talk to us during this meeting or later.

[14:52] Then in addition to the executive committee, there's standard committees. Again, on average people suggest there are different names behind this committee. One name I happen to like is called the initiatives committee, which is basically not about initiatives, but about grants and funding.

[15:17] Suggestions we have right now, we have John Wooley as chairman of this committee. And if you don't know who John Wooley is - I heard, yeah, OK.

Audience Member: [15:28] No, no, no. I was going to say you could...

Eugene: [15:29] Yeah, OK. John Wooley, his official title now is assistant vice-chair, chancellor...

Audience Member: [15:38] He's assistant vice-chancellor for research at University of California-San Diego.

Eugene: [15:42] Thank you very much. Good.

Audience Member: [15:44] I worked there for two years. It took me a year just to memorize that.

Eugene: [15:47] So he spent years in different locations, different institutions, different posts, like in the DOE at least. He's well-known. He supports RCM and he supports this committee. So we must mention that John Wooley could run this committee. [16:14] But again, if you want to help this consortium development, please step up to the plate on any of these committees, including this one.

[16:24] And usually people suggest an industry advisory committee, which means that if at some point - hopefully earlier than later - some industrial partners will be interested in sponsoring anything we do. We don't have currently any ideas about the chairs, so volunteers are very welcome here.

[16:50] Another one is publications committee. Publications committee is pretty much open as to who we can suggest for chair. So George Garrity, and I assume you already gave your presentation, right? Good.

[17:07] So then other committees are the technology committee and resource committee. Again, we don't have right now suggestions for chairs.

[17:22] Next one would be standards committee. We have suggestions for this one. It's Susanna Sansone. She's going to present later today, right?

[17:34] Then we have to have a bylaws committee, which basically if you don't know what is bylaws, Google it or whatever. But we have somebody who is line here, we need to have somebody. And we need to have some standard bylaws for the society so we can use it for our own. But we need to have somebody keep on this.

[18:02] And then with a well-established society, they have an awards committee. They basically award each other some awards. OK?

[laughter]

Eugene: [18:13] They call it nicely like S&B, but basically that's it. And we don't have a chair for this one. [18:21] Then we would have two more committees and that would be it. One is called education and training. If somebody is challenged by educational issues, step up. And the last one is the finance committee. So it's not that you're going to be chasing for grants, but at some point when we have money, hopefully sooner than later, we need to have somebody who can look after how we spend this money. Not a treasurer, which is a little bit different.

[18:57] So we aren't going to spend the time right now, but please do think those committees and think about your role you want to play in any of them. If you're a junior kind of person, you don't need to be chair. You can start learning how, the ropes or whatever, being on this committee, being a member.

[19:20] And again, depending on the size of the society, these committees could be as big or a small as possible, whatever. But usually the suggestion is at least three people for a committee.

So with this room full of people, we can field them right away. Any ideas, please let the three of us know. Thank you very much.

Nikos Kyrpides on Gene Calling Standards

Nikos Kyrpides: [0:04] So good morning everybody. So this session is, in a way a continuation of a similar session organized at API a year and a half ago, which was not as successful as I hoped. And the reason was that it was packed with talks, one talk after the other, and then we basically had no discussion and no real results. [0:29] Fortunately, Owen White stepped in and saved in a way the session by proposing something that was actually followed up. This was all the sequencing centers who post all their analyses, who post their special piece.

[0:45] And that's basically one thing that has stopped happening in the sixth journal. We have seen that already, a few special piece being published and hopefully the sequencing of the analysis center will stop posting their dictatable philosophies on how they do annotation.

[1:05] But Mike Gold was ready to come out with an agreement on certain things and this was restricted only to gene calling. So the session, the morning session has been divided into two parts. First will be gene calling, first half. This is what I will be discussing and the second part will be functional annotation.

[1:29] So rather than again having several talks, I think more or less we all know what everybody else is doing. We thought it, it's the best to restrict the discussion and keep it to one talk. And just by accident I thought it was, it is better if I give the talk [laughter] and the discussion is based on a plan. The plan is is the specific proposal both for the first section which is the gene calling and the second section which is Owen presenting on functional annotation.

[2:01] So again we'll know again more or less what everybody's doing. So we're going to take now one step forward and see how we can in a way standardize what we're doing.

[2:14] And this is in a way also follow up with discussions we had yesterday with four good and then five and my, let's say the document that I made that I am a hundred percent in agreement that we have to push forward within five. But we have, not us as a consortium but as a community, a genomic community.

[2:36] After 15 years we have not shown a single success story. We do not hear of a single success story. So maybe it's about time that we have a success story or try at least to come up with something, starting with genomics before we move to, into metagenomics. Actually that doesn't mean that first the success story in genomics should come but, I think that's easier because we have 15 years of experience and 15 years of history behind us.

[3:05] So what are the problems? This is what I will start with and then I will follow up with a proposal of what we can do. So it started with redesigning the program, this is a slide that they have shown several times in the past. I'm sorry if most of you have already seen it.

[3:21] So this is a typical exercise that almost everybody's doing, comparing closely related organisms. So in this case we're comparing mallai to comparasitive malli. So one genome was sequenced, these are both the genomes that have been sequenced several years ago. One was sequenced by Tiger, this is an obligate parasite forces, and the other was sequenced, the

sequenced mallai was sequenced by Singer Center. It's actually a free living organism. So in order to compare the two and compare the size, of course the free living organism has a much larger size 7.2, versus 5.8.

[3:58] And the number of genes of course is much larger on the free living, 5800 versus forty seven hundred. So we're curious to find out if there's a specific change on the smaller because on the smaller genome, on the parasite that cannot be identified on the larger genome.

[4:12] So we did this mathematical conversion, find genes in genome in mallai that they are not in should mallai. The result gave us 548 genes. So this was just a blast search basically. So the result was pretty strange for us, we did not expect that high number of genes. So this, 548 is more than 10 percent of the genes. So then we decided to go finally after each one of those genes. The result was really frustrating. We found out that almost 90 percent was wrong of the results.

[4:48] So 90 percent of the genes of the 550 genes where data had been added in other genome or deleted in a genome. So in this case these genes most probably would have been deleted or this gene would have been added here. We don't care really who is right and who is wrong in this case. The real problem is that the two sequencing centers used two different pipelines and the result was 90 percent error in the comparison.

[5:19] So this is extremely, it is a huge problem. So if you try to do a large-scale comparison across closely related organisms, or different studies of the same organism that have been sequenced by other, by different sequencing centers, you cannot do a mathematical conversion.

[5:34] A lot of, some analysis groups have gone to, all the way to different, they have followed a different route and this is, "I don't ask anybody. I'm going to do everything on my own, find it for myself", which actually appears really...It is quite reasonable because in order to be able to compare, you have to be able to do it with the same pipeline.

[5:56] However, then you will differentiate from the rest of the world. And that is another problem. So there is no real solution or ideal solution for this case. And why this is a problem, with gene calling, the reason is there are several different, still several different algorithms that they are producing quite different results.

[6:17] So when you are working on eukaryotic coat, and we have here, we got here from from JGI, who's leading the prokaryotic annotation group, specific work there, they usually tend to think, "Oh bacterial, it's easy." And it's easy because there are no extra genomes, which is really a big, which is really my work and increased but it should have been much easier.

[6:38] So after 15 years in the field everybody expects, oh gene calling is not a problem. We should concentrate on functional conversion, on conversion of function prediction. Which is to a great extent true but, to a great extent correct. However the gene calling in Archea bacteria aren't yet, it's still a national problem because we cannot really compare similar datasets.

[7:04] So as a conversion, or as an illustration, we have here a comparison for two genomes, the myco bacterium allow us to see, here mentha losturula and then co-bacterium, high as 67 percent, 55 percent. We'll compare the gene coding in the two genomes by five pipelines. I must

rename the pipelines because I don't want to focus on who is the best pipeline here. The real focus I believe should be on the percentage of variation.

[7:36] So what we compare here is how many genes they predict, each one of them, so you can see it starts from here, 48 up to 55 on the genes. Here it starts from 29 up to 3300, so there's 400 genes difference. This is 700, almost 700 genes difference. Seven hundred genes difference in a genome of 5, 000 genes. This is just crazy.

[8:08] Then you can compare the number for what we call anomalous genes or dubious let's say genes, actually dubious is used for, we're using as, for another example so I'll come back to that later. But genes that they seem to have a problem. And the problem may be they are short compared to all the other genes that belong to the same family or too long. And I'll give specific examples of what they are.

[8:36] So we see here a big variation again, from six percent up to 21 percent. Or in genes that have been missed. We know they are, these are real because they have very strong hints to other genes from other organisms from other genera, not just from the same strains of, from different strains of the same organism. So the genes are different. We see they rise here from eight percent up to fifteen percent. So this, this is quite a lot. And the same here, from six percent, from three percent actually up to nine, almost nine percent.

[9:08] So there's a huge variation depending on what gene encoding you're using. And these encoders are actually some of the most often or widely used methods. So if the most widely used methods produce such, so much variety standards, what can we do? We have been discussing this for years and years and we believe that there is no way, this in a way also cultural thing, there is no way to convince let's say JCVI to abandon Glimmer, they have been developing for years and adopt let's say G-Mark. Or there's no way to convince Mark Ordosky to abandon G-Mark and adopt Glimmer and vice versa. So then what is, or, it's not just the developers, it's also the users. So what could be a solution?

[9:56] So rather than going and developing one more gene coder, which definitely could have problems because there's no, we can see here there's no ideal method, we thought that an alternative would be to try to build a consensus. And what will be the consensus? Not just compare all of these gene coders and take the consensus result because we have done that as well and you end up having half of the genes because there's huge variations there. So what we did or what we felt could be a solution is to have a post processing pipeline.

[10:29] So imagine that we don't care what you're using as a model encoder, you may use Glimmer, you may use G-Mark you may use Prodigal. Anything used you can keep on using it and then have a post processing pipeline but we have specific standards and then you will get the same, the same outcome as theirs. So this is what we tried to do with a tool called Zimprint. It's a prediction improvement pipeline. So this has been developed by Rita Bati and Natali Onnova. Natali's here, Rita probably presents later.

[11:02] So what is this learning? There's a pipeline that consists of a series of computations of users that identify encoders in gene calls and mixed genes and correct the subset identified to affect the futures. So what we do is actually account, I'll give an overview of the application. So originally we developed it in order to improve the gene coding and the JGI. But forward thoughts

and after development we realized it also can be used as a benchmark for gene prediction, code it for gene prediction algorithms.

[11:32] And this is what we have on the previous slide. So you can run any gene coder you want, and then you run Zimprint and you identify, it gives you automatic output of the work identified as post work errors. So this doesn't mean that all of those are errors but they're defined, something's wrong with those genes. So what it does it identifies the genes that seem to be a hundred percent fine and then it separates the group of genes that make it a problem.

[11:58] And the problem may be, the gene is too short so all of the, the rest of the gene family is clearly another analysis that's longer or another from analysis that is shorter. Or it may define genes that have, that are appearing unique, that they don't have any hits and they are very small, very short. So these are dubious genes, maybe real, maybe not. But it doesn't mean they have been deleted but they're identified there for a specific reason and I'll come back to that.

[12:27] Then it can be used also for best for code it for combination for coordination of covers of sequencing platforms. And this has been used already by the seven or eight big, sequencing, several sequencing groups from the Human Organism Bio-Project to examine what is the result if they use alternative sequencing platforms or combinations of other sequencing platforms. What is the result on the final gene set. And those of course are group, sequencing quality because the tools identifies the pipeline and also identifies possible change.

[13:05] So we have been on it, we have been using this at JGI for the last three years so this is constantly improving and the way it improves is from, there are complicated, analysts there, this is all they do for the last three, four years actually. They manually examine all of the gene codes that are identified as erroneous and they manually correct them. Of course the goal is not to continue this interminably as a manual because it is impossible really to catch up with the, you cannot scale.

[13:39] So what we do is based on the feedback we get from the manual correction we're trying too to make that part of the correction. So the tool is constantly improved because we identify less and less problems or less and less false positives, let's say. So if it's in, it's correct, it's automated as it runs but then also the next step is that we are trying to repeat automatically the manual correction, so then the tool is already automatically correcting the errors that are identified.

[14:12] So this is more or less fifty percent, the single pattern is probably at least fifty percent done. So about fifty percent of the cases now we have automatic correction as well. We haven't implemented it in production because we're still testing it. We want to be absolutely certain that we're not going to start correcting genes wrongly, in a wrong way.

[14:33] So but overall we have examined over the last four years over a million genes and about thirty percent of this, of those, have been manually corrected. Or so in other words we have about ten percent error rate.

[14:48] So what, how the tool works? So we have, we start with any DNA sequence, you can use any tool you want, Createacall or G-Mark or Prodigal or Glimmer, you get your gene codes and then you run Zimprint. And Zimprint gives you a list of reports. So it separates the genes that are

there, OK, so you don't have to worry about those and then you get a list of reports which is that they're overlapping, short genes, long genes, unique genes, double genes, broken genes, transposed genes and missed genes.

[15:18] I don't have, I don't want to go into the data of how the whole process works but it's a series of steps so every step has specific rules. And the rules, and I think that's the important thing, the rules, have been developed by, through constant manual integration and analysis of those multi-million, manual integration of more than a million genes and it's absolutely open for debate and discussion. And that's a good thing. But it is not written in stone so if someone disagrees with the rules we can say that we will modify them.

[15:51] I want to go to some examples. For example this is a case of a sub-set; so this is the query and you see that they are all, most of the hits are more than a hundred and six longer so this starts somewhere over here. So this is, this doesn't mean that it must be expanded, but it is marked as a case that has to be examined. So what we do here is we can examine if there's an opening, if the opening can't be expanded or how we can have an analysis upstream or something like that and if there's a finding over there or the conditions are OK, are good in order to correct it.

[16:25] This is a case of a long gene, so this hit is getting more analysis upstream as well where most of the hits are starting around here. And you may see some that are much longer but then of course you pull the bench of what is the strength of the similarity here and also what, how close they are if this may be just different strengths of the same organism. Also there's another report of recommendations. And this is what makes it quite often very difficult to automate because my original you can inspect and examine several different alternatives scenarios but to automate that is become, is increasingly difficult.

[17:06] This is not a good illustration but that additional things to do is identifying new genes, but these genes that do not have any hits, they're dubious. Dubious recall genes, they're unique but also they are very short. So why does a report define those? Because quite often these may be wrongly depicted genes, these may not be real genes. So what we do is once we identify, one of the steps on the pipeline is to take all of the genetic regions and plots them, of course a most useful way to find missed genes. But this is not enough.

[17:36] So I told them the region that has a real gene is masked by a gene because we wrongly predicted. And this, in most of the cases a unique gene, doesn't have any hits so if we delete, if we mask this region we are allowing, we expand basically what is surrounding the genetic region, we allow prediction of genes that can actually be masked. And this is something we are finding not very often, but often enough to be significant. So real genes not being called, genes that have similarities are not being called because the region has been masked by another gene that has absolutely no similarities.

[18:11] So it is important to define when we find it and then we have another round of masking region and blast them again, this region. Broken genes and genes like that where they're next to each other and then there are hits, and they both hit the same gene. So this, these are obvious case of possible fractures and I believe also the fracture detection, the acpi's doing something similar.

[18:42] OK, so this is more or less all the entire presentation. So what I want to do on the rest of the time we have is discuss whether we believe this is something that can be world widely adopted, we have a common post process mechanism like that. We know already for example the Sanger Center is doing this for years. What they.

Man 1: [19:05] What are they doing?

Nikos Kyrpides: [19:09] I said Sanger Center and then was... [laughter] They're everywhere. Can you hear me? So Sanger Center has been manually inspecting their genes. And this is what is creating what then is very high quality gene codes. However, as far as I know they are not doing it in a systematic way. And this is what, this basically allows to separate the genes, the good from the bad. The genes that seem OK, based on, I'm sorry, on those are and everything else and they define the genes that are making a problem and then post on both genes. [19:50] We believe that one of the sources to the problem, one of the sources of the problem is, that there are no agreeing standards. So different groups are having a different couple or lower threshold on what is the minimum size of the gene. So I remember a group, a few years ago, a group in Japan not coding anything below a hundred in masses. Which is probably the case because there's, we know there is a large list of genes, smaller size. Other groups, give a couple of thick vinoses and again there are several, some of the good are and they are thirty, thirty five in masses.

[20:51] There are no standards on what is sequencing or whether the pseudo gene should be identified. So we see some genomes coming out with an incredibly good list of predictor pseudo genes and other organisms, other genomes, have absolutely no pseudo genes reported. So this is, this is some of the problems, you cannot relate and compare pseudo genes across different, even closely related organisms because we have to do everything. So we have done up as, I believe as a consortium with a list of standards that will say this is what we base even if it's a minimum set of standards so it will be at least there will be some sort of guidelines for the rest of the community.

[21:36] So we have here a list of proposals that actually Natalia put together and I would like to ask you to comment, discuss those. And then present those and then basically have discussion after that. Yes?

Man 2: [21:55] So how, would it be possible for that curve to genomes but then you also have a way to locate them? The problem for this fragment for [inaudible 22:11]

Nikos Kyrpides: [22:13] : Again it depends, it depends, I mean, a look at value until decided, it depends on what you are looking for, what technology you are using for. For example if you were going to use Italian, we know that there is a significant problem. This is a holy merge. So we are going to define just words consistent next with each other and things like that. Some of the parts can be used. Some other parts do not be applied or have to be largely modified. [22:38] Over and all the big bottle in this glass basically is not mere glass and everything, which we're doing in a way with all of the data genome data sets, but to the extent that can be used and applied right now or let's say, in the shortage of time from now, will allow low value, that's something relevant and its really, it's quite important.

Man 2: [23:09] The rows of extended genes is beautiful.

Nikos Kyrpides: [23:12] Yeah.

Man 2: [23:14] Do you recognize these are routine guides?

Woman 1: [23:17] I can try to answer. First of all basically Filial or any kind of gene prime title quite power, quite useful only unfortunately, can be useful only on single data or may be on the illuminant data when they will be available and long nice quantiles without reagent whatever hundreds of data types. Right? [laughter]

[23:36] Anyway, because unfortunately, that is in our experience there isn't a single ab-initio gene caller that can deal with these rate of ratios. So as of now for instance, as Nikos said, for titanium data or ulterior lot of types of hyper data we are using basically not ab-initio, but evidence based, mostly evidence based gene pool resource from Aztecs. And there, yes we take care of these atmosphere noting that whatever we can see that we have lost extra within top of several trees we will merge them. So we are sort of running it upfront without warning.

[24:15] As for single data genomes, we know that basically about 10% of the genes will be false positives. That's what the forms will be comparable assimilated data of dissimulated type of cell. And of course all these false positives, all of them are very unique genes. Actually, not quite sure because some of them will have very unique genes as compared to referenced databases, but they will have hits within the same metagenomes. Those will make the same calls.

Man 1: [24:45] Are any of those unique genes are find right now to come closely to coli?

