Bridging the Science - Information Technology Gap

Presentation Outline

- Who is the Bioteam?
- Our HPC and Storage Expertise
- Our Data Management Solution WikiLIMS
- Discussion “How can BioTeam serve Ignite?

BioTeam at a Glance

- Incorporated in 2002
- Providing high-performance computing, storage and data management service for Life Sciences
- All consultants have both Life Science and IT expertise
- Vendor Agnostics (do not accept vendor commissions)
- Global Client base of > 300
- Partners - Intel, Isilon, Illumina, ABI, Pacific Biosciences
- Channel Partners - Apple, Univa, Schrodinger

HPC & Storage to Support High-Throughput DNA Sequencing

BioTeam HPC, Storage & Data Management Services

- Research Analysis - evaluating your objectives
- Technical Assessment - formulating plans for computing and storage
- Architecture - designing capable, cost-effective solutions
- Implementation - building and integrating
- Testing - validating new and existing systems
- Training - training all users, scientific and technical
- Application development - custom software for research

HPC Informatics Consulting Practice

- Specialty practices & areas of focus:
  - Science-centric IT & Infrastructure Consulting
  - Distributed Resource Management
  - Utility/cloud computing on Amazon EC2

IT & Infrastructure Practice

- *Science-centric* HPC IT consulting & project management including:
  - **Facility**
    - Build-out, technical assessments, relocation/migration projects
  - **System Design**
    - Translate scientific need into IT requirements
    - Turn IT requirements into scalable research IT blueprints
  - **Purchase Assistance**
    - Write RFP documents
    - Evaluate vendor RFP responses & assist with vendor selection
    - Strip inappropriate or unnecessary padded items off of vendor quotes

IT & Infrastructure Practice

- IT & Infrastructure continued …
  - **Storage Practice**
    - A rapidly growing specialty practice
      - Technical storage audits for life science organizations
        - » Document requirements, estimate growth, identify capability gaps
    - Terabyte to multi-Petabyte storage system design services
    - Integration of terabyte-scale wet lab instruments
      - Confocal microscopy, ultrasound, next-gen sequencing, etc.

Distributed Resource Management

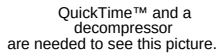
- 10+ years building production clusters & compute farms for Biotech, Pharma, Academic and Government clients
- Deep involvement way beyond “traditional” IT scope:
  - Far more than hardware setup & deployment
    - Installation, deployment & configuration assistance
    - Custom tuning & configuration to match scientific need
    - Scientific application & workflow integration
    - Custom training for end-users, developers & operations staff
- Acknowledged as global experts on Platform LSF and Sun Grid Engine in life science environments
  - Popular community blog http://gridengine.info operated by BioTeam
- BioTeam is the only company offering life-science LSF & Grid Engine training
- BioTeam is the only company offering Grid Engine training aimed at end-users

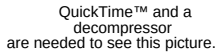
Philip Morris USA Center for Research and Technology

QuickTime™ and a
decompressor
are needed to see this picture.

- $350M total project - New construction
- $10M in IT spending (system specs confidential)
- BioTeam Role (Aug 2005 - Oct 2007):
  - Document functional & scientific requirements
  - Develop HPC compute and storage architecture
  - Contribute technical material to RFP documents
  - Assist project team and purchasing agents as ‘subject matter experts’
  - Project manage delivery, unpacking and installation within an active construction site
  - Bring all systems online; test against acceptance criteria
  - Custom training for scientists & operations staff

NASA Langley Research Center





- Providing on-going computing, Grid Engine and storage support for atmospheric research (since December 2006)
- Multi-Petabyte storage system:
- 1.8 PB raw/1.2 PB usable
- 384 fiber ports & 2560 individual disks
- IBM GPFS filesystem
- Mixed IBM server environment
- x86_64 & PowerPC
- Workflow managed via Sun Grid Engine

Naval Medical Research Center

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

- Architected, sourced, implemented and currently supporting:
  - Research IT platform for $15 million Bio-defense grant program
- Designed to flexibly support next-generation DNA sequencing from multiple vendors
  - 454, ABI SOLiD, Sanger and Nimblegen
- Leveraged best of breed commodity solutions in a highly price-conscious environment
- Features:
  - 80 Core Linux Cluster
  - A ~100 terabyte single-namespace storage system costing $150,000
- *Competitive quotes from commercial storage vendors > $1 million*

Utility Computing on Amazon EC2

- Since early 2007 ***every*** active BioTeam consultant has independently used Amazon AWS products to solve real-world customer problems
- Currently working with ISVs and client companies move software and workflows into EC2
- BioTeam's Amazon Cloud Clinets/Milesotnes:
  - 1st to publicly demonstrate mpiblast operating on EC2
  - 1st to publicly demonstrate self-organizing Grid Engine clusters within EC2
  - UnivaUD to document Unicluster/EC2 integration
  - Sun Microsystemst
  - Applied Biosystems
  - Helicos Biosciences
  - Frederich Miescher Institute
  - The Broad Institute of Harvard and MIT
  - Pfizer

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

Challenges in Managing Research Data

- High-Throughput Instruments are creating Exponential Data Growth
- New technologies, and changing technologies
- A mix of users: scientific, technical, and informatic
- Multi-platform experimentation (Illumina, 454, Microarrays, Etc.)
- Legacy data locked in out-dated systems and files
- Low volume areas of the lab that are orphaned and have no LIMS
- Data of all types: text, image, video, tabular, relational
- Personnel and conditions change, and closed software isn't maintained

What system can address ALL of these challenges?

Wikipedia - the world's most used wiki

Wikipedia is sustained by people like you. Please donate today.

Log in / create account

article | discussion | view source | history

Wikipedia

From Wikipedia, the free encyclopedia

This article is about the encyclopedia. For the different, similar terms related to Wikipedia, see Wikipedia (terminology). For Wikipedia's non-encyclopedic visitor introduction, see Wikipedia:About.

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search
Go Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

languages

Wikipedia (pronunciation) is a free,[5] multilingual, open content encyclopedia project operated by the United States-based non-profit Wikimedia Foundation. Its name is a portmanteau of the words *wiki* (a technology for creating collaborative websites) and *encyclopedia*. Launched in 2001 by Jimmy Wales and Larry Sanger,[6] it attempts to collect and summarize all human knowledge in every major language.[7]

As of April 2008, Wikipedia had over 10 million articles in 253 languages, about a quarter of which are in English.[2] Wikipedia's articles have been written collaboratively by volunteers around the world, and nearly all of its articles can be edited by anyone with access to the Wikipedia website.[8] Having steadily risen in popularity since its inception,[1] it is currently the largest and most popular general reference work on the Internet.[9][10][11]

Critics of Wikipedia target its systemic bias and inconsistencies[12] and its policy of favoring consensus over credentials in its editorial process.[13] Wikipedia's reliability and accuracy are also an issue.[14] Other criticisms are centered on its susceptibility to vandalism and the addition of spurious or unverified information.[15] Scholarly work suggests that vandalism is generally short-lived.[16][17]

In addition to being an encyclopedic reference, Wikipedia has received major media attention as an online source of breaking news as it is constantly updated.[18][19] When *Time* magazine recognized "You" as its *Person of the Year* 2006, praising the accelerating success of online collaboration and interaction by millions of users around the world,

Wikipedia

Screenshot of Wikipedia's multilingual portal.

URL www.wikipedia.org

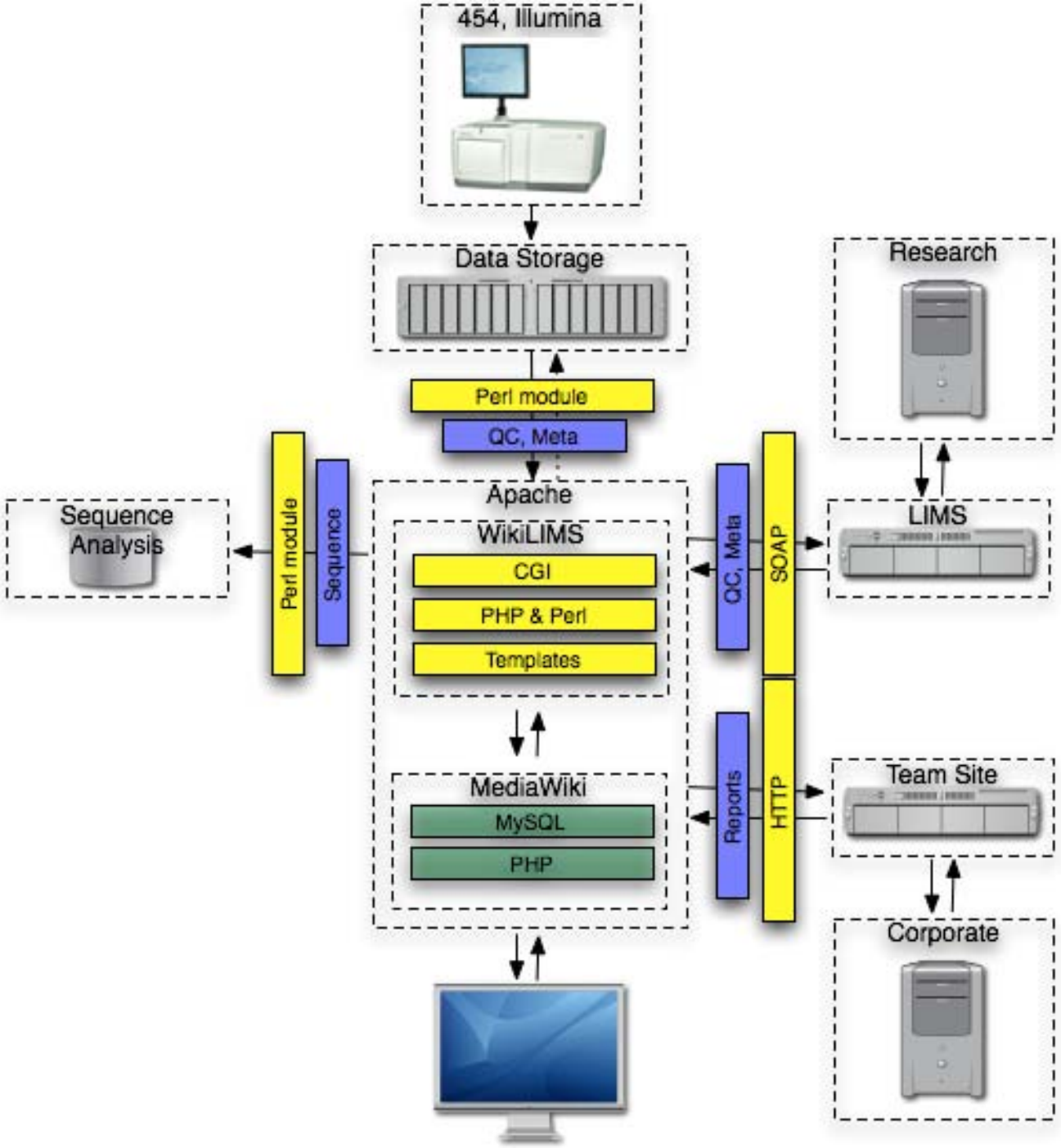
WikiLIMS Client Solutions

- A *multi-national corporation*
- The Navy Biodefense Research Directorate (BDRD)
- John Grealey Lab at the Albert Einstein Medical College
- Brent Graveley Lab at the University of Connecticut
- Cornell University
- Cold Spring Harbor Laboratory
- Indiana University, Center for Genomics and Bioinformatics
- AMDeC, New York State
- Dana-Farber Cancer Institute
- caBIG
- Pfizer

Not shown: Helicos, National Cancer Institute, EPA, CDC

“multi-national corporation”

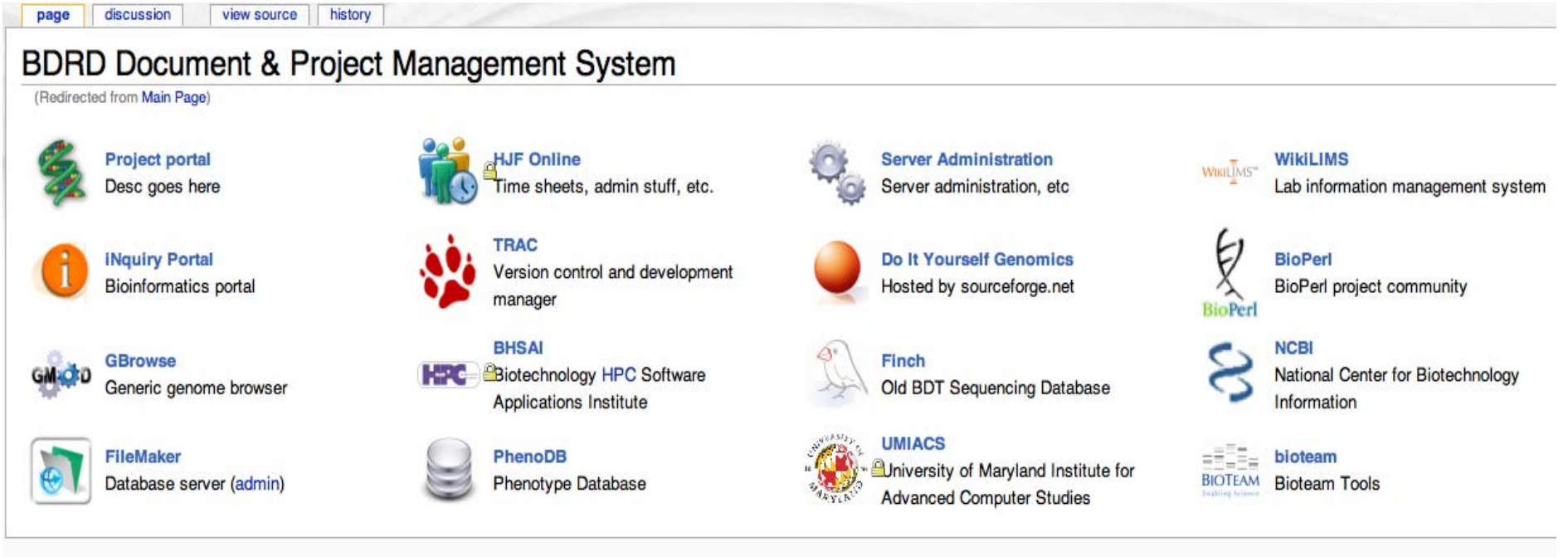
- Situation
  - 3 Roche GS and 2 Illumina GA
  - Existing commercial LIMS system
  - Existing commercial sequence analysis platform
  - Existing collaborative platforms, Web-based
  - Need to make projects visible
  - Need automatic data movement in all directions
  - Key : Projects, Samples, Libraries, Sequencing


454, Illumina
Data Storage
Perl module
QC, Meta
Apache
WikiLIMS
CGI
PHP & Perl
Templates
MediaWiki
MySQL
PHP
Sequence Analysis
Perl module
Sequence
QC, Meta
SOAP
Reports
HTTP
Research
LIMS
Team Site
Corporate

Naval Medical Research Center - BDRD

- Situation
  - An expanding collection, greater than 10,000 bacterial strains
  - Need to create rapid sequencing and annotation pipeline
  - Need to launch commands from the Wiki and get results back
  - Need to submit genome sequences to NCBI from the Wiki
  - Need to submit raw data to NCBI Short Read Archive
  - 4 Roche GS instruments in continuous use
  - Affymetrix data
  - Key pages: Strains, Cultures, Projects, Assemblies, Genomes, Runs, Microarrays

BDRD - Portal


page
discussion
view source
history
BDRD Document & Project Management System
(Redirected from Main Page)
Project portal
Desc goes here
HJF Online
Time sheets, admin stuff, etc.
Server Administration
Server administration, etc
WikiLIMS
Lab information management system
iNquiry Portal
Bioinformatics portal
TRAC
Version control and development manager
Do It Yourself Genomics
Hosted by sourceforge.net
BioPerl
BioPerl project community
GBrowse
Generic genome browser
BHSAI
Biotechnology HPC Software Applications Institute
Finch
Old BDT Sequencing Database
NCBI
National Center for Biotechnology Information
FileMaker
Database server (admin)
PhenoDB
Phenotype Database
UMIACS
University of Maryland Institute for Advanced Computer Studies
bioteam
Bioteam Tools

BDRD - Workflow

- Acquire strain, barcode, enter into Wiki (creates **Strain**)

1. Sub-culture **Strain** in the lab, create a **Culture**
2. Organize **Strains** by biological features (creates **Project**)
3. Extract DNA (creates a **Run**)
4. Sequence one or more times (creates **Assembly**)
5. Assemble one or more **Assemblies** from Wiki (creates **Genome**)
6. Annotate one **Genome** or entire **Project** from Wiki
7. Submit **Genomes** to NCBI from Wiki
8. Submit raw data to NCBI Short Read Archive from Wiki

BDRD - Strain page

Strain 34F2deltagerH | Family | Annotations | Assemble

- **Parent:** Species BAN
- **Siblings:** Strain NS1066, NS1236, NS1261, NS1262, NS1263, NS1264, NS1265, NS1266, NS1267, NS1268, NS1
- **Children:** Culture S5718
- **Grandchildren:** Sample N1288, N2264, N2265, and N2276

