NCBI Workshop

http://www.ncbi.nlm.nih.gov/education/pag2012/

- ***3:50*** *Kim Pruitt* - Primary Data Submission Portal
- ***4:10*** *Tatiana Tatusova* - BioProject, Genome, and Assembly databases
- ***4:30*** *Francoise Thibaud-Nissen* - Eukaryotic Genome Annotation Pipeline
- ***4:50*** *Deanna Church* - Connecting the Lab to the Genome: CloneDB
- ***5:10*** *Kim Pruitt* - Annual Report on Genome Sequencing Projects

Primary data submission Portal

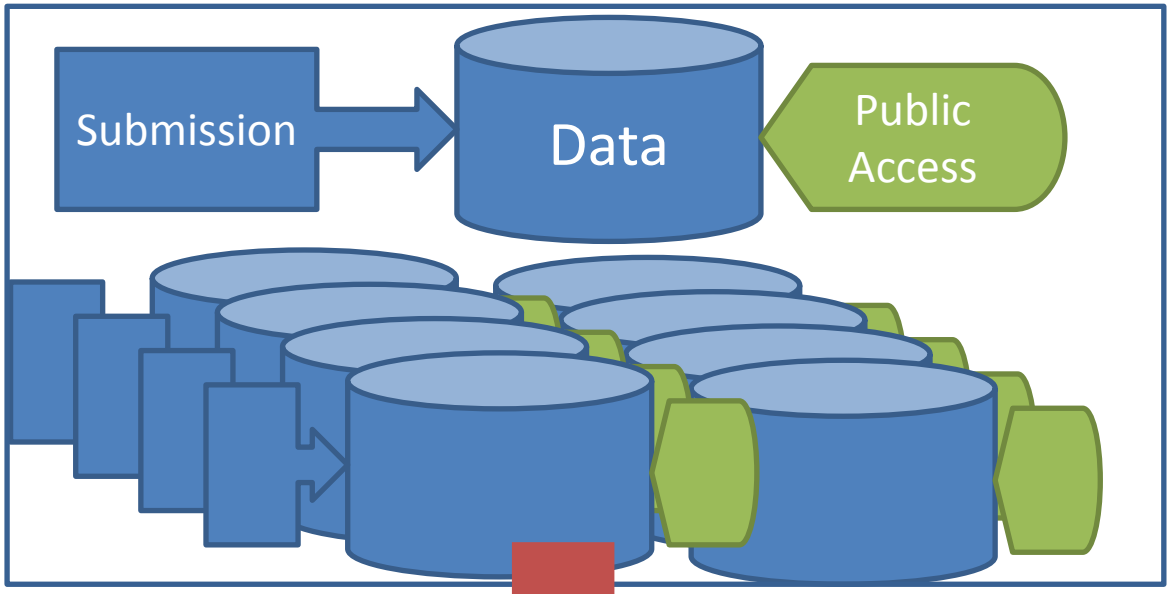
Kim D. Pruitt

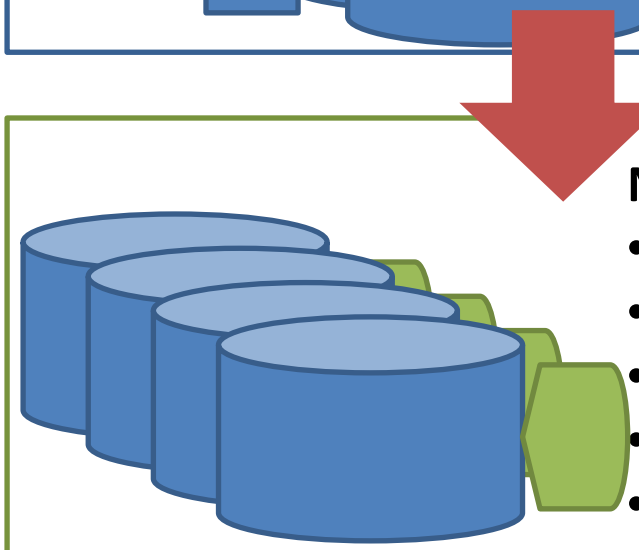
International Plant and Animal Genome XX