Woman 1: [24:51] All of them. Not all of them but.... [inaudible Cross-talking]

Man 1: [24:56] So what you're saying is that even the rule, if I see the gene twice, it's probably true, even that...

Woman 1: [25:04] Within the same metagenome, right? If you see anything in other metagenomes or in reference genomes, quite likely it is not, because basically you have...

Man 1: [25:13] Basically it's the same predictions as the same, the same regions from the same or similar strain in the environment?

Woman 1: [25:23] Again, in our experience it is highly unlikely that we will have the same strain in different environments. That's why, basically that's why you have this kind of thing that is unique, the probability will be in your favor.

Man 2: [25:34] The rule that the DVI people use to see if genes and see genes, other than in data sets.

Woman 1: [25:45] Yes. Basically on this you have metagenomes coming from very similar environments, you could have two false removals largest, that still make way the same organism, slightly different strains. We won't be leaving here for an hour saying but it is saying that you fight in standards.

Man 2: [26:03] We could see two neighboring, hundred miles apart route winters?

Woman 1: [26:09] That I won't know. That, I don't know. Anyway, so the discussion, it has two sides of it. The first one is that in a way, these requirements of all the gene calling standards and so on, they have to be not enforced, but it is suggested by the databases, the primary sequence archives, right? [26:34] Why? Because until very recently for instance if you looked at what is the requirement of gene ball, how you should use two genomes on mutations in the gene. They have I believe, at least three different choices because it is a gene feature, it has something that is a miscellaneous feature, or is just a feature, which makes a huge mess actually. We...

Woman 2: [26:56] It's just a feature... what is it? It's just a feature.

Woman 1: [26:59] I don't remember what kind of feature... [coughing]

Woman 1: [27:02] I'm positive that which is made.

Woman 3: [27:05] I think this seminar is, already, always submitted to, then to preferences is same. Woman two : [27:11] That is our preference.

Man 1: [27:12] Yeah, That's our preference.

Woman 4: [27:13] That's how the...

Woman 1: [27:14] And often the sequencing was submitting them against genes. And, I think that [laughter] I've seen in some Japanese genomes is original whatever. For some reason it's routed for some. Anyway, in this lecture the problem is if you try to after that, because we kind of, we are on both sides of this process. We are submitting the genomes to a gene bank, but we are also taking the genomes back from research, and trying to incorporate them into our ING? And that's how we find the problems that Nicholas was discussing with his slides in the beginning. All right.

Man 1: [27:45] So, in light of everything that has been said here, that, you know, there is no tried and, there is no absolute truth. We can only hope that we [noise] a certain asymptotic encoded curve towards a true gene calling. Shouldn't this process be accompanied by some sort of...either both evidence codes, and as to, you know, how these genes were called as they work, and also some sort of significance score, to tell or, you know, some sort of scoring, saying, OK. this gene has been called by letters A, B and C...

Woman 1: [28:27] Mm-hm.

Man 1: [28:29] ....and the consensus significance is an E value, or a P value, or other significance measure, or just some sort of bit score or whatever of so and so and concentrated within.

Woman 2: [28:43] But it's changing, it's not a change that ...

Man 2: [28:45] That's it? [cough]

Woman 2: [28:46] ...a significance for keeping... [cough]

Woman 2: [28:48] ...to be a continuous process. [horn sound]

Man 2: [28:48] Well developing it...

Woman 1: [28:49] If they are talking about something like blast force. I think that most gene covers are beneath gene covers, they produce their coding potential,.., or something like that, which you probably won't.

Man 1: [28:59] OK, I am saying that.... What I am saying, I am not saying how that, I am just saying that.....

Woman 2: [29:03] It doesn't have additional information, additional evidence that gives you more weight by posting it, more value, and increases the significance....

Man 1: [29:12] I made the series of scores, you know, something that we present when we have a genome, and someone wants to look at it and make adjustment calls based on the genes and where they should be, they would have some sort of idea and a set of anthologies or keywords that will tell them if, you know, how good the decision is they are making, or how risky is the decision they are making, maybe it might be a risky decision, but they still want to make it, but they will be given the choice, so I think that would be a perfect option, so I think ...

Woman 2: [29:43] The important thing is to be school-smart, so I think. [everybody talking at the same time]

Woman 1: [29:57] We have all these experiments, we tried to get all this data to validate our impulse, you can't get better than 50%, 50% of the products will be validated, but the best product [everybody talking at the same time] that you can validate without.

Man 1: [30:17] I am saying all this information is not valid, nobody knows this. They take the gene calls in the genomes as gossip, and nobody knows if these can be erroneous, and some of them can be, you know, have a low probability of being correct, you know, they can higher probability of being correct.

Woman 3: [30:36] I think this was about confidence, more on the evidence, because evidence is...

Man 1: [30:40] Two things - one is evidence -- how do we get this? And.

Woman 3: [30:46] That doesn't really apply.

Man 1: [30:48] That doesn't give them, that just tells them how this was done, some of them won't care, some of them maybe do care, but I don't know.

Woman 2: [30:55] We have more people.

Man 1: [30:57] But the other one what is important is confidence, exactly, that's important, that's out there.

Woman 1: [31:01] And right Nikos Kyrpides wants to say something about.

Nikos Kyrpides: [31:06] Thanks, I very much agree with this, I wanted to have a score, but one of the things you have to agree with is that it is pretty darn hard to create a probability, and I've

never seen any progress in that area. [31:21] And I can certainly understand the motivation of wanting to have it. But in the absence of that, having evidence codes that describe where it came from, to give the biologist a fighting chance of choosing, "OK, I want everything that was based on experimental information", or whatever.

[31:42] And also, this is usually where I put in a plug for SOPs that would actually travel with the data. Having some type of description of where this information came from, and how you arrived at it and having that attached to the gene is, I think, the best you can do in the absence of having probability.

[32:02] Because the probabilities, especially for gene finding... They're there, but a lot of times they're talking about, for example, small genes, that there is very little statistical evidence, but you do want them to come along. And a lot of times you're talking about... You're calling it based on sequence combination between small genes, or something like that.

[32:26] So, the probabilities, or what you'd be basing that score on, would be different for different classes. And I just... I don't think that you can create one uniform score. So, in the absence of that I would just advocate M codes and SOP.

Man 2: [32:40] So, there are two things. One is, yes, that you can do probability scoring investigating them, evidence score maybe, on SOPs. The other is, can you think of some sort of primitive... Like doing, maybe, some sort of menagerie of a score. Like, somewhere on scale of one to five... This is a five, this is a four, this is a three... Something, you know, which is rough. Got through...

Woman 1: [33:07] I think we're better off, likewise...I think exactly that.

Man 1: [33:10] Great. I think there's something out there. When they're down-sourcing a gene, they really don't see that, even if it's... It's rarely there.

Man 3: [33:20] I think you're partially underestimating how intelligent the users are. If we make it transparent, it's all the same, our genes all... One of the systems I designed in the past made it very clear to NGB and to me what all the users looked at. They understand that.

Man 1: [33:35] Yeah, I know.

Man 3: [33:36] You don't really need this vehicle to do the whole thing because as I've already said we haven't figured it out on the data.

Man 1: [33:40] OK.

Man 3: [33:41] One of the questions, I mean, these four, we're not very good at that. One of the things I would like to ask is, we have been contacted and harassed us by people asking, "How you're doing your gene calling, you know, XY and Z approach. Could you make this available so other pipelines could use that?" I'm pretty certain you would be willing to make that at least available if not a default another pipeline.

Woman 1: [34:05] Which one, which one? Usually only the really big ones because available basically there is a sort of public level of the server rate and some of the genomes you can. I

don't think that we have, we are really right now to basically to distribute the code and the reason is the legal department unfortunately, not because the lack of desire, right?

Man 3: [34:25] I was hoping the score of the final gene is maybe this would be a workflow.

Woman 1: [34:29] We cannot distribute any source around here because that is better here I'm sorry.

Man 1: [34:36] Isn't your only chance if you really want to have some sort of consistent annotation that you re-analyze all the genomes in the same way?

Woman 1: [34:48] That's the dream kind of thing because as of now, as Nikos said, our process is heavily manual, right? We identify the genes that look suspicious. But after that somebody will have to go and look. We are going from that, we are trying to automate whatever correction we can make, without manual operation. After that we will definitely go, that's one of the things that Nikos has been wanting for the past four years I think.

Man 1: [35:11] Because I was talking about genome scores or constant scores even if you have constant scores if you took constant scores even if you have different pipelines, it's still some comparison.

Woman 1: [35:24] Anyway as I said basically, I think that every admonished gene core comes up with some sort of measure of coding potential organism, of the region, this is good. The second thing is usually they also provide the same kind of measure for the start signs, because that's one of the potentials for these things, right? [35:42] So if you can incorporate this as evidence, it is already better than nothing. If the gene has what if no effort to work, outside of the genomes preferably, right? It is another area. And so on and so on so you can do, if you can do a prokaryotes table, wonderful except you usually don't.

Nikos Kyrpides on a GSC Global Genome Census

Nikos Kyrpides: Hello again. GSC Global Genome Census, what is it? Obviously some people have absolutely no idea. Even the name came up with discussion with Dawn a few weeks ago. The whole idea came from discussions we had with several people that are in the room. Actually a lot of people have participation or can definitely claim a part of this whole idea. Until this morning I didn't have any slides, because I was just thinking since this is something that is following over the last couple of weeks, I would just talk. But then, I tried to at least put some up. As I said, a lot of people have been talking about several different aspects of the proposal I will be presenting. For example, Jonathan Eisen, he's here. Since I first met him, he is interested in global or world domination. So this was quite challenging, a different idea.

JGI's interest over the last year or so in initiating grant proposals or grant ideas, and definitely HMP. I would definitely say what we have seen coming for the first time, this is standardization and coordination across several large sequencing centers that were in fact competing until a few years ago, working together as a team and managing to achieve standards.

So all that gave a more general idea why we would try to expand this. Those are definitely my friends, committed scientists who were also friends, kept on telling me that biologists do not think big enough, and the problems we are used to dealing with are very small. So I will start with one of the most inspiring slides, which I think, or actually journeys. Everybody can see, this is a JCVI project, the Sorcerer II expedition. This was definitely a landmark in the whole field. This is the route they followed.

But in fact it was really an incredible achievement at that point. If I remember correctly, the number of genes in GenBank was doubled by a single project. However, seen from another perspective, this is incredibly small effort. If the goal is to understand microbial diversity, the meta-genomic status, this is incredibly small. And if we follow these type of projects or these type of efforts, it would take us several hundred years, really, maybe even more, to get an idea of maybe what is happening.

We know already that the key part of understanding the microbial diversity or meta-genomic status is [inaudible 03:12] . I think everybody is getting very much excited with the challenges of the meta-genomes, and most of the community is actually focusing, or most of the experience community in genomic's is focusing right now on the meta-genome.

And I think a big part of the community is forgetting that for the first time we have a unique opportunity for the isolate genomes. And not only just challenges and a lot of unanswered issues, but also problems that they may think are not real, problems with the meta-genome that are still persistent with the isolates.

So what has been driving the last few years the whole field is the individual projects and large projects, large scale. Starting with a few years ago with the Moore Foundation funding almost 100 marine organisms. Followed by the WUGSC effort in sequencing 100 microbes from the human gut. And I believe the two largest efforts over the last three years, the HMP effort,

sequencing 1, 000 microbes from the human body, and the JGI project, which is called GEBA, which you have heard already Genomic Encyclopedia of Bacteria and Archaea.

Is there anybody who doesn't know about this project already? You know about this project. So Jonathan Eisen is leading this project. The only thing I want to say, the only slide I want to show from that is my favorite slide from the paper, which is actually also accepted in "Nature." This slide, put together by Doug.

[04:53] , shows the coverage we have from JGI genome project we have up to this point, which is the blue here. A much better perspective.

All of this is the diversity, essentially, from the actual genomes from the actual organisms at the base, and the lighter grey is from all the data in the database, so that includes the uncultured. And the red here is the coverage we increased in 16x distance basically, through the 56 genomes. But this still gives you a better perspective of how small, what is the coverage today. More or less, different versions of that slide we know already. We are covering 1% to 2% of the diversity, the cultured diversity.

Another way of looking at it is through Genomes OnLine. This is the total number of genome projects at this point. So 4, 100 bacterial genome projects representing over 700 genera and 1, 500 species. So what is the number of type species, George, at this point? Officially characterized?

George: The number of named type species is approximately 11, 500, but that's probably about 9, 000 individual organisms.

Nikos: About 9, 000. OK. And say the genera at this point is a little under 2, 000 or about 2, 000. So this seems an immediate first goal. We have to draw up the goals. And even with a process like HMP, there's over 1, 000 projects right now in GEBA, which is already around 300 genomes. We're here, 700. So we keep on increasing the number of genera that we are covering, and the species for which we have a sequence representative, but the number of isolated genomes in genera and species is also constantly also increasing. So the percentage more or less remains the same. This is an obvious goal. How can this be addressed with very small-scale projects like HMP or GEBA. The question is, where do we go from here and what can we do about it?

Another perspective is also what's happening in the entire world through the democratization, basically, of sequencing technology? This is up to date, five sequencing centers. There's JGI, JCVI, BROAD, WashU, and Baylor. They're covering more than 40% of the world's production in archaea and bacteria genomes.

Actually, I'm not sure how this is going to change. I can see it both ways: democratization would allow every single university to start their own projects, and I'm pretty much certain this would happen. Actually yesterday I heard that Sanger is sequencing bacterial genomes with $500, $500 per organism. Sequencing and annotating.

So this is incredible, with $5, 000 or $6, 000, you can have a dozen microbes. But nevertheless, every single university today can afford, not every single college of course, but a large number of universities can afford to buy formal thing or aluminum about microbial organism. So we expect

that the percentage of the rest of the whole will increase but I believe so whenever the capacity of the large sequence extenders to address big projects like, again like HMP, or like GEBA.

We get to, so what is the answer and whether we're going to see increase of thirty seven percent compared to the rest or if this will remain the same, the certainty is that these numbers will go way up and Patrick Swain had a slide before about that. I am in a way having my own prediction here, Ankleberry goes over his, so this is where we believe we're going to be by the end of this year, 2009, we're going to have our own [inaudible 08:56] by Kevin [inaudible 08:57] and about a thousand drops. Actually I think we're really, we're already there. This slide was put together three months ago and I think we're already there.

So the expectation based on that, this is the first fifteen years, I think that we can safely predict over that the next five years we're going to have three times that number, nine hundred, twenty seven hundred complete genomes and ten times more, about ten thousand. Patrick's prediction based on the current, based on linear increase of what we see today is that well we'll reach these numbers and actually higher numbers by ten, twelve, so the next three years. I think that we, that we will be both surprised that this average will be much higher.

So where this is leading? Where it leads, this is putting the whole genomic community but also the full, the GSC. How is it possible then to achieve standards across not just five or six different sequencing centers but across hundreds, or maybe several hundred different sequencing centers? Well we can go, this is where we're going to what Jonathan's talking about world domination.

So we can think about a project but it's not going to be only after JGI or the human micro-biome sequencing centers, the big five sequencing centers to deliver the rest of the world, what the reference, the key reference here are very important for any environmental studies. But we can figure about having an international corporation where we will essentially ask the whole community or everybody around the world to contribute to a much larger scale project.

What we can do, in ninety five for example, what is kick off in the next three to five years. Let's say she have the Berka's and all the organisms are in the Berka's manual where we have all the types of species, let's say I identify five to ten thousand organisms, I organize them and I say this is a list that we want to share for the next five years, three to five years and then we request ask the entire community of the world to start adopting organism.

Adopting organism is again, Jonathan's favorite phrase. So what are the success stories for that? The International Human Micro-Biome Consortium actually they have already achieved that, they have achieved something that a few years, maybe even a couple of years only it will be come but it will seem impossible. You have groups from US, Europe, Asia to have a [inaudible 11:38] of let's say math and competent schools and coordinating what they will be sequencing.

We have so much to sequence that of course the last thing we want to do is keep on sequencing the same things. This is happening all the time. We are seeing it happen all the time. There is a loose interaction between GEBA effort and X and B. [inaudible 11:56] where it overlaps. We found doing that there was at least ten, probably even more than ten percent concordance between organisms we select. So having a certain dimensional effort on who is sequencing what is definitely key in avoiding overlaps.

Another key access story is second POM. We can pull something like second POM. In this case the request, different sequencing centers for request, participation from all the sequencing centers around the world. Here's GSC, or a group from within GSC defined what is the important catalog, let's say five or ten thousand organisms, he comprised a list and then ask everybody to adopt, start adopting individual organism and contributing it to a chain in a national project like that.

Funding is definitely another operation. It is a very important issue but at this point what I'm talking about is to a great extent self-funding. I do believe that with a project like that you can definitely go and request coordination, cooperation between different funding agencies. But it can definitely start with self funding. It's really funny how often I'm hearing of people from a different university that can put sequencing in but don't even know what to do with it.

"We can sequence but we don't know what to sequence." So I don't know to what extent this is true or you know this is happening but definitely groups like that can participate and fill a spark for let's say world initiative. Countries from Asia like Japan, Korea and China, they're putting an incredible amount of resources into micro-biome genomic's. I'm sure, and I do feel and from interactions with them they feel isolated over there doing their own projects without really much interaction with the rest of the world. I do believe that such a project would bring a much tighter coordination across with everybody.

The final and I think most important part of what we're all about, well finally it seems that true standards or true, we can expand the standards we're trying to define here through everybody not just through, across four or five big sequencing centers but across all the universities that want to get into this field or all the new groups that really want to contribute and be part of that. Key partners for that is definitely the GSC and the proposal basically is to do it through GSC.

Contributing centers I will place as the first priority which have been discussing over the last few months nobody with hospital plan was also here today but also with the European Council Organization which represents more than I think fifty different cultural election centers in Europe and they are extremely eager in contributing to such a project or a much expanded particular project for free.

So grow the organism and the record to subsidy at the national level. So Hospital Claim was regulating the GEBA, the growth of public GEBA organisms for the GEBA project is a, has a successful effort and an interest in getting that far and coordinating all the cultural election centers. Representatives from grass roots projects, like GEBA fill a job as already agreed. I think, I definitely believe jobs occurring from Human Micro-Biome project are key.

People know that as well. China already has, is using this thing in contributing to that through another outside project which is called, Terragenome and this is another international effort in Human genome, in the genomic area. However again, the key part there, the key point is actual the organisms. So what are the key isolated organisms that need to be sequenced in order to support that program as an ongoing project. Of course, kick off of this has to be if we're starting to start, to kick start this [inaudible 15:53] centers and different company members. This discussed between the Russian work group. Next step, I don't know, that's up for discussion. That's all.

Audience: [applause]

Ilene Mizrachi (NCBI) on Standards and the INSDC

Ilene Mizrachi: [0:02] I'm Ilene Mizrachi, and I am the GenBank coordinator. I am going to talk about how the MIGS/MIMS/MIENS standard fit into the INSDC. And I'll also give a little bit of an update on what was the GEO projects database and has expanded the projects database to encompass other sequencing projects. Part of what we at INSDC have to do is try and capture the standards that are set forth in MIGS/MIMS/MIENS checklist. We have a couple of different mechanisms for accepting this data. At GenBank, we use what's known as a Structured Comment, which is two-column table in the middle of the GenBank FlatFile file that has the tagged value pairs.