Strain 34F2deltagerH | Family | Annotations | Assemble

- View Data Sets for Assembly (NS3818)
- View Full & MID-tagged SFFs for Assembly (NS3818)

Note: When you are doing an **Assembly** from the Wiki always start on the Strain page for the strain whose genome

BDRD - Add a Strain to a Project

Contents [hide]
1 Related topics

Unclassified Project

Description: Incomplete field

Organism: Bacillus

Reference strain: banth0001X

Point of Contact: Incomplete field

Run Annotation Pipeline (stable)
Run Annotation Pipeline (dev)

Related topics

- Tiling Bacillus genomes
- DSC2-Clade specific proteins

NS1035	ATCC_10792	rev	bthur0008	29723	527031	BTH	Bacillus thuringiensis	1 rdirs	3 pdirs	P_2008_01_01_00_00_01_454rig	bases	reads
NS2805	95/8201	rev	bcere0016	29669	526979	BCE	Bacillus cereus	4 rdirs	5 pdirs	P_2008_02_13_17_56_34_loki	121769577 bases	487766 reads
NS2969	ATCC 10876	rev	bcere0002	29671	526980	BCE	Bacillus cereus	2 rdirs	5 pdirs	P_2008_03_15_14_08_19_loki	255903816 bases	1872088 reads
NS2971	AH_621	rev	bcere0007	29655	526972	BCE	Bacillus cereus	2 rdirs	3 pdirs	P_2008_02_01_13_24_20_loki	113454102 bases	458254 reads
NS2974	ATCC 4342	rev	bcere0010	29665	526977	BCE	Bacillus cereus	4 rdirs	2 pdirs	P_2008_06_25_13_09_24_loki	108850454 bases	406807 reads
NS2996	m1293	rev	bcere0001	29657	526973	BCE	Bacillus cereus	2 rdirs	3 pdirs	P_2008_03_09_15_37_45_loki	260032027 bases	2061922 reads

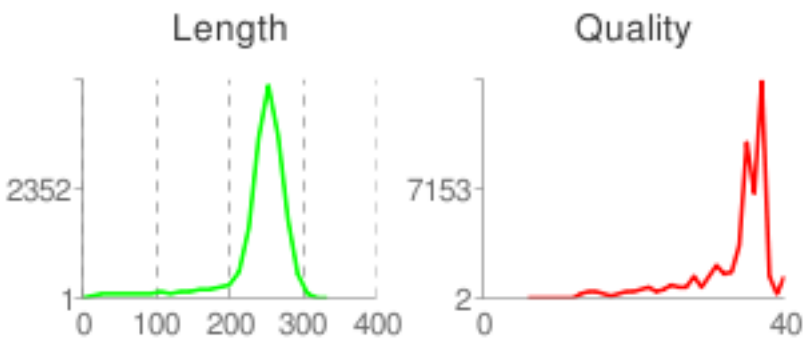
BDRD - Sequencing Run

454 Run R_2010_01_27_18_44_23_FLX01070131 | Images | Commands

- View Data Sets to Split by MID
- runAnalysisPipe on R_2010_01_27_18_44_23_FLX01070131_adminrig_NS5708xxN2413
- runAnalysisPipeAmplicons on R_2010_01_27_18_44_23_FLX01070131_adminrig_NS5708xxN2413
- runAnalysisPipePairedEnd on R_2010_01_27_18_44_23_FLX01070131_adminrig_NS5708xxN2413

Categories: 454 Run | Is a 454 run

454 Run R_2010_01_27_18_44_23_FLX01070131 | Images | Commands



BDRD - Launch Jobs from WikiLIMS

NS5646 (*Wed Nov 4 09:51:07 2009*)	*Select SFF files*
R_2009_10_27_17_41_11_FLX01070130_adminrig_NS5643XXNS5646XXNS5649XXN2337XXN2341XXN2345	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR02.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR02.MID5.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR02.MID4.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR02.MID11.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR02.MID1.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID9.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID8.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID7.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID6.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID5.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID4.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID3.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID2.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID12.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID11.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID10.sff
	D_2009_10_28_10_24_57_node003_signalProcessing/sff/F4RKQXR01.MID1.sff
R_2009_10_02_14_46_34_FLX01070131_adminrig_NS5646xxNS5647xxNS5648xxN2341xxN2340xxN2336	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ02.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ02.MID6.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ02.MID5.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ02.MID4.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ01.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ01.MID6.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ01.MID5.sff
	D_2009_10_04_12_27_28_node004_signalProcessing/sff/F3G1ZWZ01.MID4.sff

☑ Use *-large*

Assemble

... and create a Genome page

Yersinia frederiksenii ATCC_33641

Sequence	Assembly	Features	Homology	Variation	Links
Length: 4885341 bp	Contigs: 161	Genes: 4427			NS2456
GC Content: 46 %		Average gene length: 912			NCBI Project 29743
Coding content: 82 %		Average intergenic space: 189 bp			P_2006_10_06_11_11_52_runAssembly
Topology: linear		Number of overlaps: 56			NCBI Taxon 349966
		Structural RNAs: 69			Project Name: PGL1

Files

A directory | genbank | fsa (whole genome) | fna | faa | rfam | trnascan | rnammer

Genbank Submission

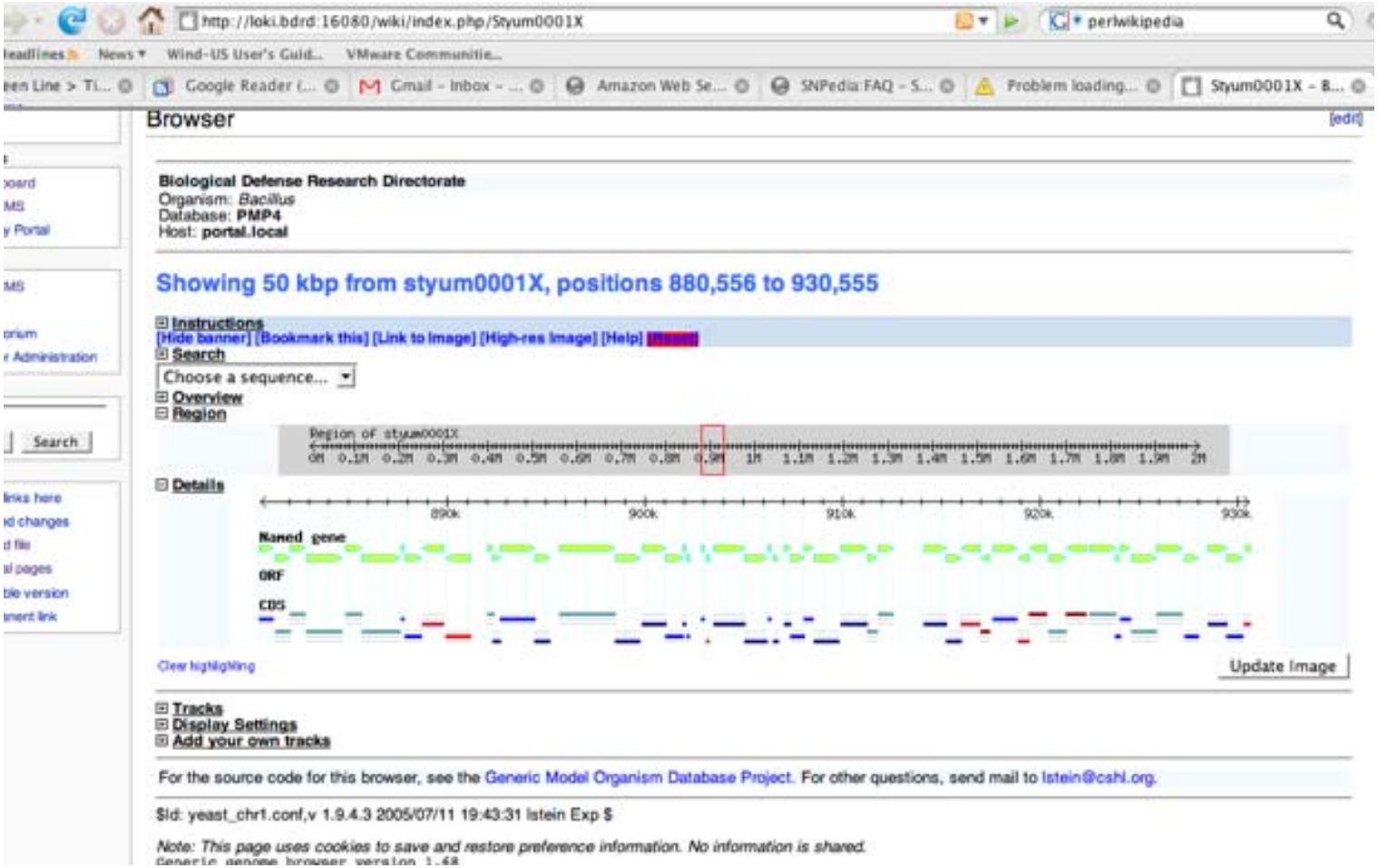
All annotation files | Genbank | ASN.1 | fasta | AGP | quality | tbl | val

SRA Submission

Submit yfred0001 Files to SRA

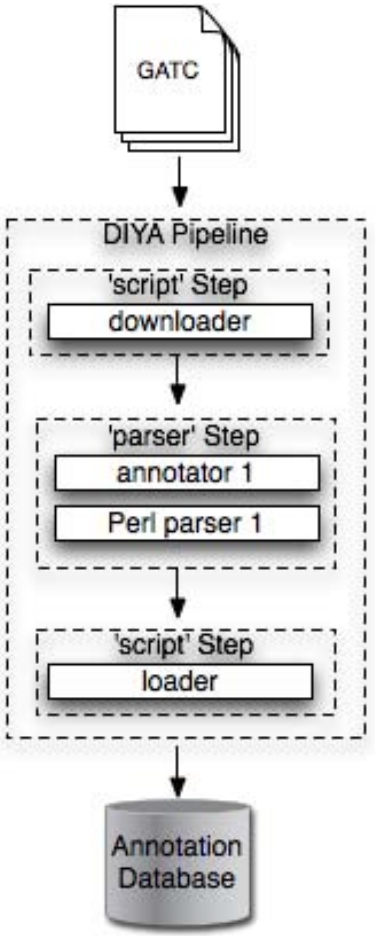
Update Yfred0001

BDRD - View genomes in the Wiki



BDRD - Launch annotation pipeline

DIYA - open source pipeline software (BioTeam & BDRD)

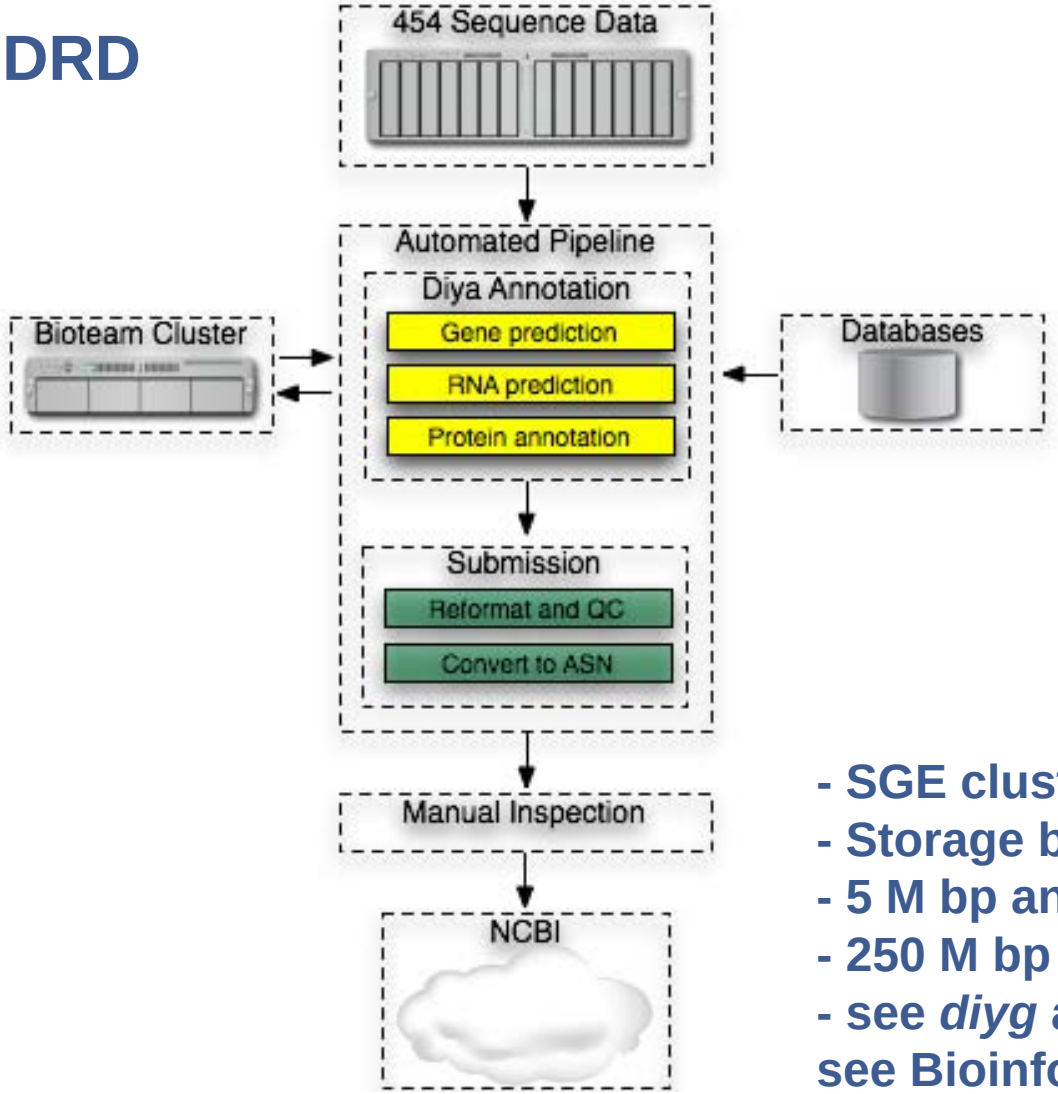



```
<script>
 <name>downloader</name>
 <executable>script-1.sh</executable>
 <command>-o OUTPUTFILE</command>
 <home>/usr/local/bin</home>
 <inputfrom></inputfrom>
</script>

<parser>
 <name>annotator1</name>
 <executable>annotator</executable>
 <command>-i INPUTFILE > OUTPUTFILE </command>
 <home>/usr/local/bin</home>
 <inputformat><inputformat>
 <inputfrom>downloader</inputfrom>
</parser>

<script>
 <name>loader</name>
 <executable>script-2.pl</executable>
 <command>-d database -i INPUTFILE</command>
 <home>/usr/local/bin</home>
 <inputfrom>annotator1</inputfrom>
</script>
```

DIYA at BDRD



- SGE cluster by BioTeam
- Storage by BioTeam
- 5 M bp annotated, 45 minutes
- 250 M bp submitted in 6 weeks
- see *diyg* at Sourceforge

see Bioinformatics March 2009

Albert Einstein Medical College Center for Epigenomics

- Situation
  - Core Facilities for Genomics and Epigenomics
  - 1 Roche GS and 1 Illumina GA, NimbleGen microarrays
  - Need to handle sample submissions
  - Need to allow external labs to retrieve their results
  - Need to reserve and schedule technicians and instruments
  - Key pages: Client Request, Samples, Jobs, Notebooks, Analysis, Billing

Albert Einstein Medical College

ad hoc client analysis

Front-end components

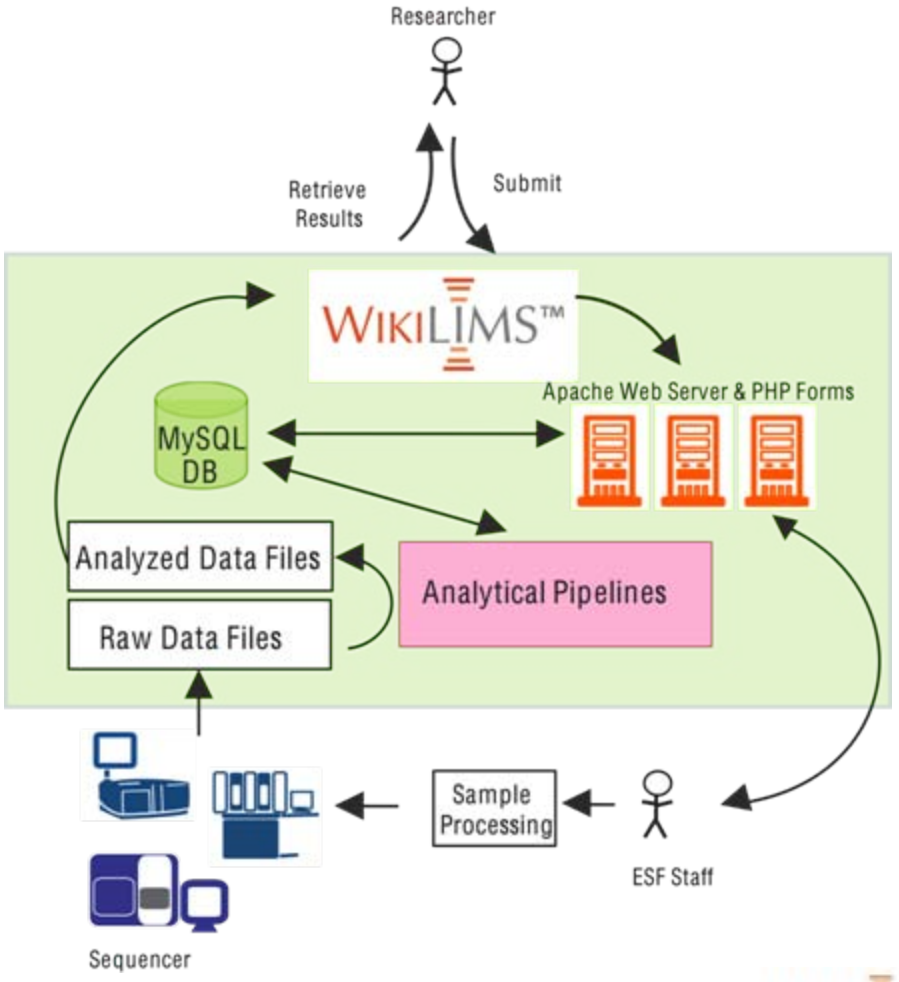
Customer request UI

Results and reporting

Data Tables

Visual Analytics

File Management



Einstein - Managing client requests for sample submission

UCSC Genome Version: | UCSC *

Priority?