January 15-18, 2011

http://www.ncbi.nlm.nih.gov/

Building resources





NON-archival resources

- UniGene
- RefSeq
- Gene
- HomoloGene
- Genome
- Map Viewer

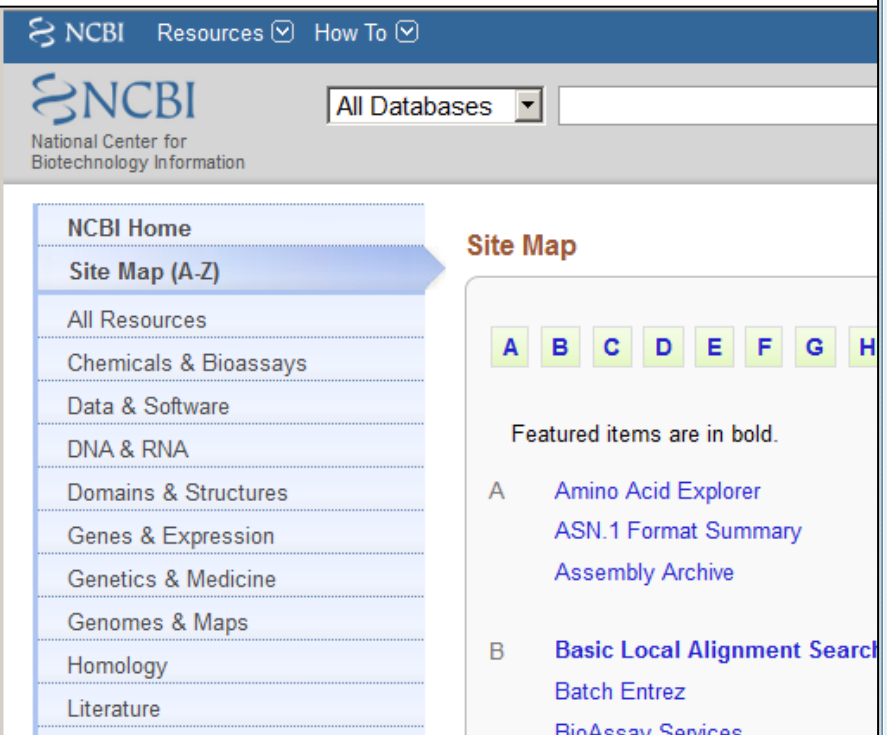
1988 Archival databases

- GenBank (INSDC)
- dbEST, GSS, PopSet
- dbSNP, dbVAR
- PubMed, PMC
- PubChem
- GEO
- dbGaP
- Probe
- SRA
- BioProject
- BioSample
- Assembly

2012

Getting organized - access

- Global query
- Site map
- Style/Format



Getting organized – submissions

NCBI Resources How To My NCBI Sign I

NCBI
National Center for Biotechnology Information

All Databases | Search

NCBI Home
Site Map (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

How to submit

Popular Resources

BLAST
Bookshelf
Gene
Protein
PubChem
PubMed
PubMed Central
SNP

Education Resources

Central point of access for help documents, teaching materials, news outlets, and other educational resources.

II 1 2 3 4 5 6 7

NCBI News

New NCBI Newsletter
01 Dec 201
Information on the new Genome Site, a new 16S BLAST database, updates to Sequin,

NCBI will continue to operate SRA
13 Oct 201

Provide a roadmap: links to resource-specific submission help

How To: Submit data to NCBI

Starting with...

SEQUENCE DATA

For guidance on the submission process for your sequence(s), please see How To: Submit sequence data to NCBI.

Your data will be submitted to one of the following databases:

- GenBank
- Sequence Read Archive (SRA)
- dbSNP
- dbVar
- GEO

Links to submission documentation and tools.

GenBank Sequence submissions

http://www.ncbi.nlm.nih.gov/genbank/submit.html

- Submitted data:
  - Sequence (fasta)
  - Annotation
  - Assembly (AGP)
- Tools
  - Details vary depending on data type
- See documentation for details

Data types:

WGS (Whole Genome Shotgun)
TSA (Transcript Shotgun Assembly)
cDNAs, ESTs
Genomic clones

Sequence Read Archive (SRA)

- SRA accepts submissions of:
  - Genomic reads accompanied by WGS submissions
  - Transcript reads accompanied by TSA submissions
  - Expression abundance surveys with RNA-seq data are submitted via GEO

Yes – SRA is accepting submissions!

NCBI News

NCBI will continue to operate SRA

13 Oct 2011

Subsequent to an announcement in February 2011 that NCBI was planning to

TSA: Transcript shotgun assembly

http://www.ncbi.nlm.nih.gov/genbank/TSA.html

- Computationally assembled from primary data
- Restrictions
  - Based on your owned data sets
  - Don't mix different datasets
- There is no physical reagent corresponding to the final transcript assembly.