Last year at GSC we said we would be willing to validate if we got a list of requirements that would be in the Structured Comment, that we would be willing to validate these structured comments and to put a tag on that it is MIGS-whatever-version compliant.

So far I don't think that we've gotten the final list of which fields are required and which fields are options. So I look forward to receiving that whoever. And Nikos they keep the genomes has been really good at providing the data to us.

The other things which is coming up soon, which was referred to by - I'm sorry I don't remember her name. Was the samples database. And right now in the short-read archive and the European short-read archive, there is sample object.

And we, with EVI, hope to expand this to make more required fields so that we can use that instead of or in addition to the Structured Comment to capture all this data. One nice thing about these databases is that a single sample might be used for a sequencing project and an expression project and other projects and databases that we have at NCBI, and we would be able to link all the different data to the single sample. And in both of these cases we can validate the compliance.

This is a sample Structured Comment on one of the aided genomes. You can see that all the different fields are available, are present, and have values. And we've agreed that we will validate for the different tags. Unfortunately, we're not able to validate for the values of the tags, but we'll just validate for the presence of these tags.

In addition to the Structured Comment for the meta-data for the source organism, we are also going to start having another part of the Structured Comment for genome assembly data.

Part of this was developed at the request of the HMP project, where they want to cite the standards for finishing the current finishing status, the assemble method and the name of the assembly, as well as the sequencing technology.

This is just a good place for somebody who's looking at a genome to find out what the coverage is and how finished it is before they want to include it in their own analysis.

Now I'm going to talk a little bit about the transition from the Genome Project, which we started in 2004, to the new projects database. We left it at the Generic Names project, because we have

expanded the scope way beyond genomes and are making it more flexible so that we can expand it in the future.

Some of the problems that we had with the Genome Project database arose when we started getting different types sequence data other than strict genome sequencing. The database was not easily extendable. We were unable to liken hierarchies greater than parent/child, and we were not able to link project to each other as peer projects.

Our submission system, you have to fill out a form on our web page, we have no batch submissions and the accessions are all stored in the project's database. And so when we thought about redesigning the database we realized that we needed to solve many of these problems and so we over the past year have been working on a new XML schema for the new project's database.

And we're, are going to allow more flexible linking between projects so you can have different projects, you can have parent-child relationships, you can have hierarchies of relationships. Rather than the project's database storing all the accessions, we are having all the primary data point to the project's database. So that a single project can have data from GEO or Ray Express and it could have short re-archived data and it could have gen-bank data.

These are just some of the requirements that we had before and the requirements in the new project's database. The new project's database has a number of different project types that we've set up for now. There's a top single organism project which is going to be as similar to I'd say a gold page, as, it's an organism specific page and it'll have links to all the different projects associated with that organism. It'll have metadata information about the organism, about the living conditions, etc., etc.

Where all, and then under this top single organism project there could be genome projects, transcription projects, multi-isolate projects which would be like viral samples, proteome projects, targeted low side which would be the 16S ribosome RNA projects so you could have a single project with thousands of 16S ribosome RNA's from a single isolation. And then we have a project type "Other" because we know that they'll be more projects in the future.

We also have the idea of the top multiple projects which is an upper level administrative project which provides structure and linking of different related projects. An example of this would be Encodereach MP we have top multiple projects for that. And actually here is an example of the HMP project. We have the, in the green circle up on top is the initial HMP project.

It's an HMP roadmap project and that's the, it would be the top level project. And below that there are three different projects. There's the reference genomes projects, the 16S ribosome RNA projects and the metagenome projects and those are also considered top level projects.

And then below the top level projects were the projects that actually point to the data. You have which are different metagenome projects, you have which are genome type projects. We have the 16S projects which are all targeted, individual targeted projects. And then the reference genome projects. So each individual genome will have its own project ID.

There are also top single organism projects that will point to the organisms within the, within the reference genome projects. So we're, there are multiple levels of linkages. You will be able to navigate up and down from many different points within the project to get to the other related data in the project. So we think that this will be a lot more flexible and a lot more easy to use. And that's it.

Male1: [8:57] That's your point.

Ilene Mizrachi: [8:58] Yes?

Male2: [8:59] Is HMP Roadmap project also able to link in like to demonstration projects, challenge projects that are also agent related? Is that how you tie it in?

Ilene Mizrachi: [9:08] We could link, if we know that they're related we would link them together. The idea is that you would link from a lower level up rather than from that top level down. So the data at a lower level project would have to say, "Hey, I'm a member of HMP".

Male2: [9:26] OK.

Female1: [9:28] Can I ask a question?

Ilene Mizrachi: [9:30] Sure.

Female1: [9:31] You commented on a single organism is that a category of strain?

Ilene Mizrachi: [9:34] Tatiana do you want to?

Tatiana: [9:36] Well there was, actually there was a change, I was basically collaborating, we had decided we were no longer, we were going to investigate whether we should stop having strain specific taxonomy notes for bank micro build genomes. And I think we're going to be moving away from that. And so I think the organ, the top single organism will just be genus, species, right?

Female2: [10:04] Right, except that we decided for some diverse strains within those specific species where we might have more than one top level note for this, their usual case would be, the average case would be genus species but in some cases we might have pathogenic versus long pathogenic [coughing] so we might have more than one top level organism and they're very different.

Jeroen Raes on the Tara-Oceans project

Jeroen: [0:11] Raes: [inaudible] Just to give you a few slices of background as to what I've been doing.. I've been working in metagenomics in general, doing all sorts of [inaudible] metagenomics to link habitat environment to functional repertoire, whether it's in TARA, in soil environments or ocean or in human metagenomes. This is just hopping from one environment to another. The last four years I spent at the Borg [sp] Lab at [Inaudible 0:54] co-developing methods to look into how microbial communities adapt to the environment. [0:58] We saw that it happens in all ways you can imagine, at the functional and phylogenetic level but also genome composition or size and evolutionary rate. They are all related to the environment. All kinds of forms of adaptation can be found. The most problems we encountered were also linked to the many pitfalls that you can find in metagenomic data analysis. In this review we tried to put everything together and we saw that everything was linked to everything.

[1:31] The whole process of how you crunch your metagenome or even from the start up sampling influences whatever you see at the very end, garbage in, garbage out. The last years were actually a big learning experience on what to look at and what to keep in the back of your head when you're analyzing these things.

[2:03] In the beginning, it was all qualitative. What is the difference between soil and water? Very, very basic, but probably far too simple of ways of doing these things. We tried to move to quantitative approaches and really trying to take advantage of quantitative metadata which is available more and more.

[2:24] One project which we did, which we started in collaboration with the Gerstein Lab was how can we integrate quantitative metadata with metagenomic data. Now the follow up of that project, the methods were developed. In the paper you see down there and in the follow up I'm now trying to do this on a larger scale.

So one of the things I did was I took the data from the global ocean sampling and tried to intercorrelate this with data you can get from all the ocean databases. So what they all do, the data that have been collected by all the oceanographers and that meteorology and that they meteorological institutes all over the world. These data are usually monthly averages. So it's not the data on site as it was measured by [Inaudible 3: [2:50] 21] but the metadata in the Gulf sampling is fairly limited. So in order to have more info a bout the CO2 and the weather and the sunlight and things like that...

Interviewer: [3:34] Just a question. The reference was already published right?

Jeroen: [3:38] Yes. [Cross talk] The first one is more of a method paper, and this is still ongoing. Basically, I gathered information from all these sites where you have monthly averages. It gives you an average idea about what the environment conditions is that these organisms from these sampling size live in. If you try to do all these correlations, you can find things. [4:21] I'm supposed to talk about TARA so I can't go into it but we found correlations between, that simple correlations between the amount of photosynthesis genes and the oxygen concentration in the

water or more complex correlations where the amount of carbon fixation genes correlates to a model that combines oxygen and temperature.

Or we can do many to many correlations which you can do with a [Inaudible 4: [4:40] 48] correlation analysis, where you try to find combinations of environmental factors that lead to combinations of in this case functional groups.

One thing we saw for example with the cost data in combination with these monthly averages, we see that the strongest player in environmental adaptation is climate. Things like nutrients, like [Inaudible 5: [4:57] 13] to nitrate they have a smaller impact on the functional composition of the environment.

[5:20] We can discuss more about this if people are interested. Other things you can is do is move towards things like bio-geography. Many people in this office have been looking at the species richness and the species diversity, and see how this links with latitude.

[5:46] You can look at it on a functional level, look at how the diversity and functions available to a community vary over a distance and look into what influenced this. There's more to this, be we can discuss this separately.

[6:04] So the thing is can we now go more... We have monthly averages, so we have an idea on general trends. But what if we want to look at very specific dynamic adaptations, not on the scale of months but on the scale of hours or days.

[6:27] In this framework, the TARA oceans project is very, very exciting to me. The goal TARA oceans is to integrate oceanography data, so things like physiochemical data and things about currents and things like that with metagenomic, metatranscript data imaging data such as microscopy and things like that to study the diversity and the dynamics in the, of the microbial in the broadest sense of the world, in the world's oceans.

[7:07] And so this is a huge consortium. There's a lot of groups involved from, many of them from France which is initiated at the EMBL by Eric Karsenti. There's a lot of French groups involved. People like Colomban de Vargas, Chris Bauer the general scope, EBI is also involved for databasing and so I am collating the bioinformatics data analysis section and the people from Maltless from MIT involved and things like that.

[7:45] It's big, it's quite a big consortium and I almost have a feeling every time we have a meeting the consortium is five members larger. But anyway, this is probably for the good of the whole thing.

[7:59] And so this is the trajectory TARA will follow so TARA has just left actually in L'Orleans last weekend and now it's about here, it will soon be in Lisbon for its first stop. And the goal is just to go if not, not only to go like in the nice warm areas of this planet but also look at the more extreme parts. And to cross multiple oceanic or seismographic problems things like that. There's also large coral components. We were also passing a lot of coral reefs. People are very much interested in looking at these things, as well.

[8:45] One of the processes that the consortium is very much interested in is ocean acidification, so the link of the amount of carbon dioxide in the air and in the ocean and the effect it has on the microbial community, on the coral reefs and vice a versa, the to, try to [coughing] how communities cope with this increased amount of carbon.

[9:17] The goal is to sample not only, do not only metagenomics on bacteria but really try to sample the whole ecosystem. So they will be, we, not solely this, [coughing] for example, also therefore the analysis of viruses. There's also be people interested in these Giruses, giant viruses. So we'll do something on bacteria, Archaea and then this is a slide through the eukoryotic sampling that I got from this form of the virus where you will use these things to select different size fractions of the eukoryotes.

The plan is to go for both metagenomes as well as 16S type sequencing metatranscriptomes, the transcriptome is mainly focused on the eukoryotes and that sequencing will be done at the general scope. And one thing I find quite exciting is this High [Inaudible 10: [10:06] 43] Microscopy, so the people at the EBL [sp] well it might just have ;moved to Dublin now, they are experts at microscopy and High [Inaudible 10:53] electronic microscopy and so the goal is to get these samples and really visualize everything we see so we try to link now morphology to metagenomes through ocean conditions and things like that. So this is very interesting.

[11:04] And on top of that so we will have on site, real time oceanographic measurements because of the involvement of several oceanographic groups, Gaby Gorsky and Stephan Pesant are doing this thing. We will also have links reaching out to people who are doing satellite measurements and things like that as well.

[11:28] And then to try to make sense of all this, there's a huge bioinformatics effort that needs to be done. I'm just saying here that I'm just one of many people who are doing this because this is an enormous effort. And we are working now very hard to get everything standardized and that we can achieve it and link all of the different data sets together.

[11:55] So the first effort which was done by Uros and Maria and Stephan was to get the sampling database. So the on board measurements linking all the sampling details to bar codes, to the GPS and these things. This data from the ship will then be sent to a central database on land which will later also contain all the data, all the microscopy data as well as the metagenomic and the metatranscriptomic data.

[12:34] And so in the end we end up with a, I call a bioinformaticians heaven dataset because we will have so much data from so many different corners with which we can play. Information on who was there, what the functionalities are, the morphology of the people as well as the physical, chemical on site data. And I'm very excited to see what we can get out of that.

[13:03] And so I'd like to acknowledge a few people because I'm just standing here as a representative of a whole team. So this part is just about, these are people who are involved in the project I told you in the beginning but the main people I have to acknowledge are the people here from the TARA Oceans Consortium especially Eric Karsenti who is leading the whole team. Colomban for many of the slides I stole and I acknowledge also here. These are the main funders of the consortium.

[13:33] You see here Agnes B., that's a clothing brand, so she, they own the boat so they are just involved in, they want to give the boat to scientific projects that are of global importance or something like that. And this here, this is funding for my own lab in Brussels and thank you for your attention.

[applause]

[13:56]

Sansone on biosharing website

Susanna Sansone: Did someone put the page up, it's not yet in the right URL but it will go live. So this is... a concept that some of us have to try you know to push in the last year in Europe that as I say before representation in the States we are doing standards and things in one domain but there is so much outside there and these will all kind of enable technology. So we need to try to pull everything together. At this stage there is not even a single Website where you can go you know access the checklist or access the ontology, access tools they use in format and et cetera. And that's what we are trying somehow to do. The reason behind this what is a genomotist Website is because Don and Seth and several other people of which are listed behind here where you don't see, put together to, the past two years put together this paper which will be published in "Science" and other next month so whatever. It's bringing together all the funding agencies so two years took to bring the US, the EU and to try to, you know because all their data policy now talk about, putting data in by domain names and the standards, having an issues repository, but somehow they don't even have a clue what their standards are or how to guide the overview to those standards and it isn't standard at all then getting progress then getting mature and so don't realize that there is still a lot to learn from that side but we also need to learn we also have to give services so they can really point their way to divide time for you know standards.

So we started to discuss with them the different data policy, we looked at different data policies so the people will cover the data policy, what they contain, how different they are and then they understand where the data policies are. So this is, this bio sharing all started because we wanted to put this data policy in a Website and make them accessible and searchable, but then we talked, OK, while writing the paper we say it's OK so we, if the data policy standard according to MIENS requirements technology why don't we start collecting this in Website and I already described MIENS it is protocol where all the checklists are there, everything is described in foundry, they're collected in total and tried to integrate.

But as I say there is not yet a single Website where you just go and find everything and then you go into the different specific pages. And of course now there are tools coming out, essential tools and technology enabling using ontology or the formats or different things and there is of course all this journal now coming out trying to use this ontology, et cetera, infrastructure like the one we use but there are many others and there's also a society, I put it here. This is not a comprehensive list of the players but we're trying to put people on this map.

So here for example is the listing that we have that by Data Society we just said, from Pascal. Here also is another link which is the Pistoria Alliance, this is a very new initiative because it illustrates all pharma people who got together and this is all kind of pre-competitive, collaboration and it's about standards and ontology. If you just look at Pistoria Alliance Website and after this we'll go live and of course there is all the data sharing policy, and et cetera.

So what we are trying here now to do is to first of all create this web page where everything can be listed. And again, this is not a comprehensive list. Our intention is to click here and go to Website. As Susie was saying this is certainly not a comprehensive list of technology tools so maybe what we're going to do is just another page where you help us build things as a list of

different tools and technology where we actually enable using the standards. And that's where all the data policy are now will go and etc.

So Tom you want to help me to put this one of the picture here? And just see it's a player behind here because you are doing the minimal requirement, you are doing ontology, you are doing tools, you have a journal, put a stakes as a journal of course, so you are there. You are just one of those community which is actually producing this map, that's enriching this map and there are data policies now that are only now describing the use of genomics data so it will impact those on your research and the search you use of port.

Woman 2: So one of the things, as you said, this is just a paper because I actually run a data site during the editing that's really dry of mixed data policy and at some point I thought it would be nice if we had some kind of publication out of it and then the other research councils in the UK started to do the same thing so we figured we knew nothing about what the other research councils had done, we'd like to put it in context. Anyway it sort of evolved up into all this and we decided in the end just to launch a Website where it's actually just a blog right now but it would be a multi-author blog just to keep people up to date with standards. Ammon has given it a completely new spin by you know putting the graphics and everything on it, but the content under it, this could just stay as a Website that just points people in key directions or it could actually evolve up into a consortium that actually starts to bring people together across these boundaries. So it might just be nice for the record, Susan Lewis is here representing Herbal Foundry. We'd like to get her interviews on whether this is the next way forward or not. Pascal is here from Biocurator, absolutely key community to bring into this connection and perhaps he'll continue to play more of an umbrella role is something to discuss with or perhaps is already done inside Biocurator. Is that George here from Sigs? It's the first time he's seeing this so I pointed it out to him before we got in here but Eugene would like to have O mix up there, UK Palmed on the UK side of the British Library is getting very, very keen on how data flows through the entire lifecycle from experiment up to, the technologies people, John Wilbanks from Science Comments is on this.

So we're already starting to fill in bits of the circle. It might just be nice to hear from Susie and Pascal and George as about whether this might be a way forward to actually maybe having a few workshops in this area.

Woman 3: Go ahead?

Woman 1: Yes.

Woman 3: Good. [laughter] While she's here.

Woman 2: I mean to be fair we've been talking, we've been.

Woman 1: Yeah we have been talking about.

Woman 2: Initiating discussion between Herbal Foundry and Mibby for example where mapping of concepts to terms. There's lots of these integration points. It's happening anyway. It would just be nice instead of all of us having to cross talk between communities. We've got Pascal here but probably every one of these groups at some point would like Biocurator and

Herbal Foundry because they have more relevance across, I mean are you guys thinking about a journal? That's Herbal Foundry, right or no?

Woman 4: No.

Woman 2: No, you're not thinking of a journal.

Woman 4: There is no journal, it's the last thing the curators need is more things to do.

Woman 2: Because Biocurator is you know thinking about.

Man 1: It's your idea man.

Woman 2: Anyway it would be nice if there were one on.

Woman 1: And also the idea of an RSS feed. Because sometimes a lot of new workshops are going on and you really don't know what's happening because there's so much happening. I think with those communities in order to say everybody putting up information on the Website when they have it. So they must try even collating information which are relevant to all those large initiatives. And... that may be it. Go ahead.

Woman 4: Yeah, no obviously in some ways it's reminiscent or echoes two things that other shorthand synopses of a project, or consortium. And one is open source projects where it's a list of certain like what does it take to have a successful open source project. And this has some of the elements of that. And the other is just you know what does it take to communicate technically. You know you need syntax, you need registration services. Just from a technical standpoint the elements for recording or exchanging.

Woman 2: So you can share it.

Woman 1: Because it's about data exchange.

Woman 4: There's all of those elements as well. You need semantic standards and proof models and all of that so.

Man 1: I think there's one other thing that's also important. How does one cut through all the chatter to make what we're interested in pushing forward much more easily discoverable?