Required result date November 2009

Sample QC Analysis:

Gel image: Upload file

QPCR: Upload file

Attach files

Antibody information:

Antibody name: *

Antibody manufacturer name:

Antibody catalogue number:

Antibody lot number:

Antibody amount used:

Samples:

See also Sample requirements

#	Name	Type	Size (bp)	Amount (µg)	Conc. (ng/ul)	A260/280	A260/230	Volume	Buffer
			200-500	> 10 ng	1-100	≥ 1.8	≥ 1.7		
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

Multiple samples

Einstein - Sequencing Job Results

Job description

- **Job Name**
  - CHP-SEQ with anti-GATA1
- **Assay Type**
  - ChIP-Seq
- **Submitted By**
  - Masako Suzuki (Greally Lab)
- **Submitted Date**
  - 07/13/09
- **Completed Date**
  - 09/18/09
- Click to Show Charts of Job Quality

Sequencing and Alignment Results

Flowcell ID	Sequencing Summary	Sample Name	Lane	Raw Data File	Alignment Result
42DCEAAXX	Click to show	ES_no_cytokines_INPUT	lane_1	Click to download	Show in Genome Browser
		ES_plus_cytokines_INPUT	lane_2	Click to download	Show in Genome Browser
FC42AHHAAXX	Click to show	ES_plus_cytokines_anti-gata1	lane_3	Click to download	Show in Genome Browser
		ES_no_cytokines_anti-gata1	lane_4	Click to download	Show in Genome Browser
		ES_plus_cytokines_anti-gata1	lane_5	Click to download	Show in Genome Browser

Peak Finding Results

Sample Name	Sample Type	Flowcell ID	Lane	Result (1 sample)	
ES_no_cytokines_INPUT	Input	42DCEAAXX	lane_1	Click to download Show raw.bed in Genome Browser Show peaks.bed in Genome Browser	

CHiP-Seq, CHiP-chip

Custom assays

Analytical jobs

Custom web reporting

Using Mediawiki API

Launch Custom Apps

Gbrowse

Jalview

Einstein - Quality Control Reports

Google Charts API

Lane by lane metrics

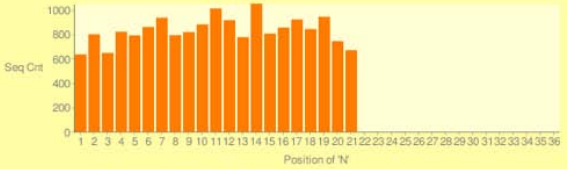
Sequence Quality Statistics

ES no cytokines INPUT (flowcell: 42DCEAAXX lane 1)

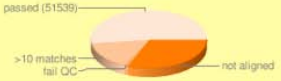
Sequences With and Without 'N'

Sequence Count vs Number of 'N'

Sequence Count vs Position of 'N'



Sequence Statistics Post Alignment

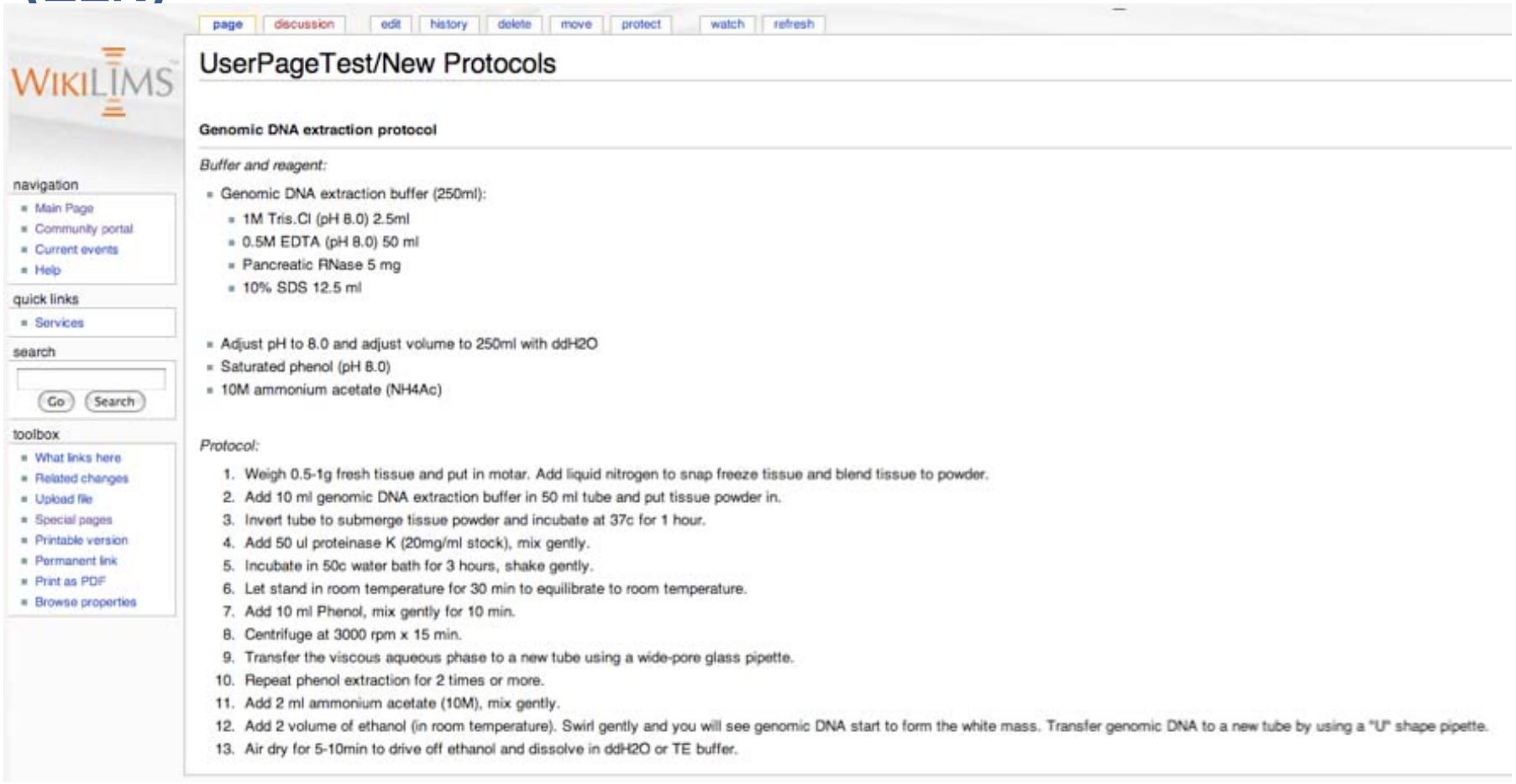


Run Quality Parameters

These quality metrics are based principally on single read 36bp sequencing of human DNA. The optimal results will vary according to experiment type.

Metric	Result	Uniformity (across tiles)	Notes
Total Yield			Result is 139730 (target is >1Gbases).
Raw Cluster Count			Result is 106688 +/- 9424 (target is >20,000).
% Clusters Passing Filter (PF)			Result is 36.46 +/- 6.93 (target is > 70%). Low % may be indicative of high cluster number (clusters too close together to obtain a clean signal in early cycles)
% Clusters PF that Align Uniquely to Reference			Result is 1.38 +/- 0.12. Optimal value dependent on read-length, genome sequenced and completeness of reference. For 30mers and the human genome, < 80% may be normal.
% Error Rate of Clusters PF			Result is 5.68 +/- 0.51. Should be ~1.5% but in any case, as low as possible.
% Phasing			Result is 0.6900. Should be ~0.5% to no more than 1% but in any case, as low as possible.
% Prephasing			Result is 0.5400. Should both be ~0.5% to no more than 1% but in any case, as low as possible.
First Cycle Intensity			Result is 15 +/- 4. Should be >1000.
20th Cycle Intensity as % of First			Result is 72.12 +/- 14.20. Should be >50%. If too high, suspect relatively low first cycle intensity

Einstein - WikiLIMS Electronic Lab Notebook (ELN)



Einstein - WikiLIMS Electronic Lab Notebook (ELN)

page | discussion | edit | history | delete | move | protect | watch | refresh

WIKILIMS

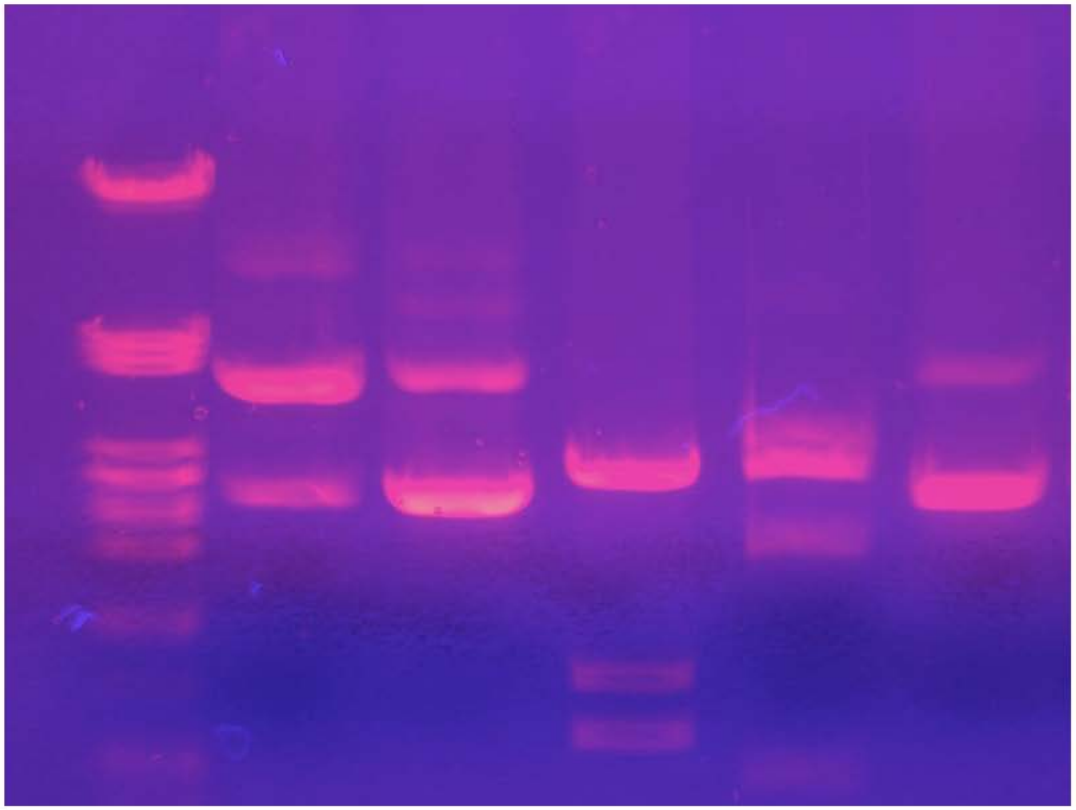
navigation
- Main Page
- Community portal
- Current events
- Help

quick links
- Services

search
Go Search

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Print as PDF
- Browse properties

UserPageTest/Presentation Images



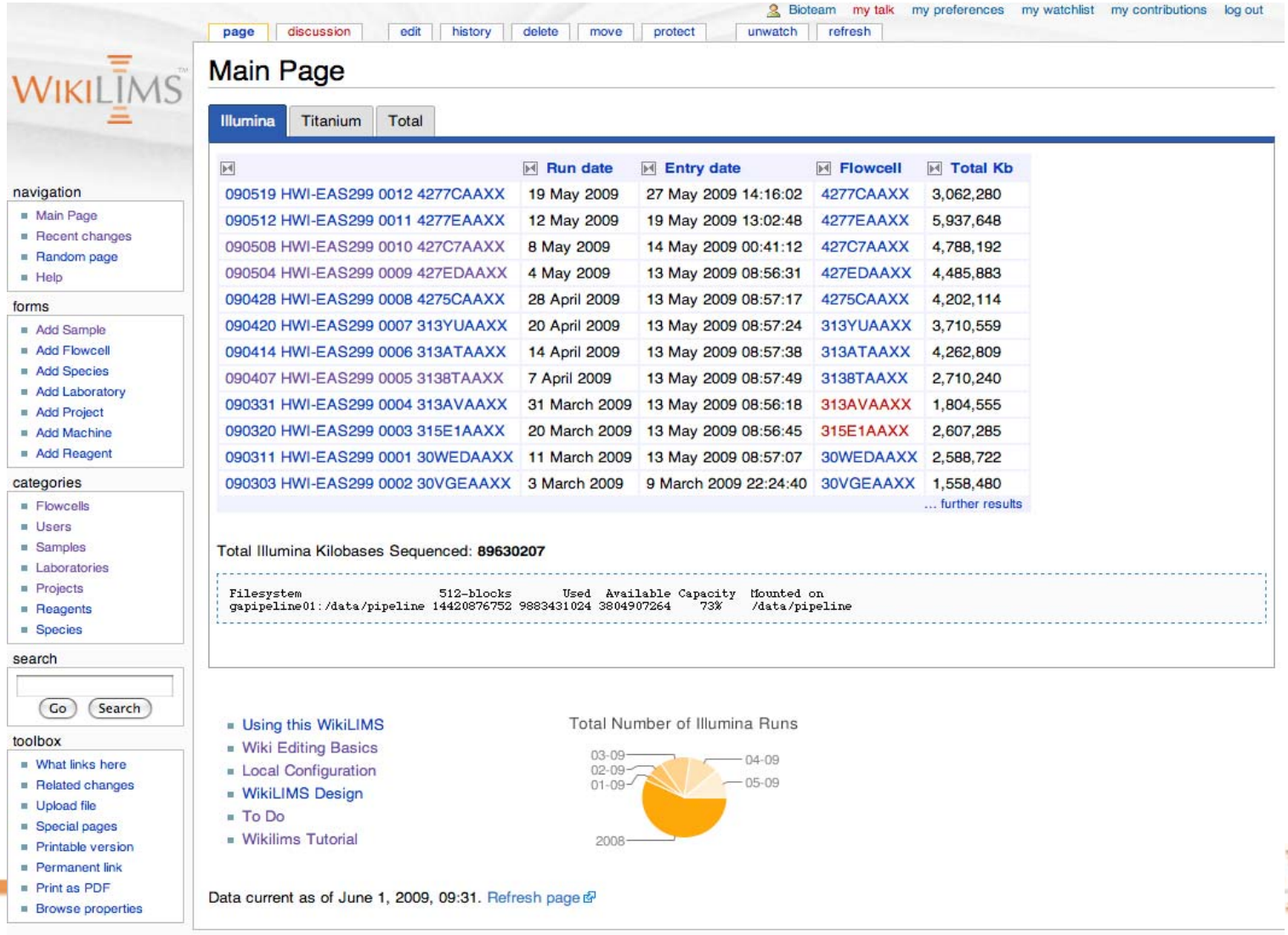
Description

Gel electrophoresis: 6 "DNA-tracks". In the first row (left), DNA with known fragment sizes was used as a reference. Different bands indicate different fragment sizes (the smaller, the faster it travels, the lower it is in the image); different intensities indicate different concentrations (the brighter, the more DNA). DNA was made visible using ethidium bromide and ultraviolet light.

