International Sharing of Biological Information in the Mouse

Hiroshi Masuya Ph.D
Technology and development unit for knowledge base of mouse phenotype
RIKEN BioResource Center
Biological information:
Another type of common-resource

Adding knowledge

New researches

Bioresources
Topics

- Introduction: publically available bio-information
- Sharing of non-sequence biological information
- Ontology
- Data standardization in International Mouse Phenotyping Consortium (IMPC)
- Ontology-based data integration for bioresources
Introduction:
Publically available bio-informaiton

- Gene
- Protein
- Protein structure
- Gene expression

- Integrated database
Gene databases (nucleotide sequence)

  NCBI (National Center for Biotechnology Information) in NIH (National Institute of Health) in USA

- EMBL ([http://www.ebi.ac.uk/embl/](http://www.ebi.ac.uk/embl/))
  EBI (European Bioinformatics Institute)

- DDBJ ([http://www.ddbj.nig.ac.jp/Welcome-j.html](http://www.ddbj.nig.ac.jp/Welcome-j.html))
  National Institute of Genetics in Japan
Protein databases


- PIR: Protein information Resource ([http://pir.georgetown.edu/](http://pir.georgetown.edu/))
Structure databases

- PDB: Protein data bank (http://www.rcsb.org/)
- SCOP: Structural Classification of Protein (http://scop.mrc-lmb.cam.ac.uk/scop/)
- CATH (http://www.biochem.ucl.ac.uk/bsm/cath/)
- FSSP (http://www.ebi.ac.uk/dali/fssp/fssp.html)
Gene Expression databases

- FANTOM: Functional Annotation of the Mammalian Genome (http://fantom.gsc.riken.jp/4/)
- Allen Brain Atlas (http://www.brain-map.org/)
Integrated database in mouse

  - is composed of...
    - Mouse Genome Database (MGD)
    - Gene Expression Database (GXD)
    - Mouse Tumor Biology (MTB) Database Project
    - Gene Ontology (GO) Project at MGI
    - MouseCyc Project at MGI
Gene in MGI

Input a gene name for your interest

Gene

Alleles
Sonic hedgehog in MGI
You can get any information!
You can get any information!

Really? Are we happy?
Standardization and international sharing of information
(Sharing of non-sequence biological information)
Not very good…. (except for sequence information..)

Information is sunk to bottom of sea..

Too big! Too complicated!!
We cannot understand all of them.
We cannot say “we share all information”
It is difficult to “share”…..Why?

Biological sequence

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>cagaaatatatctactgaccactgggt</td>
<td></td>
</tr>
<tr>
<td>gtgtgtggcagctttatctataaaa</td>
<td></td>
</tr>
<tr>
<td>ctattacaaatatattcaaaaaatcacatattccatgttattacaaaatattcaaaaatcattctctctca</td>
<td></td>
</tr>
<tr>
<td>gtttattaatgcacaaaaacatccttgttt</td>
<td></td>
</tr>
<tr>
<td>ttctctgtacaactacaggctccatc</td>
<td></td>
</tr>
<tr>
<td>Processable by machine</td>
<td></td>
</tr>
<tr>
<td>(well standardized)</td>
<td></td>
</tr>
<tr>
<td>Data analysis is easy</td>
<td></td>
</tr>
</tbody>
</table>

Higher level information

<table>
<thead>
<tr>
<th>Gene function</th>
<th>Phenotype</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>???</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We need help of analysis results by machines to understand big data.
## What is a difference?

<table>
<thead>
<tr>
<th>Nucleotide sequence</th>
<th>Gene function</th>
</tr>
</thead>
<tbody>
<tr>
<td>A symbol sequence (text) representing information</td>
<td>“GACTATCTT”</td>
</tr>
<tr>
<td>Information or content to be shared</td>
<td>“GACTATCTT” (the symbol sequence itself)</td>
</tr>
</tbody>
</table>
We need some common technologies to deal with “meaning” by computers
Gene Ontology (GO)
The first, most successful bio-ontology

- Proposed by Michael Ashburner in 1995
- Initially developed collaborative activity of three model organisms databases: FlyBase, Saccharomyces Genome Database (SGD), and MGI.
- Based on the philosophical (=ontological) procedure for classification of “terms”.
- Providing the procedure for machine-searchable annotation.
- Now developed under broad international collaborative effort.
- 35,000 terms have been included. Over 11 million annotations across more than 30 species are also openly available.
Gene Ontology (GO)

- Giving ID for terms.
- Representing meaning by machine readable tree structure.