Woman 1: Any suggestions? About how to do this?

Man 1: I beg pardon?

Woman 1: Any suggestions about how to do it?

Woman 3: You can probably create sub-groups and sign up for the stuff that you're interested in, I don't know because it's a balance, right? You need to get the information across and then you need to get the coordinate information across. And then someone needs to decide what's more important than something else then.

Man 1: Yeah.

Woman 1: At the moment we are not trying to see something, again, maybe the view is biased because we are using this for you know a ground to be writing so it's biased on what we're doing. That's why I said this is really a draft. It's intended to do really general and then maybe individual pages has more and this is a sample we don't, yeah. For the moment it would be like information collection type of thing and depending on the interest of people if they want to participate with information then it can evolve in something more.

Man 2: I think this paper actually is going to get enough attention so people interested in those kind of issues can go there, come forward and just wander, that's basically, and then you find you put even small, little box behind it that may change the whole concept. So I wouldn't worry about how to get all everything, you know the participants.

Woman 2: No, no, no. I was just saying.

Man 2: Self selection can work, I'm sure, especially after this paper going to be read.

Woman 2: We'll find out in time.

Man 2: Yep.

Woman 2: At least we have some of the people in the room.

Man 2: Sure.

Woman 2: If people are OK to expect some emails, some of the simple things would be if your own Websites could start to have RSS feeds so that we don't have to do anything to make sure, to new stuff to get it collated and as we said, this is sort of a blog system underneath and it is multi-author. Anybody who's willing to put their name up and represent any of the slates on here, you know up to 100 people, I think people involved, names would go up. It's just supposed to be an integration site. So we'll start there and then we'll see, if everybody starts thinking about it what we might be able to do with it and it will go into a CVSU Grant so there's a chance that we would get some workshop money.

Sansone on ISA-GCDML workshop

Susanna Sansone: [6:27] What I want to get across now is the point why we have to consider the metagenomics, we just can't think about genomics.

[6:35] This is example I keep using, and this is just one of the several examples that I'm sure you do have. So this is just datasets from a metagenomics and metatranscriptomics study. And this is not the first one where we'll find where there is actually another technology that we are using for sequencing. Pyrosequencing technology per set is not just used for genome sequencing, but it also used for gene expression.

[7:03] So what I'm trying to say here is that it is important that any standard put out by the GSC does link into other standards developed by other grassroots initiatives.

[7:16] I don't know how may of you are familiar - has anybody heard of MGED, PSI, or MSI? Can you put your hands up? Yeah, I know you. I know you must have. So very little. Great.

[7:29] How many of you have actually transcriptomics or proteonomics or metagenomics data in combination with genomics? Well, not too bad. And I'm sure that you're not representing just what you love, but you represent a large community. OK. Right.

[7:46] So this is really important, because if you don't use now other technology, you might use in the future. Anyway, even if you don't, what you want to do is you want to integrate your data with other types of data. So it is important that ultimately after data is served, it is annotated in a standard.

[8:02] I placed GSC up there where the genes and the genomes are. But as I say, there are initiatives which had already been working since 1991. One is the NGA, special letters society. The other is Metabiomics Standards Initiative, which is attached to UPA, which is a larger delegation that some of you might be aware of.

[8:21] So there is a system biology world, in which you want to put everything together. What's important in a standard is to be planable. Now this is not the case. The standards are fragmented. So we have these minimal requirements, these formats, and these ontology terminologies which do not fit. There is duplication, there is overlap, and there is inconsistency across those.

[8:46] This is a problem. So what we need to do is we need to overcome these fragmentations. So we don't mess up any of the people in this room, Norman, Peter, Annette. We are supportive, we need to synergize with other communities.

[9:02] So we don't waste time ourselves in another portraying community, we started an initiative community. Mainly it's about pulling together all the groups that have done the MIENS information checklist. You know about MIGS, about MIMS, so you're not the only one.

[9:18] Just look at how many checklists there are on the MIBBI website. And these are checklists in different domains. Technology trees and biological trees. And this is a problem, as I said.

[9:27] So what we did, we created this portal where all the checklists are listed. It shows checklists and has a cart like MIGS, where you can fetch the version, who has done it, what

domain is covered. The idea is to enhance and maximize the synergies across those communities so there is no overlap, there is no duplication. There will always be overlap, but at least we try to minimize it.

[9:49] What it does is an analysis across those checklists to see if there really are degrees of overlap. And naturally it can tell you there is. You can see these are the 18 highly-ranked common concepts across the checklists. And everybody has got it, everybody who did the MIGS checklist for the standard design and so on and so forth. So there is overlap, but we have to work with it.

[10:11] The second effort, very briefly, which we need to be aware of is the OBO Foundry. Suzanna Lewis, who is here, is one of the coordinators, along with Richard Scheuermann, Alan Ruttenberg, and Chris Mungall and others.

[10:27] What the foundry is, in one slide, it's large initiative which is bringing together all the community, the ontology community which has been developed, and they are all listed on the OBO portal. So the Foundry is trying to bring them together to make this ontology interoperable, like the pieces of a puzzle, so you can use them accordingly.

[10:50] The idea is that they should be in one reference ontology in every domain. Like one ultimate ontology which settles specific extensions. One ontology for this kind of sediment, one ontology for this kind of environment, and so on and so forth.

[11:04] And this is great, because actually in the GSC community, it's really working under this portal. And in fact we have the Environmental Ontology, which Norman and many other people actually lead and contribute to, which is posted under the Foundry.

[11:18] So the Environmental Ontology, along with [inaudible 11:00] , are being developed really as pieces of this puzzle. There is also the Sequence Anthology, which is relevant to this community, and this OBI, which is kind of less relevant.

[11:32] But we need to make sure that we continue to feed the terms into this ontology and we don't just reinvent things which exist. We all need to use the same terms when we annotate the data.

[11:44] The third effort is this ISATAB format, which I'm not going to go into any detail. All I'm trying to tell you here is that we have to describe standards which have several types, be it multi-omics or more general. And there was nothing out there which we could use.

[12:01] So Dawn, myself, and Philippa, which is behind you, and many other people, got together with other community-wide pieces of this standard, and we tried to create something which was somehow a top level that could describe everything, could fit any type of study, any type of assay.

[12:15] And we created this ISATAB format. It's format is not XML, but we translated XML. I'll show you afterward.

[12:22] I'm moving now to my second and last part of this presentation. It's great to do standards, if we have other species there. But what we need, as Dawn was saying, standards are just a

means to an end. We need to enable people to have better annotation, consistent notation, so ultimately we can analyze data and we can compare the datasets.

[12:44] So how do we jump from the standards that we know to signs? We have to develop tools which implement the standard. Those tools then go to the user, which are the experimentalists, those who create the data. They use the tools so the data is better notated, it's consistent, and it's ultimately going to end up in the database and serve the community, where they are consistently annotated.

[13:06] So we have to move from the theory to the practice, and it's a lot of work to be done. So we - my team in particular, like Philippe behind and other people, and with cooperation from Dawn and contributions from Peter as well - we started developing these tools.

[13:22] But these are our set of tools, and many other tools can be developed. So I'm presenting what we did, how we put the standard in action. I'm trying to put here in this is the experiments, so these are the experiments that have been created. On the other side, it's how do we comply with the standards. In this case it's the ISATAB format, the OBO Foundry ontology, and the checklist on the MIBBI, including MIGS, MIMS, and MIENS as will be posted.

[13:49] So what we created is a set of tools, a set of tools which are stand-alone Java components. They can be for local installation. They can be used in isolation or in committee. You can use one or more components along with tools you already have, or you can use them together as a unified system.

[14:05] The first component that fulfills the needs of compliance is called ISA Creator Configurator. I'll explain afterwards exactly how you configure the minimal requirements onto your ontology and onto the portal.

[14:20] The second tool is called the ISA Creator. It's the editor you use to describe the experimental meta-data following the standard which has been set. The other component is a database for combined investigation index for local installation and the other components, which is very important, because it fulfills the needs to submit to the repository.

[14:41] Some of you, of course, have a system to export forward to SRA or ENA in SRA XML. But if you are not able to, how do you do it? So we provide our community with a converter which takes the ISATAB format and exports into other formats, including the SRA XML.

[14:59] Because we are multi-omic, we also support to the Express repository, which is a MIGS-type format. We also support to the Proteomics database which uses SRA XML. So as you see, everybody's a different thing, so we have to enable this compliance to the other formats, and that's how it will work.

[15:17] This tool is called the ISA Converter, which enables these conversions. And as I say, SRA XML is where it's ported, everybody has with the SRA archive and the ENA archive. The format is in common between the two.

[15:34] Now I only wanted to go into two details here, really about MIGS and about the ontology developed by GSC community. The ISA Creator is this component that allows you to put together to standards, put together the requirements of your ontology onto the format.

[15:52] And this tool is targeted to the power user, to the curator, because we need to shield the user, the data producer, from all this complexity of standards. They don't care about it. They just want to get the data submitted and get the accession number or whatever.

[16:07] But we do help them. So a power user or a curator is the person who is able to do the configuration for them. This configuration is then read into the second tool, which is the creator, editor, which is given to the user. So the user is kind of already following the standard, is complying to it.

[16:28] Let me show you just briefly a description of how do we put the MIGS term into the configuration. This is the checklist posted under MIBBI. This is the MIGS checklist, part of the MIGS checklist. So these are all the requirements.

[16:41] What the Creator does with power users is take these requirements - let's say that you need to make sure that you use this kind of latitude. So you take a latitude, what you do here is to created from the Configurator. You put the requirement into the elements, so of course requiring an understanding of where things go.

[17:03] What you do, the Configurator allows you to define what type of value the user then has to provide for that specific field. So you can set this value to be an ontology entry, to be free text, or to be an integer, it depends.

[17:18] So lets say that I want to force my user to provide habitat information and use habitat terms from an ontology. So what I do, I make the field mandatory. So you define the behavior of the field and make it mandatory. And then in this case, you actually say I want the term to come from an ontology. This ontology being, for example, the Environmental Ontology (EnvO).

[17:42] You can use one ontology, and in the next version we are about to release, you can also sub-select a term into an ontology. So EnvO will have a classical habitat with, I don't know, 10 terms. So you can actually restrict the selection to those 10 terms.

[17:57] In this way, you are driving the user into the right direction. And that's what we want to do.

[18:03] Then everything is uploaded into the ISA Creator. This is just a screenshot from the ISA Creator that the users see. It looks like a spreadsheet and it has a lot of spreadsheet functionality. In the field there, that's actually the field the user will see. It's habitat. So because it's accepted as an ontology, what the user gets here is just a widget where they are forced to choose an ontology. In this case, the ontology comes from EnvO.

[18:29] So he would choose the term from the ontology. The tool will actually store the information from the ontology term, the ID, the source. And this is very important, because the ontologies are dynamic. Things will evolve. So you have to keep track of what term is being used, what is the ID, what is the source, and et cetera.

[18:46] All ontologies are accessed live. For the moment they are just accessed live -- there is no local storage of the ontology -- through the ontology look-up service, which is MIBBI in the portal which is the NCBO. There's just the topics for ontology.

[19:00] And there is very nice feature for ISA Creator. And if you want to know more, we can talk and sure we can explain to you more. This is just a example of an instance of the database component with MIBBI. If you're interested, we can load, for example, a junk dataset to show how things can be done nicely and annotated through these tools. You just Google [inaudible 19:27] and you actually do get it.

[19:25] And with this I just want to thank - these are the links to all the MIBBI checklist, the links to the OBO Foundry, the links for the ISATAB format. These are the community that found us and worked with us and they're all becoming ISA Hub. And these are not individual groups, these are service providers. So they have a community behind them; it's a very large community we work with.

[19:46] These are the my team members, Philippe in particular. Thank you very much, particularly to Dawn and to Peter. Thanks.

Sterk on MIGS-compliant data curation

Peter Sterk: [4:42] So I'm just going to talk a little about what we've been doing for the GOC. Basically we've been working quite hard on getting the standards together, I'm going to talk about the standards, you'll know that. Dawn already mentioned we've got an exchange language, GCDML, XML schema. And after spending years on getting these standards agreed upon and part of made by technology, I said to her, to Dawn, "It's time that we got some data so that we actually get more credible and it's not talk about six but that we start doing six." [5:36] So only this year, actually June the last, GOC meeting in the UK, Renthorough, took us through the GCML and showed us how we could use an XML editor to edit data and I looked at it and thought, "This is actually doable. It's not the most user friendly thing but if we just sit down and go straight on the job we can actually get all the genomes done." So since January we have been employed by the GSC and my main focus as of this start is the duration job.

[6:28] And my project was to protect in GCDML so we have a lot of reports now and sent to Nykos who has done quite a lot of duration goal is going to be a quick way to get his data price program to convert as much as possible into changing the amount and then in one afternoon I have caught 800 basic genomes per reports.

[7:06] But a lot of data is still missing. And then the task was to settlement so the report with more data. And then going back to the literature and read a lot of articles and try to find as much missing data as possible.

[7:31] Luckily I'm the final institute Cambridge and U.K. Has given me a desk and I'm sitting among the curators and they're going to help me with that task. And also in the future I try to get the data and to start to write a report right at the start of projects, which I think is very important to actual sitting report is just too cumbersome and you need to capture data right from the start.

[8:20] So my hands off to Seemback data into goals. I think last few months goals has gone through a lot of major updates, a lot of additional fuels have been added. I've just broken some of my scripts but now it's Nick Goldridge and author of the reports that I'm creating much richer. So this gives us gold data plus what to add as much data as possible. And as I always said, it takes a lot of time and it will often look very easy to decide what kind of data you put in the different fields.

We try and use as much as possible terms from ontologies and control vocabularies. The problem there is that not all the terms evolved to use have been defined. But I've spent time on the virtue of this here as to define as many of the terms as possible for [inaudible 9: [9:17] 52] Wiki at the appropriate level. There's lots more terms that add definitions, lots of preferred in ontology where they should go.

[9:55] So we have to work with the communities to network ontologies to actually get data into the ontologies and that's ongoing. But it starts and as I said, that's ongoing.

[10:15] So another thing is you can have the reports but you also need to have them visible. So as we are in agreement in RCN4GSC and GenBank and Vol BMDJ to start incorporating mixed data into the genome records. It's not entirely clear how it's going to be presented in the records but I've already, a lot of work is done with Guy Cofrench from Amble who has managed to take

my data and put it in X amount in the records. If you know where to find the data it's there but they're still working on the presentation.

[11:06] But it's a start and it will start in the next year. Certainly increase.

[11:17] I'm not going to say too much about adequate, because Crisanna will talk to it and it will be on the agenda today and the cart was, but we had a meeting in past two days to discuss how we could use Isitack format to put our rich data in. Created the tool Isaac creator which is much more user friendly than XML editor. And I've worked with the team to it. It's obviously very how suitable the tool is.

There are further advantages as in fast to use system because essentially I'm not going to say too much about [inaudible 12: [12:09] 25] . But some like to do say, so it works. You can get the data in that system. The importance is that the more we can use it to test how we can compare data.

[12:49] So there's a lot more to do. And I hope that from our experiences that a lot more people will start capturing the kind of data and we will work with the community to different ways to curate as many genomes as possible.

[applause]

[13:20]

Transcription by CastingWords

Tatiana Tatusova on Annotations in Refseq

Tatiana Tatusova: [0:01] So the Refseq project got started about six years ago and there was a need for some standards and some clan up on their development, on side data. So there was a need for some clean up and some standards, because, has been discussed her many times. The data cabin need to grow, have different quality, a different temptation,which is, and there is a need for the community to have some standards. Although I'm not saying that perhaps it is the perfect resource for any standards, but this was started in an attempt to have something. [0:55] So, the Refseq project is a genome oriented resource there. We have tried a represented sequence for each molecule from someone need to product. It is derived from camping grown or actually embryo DDTJ, which is to say, in NSDC. It is non-redundant. And non-redundant means that we try and have a single genome representative for the species.

[1:26] Although if there is an exception in a chromo gene where we have representing for this training. Refseq can regulate any resource, we have sometimes we have created internally. And sometimes we have a represented source by genome and we'll talk a little bit later about this. And we have little distribution through our web resources on tranplast TP and special TP and this is the agent for gene validation analysis and primary foundation for Regime.

So when I talk about redundancy I'd like to take this opportunity in this room to make an announcement and probably get some comments on your feedback on that. As I mentioned we have for other organisms we have single representatives for a species. And in my Crow bill and kill bill, we have, so far been creating, Refseq genome records for every strain.

Man 1: [2:01] I'd like to comment [inaudible 2:33] .

Tatiana: So the completely you want.

Man 1: [inaudible 2:38 same species]

Tatiana: [2:50] Right. But it depends on the species right? So we put this announcement on our website in genome project. And what it says is; we are planning to change the procurate genomes resources due to the large number of genomes being deposited. And there is a huge population studies in sequencing thousands ecoli genomes. Sequencing a lot of basals genomes. Basals, transacts that are in population. [3:23] So there is a comparative genome studies. And there are many human pathogens they have submitting many isolate and strains and sub strains for the same species. So at some point in the future, first of all there has been mention on this meeting by yesterday, that taxonomy, that we designed, start designing, we will start design a strain we have translating for every gene procurate genome to where genome sequence is available. And the Refseq project will not longer have records mentioning every genic prokriating genome. But complete but it's likely that a subset of representatives from a given species will be used as secret.

That's where impact you'll notice on the databases and see the originals are on tools, and write to TPR archives. So our approach is, we're just announcing that and we'd like to hear feedback from the community and this community that we are planning to have this single rule. That there is

one representative for the species; I will try to take this diversity into account based on birth on sequence similarity. And some importance for the research communities, so there might be some clinical isolate that is, or some experimental model organism that is there. So there could be more than one representative.

Man 1: [4:12] What is your main design process or you have any specific design is there actually [indecipherable 5:08] inception.

Tatiana: [5:12] It's both. As, some way it's saving space and time and for computing all the downstream analysis like producing brass databases, and prosting, but it also know that we found that if you have found an ecoli gene that is very similar there is also genes that...

Man 1: Do you have some early?

Woman 1: Yeah some early?

Tatiana: [5:50] For the genome level, what we thinking that we will do, several representatives of the genomic sequence that... we did come up with Costregen Coral that is 99% identical. Then we will have a pool of genes that will represent the full set of genes for this particular species. What's been called span and recover a span genome, so we will have four genes. Then eight genes foreign sample represent a complete foreign complimental species for genetic sequence will have several representatives. [6:32] We will modify it to make instantly available to obtain gene requires we just want to reduce the seriously we have and the gene level will or may the progene level. There's many progenes that are 100% identical genes but also we have already have a special versional bloods. If you do bloods now, I don't have an example but open, you will see hundreds and hundreds of ecoli heads and then rest totus.

Man 2: [7:09] So that can you someway open it up to the community, so that people working on it can get a sanction of the economy and the fact that you've made a decision on what you think the reference strain should be. It's already become increasing important to everyone. Could you say, this is an open concept?