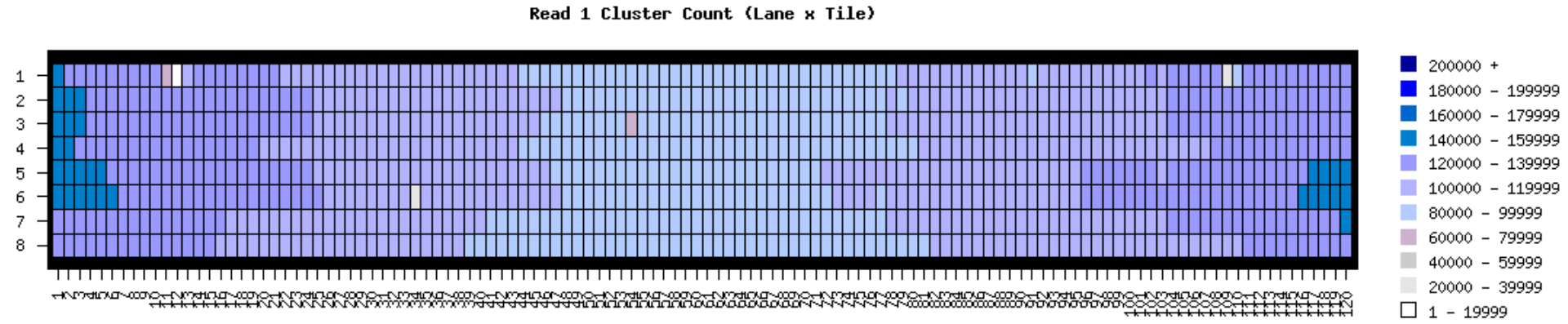
University of Connecticut

- Situation
  - 1 Roche GS and 1 Illumina GA 2
  - Multiple labs and multiple research projects (and modENCODE)
  - Need to allow data submission and data retrieval from external laboratories
  - Need to track reagent use and work by each user
  - Key pages: Flowcells, Laboratories, Projects, Samples, Reagents, Users, Species

U. Connecticut - Load runs automatically



U. Connecticut - Monitor Quality of the Sequencing Run



U. Connecticut - A Sample has Project and Laboratory data

page | discussion | edit with form | edit | history | delete | move | protect | watch | refresh

S2-DRSC Brr2 RNAi rRNA minus-1

Flowcells with Sample **S2-DRSC Brr2 RNAi rRNA minus-1**: 30B5NAAXX

S2-DRSC Brr2 RNAi rRNA minus-1	
Sample Description	**S2-DRSC Brr2 RNAi rRNA minus-1**
Sample Type	mRNA-Seq
Library Type	Paired-End
Species	Drosophila melanogaster
User	User:Liyang
Project	ModENCODE
Laboratory	Graveley
Date submitted	2009/02/01
Sample Reagent	⚠

Category: Sample

U. Connecticut - Flowcell, with User view and Flowcell details

30B5NAAXX

Lane	Sample	User
1	PhiX	Core
2	CT-2 MEF-1	Misha
3	CT-2 CM-2	Misha
4	Sexual Nonirradiated mRNA PE	Dasaradhi
5	Asexual Nonirradiated mRNA PE	Dasaradhi
6	Ago(RNAi) mRNA PE	Dasaradhi
7	S2-DRSC Brr2 RNAi rRNA minus-1	Liyang
8	S2-DRSC PS RNAi rRNA minus-1	Liyang

	30B5NAAXX
Amplification date	2009/02/09
Cluster station 1	N/A
Cluster station 2	N/A
Cluster box 1	N/A
Cluster box 2	N/A
Betaine	N/A
Lane 1 primer type	Genomic primer 1/2
Lane 1 primer lot	N/A
Lane 1 sample	PhiX
Lane 2 primer type	Genomic primer 1/2
Lane 2 primer lot	N/A
Lane 2 sample	CT-2 MEF-1
Lane 3 primer type	Genomic primer 1/2
Lane 3 primer lot	N/A
Lane 3 sample	CT-2 CM-2
Lane 4 primer type	Genomic primer 1/2
Lane 4 primer lot	N/A
Lane 4 sample	Sexual Nonirradiated mRNA PE
Lane 5 primer type	Genomic primer 1/2
Lane 5 primer lot	N/A

U. Connecticut - Track work by User

user page | discussion | edit with form | edit | history | delete | move | protect | watch | refresh

User:Misha

Email Address ⚠

Laboratory Graveley

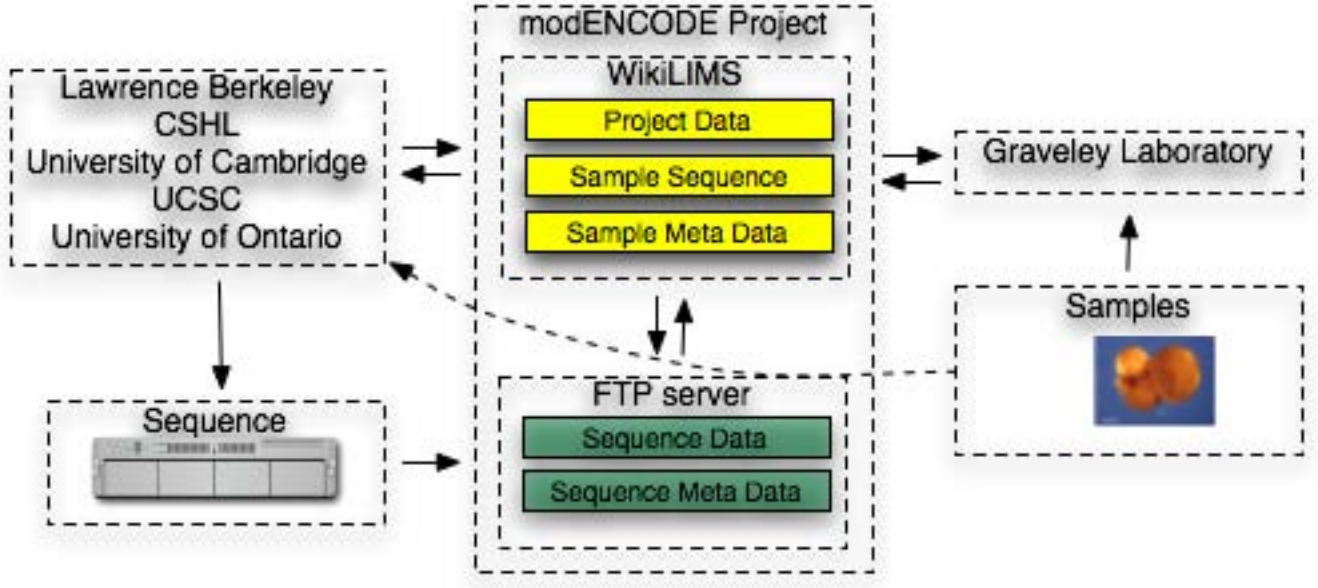
Samples

Date submitted	Species	Sample Has Laboratory	Sample Has Project
CT-2 CM-1	1 December 2008	Homo sapiens	Graveley	Stem Cell
CT-2 CM-2	1 February 2009	Homo sapiens	Graveley	Stem Cell
CT-2 MEF-1	1 February 2009	Homo sapiens	Graveley	Stem Cell
CT-2 MEF-2	1 February 2009	Homo sapiens	Graveley	Stem Cell
CT-2 TeSR-1	1 February 2009	Homo sapiens	Graveley	Stem Cell
CT-2 TeSR-2	1 February 2009	Homo sapiens	Graveley	Stem Cell
Grabel mES	1 February 2009	Mus musculus	Graveley	Stem Cell
Grabel mES Sox-1-GFP	1 February 2009	Mus musculus	Graveley	Stem Cell
H9 CM-1	1 January 2008	Homo sapiens	Graveley	Stem Cell
H9 CM-2	1 January 2008	Homo sapiens	Graveley	Stem Cell
H9 ENPd10	1 February 2009	Homo sapiens	Graveley	Stem Cell
H9 JL-1	1 February 2009	Homo sapiens	Graveley	Stem Cell
H9 LNPd17RA	1 February 2009	Homo sapiens	Graveley	Stem Cell
H9 LNPd17c	1 February 2009	Homo sapiens	Graveley	Stem Cell
H9 MEF miRNA - 1	1 December 2008	Homo sapiens	Graveley	Stem Cell
H9 MEF miRNA - 2	1 December 2008	Homo sapiens	Graveley	Stem Cell
H9 MEF-1	1 January 2008	Homo sapiens	Graveley	Stem Cell
H9 MEF-2	1 January 2008	Homo sapiens	Graveley	Stem Cell
H9 MEF-CM miRNA - 1	1 December 2008	Homo sapiens	Graveley	Stem Cell
H9 MNP	1 February 2009	Homo sapiens	Graveley	Stem Cell
H9 TeSR miRNA - 1	1 December 2008	Homo sapiens	Graveley	Stem Cell
H9 TeSR miRNA - 2	1 December 2008	Homo sapiens	Graveley	Stem Cell
H9 TeSR-1	1 January 2008	Homo sapiens	Graveley	Stem Cell
H9 TeSR-2	1 January 2008	Homo sapiens	Graveley	Stem Cell

Flowcells

	Amp date
30CYGAAXX	27 October 2008
30GGJAAXX	11 November 2008
30M66AAXX	9 December 2008
30VGEAAXX	12 March 2009
31003AAXX	12 March 2009

U. Connecticut - modENCODE uses WikiLIMS as project hub



U. Connecticut - Projects involve internal and external Labs

page | discussion | edit with form | edit | history | delete | move | protect | watch | refresh

ModENCODE

Samples for the **ModENCODE** project:

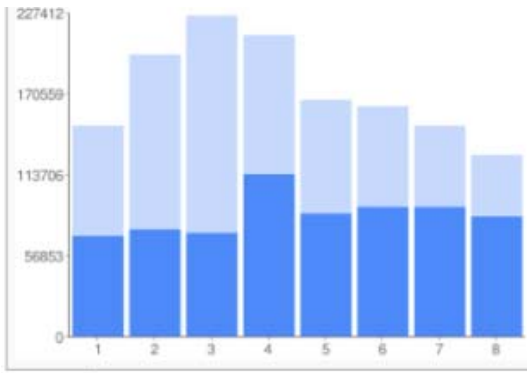
	Sample Has Laboratory	Sample has reagent	Sample Has User
CME W1 Cl.8+-60	Graveley		Liyang
CME W1 Cl.8+-62	Graveley		Liyang
D mel/D sec Hybrid - 450 bp	Graveley		Mcmanus
D. mel adult	Graveley		Mcmanus
D.mel/D.sec mix - 250 bp	Graveley		Mcmanus
D.mel/D.sec mix - 450 bp	Graveley		Mcmanus
Kc167	Graveley		Liyang
Kc167-2	Graveley		Liyang
Kc167-4	Graveley		Liyang
ML-DmBG3-c2-122	Graveley		Liyang
ML-DmBG3-c2-124	Graveley		Liyang
S2-DRSC Brr2 RNAi rRNA minus-1	Graveley		Liyang

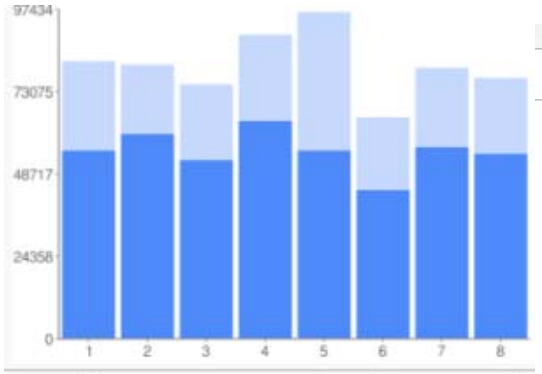
ModENCODE	
Project Laboratory	Graveley, Celniker

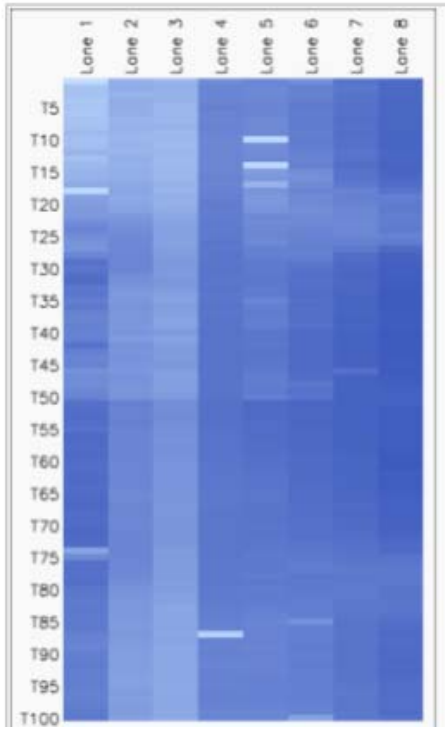
Cornell University

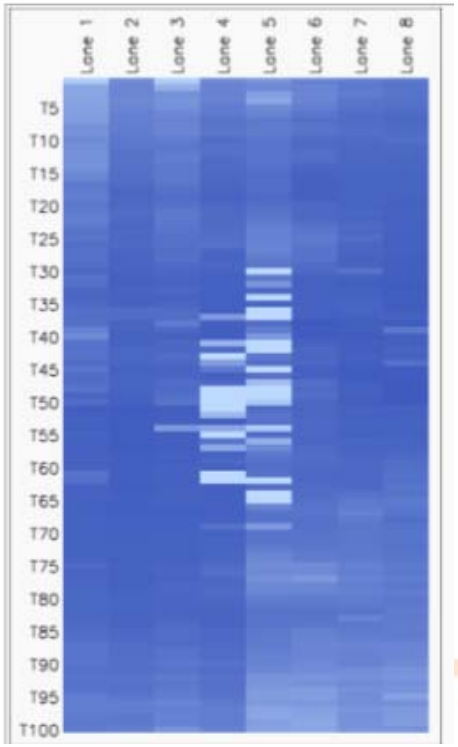
- Situation
  - 1 Roche GS and 1 Illumina GA 2
  - Need to monitor Run quality
  - Need to read customer and sample data from existing LIMS
  - Need to link to existing LIMS
  - Key pages: Samples, Customers, Illumina Runs, Roche Runs, Flowcells

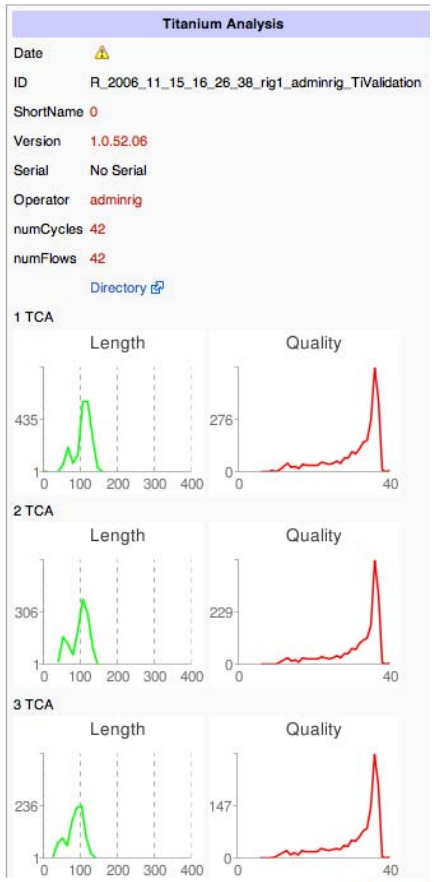
Cornell - Monitoring quality











Cold Spring Harbor Laboratory

- Situation
  - 13 Illumina GA sequencers
  - Need to run large number of instruments used by many technicians
  - Need secure environment for clinical samples
  - Key Pages: Illumina Runs, Flowcells, Libraries, PCR Reactions, Genome Amplifications, Machines, Purifications

