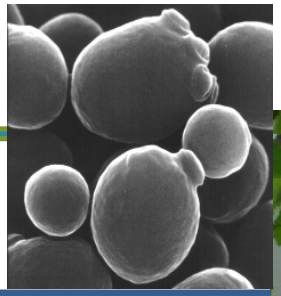
NCBI Workshop

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

January 17, 2010





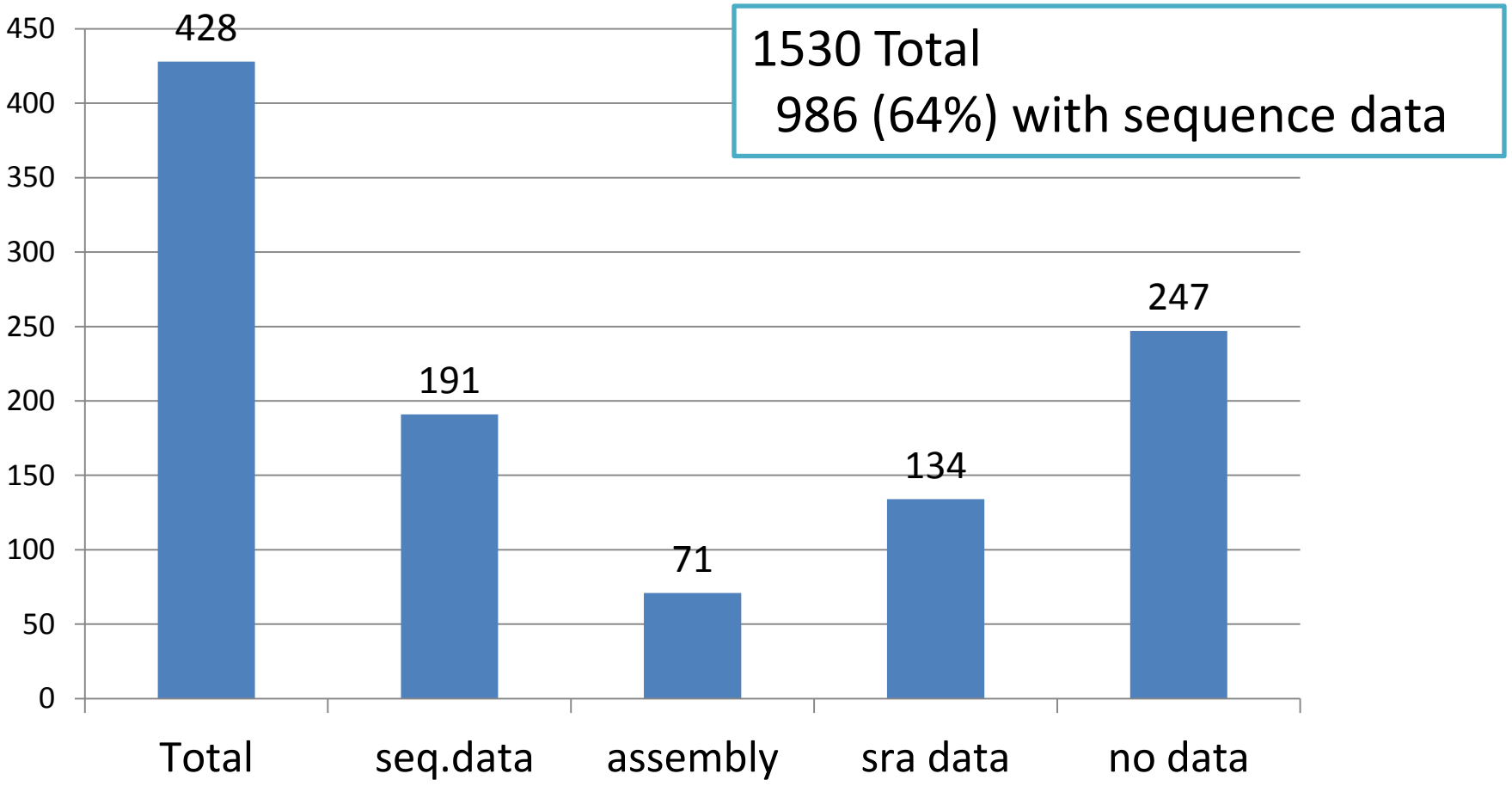

Annual Report of Genome Sequencing Projects

Kim D. Pruitt

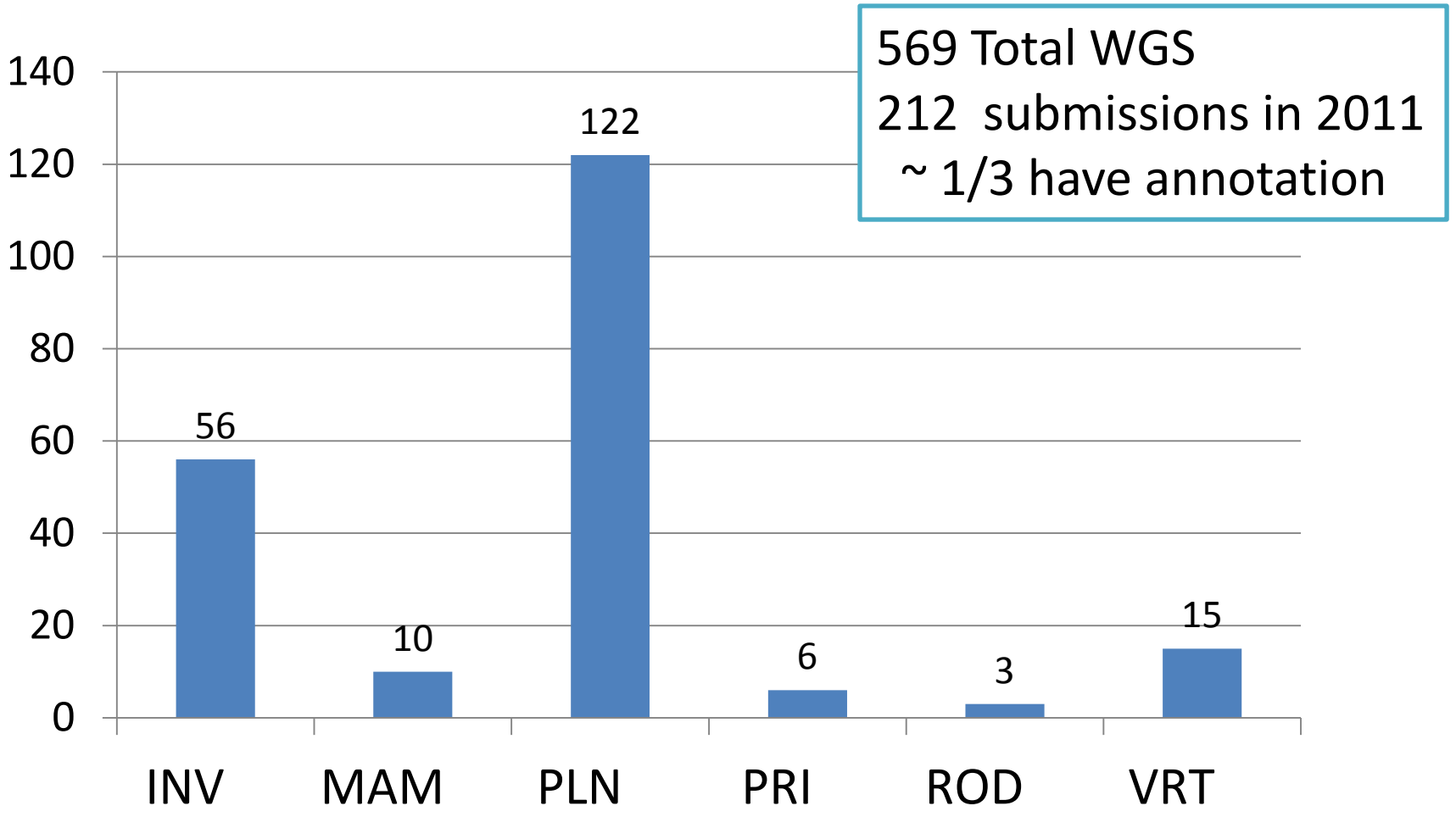
International Plant and Animal Genome XX

January 15-18, 2011

2011 BioProjects genome sequencing (Eukaryotes)



2011 WGS (Eukaryotes)



2011 Highlights - new

* NCBI whole genome annotation pipeline

WGS	Assembly	BioProject *BioSample*	Species	Common
AFTD	GCA_000223135.1	PRJNA69991	Cricetulus griseus	Chinese hamster ovary *
AFSB	GCA_000230445.1	PRJNA68323	Heterocephalus glaber	Naked mole rat
AERX	GCA_000188235.1	PRJNA59571	Oreochromis niloticus	Nile tilapia *
AEYP	GCA_000215625.1	PRJNA59869	Mustela putorius furo	Ferret
AFEY	GCA_000219685.1	PRJNA65325	Sarcophilus harrisii	Tasmanian devil
AFSP	GCA_000230855.1	PRJNA68667	Cajanus cajan	Pigeon pea
ADDN	GCA_000005505.1	PRJNA32607	Brachypodium distachyon	purple false brome *

& More!