Tatiana: [7:30] As I've said, we are announcing this. It's more like a proposal. And it might be that we seek the advice from working with a special texaniste pro so that they advice us what they want representative to be. But if there is a strong opposition to the proposal we might consider continuing doing that. So we think about this, so there is again this table that the Genbank, the source of the sequence is submitted, there are sequence, public sequence and the different terminals so perhaps you could write and edit and modify this. [8:18] There is no redundancy and we want to review the redundancy and that is also, so this is how the data flow and that is how the data is getting to the right seat and there are different resources. There are created data, there are, so there are different sources, there's one additional change that we in Refseq we have this authoritative resource that there is a model database in the way we Refseq so we are going to change this and because we will require that every genome will be in, have a version in INSCD database.

[8:56] So this slide is describing how the data get into the Refseq. You can contact us we have, make it by email but we, when you submit the data to Refseq or we choose to pick up the represented for Refseq, we have some validation annotation checks that I actually, it can be done

by submitter centers. And it's similar to this post processing that Nikos was talking about. So we have two tools, one is called, "Discrepancy Report", and another is "Praneshift" tool. And then we have just for a naming check.

[9:39] So that what is improved knowledge in genome creation? We check for the consistency of the genome features, we check for uniqueness of the local stacks, completeness of the local stack, we check for the protein coding regions, we have some discussion internal about who the genes and how to deal with that power and to identify them, annotator spends enough time on that, we check for structural RNA, RNA's for non coding regions, so discrepancy report is available publicly on our TP's or our HTP website.

[10:20] And we check for protein ID, local stack, run format and inconsistency. We check coding regions with the same gene locals as another coding region but a different product name, so we would do that. And we'll also check for coding regions that overlap, RNA teaching that Natalia and Nikos are talking about this as well, so we are basically looking for the discrepancy and give the report back to the submitters.

[10:53] Where at Refseq we might correct it to ourselves. So fracture submission check tool is more complicated because it's a whole run blast not a service so the service so the software is downloadable and you get tool that is called, "Selfcheck". It is routinely used on complete genomes but not on the double, on the draft assemblers. So this is, our goal is to identify potential fractures but also to check for RNACDS overlaps and CDCS to see as overlap. And RNA overlaps missing TRNA's and missing structural RNA's and look for truncated problems.

[11:32] So this is an illustration of how we identify potential fractures. We do find blast hits of adjacent genes that are compared and the common subjects are collected and topics analyzed for potential fracture and this is how they are, that it's illustrated and then we report it there. But the, some are one on one links so we expand the links so we could research it less than one time. And then no fractures is reported.

[12:04] This is an example of the actual fracture detection so that there is a substantial CDCS of blast hits to the same key. And this how this report is provided through this menu to the users. Suited genes, I don't want to spend more time on discussing that we already discussed this so. We prefer them if we have this converters to mistreat, but if there will be a new feature we will be happy to use it.

[12:38] Stem another check or validation that we applied is suspect for names. And again this is available on our website. So we look for things like plurals in names and like preferred, like protein names that are homolog, names that are similar to, misspellings, domain contains probably the names were there for the fragments that contain C terminal or M terminal, and the names contained truncated, is known as.

[13:14] We do that. So this is the tools that we share with Genmake and actually they're the validation checks and tools that I've applied before the data goes to archived database. To say that not necessarily all submitters agree to our to correct their discrepancies even though they know that they exist.

[13:39] So in rest in addition to that we have additional curation. We use protein clusters, we use external experts, we use the data from the expert and hope to define in the publication that the called "Genref" and we have used vertical annotation that is annotated by protein impulses and protein family and we correct all the family, or all the protein from this family in all genomes, not in a single genome.

[14:08] And in our attempt to go to the consensus ES and we started this work with private firm, in GSCIM where it made a lot of progress comparing this and this is an example of, I don't know if you can read this but with them we compared NCBI product clusters with Tigerfam and GSCIM contributors.

[14:33] We have a tool actually to do this comparison so we, that is the expression that resulted in our analysis, we found some naming that are equal and maybe it is about twenty five hundred and then we have some not equal but have same function but different names and there's also a set of different names what they're still working on resolving it.

[15:03] And in some cases we, NCBI changed the name or in some cases GSCBI group changed the name 'til we came to the same consensus. And that is I think a starting point that we want to involve other goals and compare this maybe have at the same level so we'll end up with consistent name on what are called the extension of this GCGS project.

[15:25] That is an example of our protein cluster and what you see in yellow is the same members of the cluster in many genomes that we applied and the naming in this, when we have a created a name for a cluster we promptly give it to all the members and this is the example that within the cluster you have two new genomes that are still hypothetical protein name even though that is the old concern on the memo of the same clusters so we'll be changing that.

[15:55] So this is my items for the discussion that what are the group that could participate in this consensus GCGS project and make the commitment to agree on there, to work on the resolution of the differences in naming and make sure the databases accept the name and we have the same name in all databases.

[16:22] So then we have, also working with the protein naming standards and that is part of our effort of comparison with unified naming and we want to accommodate naming guides and their comparison together. And also we had already put the slide yesterday but we already talked about this and actually formed a working group on the evidence standards that is part of this group that we just meet.

Various Speakers Flash Updates

Man 1: [0:00] I made some slides for this but we have no time for that.

Woman 1: [0:04] All right. Everybody who wants to do slides put up your slides.

Man 1: [0:07] No, I'll just [indecipherable 00:08] them.

Woman 1: [0:08] It's a little bit long but it's OK if [indecipherable name 00:11] puts up a slide if he can put up a slide.

Man 1: [0:12] He put up a slide.

Woman 1: [0:14] It was supposed to be verbal, but go, go, go.

Man 2: [0:16] It was not slides, [indecipherable 00:17] this.

Man 1: [0:19] So OK. I'm without slides. So the [indecipherable 00:23] database that I was supposed to present is currently being rewritten. Just to remind you for those don't know what it's good for it was meant as a database, as a portal for marine ecological genomics. What they try to doing there is to integrate sequence data with genomic information from sequencing genomes, [indecipherable 00: [0:28] 44 to 00:45] and the metadata and diversity data, all to get Integrated, you want this therefore standardization was or is still very important in this respect. Because only with this standardized data we can use integration.

[1:00] We choose our genome referencing as our metadata for integration and if you like to know more about this just look at www.maxoutlet. Don't do it today, do it next week on hopefully Wednesday that is when the new version will be online. And then you can just play around and see what is going on in there I guess, I will stop.

Woman 1: [1:26] And you're based on GCDML [name spelling 01:29] now, whereas ...

Man 1: [1:29] Yeah, sure it's all the backbone there is GCDML [name spelling 01:44] as expected that's as good. Go to the next?

Woman 1: [1:38] OK. No we can't [indecipherable 01:44]

Man 1: [1:45] You can't.

Woman 1: [1:47] Just a minute.

Man 2: [1:53] OK. I just wanted to show one thing. So IMG and IMGM are reading set from the old database and what I want to show you is how does this work. How difficult it is to actually handle the metadata. So both IMG and IMGM [indecipherable 02: [2:15] 18 to 02:21] collected through submission site. Everybody who submits their own data says to IMG or IMGM have to enter metadata. And it goes a little framing back and forth so convincing people to do that.

[2:34] What I want to show you is that in the end, there's a little that needs to be organized as for meta genome data says this is really a struggle. Everything you see here is work done manually by Nikos and several of his colleagues in the Genome Biology Program Group. And each time we get a dataset they sit down and figure out how to organize this data.

So, you see that this is a hierarchy. That is describing the last whole paper that was published in 2007 and it is still a work in progress. It is a lot of work to actually do the editing of each one of the data sets and figure out what is a reasonable name. Also, to make it clear as well as to fit it into something that is close to an ontology, it's not quite [indecipherable 03: [2:56] 27] . [cough 03:28]

[3:26] So, this is most of the work. This is essential to actually make, allow users to browse intelligently through a very large dataset over hundreds of datasets.

And in the end, what we have is the ability to search on all the metadata fields, as many as they are and you know, do meta database type of analysis. But again, I want to end by saying that most of the work is actually collecting this data. Scientists do not voluntarily provide this information. It is the work of people like Nikos going and figuring out how to get the data. Prompting people to get this data. In a very, very time consuming and a [indecipherable 04: [3:38] 14] cycle to get the information.

Woman 2: [Indecipherable 04:22]] [4:22] Is the data available from the publication or is this actual communication with the author?

Man 2: [4:28] So, all the datasets that we are getting, so these are things that we collect that are not published yet, so, it is through cycles of interaction with people [indecipherable 04:38] provides the data, emails, you know and these are highly motivated people because they want their datasets to be in our system so they can analyze them so they do respond. In other cases I don't know how Nikos does that every weekend he sits and browses through papers and collects this information. [Pause 5: [4:49] 00 to 5:07]

Actually, before I say anything about [indecipherable 05: [5:06] 12] , I 'm not sure our talks from this morning saw the beginning of [indecipherable 05: 05 to 05:21] paper, another whole extraction. So this is one issue, one of the papers that was published in the first issue. And if I just browse through it, there is an introduction.

Actually I would say more than 80% of the paper, articles, maybe we will get by [indecipherable 005: [5:33] 39] and his group. This has been instrumental. They provide all the information, all the biology part of the work. So it is quite an extensive introduction. And this is extremely important because for most of this organisms, there is no papers or literature that have been very, they have been described very minimally if any at all.

So if we move down to classifications in the future, there is, for genetic trait it shows the position of the organism. And you have a picture, [indecipherable 06: [6:03] 21] , and then we'll go into the mixed compliant fields.

So this is table one with their information about the organisms. And you will see on the left here the mix of ideas, which are coming from, which are the mixed fields that were published in

[indecipherable 06: [6:25] 38] mode. So not all of the fields have an equivalent of mixed ideas. This is probably going, [indecipherable 06:41 to 06:42] development and of course I would want to expand the number of links ideas or fields that are captured by mix.

[6:51] Which actually is not, is coming, we're going away from the minimal information now to more expanded information. But that's this is a better plan.

So this is table one and then we'll go, which is general information about the organism. And then we'll go into table two which has information about the genome project, [indecipherable 07: [6:59] 09] quality and this [indecipherable 07:11 to 07:12] standards, presented before. Then library is used, secrecy platform, [indecipherable 07:18 to 07:19] assemblers, decoding method, [indecipherable 07:23] and other ideas.

[7:25] It is, the point I want to make here is that in most of those cases this data is not extracted by code. But in most of those cases this data does not exist and you can't go and collect them in order to publish the paper.

[7:38] So eventually what's happening, originally the goal was to have a place, this was the sixth journal, to pool all this information in a single place but now actually it's becoming, its moving to a stage that it's actually trickier for us to go and collect the data in order to publish it.

And I think it would be great to see the whole community start doing the same thing. And start posting the papers into six. Which will insure the standard, the standards, in a [indecipherable 08: [7:58] 12] compliant in that same manner.

[8:15] So that's about it for the, for six.

[Background noises 08:20 to 08:51]

[8:51] So the new release of code that came out actually yesterday, especially UI improvement. So we kept the Google maps and this our cluster. You can use our cluster and then you will have individual, all the individual what is getting quite massive if you just use unclustered.

[9:16] You can click one of those and hopefully will get, we're in the process of collecting the images for as many as we can. Or you can just browse from here. And define the organism you're looking for here.

Woman 1: [9:33] Nikos,

Man 2: [9:34] Yes?

Woman 1: [9:35] That's the geospatial information regarding the isolate that was sequenced where the genome came, where the organism came from that was sequenced. Is that correct?

Man 2: [9:44] Yes. The isolation, yeah, the position of the isolation of the organism.

Woman 1: [9:47] OK.

Man 2: [9:48] Now you will notice here it says 659. So this is about out of more than 4000 [indecipherable 09:58] projects. Actually there are some on [indecipherable 09:55] as well, up

here on this information. But the majority, I would say 99.9% there is no accurate description. [10:05] So they were describing the setting or and that's what you start adding. So there is very minimal information. So the hope is that we will start getting much more exact coordinates on the location.

So this is one. The other development is the classification for all the [indecipherable 10: [10:19] 30 to 10:32] bacteria and you can, where you break them down to phylum, class, order, etc. so you can see at any point, for example here you have 4100 bacterial organisms and 700 general order, and three hundred species.

So the next goal which actually will be within the next month or so, Dennis [name spelling 10: [10:45] 49] try to get all the information from [indecipherable 10:52] so then we will have how many. We have seven phyla out of how many, OK, or 29 let's say out of 40, with isolates, for which we have projects.

So we thought, we are going to identify the remaining targets. So we have 752 general, OK, but how many general, we have been struggling with, for how many general we have an isolate at this point. So this is something we have been doing for the [indecipherable 11: [11:01] 17] project. It was quite painful going through the thing. This will give a much easier whole list of targets that we should focus on.

I don't know why we cannot see the [indecipherable 1: [11:26] 28] here.

Man 3: [11:31] Question?

Man 2: [11:31] Yes.

Man 3: [11:33] Why are the [indecipherable 11:34 to 11:35] ?

Man 2: [Indecipherable 11:41] [11:40]

Man 3: [11:42] What's that?

Man 1: [11:44] I don't answer. [laughter] [Indecipherable 11:48 to 11:50] At that point, in this case the difference also with the project with [indecipherable 11:56] CPI, they were going aggressively, let's say in collecting all the information that's out there. And this is different. Basically the projects to the [indecipherable 12: [12:01] 02] as I understand, have to be submitted. So it's [indecipherable 12:06] go out and collect all the information. So if somebody, somebody posts it to CPI, it is released.

OK. Now to saying what we are doing here with the [indecipherable 12: [12:14] 16] . The difference also here is that from the moment a post appears from JZI [name spelling 12:21] or from other places, we're trying to put it into the code, so we'll discourage, in a way, other people to submit with the same wording. That's one of the main goals. With viruses [indecipherable 12:34 to 12:36] .

Man 4: [12:40] We've done a quick survey for viruses [indecipherable 12:43 to 12:50] . But these are probably not the viruses that haven't been published in the journal and it makes it very hard to make some kind of [indecipherable 12:55] .

Woman 3: [Indecipherable 13:02 to 13: 08] [laughter] [12:56]

Man 4: [13:07] There are several problems getting all the various information [indecipherable 13:11] on influenza.

Woman 3: [Indecipherable 13:12 to 13:13] [13:10]

Man 4: [13:12] The reference, the reference ...

Woman 3: [Indecipherable 13:14 to 13:16] meeting with Isaac to get the description on that. The problem is, it's not only that they'll be eyesores but that some of them will be, [Inaudible 13:28 to 13:33] and that's it's not only [inaudible 13:36 to 13:44] [13:14] .

Man 2: [13:44] So, the best kind of report there, is to just put it in the database, but why. We cannot add any additional information so there's no reason for us to do it.

Woman 3: [13:53] Google will do this operation for us. You can point to site.

Man 2: [13:57] Yeah, exactly. Just have a link that goes,

Woman 3: [13:59] A link to their site. Don't duplicate anything.

Man 2: [14:02] Exactly.

Woman 1: [14:03] Full [indecipherable 14:07] . We have three more to go through. You promised [indecipherable 14:06 to 14:09] that [indecipherable 14:11] was really code compliant. Didn't you?

Man 1: [14:12] Yeah.

Woman 1: [14:13] OK.

Man 1: [14:14] So I have a [indecipherable 14:20] prepared memo because I thought this was going to be without slides. So, if this breaks then it's all my fault.

Woman 1: [14:23] So we're eating in to your M5 discussion time.

Man 1: [14:25] Oh Good. Just very quickly, there is more than 4000 meta genomes now on, [indecipherable 14:33] . So metadata is missing and we're really hurting, so we've noticed that. It's working beautifully. [14:39] So we've built in a bunch of tools. We're curating, among other things, locations. There's, I'll just give you that slide, but you've seen the data, right? You've probably guessed amid whose voyage that is.

We have all that metadata integrated now, we have a metadata editor, in the future, and I have sort of revised my position on that. If you want to share your metagenomes with a friend, in [indecipherable 15: [14:53] 02] or share it with the public we will force you to enter all the appropriate metadata.

And, after learning what I've learned today, we'll probably not make it [indecipherable 15: [15:07] 13] we'll make means or something like that. So we need the standards to be finalized so we can release [indecipherable 15:16] . That's another message.

[15:18] With all these standards discussions, for us building tools and having, I mean this is like two and a half thousand users. We need the standards to be there and stable, otherwise I cannot possibly force users to fill out these things. If we change them too rapidly they'll tell us next time around, "No, I just won't do it."

[15:36] All right, then just as a reminder, we have to get that right. There's metadata editor, there's all sorts of things. I like this tool, it's pretty cool. It's built by a student from San Diego. So if anybody has other nice tools, bring them to us.

[15:50] And by the way, collecting metadata for existing data that's in there is a big pain in the neck. If anybody has a good set of metadata, and I know some people who have, we should get into a discussion and maybe codify and come up with an exchangeable format so that I can export from here or from other sources into this sharable format so that we have the metadata and we don't have to go around in circles and everybody has to do the same thing.

Woman 1: [16:17] Can you just define that you are GCDML [name spelling 16:20] and that ...

Man 1: [16:21] Oh, right, right, so the one thing I don't worry about you changing. What we do is part of the GCDML file we came up with our own sort of [indecipherable 16:30] , I guess is the word you guys are using there. Tab-like form, the tab form that we have inside the application. But the actual interface is generated by having you guys in schemo [name spelling 16:42] or whatever the current thing is automatically parse and turn it into a user interface. So if you come up with a new standard we can have an interface that our data curation, in five minutes. I just hate to do that because it's going to frustrate the users. [16:54] So what we're going to do once you log into this new system and the new metadata is available and you have shared metadata or you have the datasets in there, we will ask you to fill out all the metadata forms and I'm looking at Jack you know probably how painful that all is.

Man 4: [17:07] How are we doing that?

Man 1: [17:08] Yes, you are actually our one showcase user.

Man 4: [17:11] I'm nervous.

Man 1: [17:13] And what we've learned from Jack is that we need the ability for you guys to upload, the users, the users to upload files that have metadata in a variety of forms. So we're actually trying to be very helpful. So that somebody who owns 100 metagenomes, there's few people like that. But many people have five, fifteen or twenty. They can upload the stuff in some more efficient format. So, again ...

Woman 1: [17:36] Good. I think what we'll do is to get a bit more profile to all these databases and activities, because hopefully the next two presentations are verbal, we'll make sure there's a Wiki page that starts to discuss these and has links out there, especially early about this. So that

the next few talks about [indecipherable 17: [17:49] 54] , so means, so Linda if you just want to give the update and then Nicole, about the IDP.

[Background noises 18:04 to 18:13]

Woman 5: [18:11] So these flashcards worked really well at the end of your meeting in [indecipherable 18:16] at the poster sessions, as opposed to now, but it's important. OK. This is just the, this is the front page for ICom [name spelling 18: [18:45] 48] and what I want to talk about briefly is just some of the progress that Microbis has made. Microbis is a microbial, ocean-observing information system equivalent. Opus is the database that serves the census of marine life, and Microbis is our microbial equivalent of it.