CSHL - Create and edit Library pages

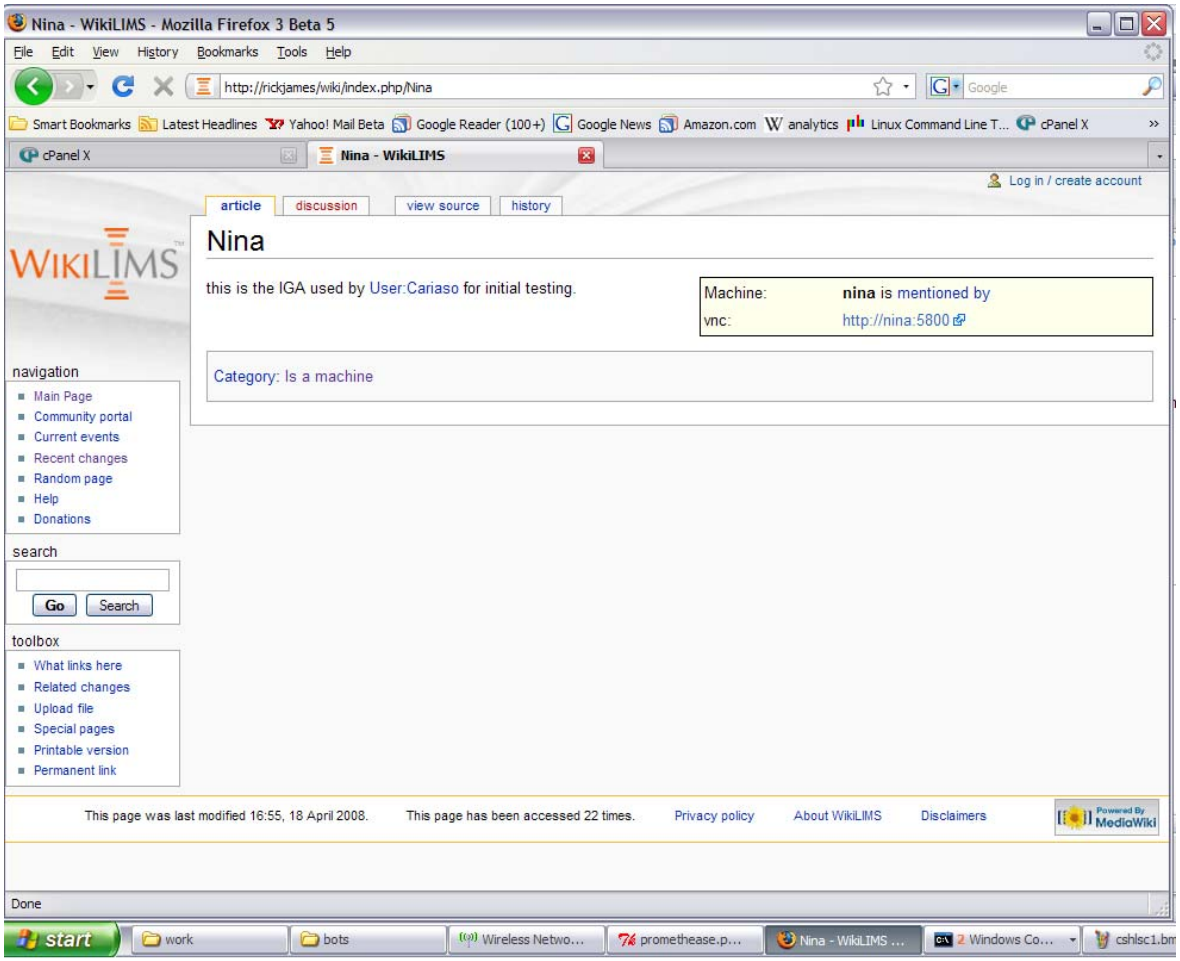
page | discussion | edit with form | edit | history | move | watch

Edit Library: LID2301

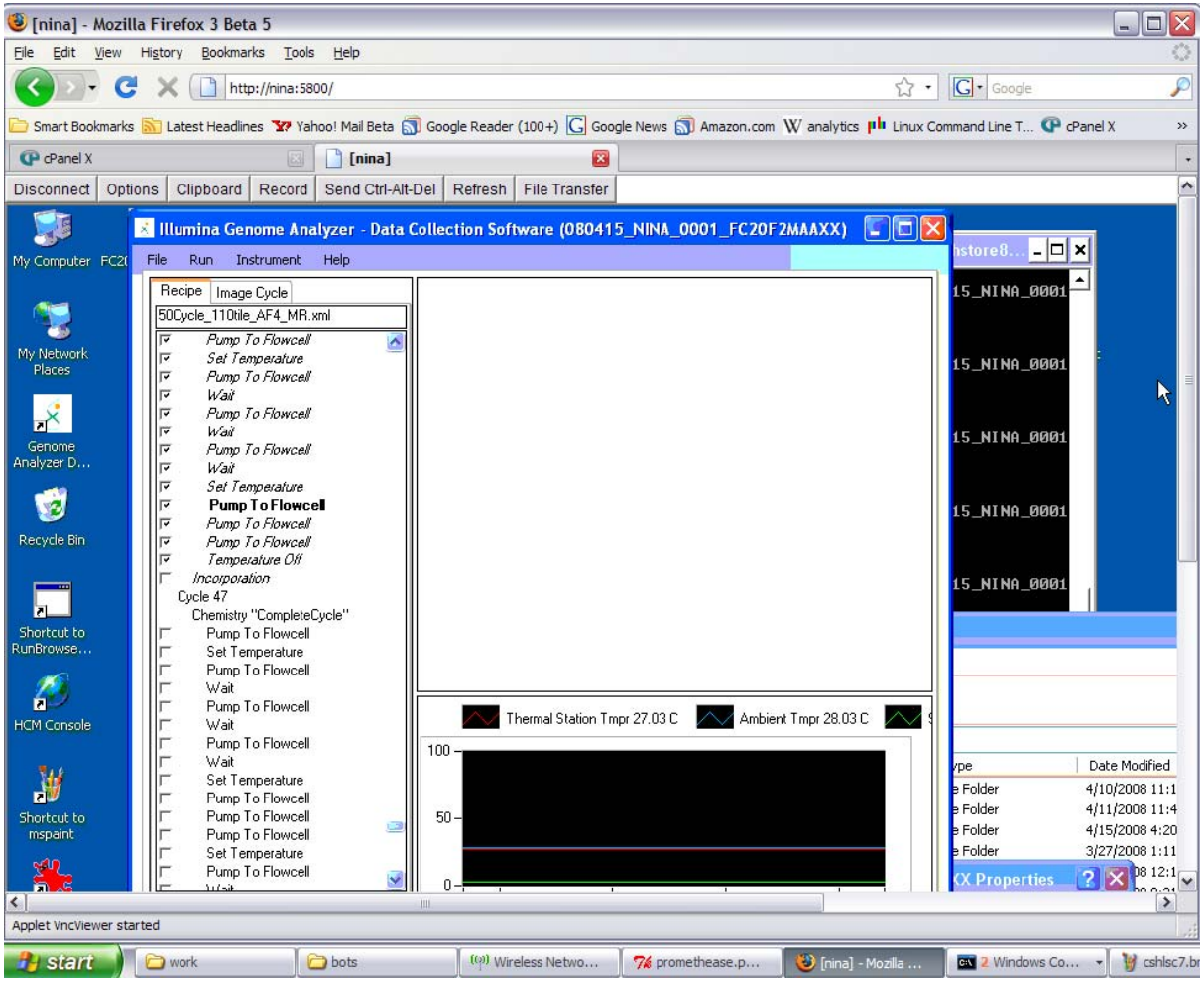
Library

Date:	March 16 2009 3 : 26 : 28 PM
Sample_id:	SID1621, SID1622, SID1623, SID1624, SID1625, SID1626, SID1627, SID1628, SID1629, SID1630, SID1631, SID1632
Dna_input:	E05; LBC360083/Index1, F05; LBC360049/Index2, G05; LBC360127/Index3, H05; LBC360002/Index4, A06; LBC360009/Index5, I
Contact:	Cardone
Constructor:	Mavruk/Cardone
Originator:	Ian Deary
Type:	Custom
Post-enrichment concentration:	
Working dilution:	10 nM
pM to load	8.0pM
Shearing Method:	Covaris
Pressure	
Duration	90 seconds
Adaptor:	Illumina Index 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
Quantification method:	Nanodrop
Size:	300bp
Primer:	

CSHL - Remote Instrument Operation



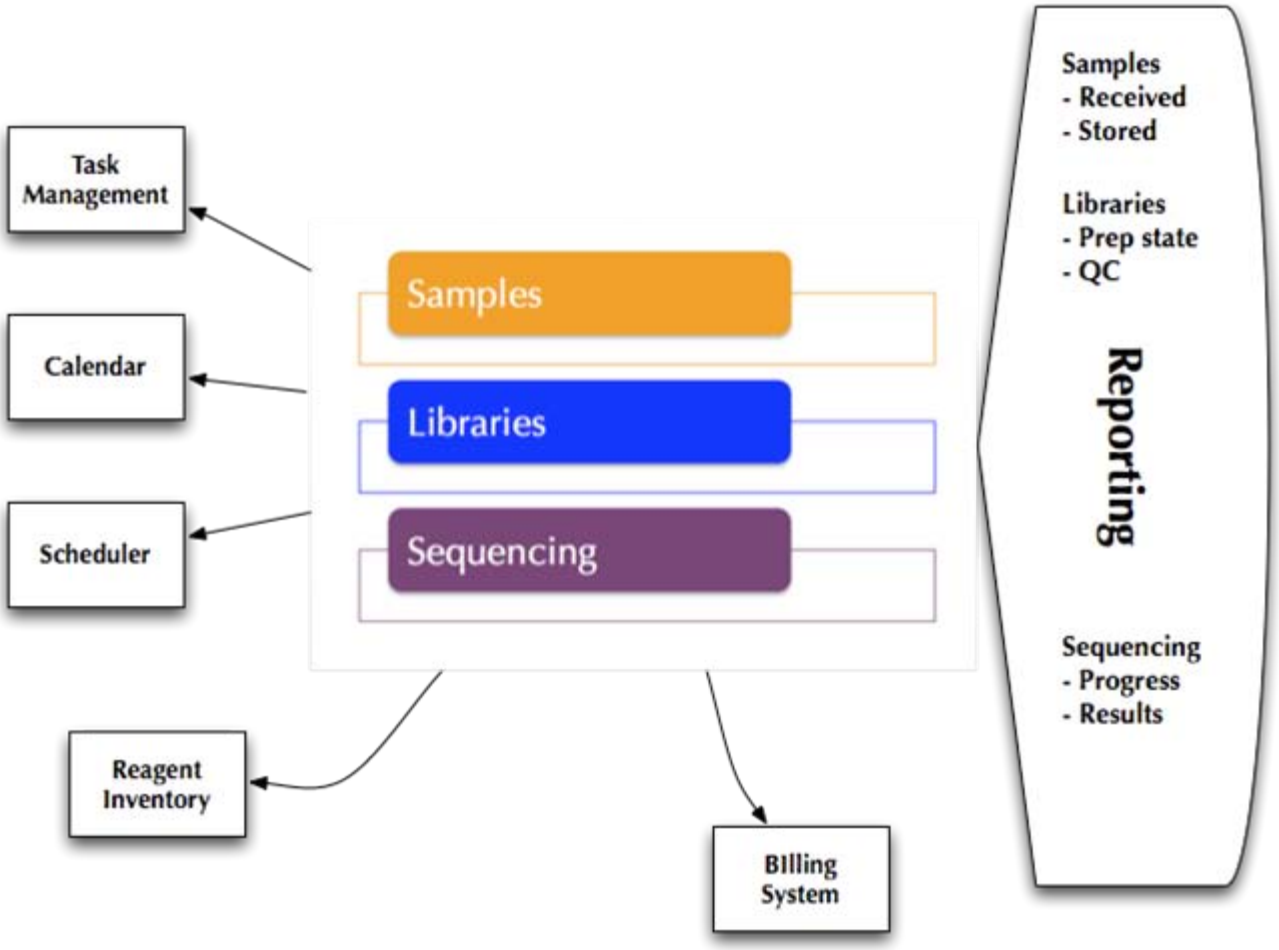
CSHL - Remote Instrument Operation



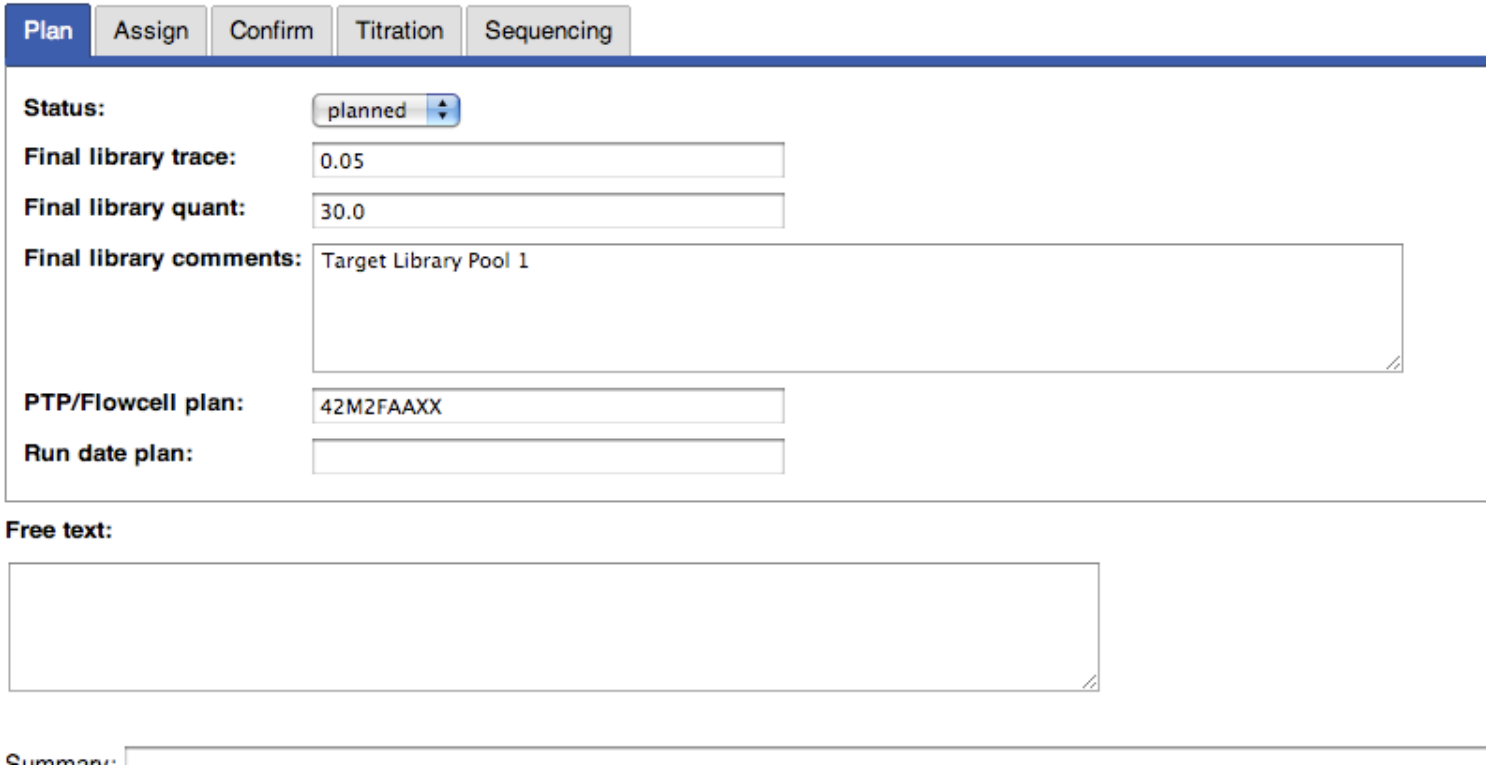
Indiana University Center for Genomics and Bioinformatics

- Situation
  - 1 Roche GS and 1 Illumina GA, NimbleGen microarrays
  - Need to track runs, samples, reagents, and group by project
  - Need to track task-level and job-level provenance data
  - Need to send notifications and email alerts
  - Need to carry projects through all the way to billing
  - Key : Projects, Samples, Libraries, Sequencing, Reagents