4,816,402 eukaryotic records

```
LOCUS       JL968987                 967 bp    mRNA    linear   TSA 30-OCT-2011
DEFINITION  TSA: Acrasis rosea EUBAC_GENE31 mRNA sequence.
ACCESSION   JL968987
VERSION     JL968987.1  GI:339522472
DBLINK      Project: 68319
KEYWORDS    TSA; Transcriptome Shotgun Assembly.
```

GEO submission options

http://www.ncbi.nlm.nih.gov/geo/info/submission.html

Gene Expression Omnibus

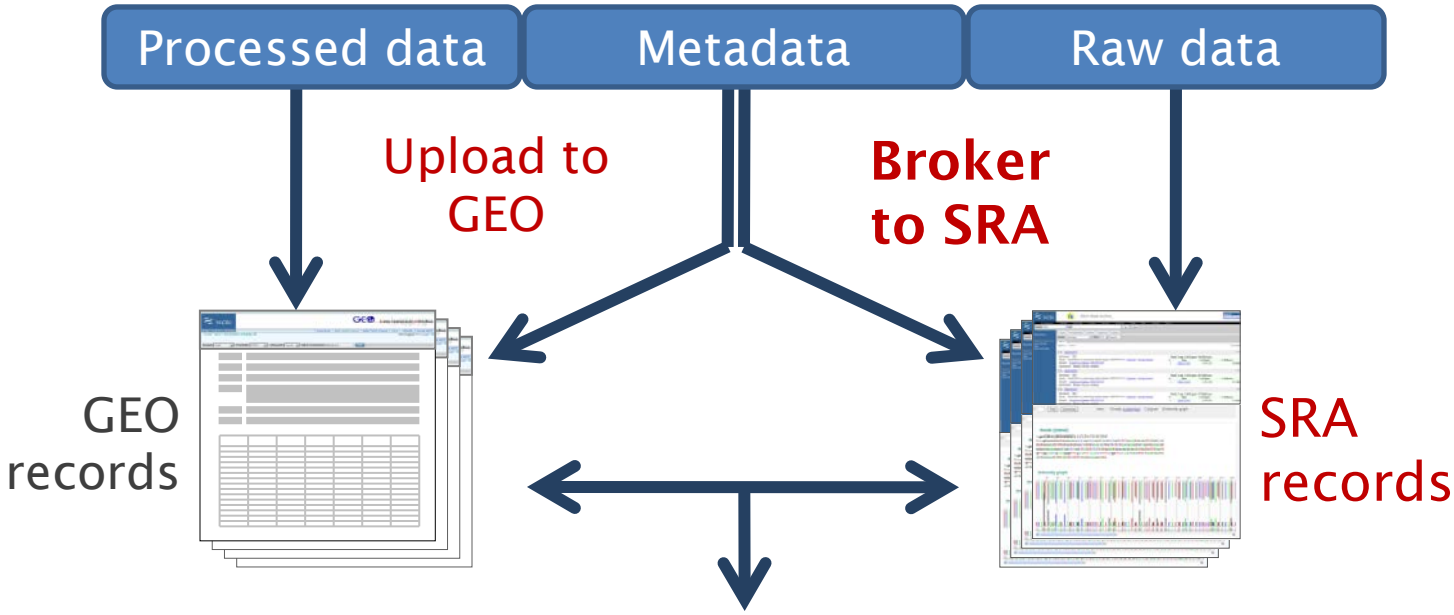
- **Submitting array data**
  - Array deposit options

    All the array deposit options described on this pa
    of array data. However, submitters who use
    arrays are advised to see these additional guidelin
    - Affymetrix
    - Agilent
    - Nimblegen
    - Illumina
  - Basic requirements for array submissions
  - Fast facts about array data submissions
- **Submitting real time PCR data**
  - RT-PCR deposit instructions
- **Submitting high-throughput sequence data**
  - High-throughput sequence deposit instructions

FORMAT	
web forms	Bench biologists
spreadsheets	
plain text	Large data producers
XML	

GEO <-> SRA relationship

GEO handles functional genomic next-generation sequence submissions, including RNA-seq, ChIP-seq and methyl-seq studies



- Public resources
  - Epigenomics Resource
  - SRA BLAST

    …etc.