[19:08] So in addition to geospatial data we are also keeping track of other kinds of metadata associated with the sorts of data that we're collecting.

[19:16] For those of you aren't familiar with Microbis, Microbis actually collects different kinds of data. We collect legacy data and those are all shown, sorry, those are shown in blue. In red is the 4-5-4 data that we've been collecting for a Keck project that was funded to sequence projects from community proposals that were submitted a few years ago. And we also have the original Keck project's in green. And in yellow we actually have lipid data. So we have different kinds of data that we are collecting as part of the international census of marine microbes.

[19:54] In addition, each of our projects that are involved with the 4-5-4 community sequencing has metadata associated with it in the form of project pages. I'm not going to step you through those but if you go to I.E.M. you can find all the projects that we've been working with. One of our challenges right now is taking the metadata and actually incorporating it into the S.R.A.s, which are now very timely because people are starting to publish a lot of the data that has been generated as part of this effort.

[20:24] And just to give you an update, for example, we have about 12 publications in progress from various journals. What we're hoping to do is to have people take the data that is in Microbis and incorporate it into S.R.A.s, but we realize that we need to facilitate that. And right now, we're working on scripts to export Microbis metadata to XML text files so that they will be readily uploaded into NCBI. We're working with Chris O'Sullivan for the proper formatting of these XML files and that's all in the works right now. So that's about it.

Woman 1: [21:05] Jim, do you want to say something about R.D.P. [name spelling 21:09] You can come up if you want to.

Jim: [21:09] I can just do it from back here. I think I'm loud enough. So, the R.D.P.,I guess I'm supposed to say a minute on R.D.P. and Meems [name spelling 21:14] . One thing that we've done is, if you do have the multiplex and [indecipherable 21:26] type data, and I know a lot of people are having trouble getting it into S.R.A. or E.R.A. We have an online tool that helps you format the data into the right format so that you gather it into the submission that you then send to S.R.A. or E.R.A. Right now we have to have pre-loaders with the M.I.M.s attributes but, of course, we'll change that to Meems when that gets finalized.

Woman 1: [21:48] Good.

Vaughan on INSDC submission and Kaye on Sequencing Pipelines

Bob Vaughan: [0:02] I'm actually going to try and do a live demonstration of our new submission system. The first thing to know about submission is that it's actually really very difficult to submit data to us. We're aware of this.

Audience: [laughter] [0:16]

Bob: [0:17] The reason why it's difficult to submit to us is the point you see here. It's a list of our features, and some of the features may be qualified. I'll just quickly scroll through. You can see there's an awful lot of them, and there are relationships between them. Keeping track of all these and knowing which ones are appropriate for your data is enormously challenging for anyone. So building a submission system which makes it easy to work out how to figure data has been one of our key challenges.

[0:50] With a view to that, we have over the past couple of years been building a new submission system. The first thing to note here is we actually have a common login system, so it's not that in submitting data you have to, at each submission time you have to put in your name, put in your address, put in your contact details, put in the rest of it. We've now created a login system which is shared with our project submission system and also with our SPIN, which is Unicross submission system. So I suggest, I'll try and log in. Here we go.

Audience: [laughter] [1:34]

Bob: [1:57] Then I'll bring it back up. This is not a function of the submission system, this is an issue with I connecting to the API to show you this, as this isn't yet online yet. We're actually looking at an internal site here. In this case, we're going to come back to the MIENS standard and look at how you would go submitting this as DBR. If you go to create a new submission, this is very much a new concept. Rather than present you with a very, very generic submission field where you could put in any kind of data that you can think of.

[2:33] What we intend to do is provide prepared templates for specific types of submission. These will be more or less granular. It might be something like a single coding region. It could be a very, very specific template, as this one is, which follows a particular standard. In the future, we hope all of these will have MIGS, MIMS et cetera available. Obviously this list is going to get much longer over time. If we drop into the next bit, the first thing is for the collector.

[3:15] How many entries are you going to submit to this? Rather than our previous template system where we expected at least 25 entries, now we're saying whether you're submitting one entry or 10, 000 entries, you're using the same system. Moving through the system, we now invite people to add citation data. This looks a tad untidy for now, but you can see the rough concept of how it works. You put in your title, author, jump from box to box. Here we go.

[4:21] Now with the citation, the nice thing about the INSDC's FlatFile is quite easy to assemble them in pieces. If you imagine there's a top section of the FlatFile which is purely heading information, so it's articles with are linked to it, it's names and addresses of submitters, et cetera.

Then you've got the source information at the top, and then finally features and sequence. So we can easily assemble this. Moving on to the next section, this were people define in advance what fields should be used, or can be used, within a given submission.

[5:03] So in this case, at the moment, all of the general MIENS fields are marked as compulsory here. You can't turn them off, you're obliged to supply some information. In fact, in the future we'll be taking this out because some of it will be potentially off. You can see here we've got little mouse operations that bring up the descriptions. You've got descriptions that are provided from the MIENS checklist. We scroll on down, for each of the individual environments, we can pop up a box.

[5:43] You can see the rest of the fields that are required. You can see, yes, we'll have altitudes. We will be able to provide barometric pressure, et cetera. Again, these can be set to be mandatory or optional as you like. You can just pick up a handful there. Now if we jump onto the next bit, this is where it starts to get much more powerful. On this page, you can specify the fields which will be common across all of the entries which you've submitted.

[6:17] So effectively if they're all from the same organism, for example uncultured bacterium, which is a favorite, you can drop it in there, and that's it across all your entries. You don't have to fill that in every single time. You're only doing that once. Anything which is set to be a variable field, and for the moment I'll simply leave them all set as variable fields, you'll be able to come and X that.

[6:52] Jumping onto here, you have a table which you can fill in. If you want to cut and paste into it, that's fine. It should be noted, by the way, if you're only doing a single submission, you would never come to this page. You'll fill in all of your data on the previous page, because everything would be a single entry. So you can fill that in there, but obviously with large numbers of entries that would be a real pain to try and cut and paste into all those boxes.

[7:19] So one of the options that we offer for bringing in this data is that you can pick out the template you've prepares as a CSV file, which you can then open up in Excel, OpenOffice, whatever you like. The headings for each of the columns are shown there. You can paste in your data, bring it in from your own spreadsheets. Once you've done that, you just upload it back into the system and that's where your data is.

[8:06] Finally, sorry, unfortunately the redraw on this is not particularly good. But never mind. But you can see that you've completed the "add citations." You've set your date for your data. You've filled in the common fields. Obviously in this case we haven't filled in the variable fields, so you can see that part is incomplete at the moment. Once you've filled in the variable fields, the final step is you click the "validate" button. Once it's done with validation, it depends upon effectively what we're told by the community and what we can implement.

[8:51] We can certainly check values in MIENS and control vocabularies, steps like that. We can in some cases do more statistical and logical validations. Therefore we can say if you have a value in one field which negates a value in another, then we can make those kind of comparisons and tell you at this stage, before you've left the submission system, that we regard there as being potential problems with your data. Either things which actually represent errors, or things which would represent warnings, which a curator would then contact you about.

[9:27] And hopefully, this will make it much, much easier for you to submit data sets. To give you an idea or what kind of things are at the back end of the system and why we feel that this will be adapted very quickly to anything that people want to submit to us, we take a look at this. I can't imagine it's going to be particularly pleasant to read.

[9:50] But this is basically just a very, very rough XML file which completely defines variable fields, descriptions of fields, all the information which you saw. At the end, as an example of files into which those fields are sorted. These take 10 minutes. It's done very, very quickly. So it's very easy to make them. It's very easy to modify them, and hopefully it's going to represent the good submission system that we want. Thanks very much.

Audience: [applause] [10:28]

Audience: [10:35] Maybe I missed something. How about stand-alone submission systems? Is there something in the works where you can create an XML file today and shift it?

Bob: [10:47] OK, so what we did imagine happening, because you can be using much the same template regularly, in the future we may have some kind of "my templates" system where you could have a pre-build template which was the specific one which you tended to use.

[11:02] So at that point, you'd just be able to pick out the Excel spreadsheet, if you now that this is the spreadsheet format you're going to employ for your data. So you can effectively fill that up as you're going, pop it into the system, say, yes, I want to use this template, put in my spreadsheet, and you're done.

Audience: [11:21] Since you just mentioned Excel spreadsheets, is there any explicit connection yet between this and the ISOTABs work?

Bob: [11:30] No. This came up in the discussion yesterday. It's probably worth noting that where ISOTAB comes into its own is when you're talking about things which cover a very large range of resources, whether it's DBI or NCBI. So if you're talking about projects which are extending across nucleotide databases, GEO ArrayExpress, et cetera. At that point, ISOTAB submission is probably the best way to go. Obviously because we have our own community of submitters, we're looking at our own specific submission system.

Audience: [12:06] Just completely out the window, we do have stand-alone, so from the ISOTAB conversation yesterday, he told us, he mentioned afterward.

Audience: [12:15] Just a quick comment. Both [indecipherable] and CPI accept similar spreadsheets for submission to GenBank.

Audience: [12:24] Why are your templates not in HTML? Because it looks the same, only it's in XML.

Bob: [12:32] I'm sorry?

Audience: [12:33] Your templates, why is not HTML?

Bob: [12:37] Because it's very, very much more generic. This isn't just purely for this kind of thing. This is for any type of submission. Any type of nucleotide submission.

Jon: [12:50] Kaye: So I guess this is sort of starting off with introductions. My name is Jon Kaye, I'm a program officer at the Moore Foundation. And like John said, Mary Matson is unable to be here, and is on her way out of the foundation actually at this point. I wanted to briefly describe three of the sequencing pipelines that are currently active at the foundation. It may help give a sense of the amount of information which is coming down the pipe, so to speak. And also talk briefly about the meta-data requirements for those pipelines.

[13:22] So there are three. The first you're probably maybe most familiar with, which is the Moore Microbial Genome Sequencing Project at JCVI, and I think Saul Kravitz is here somewhere who can perhaps speak to some the meta-data questions coming up. So that was a community-wide cull. At this point, the original idea was to get about 130 genome sequences, and it's now up to upwards of about 170. 175 are expected.

[13:54] A second community-wide sequencing pipeline has to do with phage, virus, and viral meta-genome sequencing, and that's a grant that was made to the Broad Institute. It's expected that there'll be about 320 genomes and viral meta-genomes that are coming through that project. I think just the very first genomes have been completed and sent to CAMERA at this point. Then there's a third sequencing pipeline, which was not a community-wide one, but it's just with the MMI, the Marine Microbiology Initiative investigators.

[14:28] That's focused on and meta-genomics and metatranscriptomics. That's a grant to Penn State and it's done on Schuster's Lab. The three different projects have different types of meta-data requirement lists. It's not standardized, but then what our expectations are are what we mapped out at the beginning at the foundation in part because these projects started at very different times. I tried to take a crack at comparing, at least for the genome sequences, the MIGS checklist with what we've required, and I couldn't quite do it.

[15:08] But the JCVI project has, I would describe it as, a minimal set of environmental information, which is less than on the MIGS checklist. And the Penn State pipeline has also just about eight environmental meta-data fields. The Broad Institute Phage, Virus, and Viral Meta-genome Sequencing Project actually has a pretty lengthy set of meta-data requirements, which you got about 30 to about 45 different fields. It actually at a quick glance looks like it maps OK to the MIGS checklist.

[15:52] I'll make a brief comment about the data sharing philosophy at the Moore Foundation, where we try to be as progressive as possible in terms of rapid open access to this information. All the sequence data and associated meta-data is required to go into CAMERA and CVI. The researchers have a six-month data embargo where they have exclusive use of that information. But after that, the data is publicly released. So that's just a quick overview, and I brought with me the meta-data lists that we have for those three different pipelines if anyone wants to take a look and see what those are.

Audience: [16:38] Thank you for your overview. I have two short questions. Number one is can you give us some specifics about this third project, metagenomics and metatranscriptomics? And

what do they mean by the second part of it? The second question is what is your long-term view on CAMERA? Support of CAMERA?

Jon: [17:02] OK, so the first question was the Penn State pipeline? OK, so the way that came about was a need identified by the MMI investigators, and there's about a dozen of them, for some sequencing banquet. So the grant was made to Penn State to, I think it was about 68 454 plates were allocated in the project.

[17:28] The investigators could submit individual projects or sometimes we require them to be collaborative between the investigators. We figured out which ones would be accepted into the queue, and those samples I think are just about all at Penn State at this point. Many of the data has already been sent to CAMERA. So your second question in terms of...

Audience: [17:51] What is your support of CAMERA?

Jon: [17:54] The plan is to enable CAMERA to become a financially self-sufficient entity. We're still working on figuring out how to do that exactly, but that's the long-term view. As with most things at the foundation, the foundation tries to enable organizations or entities to become independent organizations and entities.

Man 1: [18:26] Do you have any specific ideas? I'm just curious. Time to learn, please.

Jon: [18:32] Specific in terms of how it's going to...

Man 1: [18:34] How [indecipherable] ?

Jon: [18:37] I don't know. I may learn more at the CAMERA users next week.

Audience: [18:42] I was at the CAMERA three weeks ago. So basically the idea is to get it up and running with federal money mostly. That would be the long-term view of CAMERA, to get large fed grants to support as a resource. As to the advice of the house, I should ask Mark Edison who's now the active PI at CAMERA. Mark Edison.

Mark Edison: [19:15] Just a comment about what you said about standards and CAMERA. So CAMERA, at least up until three weeks ago, though the government work wasn't specifically involved in it, is completely makes MIGS compliant in terms of submission and the way things are filtered in. CAMERA actually has [laughs] a submission system that's pretty much analysis of what we've seen here at EBI. The inner workings are definitely based upon sequencing.

Weinstock on Human Microbiome Project

George: [0:05] Weinstock: So I am here to give you a whirlwind view of the Human Microbiome Project. So let's start with who the players are. All of this stuff is online at the HMP website. This is NIH Roadmap Project, so they have a... [0:27] And, in particular, if you want to see all of the projects that are going on there, look at the 'funded research' subheading at that site. That will take you to a page that has all the grants that are listed and who all the PIs are and what all the projects are. And as you'll hear in a second, there's a fair amount of that.

[0:48] One of the major groups is the HMP centers. These are big genome centers that are funded. So that's Balyor, Broad, JCVI and WashU. And as you'll hear, all of those centers are engaged in very, very similar activities to what you're doing, because in order to do a project of this size - this size being $150 million - it really forced these centers to come to grips in terms of consistency, uniformity and standards for all the genomics that they're going to do in this project.

It strikes me that they're not fully represented in this group. And what this group has done - which is just fabulous, I am very impressed with everything I've heard here. But it's going on in parallel with something that also is going to have a lot of weight because of this [inaudible 01:[1:26] 41] which is being done in even in the microbiome project.

[1:44] So somehow I think one of the future steps that would great for this group to do is to figure out how to engage the HMP centers more in your things, and figure out what the overlap is, and figure out how to emerge or make consistent between all your efforts and all of their efforts.

[2:05] There is a data analysis and coordination center, the DACC, which Jennifer is going to talk about. But many member of that group are here. I counted, I think, half a dozen or so. So I'm not going to even go there in terms of saying anything about that. And they're charged with an enormous responsibility, which is being the informatics hub.

[2:29] But maybe the biggest responsibility they have is trying to figure out what they're supposed to do, because all of these big genome centers had already, for a year or so, have been doing large scale informatics on all of the activities that the Human Microbiome Project before the DACC grant was even awarded.

[2:47] And so, there is a certain amount of catching up, certain amount of integrating, and a certain amount of figuring out what are all the things that weren't being done that really need to be addressed.

[2:58] There are a whole bunch of centers that are a part of what are called demonstration projects, which I'll tell you about in the next slide. And in addition to the genome centers, three of which have grants for the demonstrations projects, there's all these other places that have grants. And so, you start to get a feeling that the consortium of the Human Microbiome project is actually very, very large.

[3:24] It's - I don't know how many - but 20 or so different funded units, all working on different parts but all under the same umbrella and, therefore, were all having to come to grips with issues of consistency and uniformity across the program. And that's what I meant when I say there's a

huge effort going on in the HMP to solve many of the same issues that you all have been trying to and making a lot of progress in solving.

[3:54] And there are - which I won't say anything more about - development projects. There is computational ones, there's technology ones, and there's yet other groups that are involved in those.

[4:04] And so, here's the overview of the different funding boxes that the $150 million has gone to. The jump start program, this one over here, was what was done in the first year. A thing about these roadmap projects is that they're five years - those $150 million has to be spent in five years. But it starts from a dead stop.

[4:28] So the first year was putting out our phase and figuring out how you're going to dispense the $150 million. So what was done is, euphemistically called, the jump start phase. That was some supplemental money that was given to the big genome centers to do one year's worth of stuff.

[4:44] And, in that one year's worth of stuff, which obviously is going to take longer than one year, 500 reference genomes. And then, this bit here, human sampling, is the activity of roughly 375 people who will be sampled at 15 or 18 body sites, depending on whether it's a boy or a girl. And may I add, it will be sampled more than once in time. So, that adds up to something like 12,000 specimens. Each one of those specimens is microbial community which has to analyzed.

[5:20] And so, that's one of the things that's been done. I would say it's probably half way done, in terms of the amount of samples that have been collected. But the sequencing and the analysis has really only been in the pilot stage, because there's still techniques to be worked out and the standards and uniformity between the centers to be developed. But that's going to be a massive amount of analysis sequencing and data from 12,000 specimens.

[5:53] There is active discussion about the metadata. It should be included without... I think there's something like 200 fields, or something like that, that are collected in terms of information about the subjects. So a lot of stuff going on there.

[6:12] And then, this bottom part, metagenomics, is the analysis of those specimens, which is mainly going to be largely zonal and age sequencing and some shotgun sequencing. And we'll talk in a minute about the strategy there.

[6:29] Then this thing the middle, the HMP centers grants, that's when the first big grants were awarded, and that was awarded to the four genome centers that I mentioned on the previous slide, the HMP centers. They're going to do another 400 reference genomes.

[6:44] So these 900 reference genomes in these two projects have required the centers to come to grips with the definition of when the project is done in terms of sequencing and assembly and quality. You heard Patrick present what was an outgrowth of our definitions of the different levels of quality that we want to achieve.

[7:09] Annotation is a moving target, because everybody wants to try to annotate it in as uniform way as possible. So anyway, all that stuff is some of the standards that have been hoisted on the project because of the scale and the different centers that are working on it.

[7:26] There will be more metagenomics, and the metagenomics is going to be additional analysis of these 12,000 specimens from the jump start phase. Don't forget about viruses - there's the virum. And this strikes me as another area that I haven't heard too much about here, which is, there's going to be a huge amount of virus discovery - next generation sequencing, which is mainly to find viruses.

[7:54] And so, there are a lot of specimens. There are all going to be used for virus discovery, not to mention sequencing many, many, many independent versions of the same virus that define variability in the viruses in the human body. So there's a whole other area where standards are going to be required, and for this there's going to be going to be a lot of data.