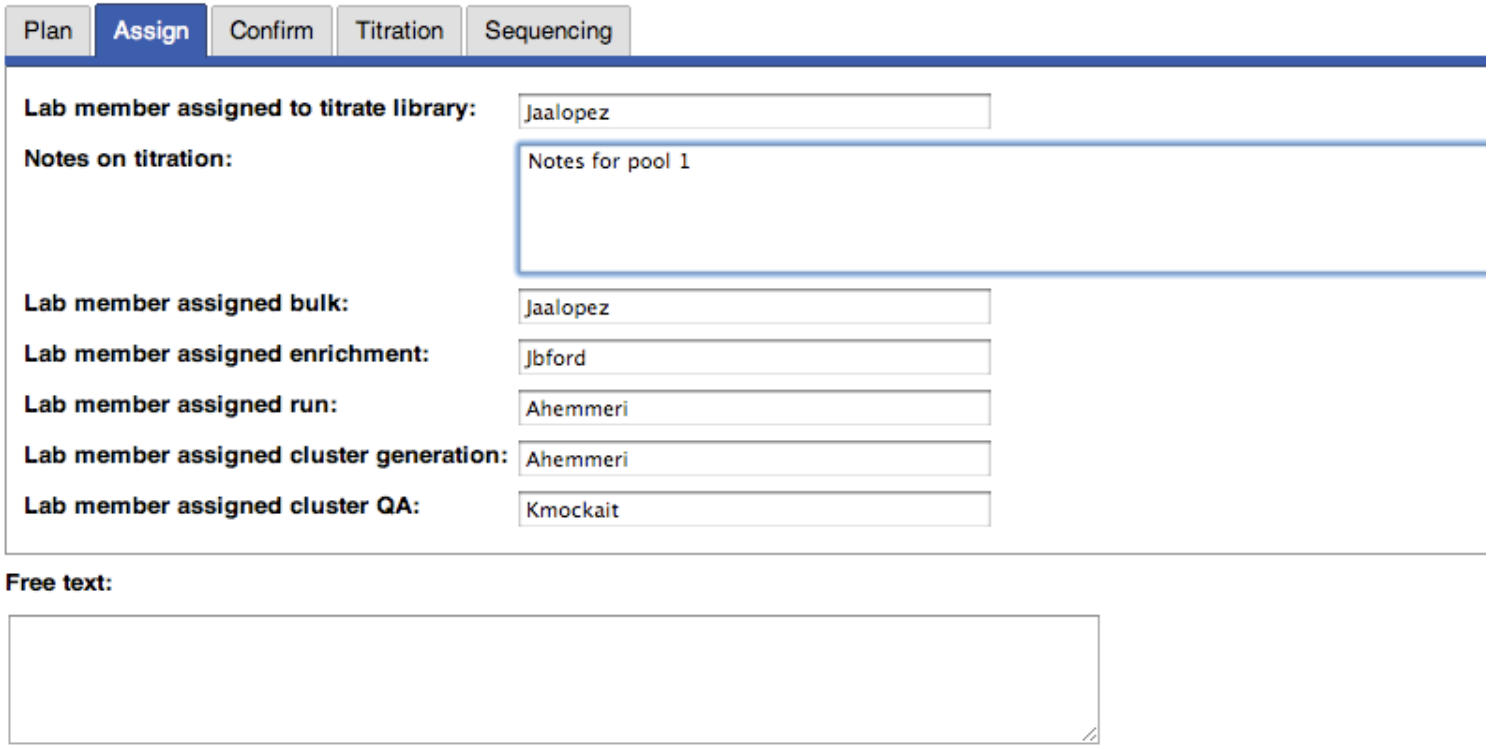
Indiana - Samples, Tasks, and Reporting



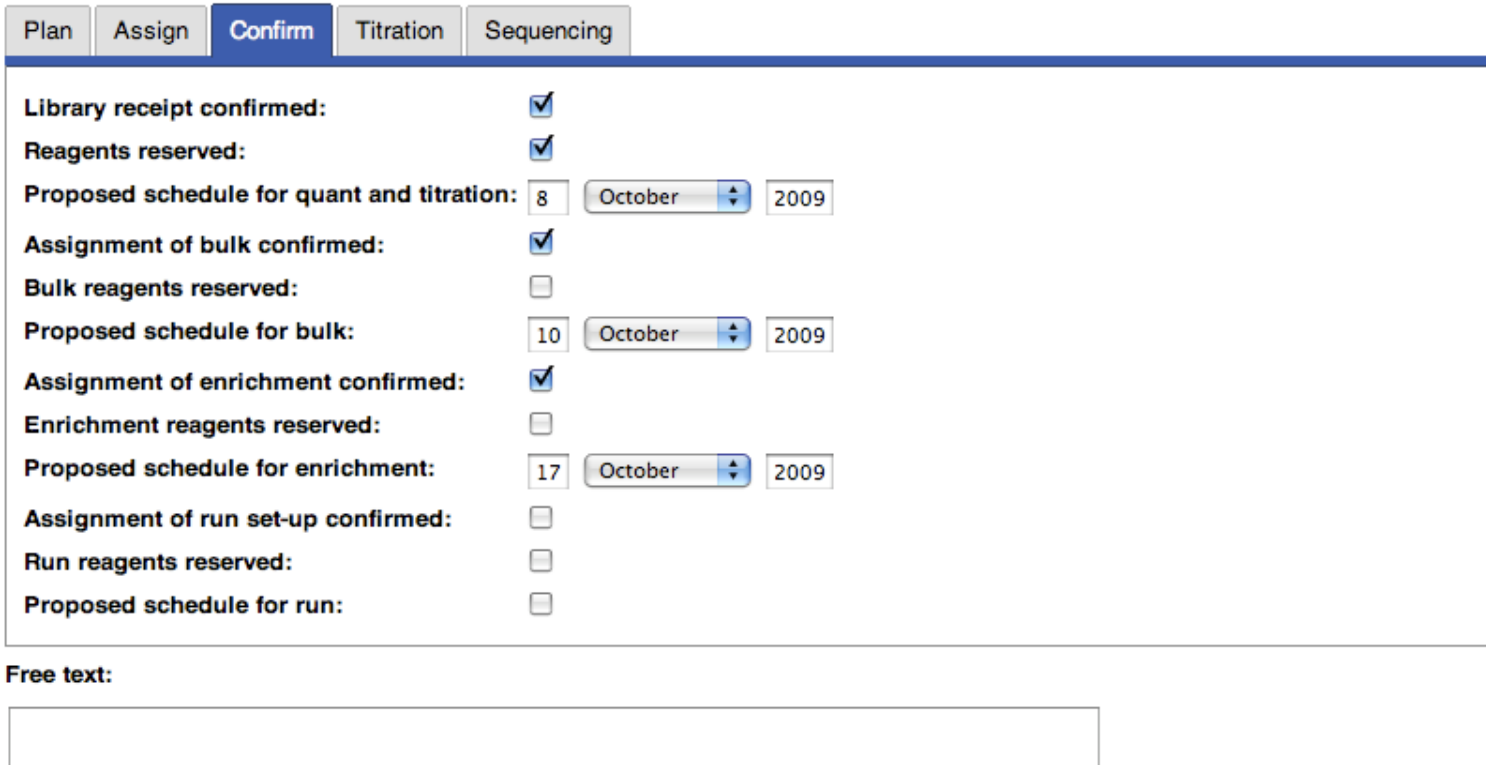
Indiana - Managing workflow tasks



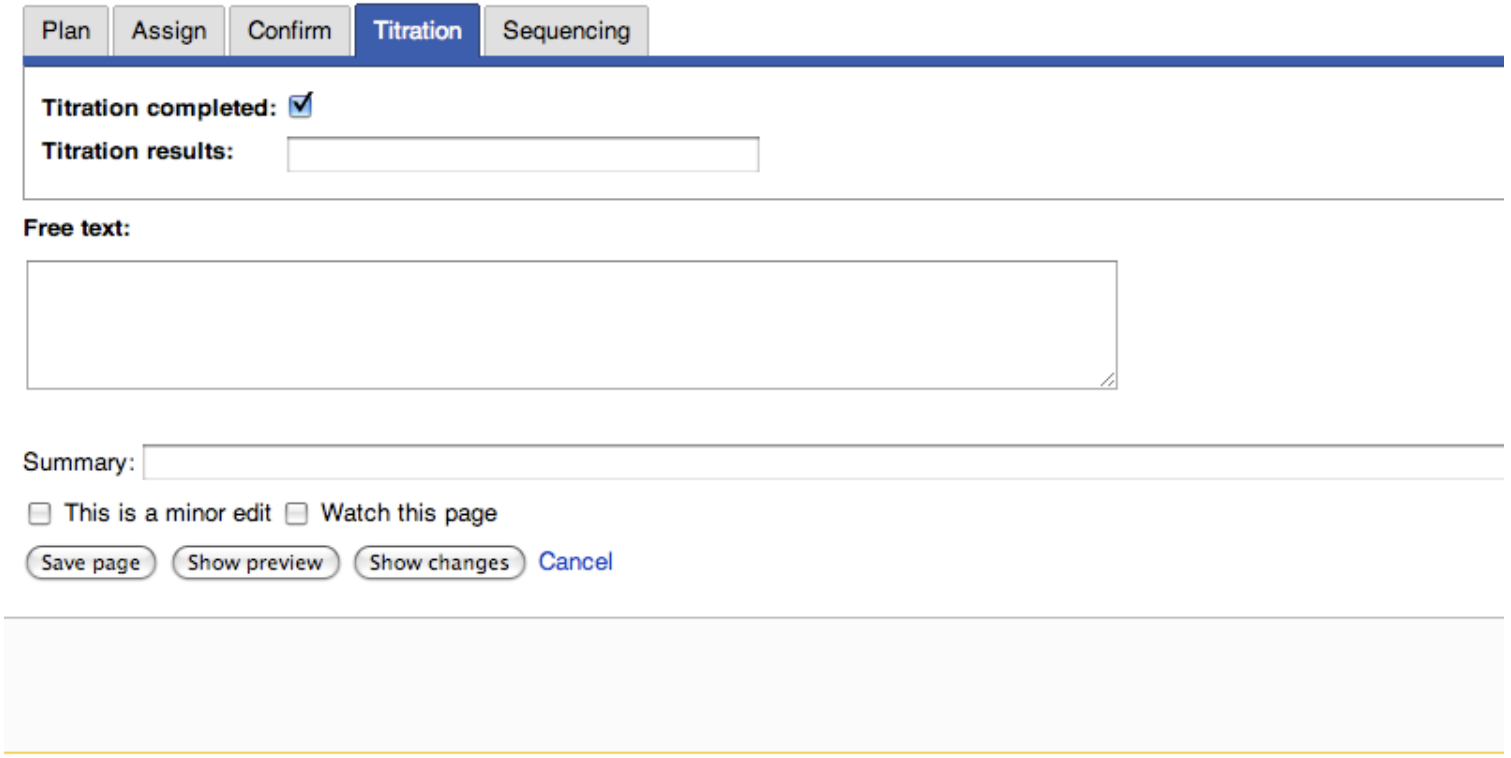
Indiana - Managing workflow tasks



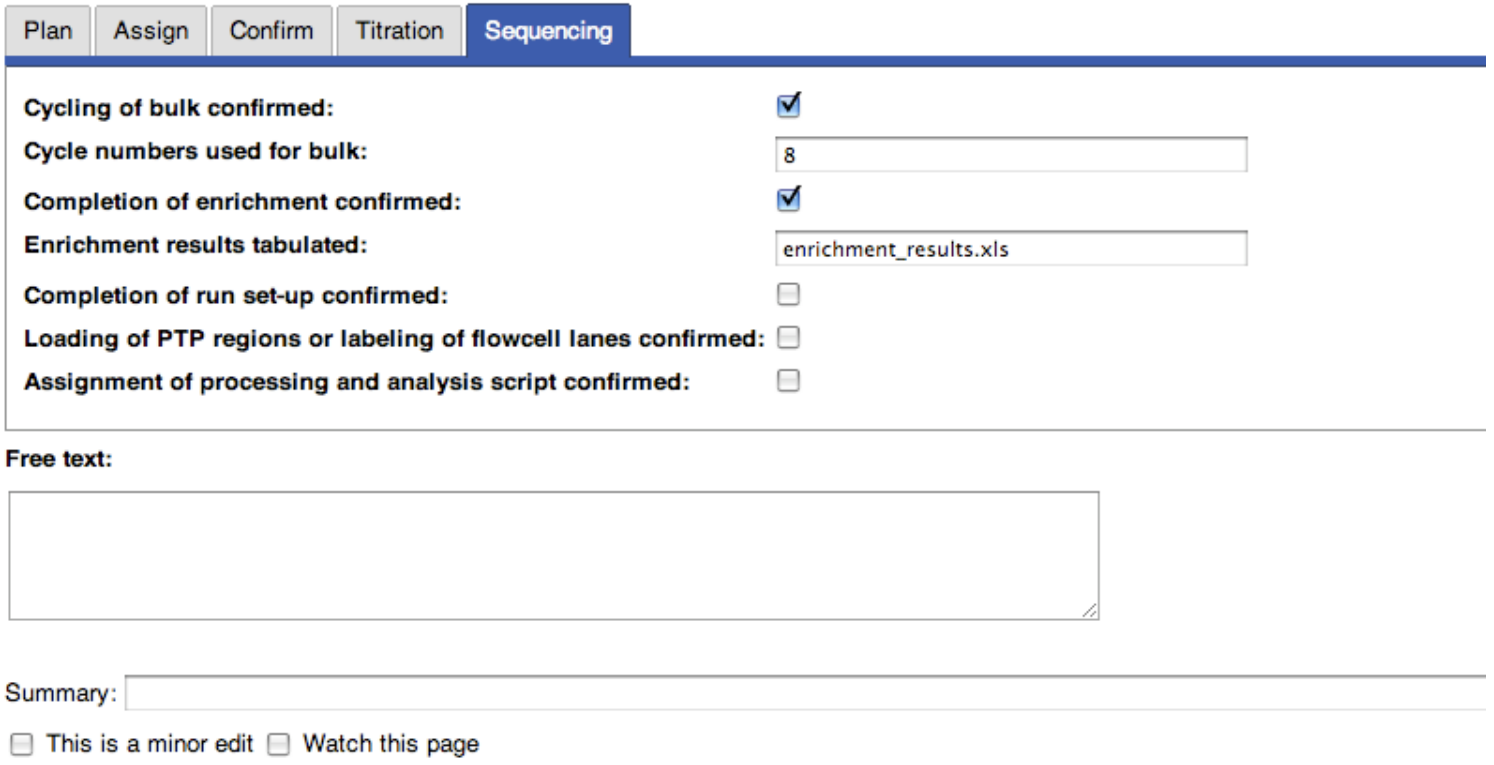
Indiana - Managing workflow tasks



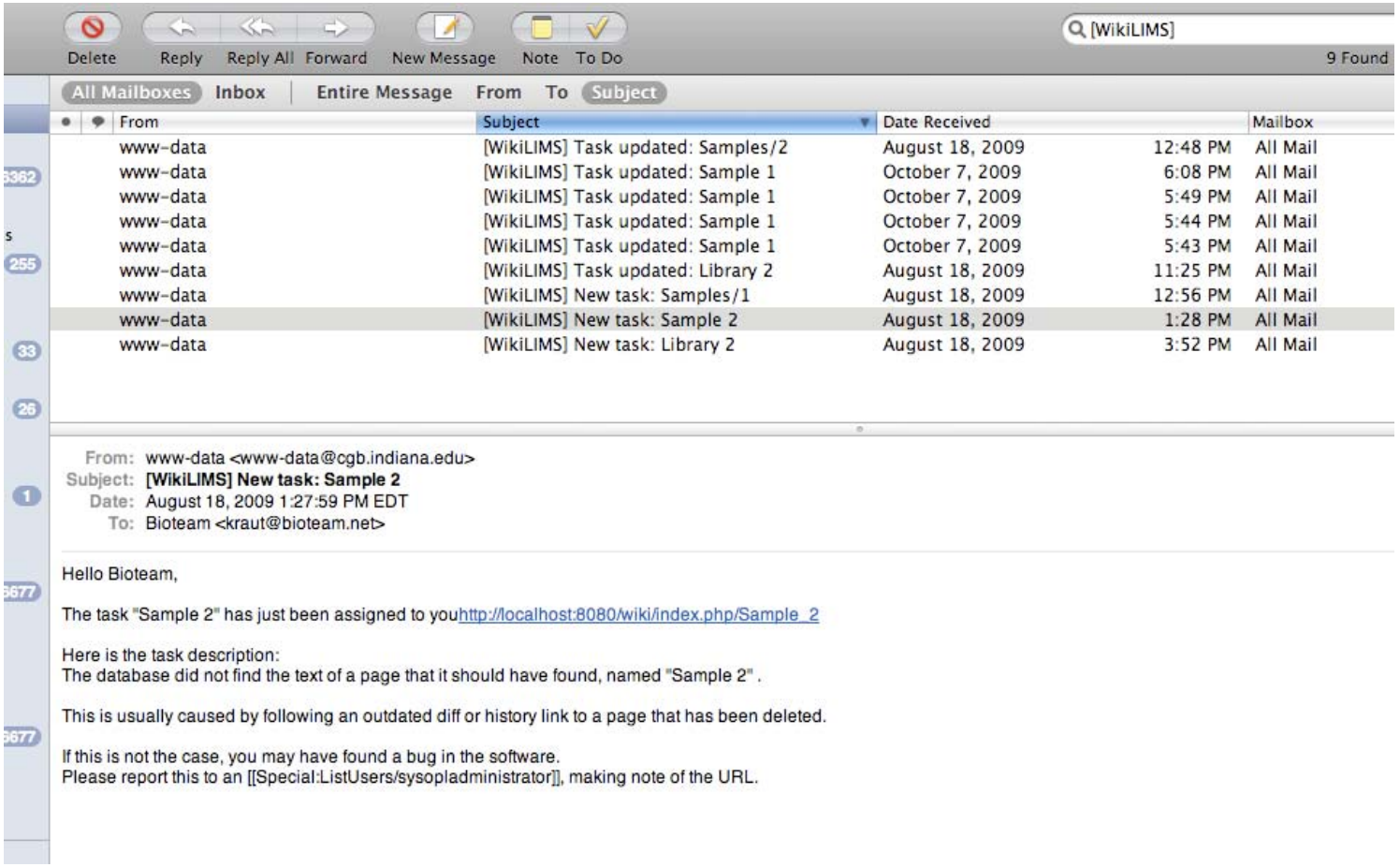
Indiana - Managing workflow tasks



Indiana - Managing workflow tasks



Indiana - Sending e-mail notifications



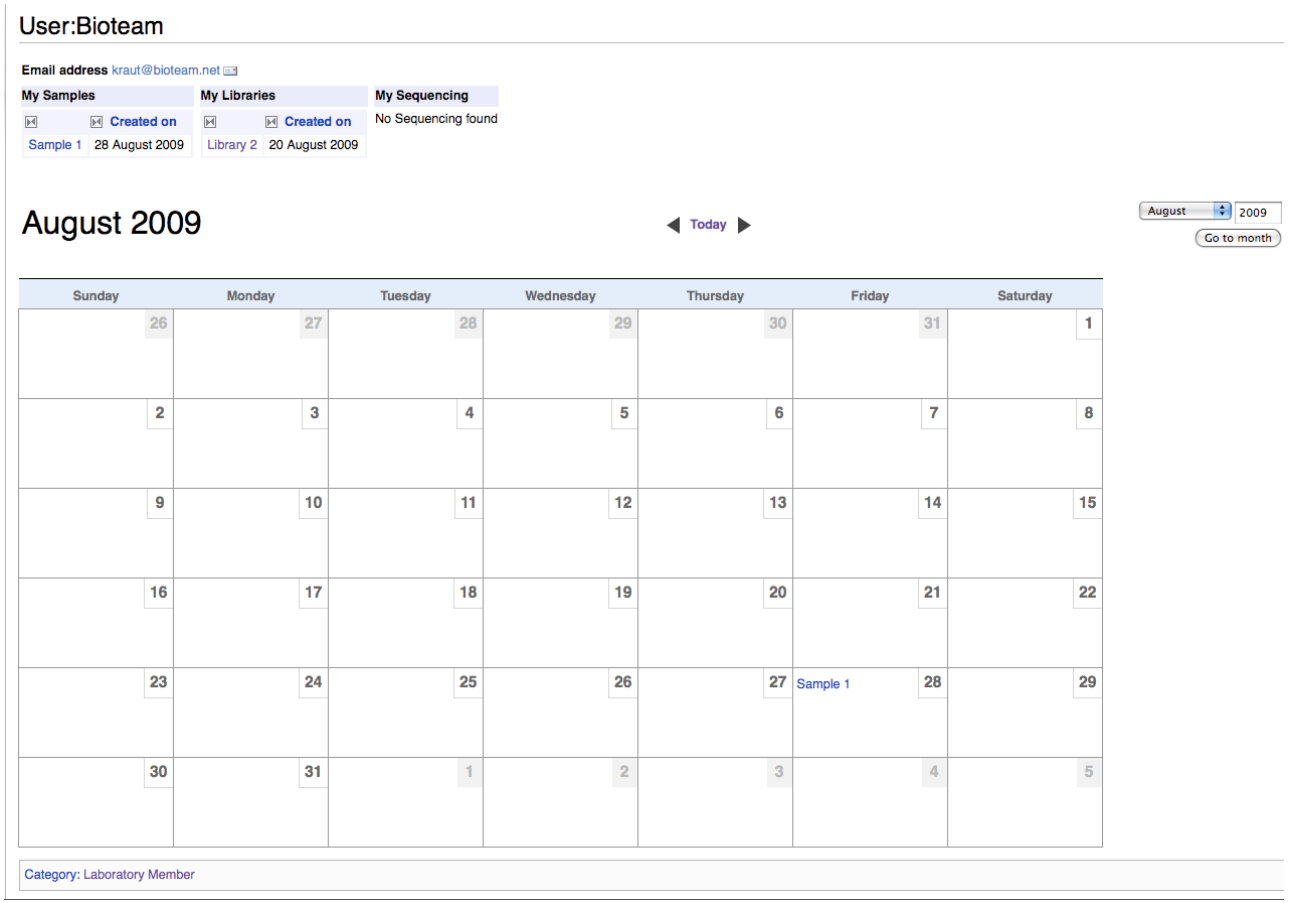
Indiana - Simplified tracking information

← Older edit

Line 5:	Line 5:
|Library type=pool	|Library type=pool
|Lab member assigned to prepare library=Bioteam	|Lab member assigned to prepare library=Bioteam
	+ |Sample receipt confirmed=Yes
	+ |Sample QA confirmed=No
}}	}}

- Email and RSS notifications for every step in workflow
- Wiki Revision Control explains ***who*** did ***what*** and ***when***
- Lab Managers can revert and undo tasks

Indiana - Managing tasks by Lab User – Calendar view



AMDeC

AMDeC Core Facilities | Core Services | Core Instruments

	City	Category	Institution
Albert Einstein College of Medicine Histotechnology and Comparative Pathology	Bronx	Pathology	Albert Einstein College of Medicine
Columbia University Genomics Shared Resources	New York City	Genomics	Columbia University
Mount Sinai School of Medicine Mouse Genetics Shared Resource Facility	New York City	Animal	Mount Sinai School of Medicine
Stony Brook University Genomics Core	Stony Brook	Microscopy	Stony Brook University
Weill Cornell Medical College Crystallization and X-Ray Diffraction Core Facility	New York City	Crystallography Proteomics	Weill Cornell Medical College

DFCI - Managing a large-scale, multi-year clinical collaboration

- BU, NJ Health Center, Dana Farber, Clinical Trials and Surveys Corp., U. Michigan Health Center, UCD, UPMC, National Heart Lung and Blood Institute
- 1000 human lung samples in a study of COPD
- 75 assays
- Monthly reporting