Dolphin

Bat

Gorilla

Fungi

Drosophilids

Sorghum

Bees

Armadillo

2011 Highlights - updates

* NCBI whole genome annotation pipeline

WGS	Assembly	BioProject	Species	Common
AAEX	GCA_000002285.2	PRJNA13179	Canis lupus familiaris	Dog *
AEMK	GCA_000003025.4	PRJNA13421	Sus scrofa	Pig *
AAFC	GCA_000003205.4	PRJNA12555	Bos taurus	Cow *
AADN	GCA_000002315.2	PRJNA13342	Gallus gallus	Chicken *
AADG	GCA_000002195.1	PRJNA10625	Apis mellifera	honey bee *
AAGJ	GCA_000002235.2	PRJNA10736	Strongylocentrotus purpuratus	Sea urchin
AAQR	GCA_000181295.3	PRJNA16955	Otolemur garnettii	Small eared galago
ACYX	GCA_000181215.2	PRJNA40349	Phoenix dactylifera	Date palm
AEKE	GCA_000188115.1	PRJNA119	Solanum lycopersicum	tomato

WGS	Assembly	BioProject	BioSample	Species	Common
AERX	GCA_000188235.1	PRJNA59571	117560	Oreochromis niloticus	Nile tilapia

```
LOCUS       AERX01000000            77754 rc    DNA     linear   VRT 10-FEB-2011
DEFINITION  Oreochromis niloticus, whole genome shotgun sequencing project.
ACCESSION   AERX00000000
VERSION     AERX00000000.1  GI:3[illegible]0434504
```

Global statistics

Total sequenced bases	927,742,539
Gaps between scaffolds	0
Number of scaffolds	5,901
Scaffold N50	2,802,423
Number of contigs	
Contig N50	
Total number of chrom[illegible]	

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide	48918
SRA Experiments	39
OTHER DATASETS	
BioSample	1

▾ Genome assemblies, organelles and plasmids:

Name	GenBank
Whole Genome Shotgun Assembly	AERX00000000

Tilapia Nile tilapia Genomic DNA sample 000638D3DF

SRA:SRS119092;

Oreochromis niloticus (Nile tilapia)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Eutele
Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleos
Acanthomorpha; Acanthopterygii; Percomorpha; Perciforme
Cichlidae; African cichlids; Pseudocrenilabrinae; Tilapiini; O

sample type Genomic DNA
sex Other
strain Nile tilapia

sample IDs 000638D3DF - extreme low conc, 000638
HMW >50kb, 000638D3DF HMW >50kb p
conc to 101ng/ul

sample collection dates 20-10-2009, 2010-01-10

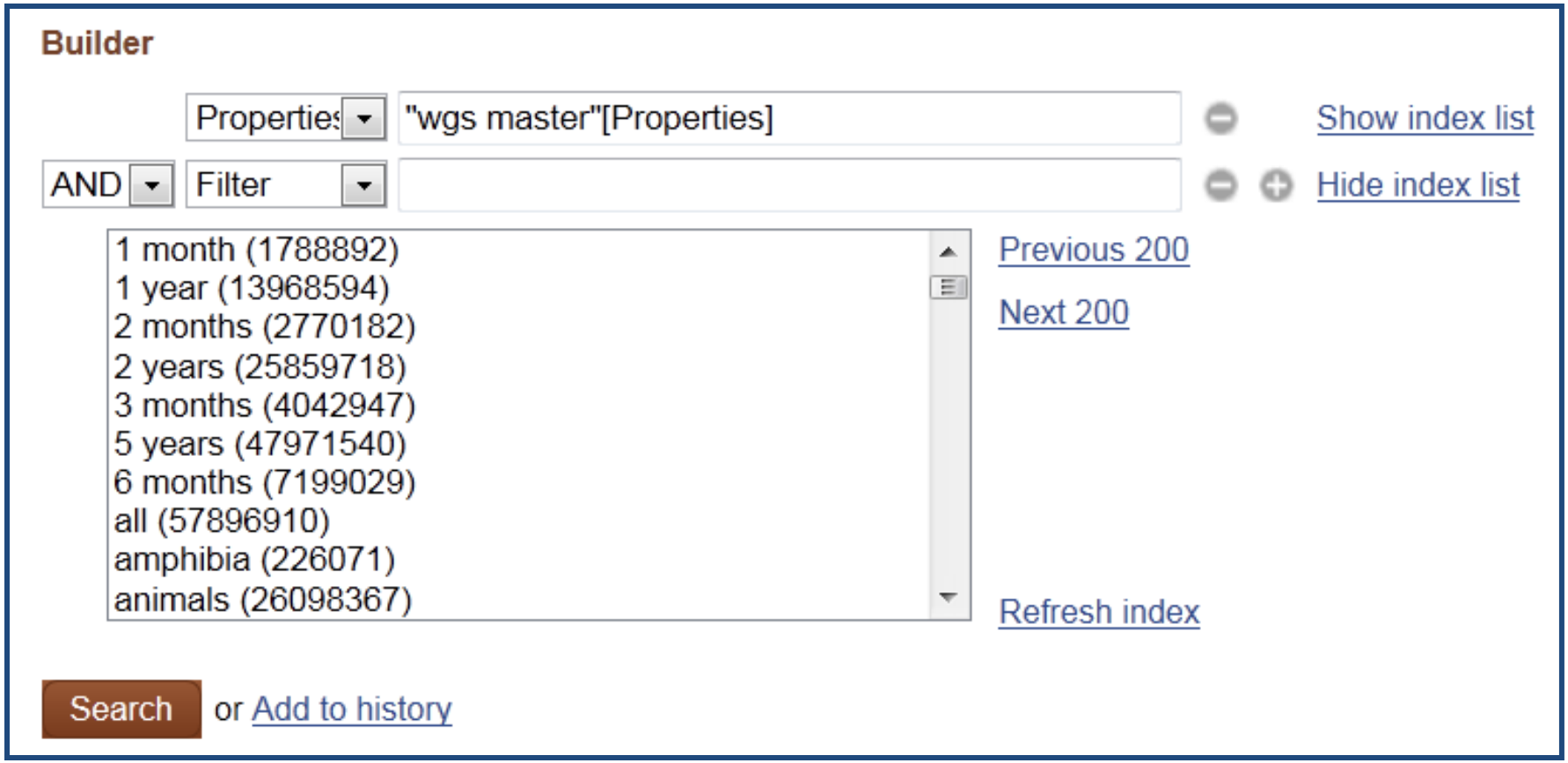
tissue sources Blood, DNA from agarose plugs