[8:16] Eukaryotic microbes, fungi, things like that, that's part of this too. And then, there's sort of another phase of the centers RFA, which usually is being encouraged to be doing transcriptional analysis of the metagenomic samples, which most of the centers are going to approach by doing cDNA sequencing with next generation platforms.

[8:41] Then there's all these demonstration projects. I said there were 15 of them. Fifteen of them have been selected and they will get funded for one year. So they have to go out, and in one year, they have to generate as much as data they can about their disease. And then, five of them will be picked for scaled up funding for the next three years after that.

[9:01] So this is musical chairs, where the number of chairs will be very small compared to the number of people circling around them. And so, lot more data generation this year. I think only 10 of the projects or so involve the large centers. So there's a bunch of other people doing large scale sequencing as part of this project that are not necessarily hooked into the standards and things like that of the centers and so there's a lot more discussion with them and efforts to come to agreements about how they're going to be as uniform as the big centers are as well. So you can see that this thing has tentacles that start reaching into other data producers besides just the big centers.

[9:48] This is mainly metagenomics and frankly it's mainly ribosome RNA sequencing. I don't think there's very much shotgun sequencing except for there is one virus discovery project. And some of the projects have whole genomes. When there's a particular organism that is believed to be important for that disease, bang out a whole lot of whole genome sequences of independent isolates of that organism as well to contribute more genomes in that.

[10:16] So, that's sort of the lay of the land for the major activities. As I said on the bottom here are the development areas and the LC, which is always part of a NIH projects and then the DACC, at the bottom which really is, it should be at the top, kind of like the umbrella over all of this.

[10:36] Like I say, this is the part that's well coordinated, the things that involve the big centers that have been working together now for a year and a half or so, but there's all kinds of other players that are coming in from the other projects.

[10:49] So, what's the strategy for doing all of this, how's it being coordinated? The nine hundred reference genomes that are being done, there is a master list of well over 900 genomes at this point that was put together by working groups at NIH put together, community recommendations and also the centers initiated their own collaborations and had some legacy projects as well. So, we're not sure of genomes that do, and those are just sort of individual genomes that's not taking important genomes that you now want to sequence fifty different isolates of in order to look at diversity.

[11:26] And this is also, the 900 is based on sequencing costs from two years ago when this RFA was written. Sequencing costs have come down so in all likelihood the project will do several thousand genomes before it's done in the next few years.

[11:44] How do you sequence 12,000 specimens of microbial communities in a way that's meaningful? So the strategy is to do some kind of scan of all of them which will probably be 16S sequencing with 454. We all know that there's inaccuracies with that approach, but the hope is that it will be good enough to bin things. In other words, if for the samples coming from the nose, there are really only six different community clusters that you find when you look at 375 people, then you only need to take representatives of each of those six different bins and sequence them more deeply rather than having to sequence all 375.

[12:29] So there's a vetting and then there's an as yet to be determined algorithm for defining the microbial communities more deeply. Certainly deeper, 16S sequencing and shotgun sequencing as well will be involved there. Not clear on how deeply the viruses or eukaryotes will be studied and as I said, the demonstration projects are now much more variable because they're new and they haven't benefited from all these other discussions.

[13:01] This is just to give you a feeling for the data sets to be generated. As we said, well first of all, everything is rapid submission but because these are clinical samples here's another area that I haven't really mentioned here, controlled access data bases. What are the standards and how are you going to handle those? Because a lot of things you're talking about get more complicated when you have to do that. This is one of the things that the DACC has to come to grips with.

[13:30] Won't say much about cold genomes, we all know about that, lots of data sets from 16S sequencing and many hundreds of thousands if not a million full length sequences to be dealt with there. And in terms of next generation sequencing, it's, we just heard the discussion of M5, it goes through the roof.

[13:52] We can generate data in a year or two that with current computing infrastructure, even using hundreds of thousands of cords is going to take us five times as many years in order to compute on. So we're sort of coming to grips with that but it's not stopping us from getting the data.

[14:13] There's the DACC. It's an upset.

[laughter]

George: [14:21] Let me just say that there's a whole lot of other projects. This is the last slide. There's a whole lot of other projects that there's just going to be more of this; it's not just the HMP. There's other RFAs from most of the other institutes at NIH that are not on this scale but they're not trivial. And then there are these international projects that we heard mentioned earlier, not quite the same scale as the HMP but certainly ten to twenty percent of it multiplied by a number of other parts of the world so it adds up to quite a lot more stuff. [14:57]

Owen White on Consensus Annotations

Owen White: [0:03] ... to possibly make it possible to combine the annotations from different sources. And I would give us a reasonable grade on developing consensus in this group, but I'm nervous. Do you know the word game called, "Hangman," where you come up with a fake word and you guess what they are, right? [laughter]

Owen: [0:37] So, the idea is, you draw a little hangman. And once you get to the complete hangman, you've lost. So, I'm going to just make some suggestions, here. And if I get dings and I get down to the hangman, then that's it. We're done. I'm not pushing this any further. [0:56] So, I'm entertaining the fantasy that we could adopt some relatively simple procedures to describe our data, and also, evaluate that data. And then, based on those evaluations, we may be able to bring that data together from different sources. So, that's why I refer to this as kind of a consensus annotation system.

[1:26] So, there's multiple annotation systems. We've just heard about some. There are lots of different providers. What the picture there is showing is that this is some state of the art annotation system. And there are lots of contributors, and I'd like to see them work a little bit more effectively.

[1:47] Now, I'm going to insist that we engage in appropriate dialogue, here. And I may moderate some of your comments, and just request that you let me finish, just for the purpose of getting through the talk.

[2:01] There's some different ways that we can go, but if we bought into this idea, I think there might be some really amazing dividends.

[2:08] One question is; can we evaluate the annotation data? And we've already seen examples of this, OK? So, let's take, for example, REF SEED approach, is that they're saying that they've got better quality assignments of function. And other people would say that they have good quality assignments of function.

[2:26] And I'm just going to say, let's develop a gold standard, and whatever gold may mean that particular week. And see how different annotations are comparing. Then decide, maybe we will go with one versus another.

[2:40] Now, I'll point out that we've already done this in a couple different areas. So, this is what I mean by, the future is actually now. For the HMP projects, there are several centers that are contributing annotation. And we have developed a few different metrics for evaluating the consistency, here.

[3:00] And we're just taking a look at the quality and we may agree on a common theme. What that would mean more is that if one person's got a nice system, and one person's got a nice system; let's just choose to use that, as opposed to a consensus annotations system. We've looked at the quality of a few different pipelines, and I'm not going to report on that. But I will give you some other examples.

[3:24] So, one of the things that this creates is that you've got to make a decision about what the gold standard is. OK? And again, as I was saying earlier, I think this is really about is, everybody understands there's some peril with creating a gold standard.

[3:40] And there's definitely... once you establish what the gold standard is, you can game the system. You could even say, "OK, I'm going to use all your assignments from your gold standard."

[3:49] But what we did, was we, in the examples that I'm going to show you, we took TIGRFAM HMMs, which are pretty good quality families. And they have, somebody used the term, assertion, and try to be very clear about what that means.

[4:05] They have assertions like, this TIGRFAM gives you a potential functional name, or it give you an EC number, or it gives you a genetic name.

[4:14] And these are assertions, OK? What we've mostly been talking about, or what we're talking about the regression, we're talking about functional names. Obviously, there are many other assertions.

[4:22] And TIGRFAMs have these assertions. So, what I'm going to do is, I'm going to probe my TIGRFAMs against your annotation system, and just look at the results, and ask the question.

[4:35] For example, one of the things that you could do is just say, "For all these potential assertions for your annotation pipeline, did you make all these assertions, and did you do it consistently, OK?" There could be plenty of other measures that we use.

[4:49] But we just started out with this simple one. And when we did this exercise, this was back when there was several different Bioinformatics resource centers that we funded by NIH, to be creating annotation, and we grabbed up the data from those sites.

[5:03] And here's the overall volume of things that we looked at; the total number of organisms, the total number of genes.

[5:10] For every gene, if had a match to TIGRFAM, how many could we actually evaluate? So, it's like the yield of what we could evaluate with, let's say, hovering around 13% for these different sites. And one of the questions that we asked was, did we assign it consistently?

[5:28] And this was a formula that Sam came up with for, just asking a question, if you've got a gene product name, what was the frequency where you got the same identical name?

[5:41] And this is really down at the level of like, you remove all spaces and all commas, and just ask using Perl does this string equals this string?

[5:49] And here's an example of one annotation center that's annotated HisA from lots of different organisms, and you can see what you'd like to see, which is that there is nice consistency across all these names.

[6:02] And here's another example, the site where they had looked at this HisA and you can see variation, subtle variation. These aren't necessarily wrong, these are conveyed pretty much the same information as biologists.

[6:15] So you can see there are some weaknesses to this gold standard, and I think that there's lots and lots of room for how you can define what consistency means, and I'm not saying this is the best way to go.

[6:28] So, we also asked the question, if there was a potential assignment, you know, how complete were they? If you could have made an EC number assignment, if you could have made a genetic name assignment, did they do that?

[6:41] So for all the genes that could receive an assertion what was the percent that got an assertion? And here's an example of if I was looking at GOA assignments, for one site there was a possibility based on using our TIGRFAMs that a total of three thousand genes could have gotten to the site, but that got an assignment from that site they only gave 896. OK?

[7:06] Here's another example for, EC numbers, and this is, I think one of the things that, sort of highlights this issue which I think is just fine, is that some sites don't do EC numbers. Now, that's absolutely fine, OK? Some sites do, some sites don't, it's not a big deal.

[7:20] And now getting down to looking at an individual gene, I'm looking at what people did for HisA, and you can see that it's sort of spotty.

[7:28] There are some places where cells are filled in, you know there was potentially thirty-six genes that could have gotten an assignment and this is how many got an assignment.

[7:37] OK so, are these spotty results bad? What I'm going to argue is, no, there's really no problem with this at all. Let me just walk you through this.

[7:47] This is again, a sort of a summary of, here's the source of the data, here's the assertion that we were looking at, here's a GOA assignment, EC numbers, and I've got numbers for completeness and consistency.

[8:01] OK, for lots of these different places. And I think my animation's going to kick in here. So what I'm looking at, at GOA assignments. What I'm going to do, in terms of, approaching if I've got different assignments for these different assertions?

[8:18] I've got GOA assignments for example. Well, what I could do is I've got completeness and I've got consistency, and I can just rank them, for seeing who did better with consistency or completeness.

[8:29] And the idea is that NEPDR, when they made assignments, they made very good GOA assignments. They made great, consistent, GOA assignments. But when you look at this, you can also see that there were some cases where they didn't make all the assignments.

[8:47] OK, how to deal with this? This is a relatively straightforward idea. I could go ahead, and take these NMPDR assignments for all places that they were there, and I could basically go to the next one down and grab the rest of their assignments.

[9:05] So here's the idea, OK, I've got lots of different genes, suppose we're just sort of thinking linearly along the genome, OK, and I've got different sources, and I've got different levels of quality, OK?

[9:20] So I've got the best quality here, I've got the best quality here, I've got the best quality here, but it's not complete, so there's still some empty cells, so first I'll just grab these, this is Owen White's algorithm and if it works in PowerPoint, obviously we've been encoded in Perl, OK?

[laughter]

Owen: [9:36] So, we're going to grab all the genes that are best quality assignments, but that doesn't really fill everything up, and we're going to grab the next level of assignments and even go down to compromising. [9:48] If we sort of say, OK, there's some level of cut-off that we trust, and this is consensus annotation, where we can think of it as data type saturation. This is sort of, what I'm after. OK?

[10:01] So I'm going to take this a little bit further, and this is where it would be really nice to get agreement across groups is that we can start to ratify certain assertion types. I don't think anybody, would really fight this too much, OK?

[10:15] We have something like a common name or a gene product name, functions, a function assignment. We could have an assertion of EC number and genetic name. We could also have GOA assignments as potential assertion types.

[10:31] I don't think people would challenge that too much. And it's fine if we want to go into different types of assertion types; there might be regulation, there might be pathway assignments.

[10:41] The idea is, we try to develop a nomenclature for a common set, a common control category for assertion types. OK?

[10:50] And this is the same thing, I've got my different assertion types like, the description of a DNase quality, easy thing and for every one, G1, the same deal across all of these things.

[11:02] I've got different types of assertions. Some of them are high quality, this is the best and this is the best. Same thing. Pull out my magic PowerPoint algorithm, wave my hand over it, and grab this stuff up. And it's the same thing.

[11:14] This is what I mean by data-type saturation. I could get this from different sites. But the thing that's required is that we all have to agree on the concept of assertion type.

[11:23] We have to agree on the concept of GOA assignment, which isn't too big a deal, I expect. So I'm going to skip this. The next thing I want to talk about is these assertion types like the EC number has evidence behind it.

[11:42] OK? And when we started to encode the evidence behind this with the BRC program, we just drew on what was already out there.

[11:50] There's a bunch of evidence codes that have been created by the GOA Consortium. And they're listed here. They're just three-letter codes, and they're introducing essentially concepts, and there's lots of documentation behind this, where ISS is really what all the manual curators do.

[12:07] It's called "Curated from Sequence Similarity," but the idea is that there's a pair of eyes who's a lot of experience with manual curation and they're looking at the matches from sequence similarity.

[12:19] And they're looking to see what the information is behind all those sequence similarity matches and how much we can trust that, and they're making an assignment.

[12:26] This is some of the richest assignments that are out there, actually. EXP, inferred from experiment. Literature, inferred from looking at literature, and we had slightly nuanced definitions here to make it work for all the BRCs.

[12:40] There's just a little bit different from the GOA Consortium, but it's not a big deal. Fully electronic annotation, so you're just scarfing BLAST matches, you're not really looking at it.

[12:49] But then we threw in a couple others, and Michelle Gwinn has been leading the charge with actually getting these evidence codes adopted by the GOA Consortium. So we got a little trick here. We threw in some things like Inferred From Genomic Context. So NMPDR was doing this great job of looking at clusters of genes that are neighbors to each other.

[13:06] And based on that clustering in the individual genome, they knew that this was a particular operon, and they developed the algorithms to say, "OK, if this guy hasn't gotten an assignment but is in this operon, I'm looking at the genomic context and I'm making an assignment from there."

[13:19] Or presence in cluster, the idea about the cluster was that clusters actually could be genes that come from lots of different genomes, like TIGRFAMs, OK?

[13:29] And you could be looking at if it had similarity to something like TIGRFAMs. So here's these evidence codes, which are different than assertions. You could make an assertion of an EC number, but it would be based on these different types of evidence.

[13:42] And what I'm going to work up towards is if we adopted these, there's ways that we could combine pretty rich annotation.

[13:47] So same story, we had lots of these assertions, lots of different types of evidence, OK, for individual genes, that are for more than, one gene, you could have lots and lots of rows of evidence that was associated with this.

[14:01] And this was really great, this is rich data. And you're starting to deal with the fact that some people have very great types of evidence.

[14:09] So for some particular site where I'm very impressed by information that's coming from the literature. OK? And for some sites, they had a whole lot of literature assignments. Their function was based on a match to literature.

[14:26] And some places didn't put a lot of emphasis on that, that's not where they were placing their priority, and that's still a beautiful thing, because they were developing other types of systems for gathering up evidence.

[14:39] And what I'm trying to get at here is this, actually, when you start approaching the world this way, you start to celebrate that different people have different approaches and place their priorities in different areas.

[14:51] So now again, the magic PowerPoint algorithm, here. In the past, if I've got different assertions, OK, without thinking about the evidence, I could go ahead and just grab them up.

[15:05] That was great. OK? But I could also, just for GOA assignments, now I'll say assertion dot EV code, GOA assignment dot inferred from sequence similarity, GOA assignment inferred from looking at clusters.

[15:19] There was one site that did a great job of looking at cluster information. OK? So if I want to grab this stuff up, I've ranked them, I think some guy's good, some guy's better, and I can just grab these up as well.

[15:34] So for the same gene, you can have an assertion dot EV code, and you can have multiple lines of evidence coming at this thing. And you can combine them together, make it available to the user. OK. So that's where we could go with this concept of consensus annotation.

And the biggest thing that's really preventing us from doing that, is really agreeing on these two things: [15:55] a description for EV codes, which is published, which is out there, which is documented.

[16:07] And we could have a small feat food fight about, if we could have some nuanced things that we want to do, but if we all agreed on this, we could start taking this approach.

[16:17] So that's what I'm out here for. The main thing that I really want to tell you is that we've got the potential for making much more richer data-type descriptions.

[16:30] One of the things that's nice about these data-type descriptions is that I think that they present a potential audit trail. We talked about evidence codes earlier. It seemed like that was a good thing. The other thing that I like is that it starts to celebrate the diversity of different centers producing different types of annotation.

[16:50] And instead of having this sort of competitive model of REF SEED arguing that our method is better than your method, so we wipe all your data, all your different types of assertions, all your different EV codes.

[17:04] We just go with ours. We start to work in a much more cooperative way. OK. So that's me. I'm like Joe consensus builder. And the idea is that we might make some progress.

[17:16] And the other thing is, we know by definition, that if you're ranking this data by its quality, and saturating the number of cells for anything, you know by definition you're improving the quality.

[17:33] OK, there's no real harsh argument there. And there might even be some other ways that we could make decisions other than just ranking based on something as simple as, or just sorting based on completeness.

[17:47] And if we got sensuous about how we actually evaluated the data, I just came up with two methods, consistency and completeness, but there really might be other ways to do it, then we also know that we stand a better chance of making a wise decision about this.

[18:06] So, I have this belief that if we just develop some more complex methods, we could combine this data.

[18:12] So I want to recognize Heather and Michelle in my shop that's been working with the REF, HMP Annotation Group. And they have started to develop this. Ramona is here.

[18:22] There's other contributors on the HMP side for people who are starting to evaluate this, and trying to arrive at a common pipeline. It's just remarkable watching people work together on this.

[18:35] I think it's just a wonderful thing. Sam's here. Him and another fellow did a whole lot of these evaluation methods for the Bioinformatics resource centers. I want to thank the GOA Consortium, because they're basically the ones who have created the evidence codes. And be advised, the evidence codes for GOA were devised for GOA assignments.

[18:59] And I'm doing something relatively heretical. I'm saying, we could even use the same evidence codes for something like EC numbers. There's no reason not to reuse them. OK?

[19:09] And in the documentation on the GOA side, they say, these are about GOA assignments. Yeah. Thank God. Suzanne is in the back. She's shrugging, she's saying I could live without, that's totally cool.

[19:19] All right, so there are these guys. Now. This is a mail message that I got recently. This is coming from Susan Gregurick, who's the program manager at the DOE. And I just want to highlight a couple things.

[19:34] She and I and Dan Drelick, have been talking about essentially setting up an annotation competition. And when I use the term competition, it means I want people to elect to participate in this.