- Automated data loading

DFCI - Detailed Sample Tracking over

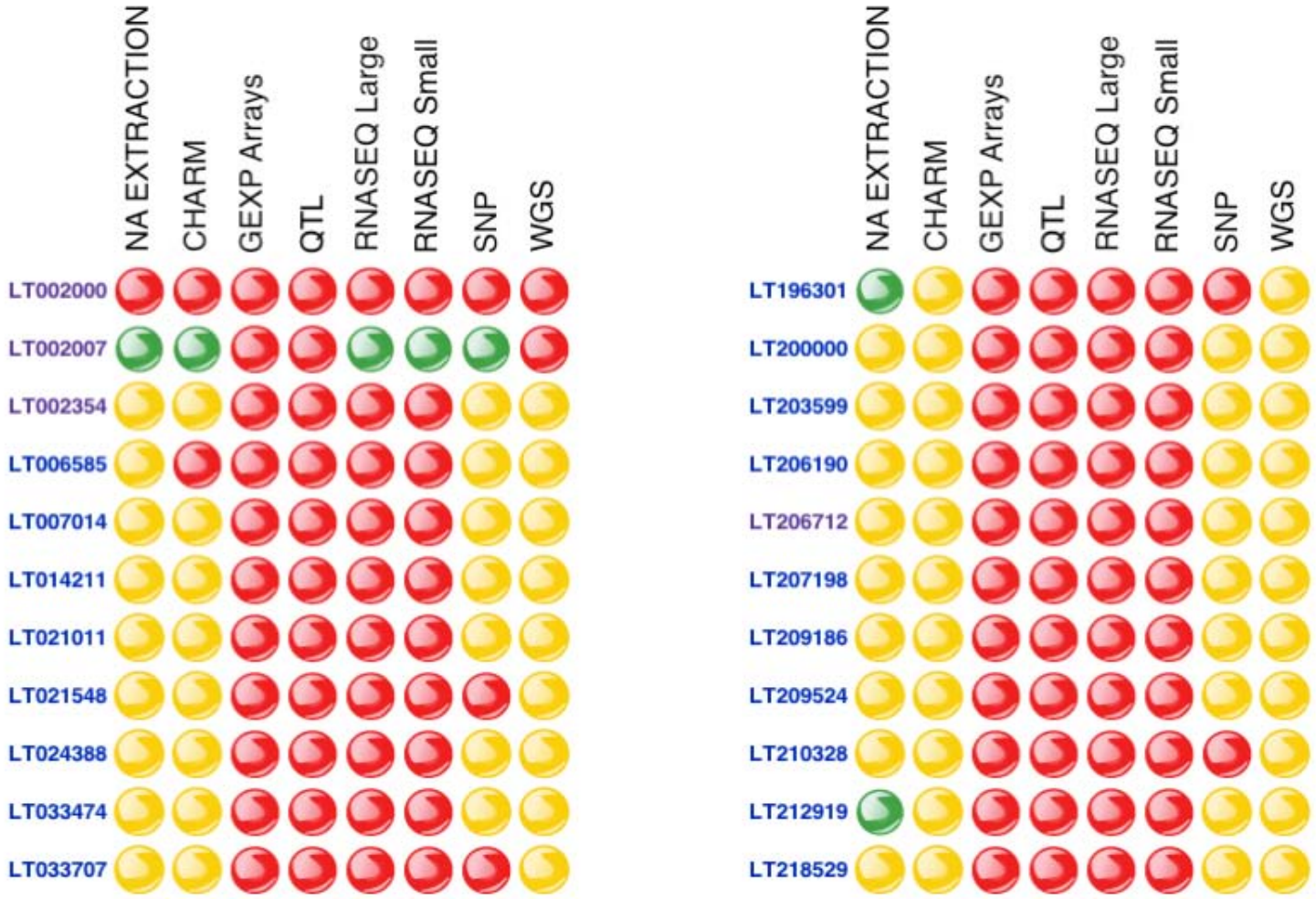
LT002007

CHARM | GEXP_Arrays | NA_EXTRACTION | QTL | RNASEQ_Large | RNASEQ_Small | SNP | WGS

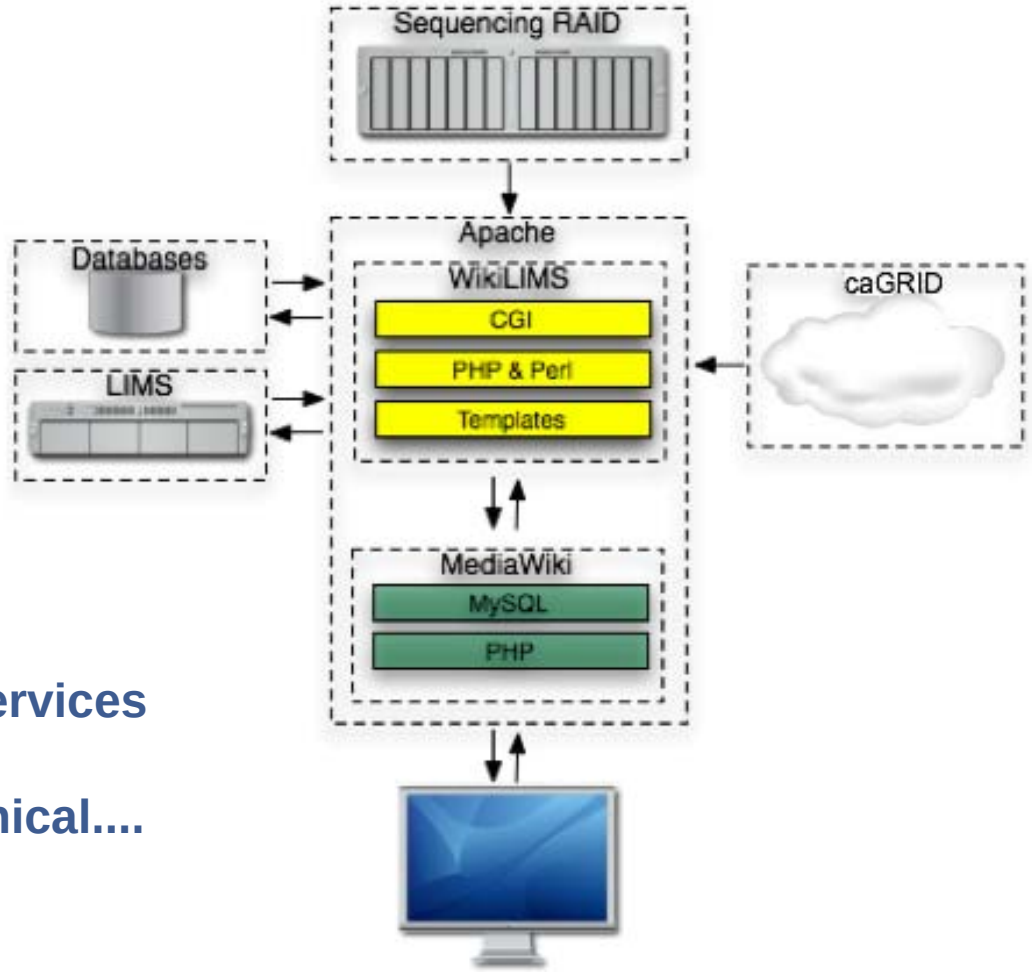
	Start	Completion	Assay	Order
LT002007 DNA received from Pittsburgh	12 March 2009 00:00:00	22 March 2009 00:00:00	CHARM	1
LT002007 Hybridization performed	2 May 2009 00:00:00		CHARM	2
LT002007 QC and normalization performed (charmR scripts)	15 October 2009 00:00:00	24 October 2009 00:00:00	CHARM	3
LT002007 CHARM data sent to DFCI			CHARM	4
LT002007 CHARM data received by DFCI	8 September 2009 00:00:00		CHARM	5
LT002007 CHARM data processed and available	15 October 2009 00:00:00	24 October 2009 00:00:00	CHARM	6

Category: Sample

DFCI - Global Sample Tracking over Assays

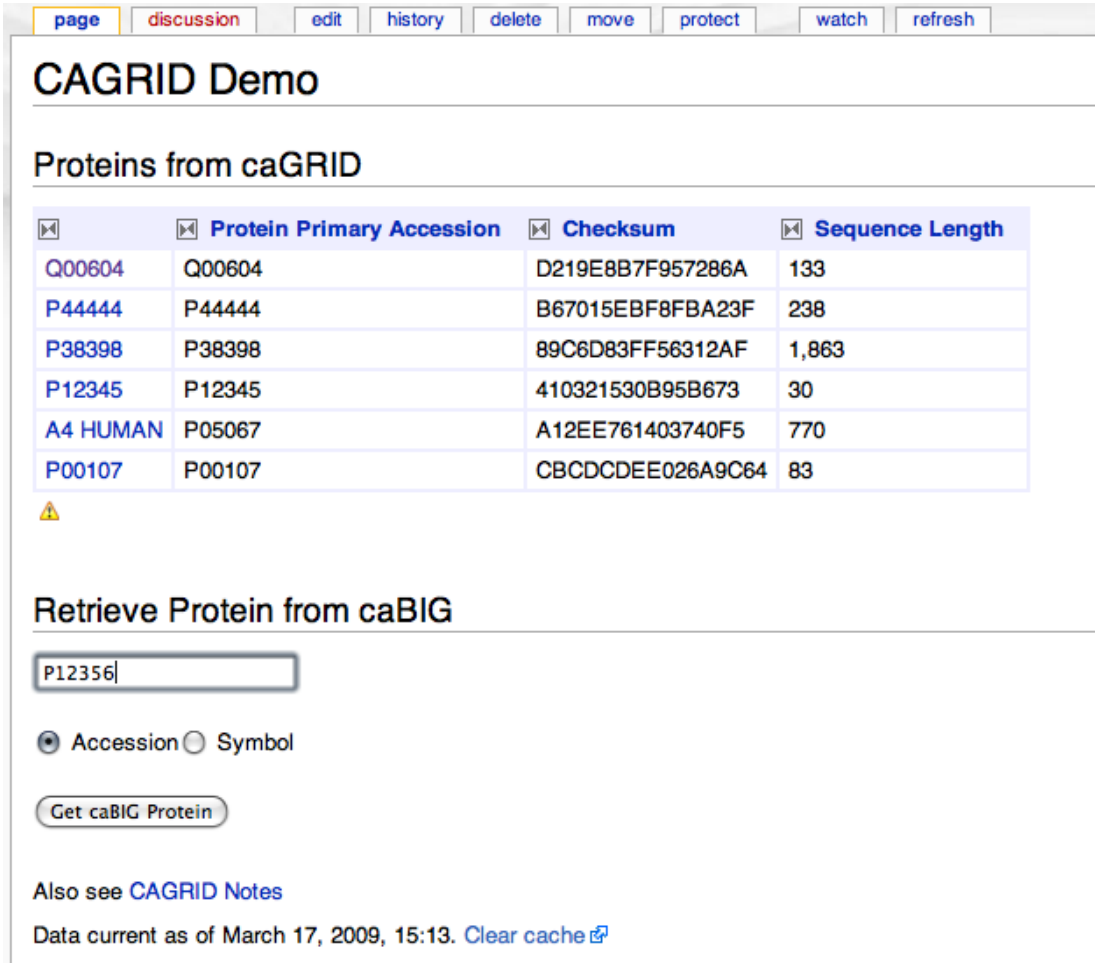


caBIG Integration



- WikiLIMS queries caBIG *via* Web Services (SOAP)
- Gene, protein, microarray, SNP, clinical....

caBIG Integration



caBIG Integration

BRCA1

```
TBLASTN 2.2.10 [Oct-19-2004]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= BRCA1_HUMAN
         (1863 letters)

Database: NTs.fa
           9 sequences; 567 total letters

Searching.........done

                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

77209                                                                  49   1e-10
1750                                                                   49   1e-10
9911                                                                   46   7e-10
9082                                                                   46   9e-10
305                                                                    45   2e-09

>77209
          Length = 65

 Score = 48.9 bits (115), Expect = 1e-10
 Identities = 21/21 (100%), Positives = 21/21 (100%)
 Frame = +1

Query: 371 PWITLNSSIQKVNEWFSRSDE 391
           PWITLNSSIQKVNEWFSRSDE
Sbjct: 1   PWITLNSSIQKVNEWFSRSDE 63
```