Standardized procedure to annotate gene function.
(MGI00000xxx, GO:00001833) (binary relationship)
= gene A has function in inner cell mass cell proliferation
Fruits from GO annotation

• Enabling comprehensive search robust to the unevenness of vocabularies. (improving inconsistencies in search results)

• Providing the simple standard procedure for annotation of gene function across databases.

• Databases are connected via GO terms across species.

• Vast annotation data were accumulated across databases and species.

• Enabling reuse of existing knowledge: statistical analyses of biological annotation applicable for omics study (e.g. microarray analysis).
Why these genes are highly expressed in Experiment A?

List of gene

List of GO annotation

molecular function

cellular component

biological process

Transcriptions and metabolisms are activated by experiment A...!
Ontology for “phenotype”

To search “hairless” mouse, what vocabulary do you use?

<table>
<thead>
<tr>
<th>Hairless</th>
<th>Nude</th>
<th>Bald</th>
<th>Hair loss</th>
<th>Loss of hair</th>
<th>Furloss</th>
<th>Loss of fur</th>
<th>Composite 'true' result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A1</td>
<td></td>
<td></td>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>Aldoc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>alp</td>
<td>alp</td>
<td></td>
</tr>
<tr>
<td>ao</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ao</td>
<td>ao</td>
<td></td>
</tr>
<tr>
<td>Bda</td>
<td>Bda</td>
<td></td>
<td></td>
<td></td>
<td>Bda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blai2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bmp6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Btk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ctsl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ctsl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dsg3</td>
<td>Dsg3</td>
<td></td>
<td></td>
<td></td>
<td>Dsg3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ebp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eda</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Eda</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Edar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edaradd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Edaradd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>exf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>exf</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fgf7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foxj1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foxn1</td>
<td>Foxn1</td>
<td></td>
<td></td>
<td></td>
<td>Foxn1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Frl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>hl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoxb8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hr</td>
<td>Hr</td>
<td>Hr</td>
<td></td>
<td></td>
<td>Hr</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search by different term results different results to looking for “hairless” mutant in mouse locus catalog (Smith et al. 2004).
Mammalian Phenotype Ontology (MP)

Hierarchical classification of phenotype

Phenotype annotation to realize comprehensive retrieval

<table>
<thead>
<tr>
<th>Allelic Composition (Genetic Background)</th>
<th>Annotated Term</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aars^{R1}/Aars^{R1}</td>
<td>dishveled coat</td>
<td>b:110647</td>
</tr>
<tr>
<td>Adipoq^{tm1fsV}/Adipoq^{tm1fsV}</td>
<td>abnormal coat appearance</td>
<td>b:111346</td>
</tr>
<tr>
<td>Doptilm1f/Ear opartillicer</td>
<td>(involves: 129S4/SvJae * C57BL/6)</td>
<td></td>
</tr>
<tr>
<td>F12{/Ar{+}</td>
<td>(involves: 129S1/Sv * C57BL/6)</td>
<td></td>
</tr>
<tr>
<td>Amigo^{tm1S/Amp^{tm1V}</td>
<td>dishveled coat</td>
<td>b:81166</td>
</tr>
<tr>
<td>(involves: 129S1/Sv * 129X1/Sv) * Black Swiss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eo/a</td>
<td>sparse hair</td>
<td>b:5122</td>
</tr>
<tr>
<td>ep/ap</td>
<td>sparse hair</td>
<td>b:39747</td>
</tr>
<tr>
<td>Apert^{tm1l/Bm} /Apert^{tm1l/bm}</td>
<td>abnormal coat appearance</td>
<td>b:38450</td>
</tr>
<tr>
<td>(129P2/OlaHsd)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apert^{tm1l/Bm} /Apert^{tm1l/bm}</td>
<td>dishveled coat</td>
<td>b:38450</td>
</tr>
<tr>
<td>(involves: 129P2/OlaHsd * BALB/c)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apert^{tm1} /Apert^{tm1}</td>
<td>dishveled coat</td>
<td>b:33255</td>
</tr>
<tr>
<td>(involves: 129S2/SvLw * Black Swiss)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apert^{tm1} /Apert^{tm1}</td>
<td>dishveled coat</td>
<td>b:33255</td>
</tr>
</tbody>
</table>

Back to entry page
Standardization of information/data is beneficial

• (Ontology is a tool for standardization of information by giving unique IDs and mutual relationships)

• Making data collection comprehensive.