[19:47] It's not really going to be about us going on a witch hunt to find out who's got a high quality and who doesn't. OK? But Susan reported that it was fabulous to see me. I always like that. And she says that she has been talking to people at NIH.

[20:04] And we may have a funded initiative, OK, so money for, potentially, a meeting to get together and start talking about CAFE, which is the term that she coined for "Critical Assessment of Functional Annotation Experiment."

[20:21] So they're looking at it from the angle of, back in the day, a lot of protein structure people were arguing about how well their software and their methods were working for making determination of function from structure.

[20:38] And they started a competition called CASP. And if any of you have seen any of this, CASP changed everything. OK?

Participant: [20:48] We started CAFE already.

Owen: [20:48] What's that?

Participant: [20:50] CASP already organized the CAFE.

Owen: [20:51] OK. Cool.

Participant: [20:52] We organized the CAFE in 2005.

Owen: [20:53] OK.

Participant: [20:54] It's not a new name.

Owen: [20:55] Oh, my bad.

Participant: [20:55] We did it before.

Owen: [20:56] This is entirely my idea. OK? Right up here. Oh, all right. Fine. One. Dun, dun dun dun. But CASP changed everything. [21:09] Now what was the social engineering that happened with CASP? There was an interest in people participating in CASP. They wanted to participate in CASP to showcase how well their system was working.

[21:22] OK. And they started to develop fair and open evaluation methods to really look at the system. And then after a while, there really wasn't a publication that came out that didn't talk about how their system compared against CASP.

[21:39] And it started to be a reasonable standard. And I humbly offer that we have an opportunity here. We could get together. We could talk about what the measures would be.

[21:53] What would be the evaluation process? I think one of the coolest thing is if you could take a totally anonymous genome that was sequenced out of the JGI, but everybody got it without it having gone through any annotation process. And they could run it through their systems and we could evaluate how their data comes out.

[22:10] We could develop other types of gold standards. We just used TIGRFAMs. You could completely habituate the system really quickly. Everybody could start using TIGRFAMs and everybody would get near unity in terms of exact matches.

Field on RCN4GSC and Sterk on MIGS-compliant data curation

Dawn Field: [0:02] So as I said before, this meeting's actually being funded by this NSF grant now that we have this research coordination network. And John Wooley was supposed to give this. This was supposed to be the big launch you know the next five years where we start to pull the steering committee together, decide where we want to go as a community but unfortunately John isn't here. [0:18] He has sent me just a few slides. At this point, again, it's simply to say that we have the funding to get this onto our horizon and to talk a tiny bit about how this is a special part of what the GOT is doing because I think the main work to be done with this is to connect with some of the other NSF funded communities in the US.

[0:37] So Nykos is here for the long term ecological research stations and networks, they're starting to do more and more genetics and metagenomics. We want to make sure that that ties in with traditional ecological work with contextual data and also moving more closely allied to the biodiversity community.

[0:54] So I don't know how many people know about research coordination networks but they give you about five years of funding, it's a hundred thousand a year, specifically for networking. So creating a nucleus around a particular topic, a lot of people that come to meetings, exchange staff between groups and it seemed an ideal mechanism for the GOC.

[1:12] We had Matt Kane at early meetings who sort of helped us think about how we would pitch the proposal but when it came down to it I came and talked to Nykos and Victor at the J-Jine, they had just come back from GSE-5 and they only wanted to talk about cross counsel that, not cross counsel they're in the UK but inter-agency, excuse me, inter-agency activity and they figured the one person who could start to engage in some of these discussions was John Wooley and he successfully took it forward.

[1:41] So the research coordination network is to continue the GOC, as we said it's about half a million for five years which seems like a lot but on the other hand it's not that much, and I'm extremely thankful for it. We need to focus it more on outcome based meetings which is why we're splitting out into working groups here and it'll probably end up, we're not sure how many big meetings we'll end up supporting versus meetings that satellite onto other meetings whether it's ISME or ISMB as we've done it, it'll be a PSB satellite meeting in Hawaii this year with people being funded to go.

[2:13] Four F working groups will actually be funded to meet as smaller groups and also to promote this exchange of early career scientists. Lorenzo's on the line. Many of you know Lorenzo Coppman but I think he'll be one of the first people who will come to the US and work with a group or sets of groups here on GOC issues.

[2:32] There is a short six paper. So George has now given us a platform for speaking to the wider community and John did pen something, I think Nykos and I are on there, which simply directs the community to the fact that it's there. As I said the two key things which are to engage further with other NSF communities but it does cover all the GOC core projects and I realized in

the agenda, we're not even going through these core projects anymore in detail because they're sort of just part of the fabric.

[3:00] But again, they are all described on the Wiki so it's GC demo exchange language genetic preset standards which is mapping of identifiers across databases so you can go from goal to IMG to seed NTPI and back again [overlapping whispering] , OK, habitat and life, how we would describe habitats which is really an extension of the environment oncology and the status journal. I had this idea that the genome journal could talk about more.

[3:29] So we do have a governance body which is the steering committee. So John is the PI but many people from the board and beyond, people needing core projects are all part of the steering committee. Another thing to notice that there is a line in here that says, "Other steering committee members accordingly will be selected to further expand the efforts in ecological environmental and biodiversity efforts", so it's still an open slate.

[3:50] It's very, very early days. And hopefully they'll be more activity about this after the mainstream, there's a lot of it here. We can certainly start thinking about how to list effectively in a case of time. So thanks again to John Wooley and I'll end there so we keep a bit more time on the agenda. The reason why I was here has new hyperlinks to the original award and I hope we start on that very soon.

Female1: [4:13] OK, thank you Dawn and it is too bad that John Wooley can't be here. The, we're going to jump ahead in the agenda because Eugene Koker has not arrived yet so we're going to postpone that and the next person will be Peter Sterk who has been doing the duration of mixed data. He'll talk about that.

Peter Sterk: [4:42] So I'm just going to talk a little about what we've been doing for the GOC. Basically we've been working quite hard on getting the standards together, I'm going to talk about the standards, you'll know that. Dawn already mentioned we've got an exchange language, GCDML, XML schema. And after spending years on getting these standards agreed upon and part of made by technology, I said to her, to Dawn, "It's time that we got some data so that we actually get more credible and it's not talk about six but that we start doing six." [5:36] So only this year, actually June the last, GOC meeting in the UK, Renthorough, took us through the GCML and showed us how we could use an XML editor to edit data and I looked at it and thought, "This is actually doable. It's not the most user friendly thing but if we just sit down and go straight on the job we can actually get all the genomes done." So since January we have been employed by the GSC and my main focus as of this start is the duration job.

[6:28] And my project was to protect in GCDML so we have a lot of reports now and sent to Nykos who has done quite a lot of duration goal is going to be a quick way to get his data price program to convert as much as possible into changing the amount and then in one afternoon I have caught 800 basic genomes per reports.

[7:06] But a lot of data is still missing. And then the task was to settlement so the report with more data. And then going back to the literature and read a lot of articles and try to find as much missing data as possible.

[7:31] Luckily I'm the final institute Cambridge and U.K. Has given me a desk and I'm sitting among the curators and they're going to help me with that task. And also in the future I try to get the data and to start to write a report right at the start of projects, which I think is very important to actual sitting report is just too cumbersome and you need to capture data right from the start.

[8:20] So my hands off to Seemback data into goals. I think last few months goals has gone through a lot of major updates, a lot of additional fuels have been added. I've just broken some of my scripts but now it's Nick Goldridge and author of the reports that I'm creating much richer. So this gives us gold data plus what to add as much data as possible. And as I always said, it takes a lot of time and it will often look very easy to decide what kind of data you put in the different fields.

We try and use as much as possible terms from ontologies and control vocabularies. The problem there is that not all the terms evolved to use have been defined. But I've spent time on the virtue of this here as to define as many of the terms as possible for [inaudible 9: [9:17] 52] Wiki at the appropriate level. There's lots more terms that add definitions, lots of preferred in ontology where they should go.

[9:55] So we have to work with the communities to network ontologies to actually get data into the ontologies and that's ongoing. But it starts and as I said, that's ongoing.

[10:15] So another thing is you can have the reports but you also need to have them visible. So as we are in agreement in RCN4GSC and GenBank and Vol BMDJ to start incorporating mixed data into the genome records. It's not entirely clear how it's going to be presented in the records but I've already, a lot of work is done with Guy Cofrench from Amble who has managed to take my data and put it in X amount in the records. If you know where to find the data it's there but they're still working on the presentation.

[11:06] But it's a start and it will start in the next year. Certainly increase.

[11:17] I'm not going to say too much about adequate, because Crisanna will talk to it and it will be on the agenda today and the cart was, but we had a meeting in past two days to discuss how we could use Isitack format to put our rich data in. Created the tool Isaac creator which is much more user friendly than XML editor. And I've worked with the team to it. It's obviously very how suitable the tool is.

There are further advantages as in fast to use system because essentially I'm not going to say too much about [inaudible 12: [12:09] 25] . But some like to do say, so it works. You can get the data in that system. The importance is that the more we can use it to test how we can compare data.

[12:49] So there's a lot more to do. And I hope that from our experiences that a lot more people will start capturing the kind of data and we will work with the community to different ways to curate as many genomes as possible.

[applause]

[13:20]

Wortman on the HMP DACC

Jennifer Wortman: [15:05] Sure. Don't worry I don't have slides, but I do have talking points. I'm afraid I'll go blind if I use my computer. Thank you. [15:25] So the DACC web portal is up and running at HMPDACC.org and I hope everyone will take a look at that. So I want to acknowledge the DACC members that are here at this meeting. Of course, Owen is the PI for the DACC at the University of Maryland School of Medicine. I work with him on trying to help manage the project and we collaborate with Nikos and Victor here for the various aspects of the Repair Genome Project Catalog, and genome and metagenome representation through ING and INGM and with Todd Santos and Rob Knight for the 16S component of the project.

As George sort of alluded, the project, the [inaudible 16: [16:17] 23] overall has a working group organization so the various parts of this, the reference genomes have both a strain selection working group and an annotation working group. And then there are working groups for the data analysis and generation and data benchmarking, and coordination of demonstration projects and so we work across all these working groups and try to come up with the best way to support those working groups in their data integration and visualization analysis needs.

[16:54] As far as interaction with the GFC and the standards, the reference genomes are currently available listing of those through the project catalog which is powered by the gold database and so all of the available metadata is available through gold, through the project catalog. And I think that, over time, we'll be trying to fill in any gaps that exist in those genomes in the metadata through curation in new business group and coordination with resources at the University of Maryland.

And as George said, these will mostly be draft genomes. Some of them are going to be, going to higher level finishing so we've actually got an extension that Eileen mentioned in her talk to try to deal with the fact that we will have both a finishing status and potentially a finishing goal which could be different from the current [inaudible 17: [17:29] 49] DACC. So we've been trying to add extensions to the normal information requirement.

[17:56] And one of the reasons I'm here is to really help determine where the genomes, where all of this maintenance means data should go eventually in Gen Mag and so I'm hoping we can talk some offline about what should go at the project level, what's going to be in the sample objects, what should be in structured comments and what's the current advice versus what the transition will be going forward.

[18:20] And I have the 16S sequences obviously working with Todd and Rob and they've been assigned to make sure that the centers are aware of the means developments and we are collecting nucleic acid prep and library metadata and providing the centers with exemplars and guidance for submissions to SRA.

Yilmaz on MIENS specification

Pelin Yilmaz: [0:03] So now I'm just going to talk about the MIENS. So I'll try to give you an overview about the activity of the MIENS or the meaning for information about environmental Nykos' working group and what you're going to see in this talk is structured in six main points. [0:18] So first we'll have a brief introduction for those who weren't in the MIENS working group and then we'll specifically see how we make the MIENS, so this is second, third and fourth points. And I'll present you the checklist and an example of compliant recourse and finally I'll close my talk by closing the cycle with the INSDC partners or in terms of mixed complying contextual data.

[0:43] So I'll start with the short history of MIENS, so the proposal was given by Frank Oliver at the sixth GSC meeting in 2008 and it grew there. And it is designed to be the reporting standard for marker genes coming either directly from the environment or from cultured organisms and regardless of the sequencing platform.

[1:03] And there is a, it was unclear in the beginning whether MIENS standards were for the Migsman standards so I just want to say that MIENS is actually a continuation of the Migsman standard and it inherited the Migsman standard so it's nothing separate from that and yeah, this is the short history. And we started the active dual period after the entry period in Sweden this year and so the checklist was made nice within a series of teleconferences over the summer with the following group members.

[1:38] So again before jumping into the specific parts of making of MIENS I'm just going to show you a simple schema of how it was done and again say, showing that, our organic relationship with the Migsman's checklist. So we practically inherited the Migsman checklist and extended them to cover requirements of marker genes by analyzing some specific community surveys and analyzing publications, INSD resources and of course incorporating the expert knowledge from our working group members.

[2:11] So the first part will be community surveys. So up to date there were four surveys around to make by, made by different groups and we merged them into this metasurvey within our working group and analyzed their results to get their consensus.

[2:25] Now I'm going to talk about what they were focusing on and who made them. So the first one was the JPI survey in 2005. And it was, it focused on general descriptions for marker genes. The RDP survey was the second one. This was rather different. And it focused on the potential uses of high level habitat terms for sequence retrieval searches by researchers, by users or by RDP. The third survey was the Silva survey and it also had a general focus of, for description of marker genes. And the fourth, the last survey we just got the results from was the Perredino survey. It was also a specific, very specific survey. The importance and need of obtaining of a number of descriptive for several method genome studies.

[3:14] Now I had a feeling that the results of all these surveys were presented over and over by different people in various vocations so I'm not going to go through them all over again, I'm just going to show you what we have learned from them. So once again we approved the increase

from the community who does the marker gene studies, they wanted to have the contextual data. And of course we had great ideas about the fields that should be included that are important for people out there, for MIENS. And since the habitat or environment part is important for the new information about the environmental sequence, we got aid in selling into these big groups of environments.

[3:57] The second part of making of MIENS was making analysis and it starts with publication mining. Why do you, simply the publication mining is because the surveys gave us some great insights but they can also be biased in terms of information because the users or respondents can say that something is relevant even though it can mean a lot of things for in real life or it wouldn't really apply in real life so we tried to read early publications and from, coming from different habitats focusing on different environments and tried to extract some so-called environmental premises for our environmental packages in the MIENS.

[4:34] So, so far we have thirty nine publications. The first pass of these publications, the first pass scanning was really only the publications that summed up the most sequences. But then we saw that we were biased how was the soil and the organism is in differentiated habitats and we did a second pass scanning for to get specifics of habitat such as I don't know, environmental events or micro mass or bio fields and eventually we extracted some two hundred parameters like these here. And I guess most of them are included in the environmental packages of MIENS.

[5:12] The last part of analysis is the INSDC resources. We tried to get user statistics from the source of information parts from INSDC by Silva and we tried to get ideas about we feels as such already have been used. I'm again not going to focus on our results because in terms of getting new ideas it was a bit unsatisfactory. But in terms of approving our activities it was good because we saw that the contextual things that we want are not there and there is a need for standardization in that sense.

[5:50] So now I'm over with the making of MIENS probably and I'll try to give you an overview of the checklist. So it's really hard to read so I divided it into different sections so I'm hoping that some of you will see something. So for this slide and the other slides the general idea is that on the first column you see the item made. On the second column you see the definition of this item which is something also new and will be applied to Migsman's checklists. And on the very last column you see whether this item is mandatory or recommended or not applicable. And yeah, in gray highlighted you see the core items of MIENS, meaning that it's mandatory for uncultured and cultured parts.

[6:34] And of course we have uncultured and cultured because as I said, MIENS is for environmental and uncultured organism sequences and the requirements are rather different. And after seeing the whole section, I'm going to walk you through the new things introduced with MIENS. So new being brand new or change from the old MIGS/MIMS fields.

[6:58] So here we have the Investigation and the Environment section. From the Investigation we have three core items, submissions to INSCC, the investigation type and product name. From the Environment, we have geographic location, time of sample collection, and the habitat. The new items are submission to INSCC, which used to be submit to NSCC and submit to trace archives in MIGS/MIMS, but we merged them. And the experimental factor to explain the experimental design of the study.

[7:32] The second section is the MIGS/MIMS extension, which used to be the MIMS extension only. Now here is the most changes because if you remember, we only have the water body in the MIHS/MIMS checklist, and this was a criticism saying that this was sort of habitat bias. To overcome this we increased the habitats. Now we have six more as air, extreme habitat or in an associated sediment, terrestrial and the wastewater sludge. So here you see all the work done and all the information pulled out from surveys and publications and all the legacy data we got from our group members, thanks to Rodni and Linda.

You will here see a better view of how they look. So the next section will be a nucleic acid sequence source. We have again four core items here, specific host, the sample collection device or method, sample material treatment, and library screening strategy. And the following are the new items: [8:14] sample collection device or method, this is brand new. And the sample material treatment was biomaterial treatment in the MIGS/MIMS, and we changed it to clarify sort of what this field was asking. The amount or size of sample is also inherited from MIGS/MIMS but changed. It was volume of sample, so we wanted here equals mass here obviously. And library screening strategy is also changed from MIGS/MIMS as it was screening strategy I think only and again to clarify this field.

[9:11] So now the final section is the Sequencing. So we have 10 core items here. I'm not going to go through them all and showing the new items. We'll go through most of them. Field Extraction and amplification used to be one field, so we split it to make it more atomic. The target PCI primaries and PCI coalitions are here for obvious reasons. And the sequence quality check and the primary check, yes quality check specifications which are now mostly applying to standard sequences but which might be extended to cover next-generation sequencing as well.

[9:44] Yes, now just to convince you that this checklist can actually be done, we present an example compliant report. I guess it's really hard to see, but you have to believe me that it is from an actual study. It's from a marine study from our institute. It's the result of a typical RNA approach. So get the RNA genes from the environment and Sanger sequence them. And yes we tried to make this compliant with the researcher and she said it was rather easy to comply with this checklist because she was writing the paper already and she had all the fields we asked for. Just for your information.

[10:19] And the very final topic will be the closing the cycle. Here the discussions are rather incomplete, but I'll try to give you what appears to me so far. I am on the INSCC so I don't know how they will accept our contextual data. So we have a sequence, and the MIENS-compliant report or contextual data on my hand. So the researcher may choose to merge them or make them submission ready himself or he can get aid on all the tools that are out there.

[10:52] I am just focusing on the contextual data life cycle the sequence is nevertheless INSCC partners so that's not a problem. If it goes to GEM Bank or NCBI, as secret system, it is proposed that it ends up as a structured comment. If it goes to MBIL via the NYS or the new submission system that NVL is preparing it can either ed up as a structured comment or a sample object, and the format is not really important for them. And the next generation sequencing has always a different submission system and in this case it will most probably if not definitely end up as a sample object.

[11:36] And closing the cycle, all these contextual data will be parsed back by the secondary databases and will be ready for review researchers and users of this contextual data to search and index and incorporate in their research.

[11:50] And I'd like to finish my talk by thanks for your attention, thanks to all my Bremen colleagues and the MIENS working group.

[applause]