Run TBLASTN

Symbol	**BRCA1**
Accession	P38398
Protein	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQCPLCKNDITKRSLQESTRFSQLVEELLKI
Species	Homo sapiens

Facts about BRCA1

CAGRID Gene Protein	MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHI ... HAIGQMCEAPVVTREWVLDSVA
CAGRID Gene Symbol	**BRCA1**

What They Are

- Inflexible
  - Unable to adapt to changing research processes & organisations
  - Can never capture everything
  - Difficult to administer
- Culture Defiant
  - I'm not ready to share
  - What benefit do I get
  - It's not my solution
- Traditional
  - Software Development cycles are too long – even agile ones
  - Too expensive

What They Need To Be

Pfizer

- Useful — return more benefit than cost of construction
- Intuitive — low adoption barrier
- Fast — rapid prototyping
- Flexible — modular, able to retool to new uses
- Talkative — communicates easily with other technologies
- Cheap — design cycle not limited by cost of development
- Open — open source, transparent & easily modified
- Adopted — Community of developers expanding functionality

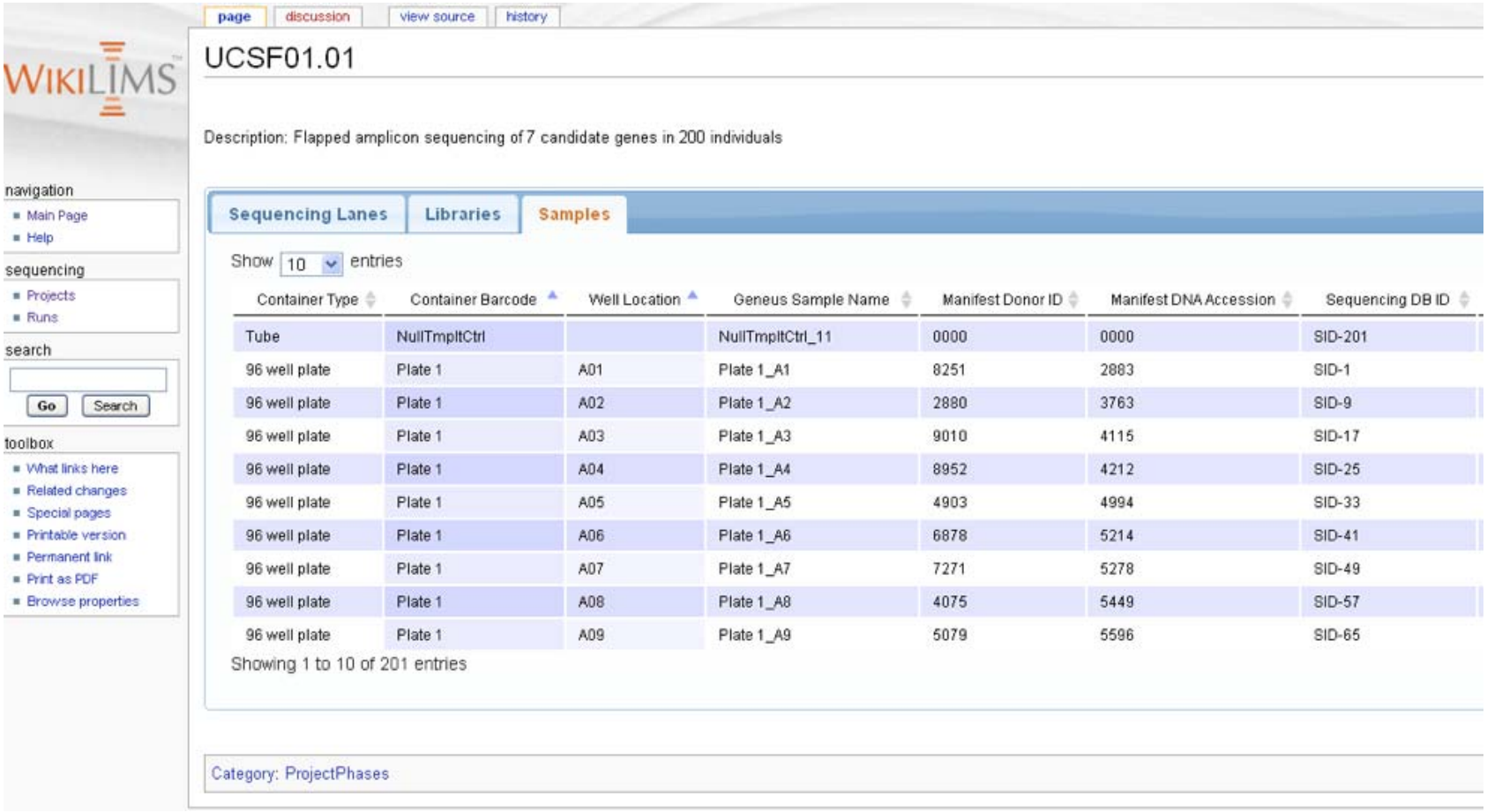
Informatics must co-evolve with scientific innovation

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.

QuickTime™ and a
decompressor
are needed to see this picture.



page | discussion | view source | history

WIKILIMS

UCSF01.01

Description: Flapped amplicon sequencing of 7 candidate genes in 200 individuals

navigation
- Main Page
- Help

sequencing
- Projects
- Runs

search
Go Search

toolbox
- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Print as PDF
- Browse properties

Sequencing Lanes | Libraries | Samples

Show 10 entries

Container Type	Container Barcode	Well Location	Geneus Sample Name	Manifest Donor ID	Manifest DNA Accession	Sequencing DB ID
Tube	NullTmpltCtrl		NullTmpltCtrl_11	0000	0000	SID-201
96 well plate	Plate 1	A01	Plate 1_A1	8251	2883	SID-1
96 well plate	Plate 1	A02	Plate 1_A2	2880	3763	SID-9
96 well plate	Plate 1	A03	Plate 1_A3	9010	4115	SID-17
96 well plate	Plate 1	A04	Plate 1_A4	8952	4212	SID-25
96 well plate	Plate 1	A05	Plate 1_A5	4903	4994	SID-33
96 well plate	Plate 1	A06	Plate 1_A6	6878	5214	SID-41
96 well plate	Plate 1	A07	Plate 1_A7	7271	5278	SID-49
96 well plate	Plate 1	A08	Plate 1_A8	4075	5449	SID-57
96 well plate	Plate 1	A09	Plate 1_A9	5079	5596	SID-65

Showing 1 to 10 of 201 entries

Category: ProjectPhases

QuickTime™ and a
decompressor
are needed to see this picture.

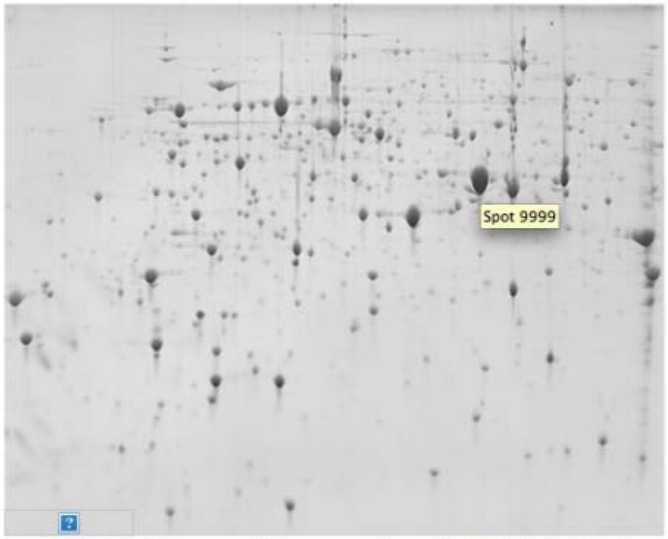
Future Directions: Proteomics

article | discussion | view source | history

GEL1233

This page is all about a 2D gel image. Try moving your mouse over some of the spots. Remember I cooked this one by hand, so not all spots are active. Everything scales, so feel free to edit this page and change, or remove, the size parameter.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Fusce pellentesque odio porta ipsum. Nunc nisl odio, vehicula a, feugiat sit amet, tincidunt sit amet, metus. Sed neque. Donec venenatis vestibulum purus. Duis auctor augue eget metus. Quisque sit amet erat. Suspendisse a urna. Nunc cursus magna tincidunt arcu. Morbi augue. Suspendisse accumsan odio eu risus. Phasellus fermentum, dui in consequat tempor, ligula magna rutrum ante, in rhoncus tortor metus eu dolor. Sed non dolor et purus vehicula viverra. Suspendisse pede ligula, laoreet et, volutpat eget, dapibus ut, urna.



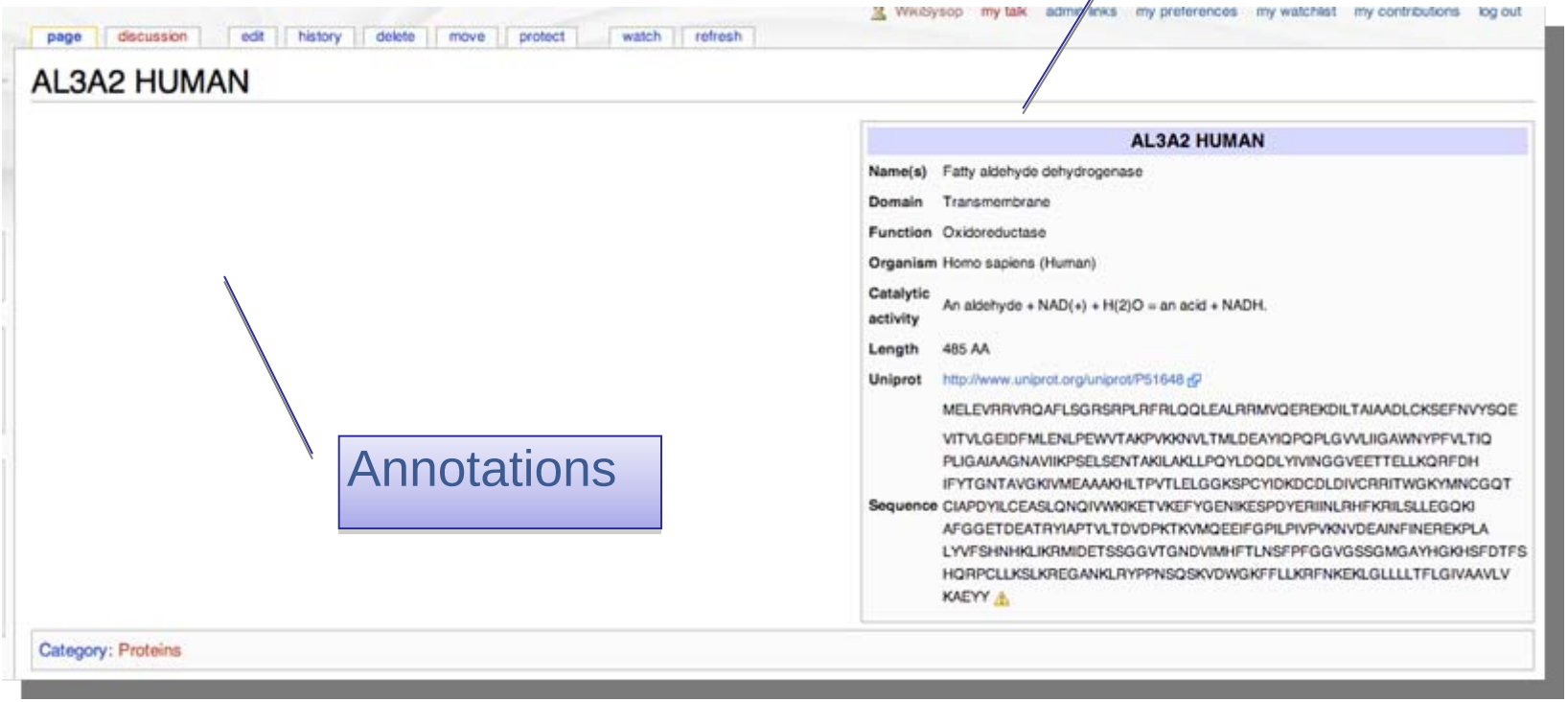
Spot 123456 is the biggest one. This represents the protein TCF7L2 in the Pseudomonas Carolinis. You may also be interested in Spot 987654 which was later determined to be a contamination error.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Fusce pellentesque odio porta ipsum. Nunc nisl odio, vehicula a, feugiat sit amet, tincidunt sit amet, metus. Sed neque. Donec venenatis vestibulum purus. Duis auctor augue eget metus. Quisque sit amet erat. Suspendisse a urna. Nunc cursus magna tincidunt arcu. Morbi augue. Suspendisse accumsan odio eu risus. Phasellus fermentum, dui in consequat tempor, ligula magna rutrum ante, in rhoncus tortor metus eu dolor. Sed non dolor et purus vehicula viverra. Suspendisse pede ligula, laoreet et, volutpat eget, dapibus ut, urna.

Future Directions: Proteomics

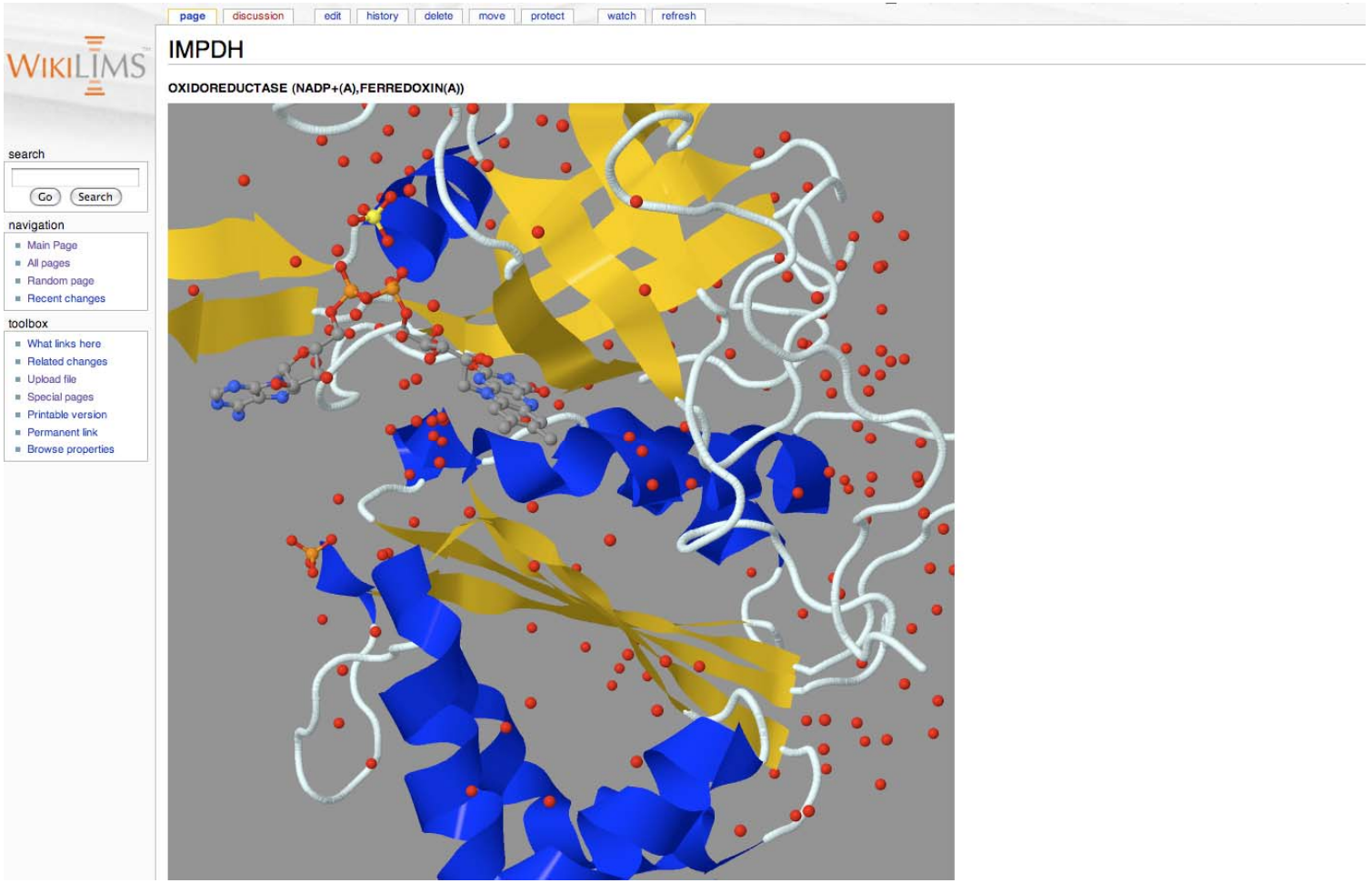
Future Directions: SOAP and RESTful Web Services

External Data



Annotations

Future Directions: 3D Structure Viewing (Jmol)



Future Directions: MALDI-TOF Data (R, Gnuplot)