Report availability

- Genome reports
  - http://www.ncbi.nlm.nih.gov/genome/browse
  - ftp://ftp.ncbi.nih.gov/genomes/GENOME_REPORTS/
- Assembly reports
  - http://www.ncbi.nlm.nih.gov/genome/assembly/organism/

Nucleotide Advanced Search

"wgs master"[Properties] AND "eukaryotes"[Filter] AND "1 year"[Filter]



Working with research communities

- Meetings and workshops
- Feedback via NCBI help desk

- We value feedback !! – it may result in
  - Updates or error corrections for your archival records
  - New or improved resources
  - Improved data in non-archival resources

NCBI Microbial Annotation Workshop

www.ncbi.nlm.nih.gov/genomes/AnnotationWorkshop.html

- Engage the community to improve guidelines for submitted annotation
  - Names
  - Genome annotation goals (RNAs too!)
  - Evidence

Stand Genomic Sci. 2011 Oct 15;5(1):168-93. Epub 2011 Oct 1.

Solving the Problem: Genome Annotation Standards before the Data Deluge.

Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrachi I, Pruitt KD, Tatusova T.

Abstract

The promise of genome sequencing was that the vast undiscovered country would be mapped out by comparison of the multitude of sequences available and would aid researchers in deciphering the role of each gene in every organism. Researchers recognize that there is a need for high quality data. However, different annotation procedures, numerous databases, and a diminishing percentage of experimentally determined gene functions have resulted in a spectrum of annotation quality. NCBI in collaboration with

Annotation workshop

Results The 2010 workshop attendees agreed to the following minimal standards for complete prokaryotic genomes

- 1. MINIMAL GENOME ANNOTATION SHOULD HAVE
  - a. rRNAs (5S, 16S, 23S) and corresponding genes with locus_tags
  - b. tRNAs and corresponding genes with locus_tags
  - c. protein-coding genes with locus_tags (see below) and corresponding CDS
- 2. ANNOTATION SHOULD FOLLOW INSDC SUBMISSION GUIDELINES.
  Annotation standards should follow INSDC submission guidelines (GenBank/EMBL/DDBJ) part of which were documented as part of the workshop.
  - a. prior to genome submission a submitted Bioproject record with a registered l ocus_tag prefix is required according to accepted guidelines
    http://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf
  - b. the genome submission should be valid according to feature table documentation
    http://insdc.org/documents/feature_table.html
- 3. METHODOLOGIES AND SOPS (STANDARD OPERATING PROCEDURES).
  Genomes should be linked to the SOPs used to create the annotation and with the evidence used to create annotation.
- 4. EXCEPTIONS.

Finding announcements & help

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

http://www.facebook.com/ncbi.nlm

http://twitter.com/ncbi

http://www.youtube.com/ncbinlm

info@ncbi.nlm.nih.gov

Acknowledgements

- Tatiana Tatusova, Ilene Mizrachi, Kim Pruitt, Karen Clark (BioProject)
- Tatiana Tatusova, Kim Pruitt (Genome)
- Mike DiCuccio, Paul Kitts, Francoise Thibaud-Nissen (genome annotation pipeline)
- {? Include assembly db counts??}

NCBI Workshop

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

January 17, 2010