A new model – centralized submissions

THE PROBLEM…

- Tons of data
- Many archival databases
- Sites have distinct user interfaces
- Submission sites can be hard to locate
- You may not know what data should be submitted

Submission Portal plans

THE SOLUTION…

- A common submission portal system
- Advantages include
  - Single start page
  - Secure login
  - Wizard-guided interface
  - Integrated help, error checks
  - Consistent interface
  - Access to previous submissions and reports

Submission portal home page

https://submit.ncbi.nlm.nih.gov/

NCBI Site map All databases Search

Submission Portal

Submit data to NCBI

SEQUENCE DATA

GenBank

Genetic sequence database, an annotated collection of all publicly available DNA sequences.

WGS

Draft or incomplete genomes that are not yet completely sequenced, so contain NNNs or gaps in the sequences. These genomes consist of sequence contigs assembled from overlapping sequence reads and/or cloned sequences such as BACs. They often also include higher-level scaffolds or chromosomes that have been assembled from the sequence contigs and/or BACs.

Complete Genomes

Collection of genomic sequences that are used to represent the genome of an organism.

TSA

Computationally assembled sequences from primary data such as ESTs, traces and Next Generation Sequencing Technologies. TSA sequence records differ from EST and GenBank records because there are no physical counterparts to the assemblies.

SRA

The Sequence Read Archive (SRA) stores sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, Helicos Heliscope®, and others.

dbSNP

PROJECT DATA

BioProject

A collection of biological data related to a single initiative, originating from a single organization or from a consortium.

MICROARRAY DATA

dbGap

Microarray data from clinical studies that require controlled access.

GEO

A public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data.

MANUS

An electro
manuscrip

CLINICA

Genetic te
variations,
arrays and

Data from
interaction

Current support:

- BioProject
- Links to archival DBs

Goals and Scope:

- Archival databases
- Options:
  - Wizard-guided
  - File upload
  - Programmatic
- Collect meta-data once
- Submit primary data

BETA version

Submission Portal | Home | Submissions | Files | Messages | Kim Pruitt | Log out

Submit data to NCBI

SEQUENCE DATA

GenBank

Genetic sequence database, an annotated collection of all publicly available DNA sequences.

Genomes (WGS) 1

The Whole Genome Shotgun (WGS) database accepts prokaryotic and eukaryotic genomes that are draft or incomplete. A WGS submission comprises a set of contigs, assembled from overlapping sequence reads. The submission can also include hierarchical

PROJECT DATA

BioProject 3

A collection of biological da related to a single initiative originating from a single organization or from a consortium.

BIOLOGICAL MATERIALS

BioSample

Descriptions of biological s

NIH

NCBI Primary Data Archives Login

NIH & eRA Commons

An electronic version of your peer

Coming Soon!

BIOLOGICAL MATERIALS

BioSample

Descriptions of biological source materials used in experimental assays.

SEQUENCE DATA

Genomes (WGS)

The Whole Genome Shotgun (WGS) database accepts prokaryotic and eukaryotic genomes that are draft or incomplete. A WGS submission

Submission review options

NCBI Site map | All databases | Search

Submission Portal Home Submissions Files Messages Kim Pruitt Log out

Upload files
Review messages

BioProject New submission

Start a new submission

Note: to update an existing record or recent submission, please email your request.

Search

Submission title	Status	Updated
New	**Unfinished** at the General info step Delete	24 weeks ago
asdf	**Unfinished** at the Overview step Delete	24 weeks ago
docsum title	**Approved** 65291	24 weeks ago

Finish an incomplete submission

Review a prior submission

BioSample

http://www.ncbi.nlm.nih.gov/biosample

- Metadata for the sample
  - Source
  - Other IDs (stock center etc.)
  - Sample type & method
  - Organism, Gender, Pathogen
  - Captive vs. wild
  - Free text

BioSample

Submission: SUB002213 > **BioSample** New

* Specify when this submission should be released to the public

- (•) Release immediately following curation
- ( ) Release when referenced data is published
- ( ) Release on specified date

* Specify if you are submitting a single sample or a file containing multiple samples

- ( ) Batch/Multiple BioSamples

For a Batch/Multiple BioSamples submission, you will upload a tab delimited text file containing the Organism name, local ID (sample name) and attributes for each of your multiple samples. The Attributes page contains a link to a downloadable template file with required and suggested attributes specific for each sample type. (Help on Attributes File)

- (•) Single BioSample

For a Single BioSample submission, you will first manually enter the organism name and identify the sample type. On the Attributes page, you will be asked to either fill in attributes manually, or to upload a tab-delimited text file containing the attribute information. That page contains a link to a downloadable template file with required and suggested attributes specific for each sample type. (Help on Attributes File)

* Organism name

Escherichia coli (taxid:562)

Strain, breed, cultivar: ABC123

Isolate name or label:

BioProject: Escherichia coli (PRJNA32773) — Delete

Project Data Type: Genome sequencing, Organization: University of Minnesota, Accession: PRJNA32773

Add another BioProject

BETA version

Submitter | General info | Sample type | Attributes | Publications | Comments | Overview

Sample Type

Core package

- General Sample
  - Use for any sample type. These samples are described using common core attributes and submitter-supplied custom attributes.
  - Non-Pathogen
  - Pathogen
- MIxS Sample
  - Use for genomes, metagen
    attributes that have been c
    formally describe and stan
    marker sequences. The sa
    the required core attribute

Continue

Attributes

- Upload tab-delimited text file with sample at
  - How do I create a sample attributes file? Download BioS
- Enter attributes manually

collection_date

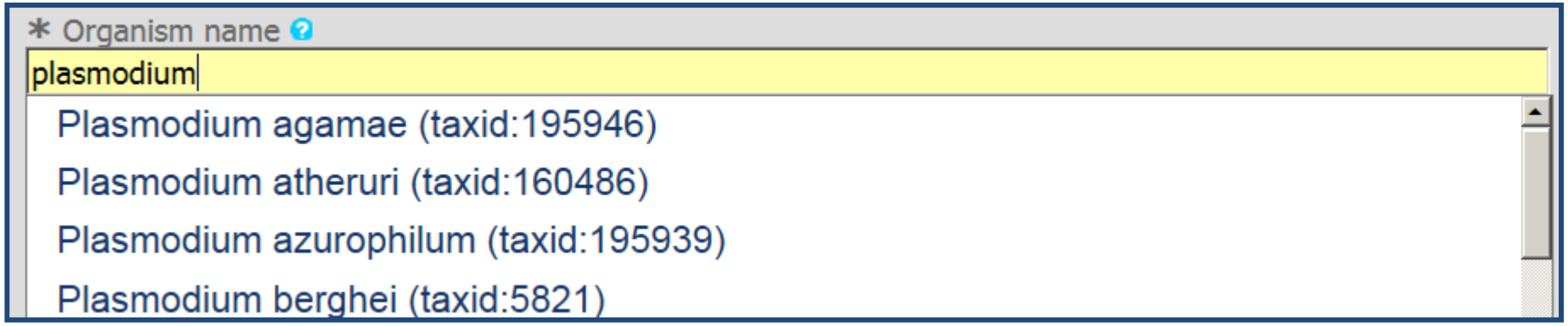
geo_loc_name

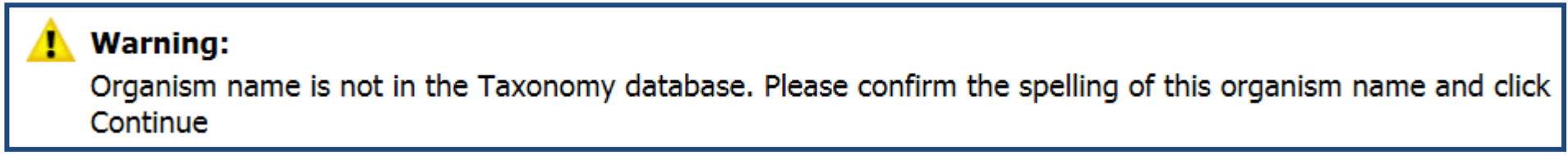
age

The interface suggests attributes to provide based on an internal dictionary. For vertebrates, we may adopt the Genome10K Sample definitions. Suggestions welcome!

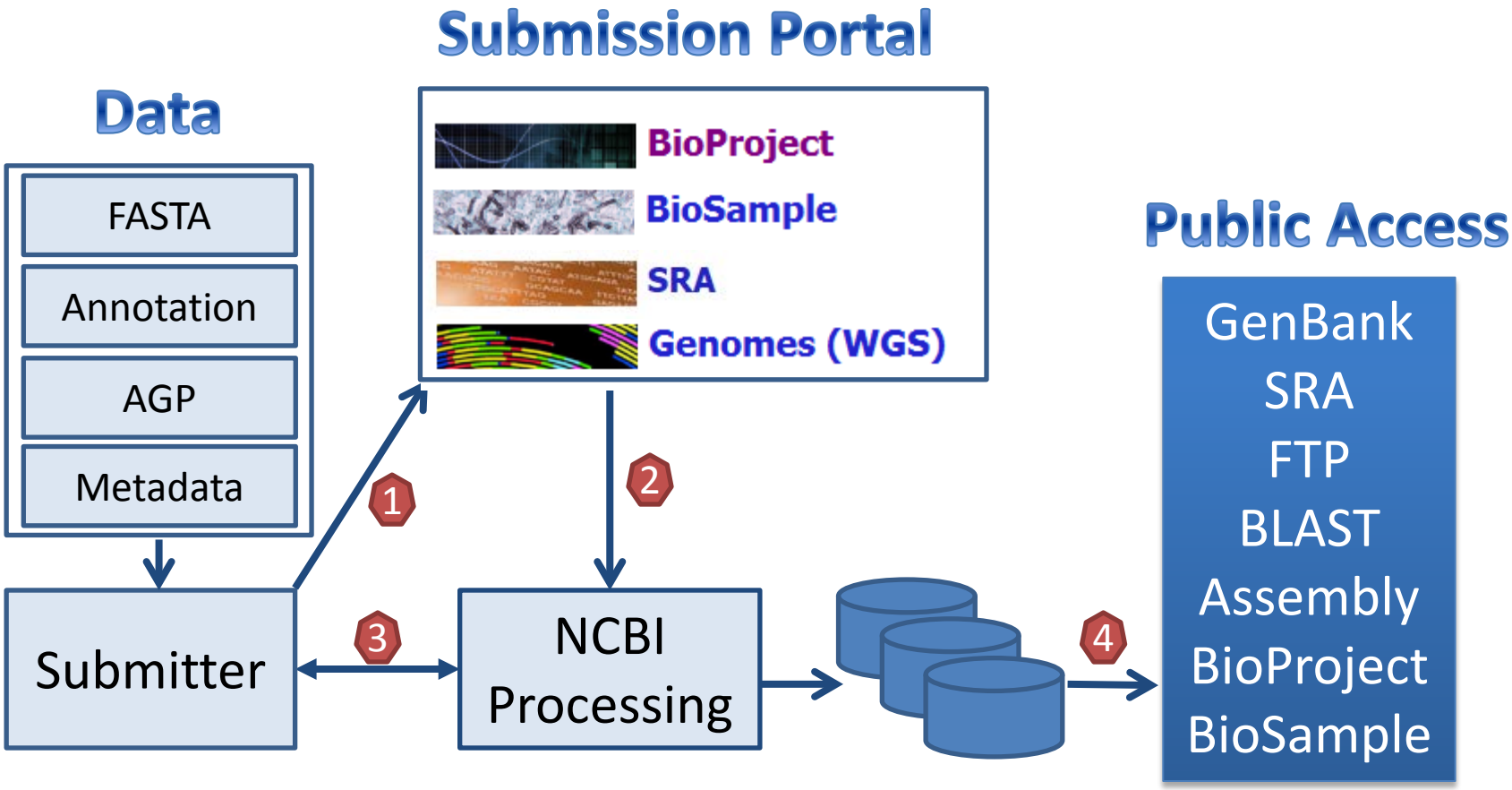
Integrated QC

- Connected to other data
- Clear warnings and error reports





The Plan: Submitting Genomes



Acknowledgements

- Ilene Mizrachi (Submission Portal)
- Karen Clark (WGS Genomes)
- Tanya Barrett (GEO, BioSample)
- Martin Shumway (SRA)
- Tatiana Tatusova, Ilene Mizrachi, Karen Clark (BioProject)

NCBI Workshop

http://www.ncbi.nlm.nih.gov/education/pag2012/

- ***3:50*** *Kim Pruitt* - Primary Data Submission Portal
- ***4:10*** *Tatiana Tatusova* - BioProject, Genome, and Assembly databases
- ***4:30*** *Francoise Thibaud-Nissen* - Eukaryotic Genome Annotation Pipeline
- ***4:50*** *Deanna Church* - Connecting the Lab to the Genome: CloneDB
- ***5:10*** *Kim Pruitt* - Annual Report on Genome Sequencing Projects