• Making possible large-scale data analyses. (Promoting paradigm shift in biological research from hypothesis oriented to data analysis oriented.)

• Enabling global integration of biomedical data.

• Further progress of ontology-based technology will be required...
(A trial of) ontology-based data integration for phenome of bioresource
Project Overview

Bioresources (experimental materials)

USE

Human disease researches

Bridge by information

Linked Open Data (LOD)

Common role for data integration

Distributed development of LOD in RIKEN

Phenotype annotation
(by RIKEN BRC)

Genome information
Disease information
(by RIKEN BASE)

• Open for public
• Globally standardized by RDF
• Easily integrated mutually

By institute A
By institute B
By others....

Mouse strain A URI xxx
Phenotype Y URI xxxx
Human disease C URI xxxxxxxxxx
Gene B URI xxxx
Allele X URI xxx
Pathway Y URI xxx
Common representation of phenotypes with ontologies

Common schema for phenotype

Phenotype data

Body part

Trait

Attribute

Value

is_a

Plant

cells

mice

(Substitution with ontologies)

Table of phenotype data:

- **Phenotype**: M100702
- **Trait**: increased white blood cell distribution
- **Value**: increased white blood cell distribution

(Example of mouse phenotype)

生物横断的に特性情報を格納できる汎用スキーム（各生物の特性情報 is_a 表現型データ）
EAV annotation of mouse phenotype

**Body part annotated by MA ontology**
- neural: 10 strains
- Skeletal: 3 strains
- Non-long bone: 5 strains
- Axial skeleton: 12 strains
- Joint: 3 strains
- Integument: 87 strains
- Connective tissue: 17 strains
- Body fluid: 102 strains
- Urine: 6 strains

What is affected in “Long bone”
- size: 20
- short: 13
- thin: 6
- dysplasia: 1

Morphology: 35
- shaoe: 15
- Non-morphology: 1

Genes controlling length: 13 strains
Controlling thickness: 6 strains

**What mice shows phenotype in bone?**

Filtering with body part

<table>
<thead>
<tr>
<th>Name of mouse strain</th>
<th>EAV annotation of mouse phenotype</th>
<th>mammalian phenotype ontology</th>
<th>Reference</th>
<th>Genotype</th>
<th>Initial</th>
<th>Phenotype</th>
<th>Appearance</th>
<th>Genotype</th>
<th>Initial</th>
<th>Phenotype</th>
<th>Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>phenotype of M100451</td>
<td>M100451</td>
<td>long bone epiphysis</td>
<td>morphology</td>
<td>abnormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100866</td>
<td>M100866</td>
<td>long bone epiphyseal plate</td>
<td>physical object quality</td>
<td>abnormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of 6.1295-</td>
<td></td>
<td>common myeloid progenitor</td>
<td>morphology</td>
<td>abnormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100656</td>
<td>M100656</td>
<td></td>
<td>pelvic girdle bone</td>
<td>abnormal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M121679</td>
<td>M121679</td>
<td>cranial</td>
<td>decreased diameter</td>
<td>decreased anterior-posterior diameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100966</td>
<td>M100966</td>
<td>cranial</td>
<td>decreased diameter</td>
<td>decreased anterior-posterior diameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M121132</td>
<td>M121132</td>
<td>cranial</td>
<td>decreased diameter</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of PIC/Nem-Pdi+/</td>
<td>PIC/Nem-Pdi+</td>
<td>bone</td>
<td>attachment quality</td>
<td>attached to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100576</td>
<td>M100576</td>
<td>cranial</td>
<td>height</td>
<td>decreased height</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100117</td>
<td>M100117</td>
<td>cranial</td>
<td>curvature</td>
<td>domed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M121682</td>
<td>M121682</td>
<td>cranial</td>
<td>decreased diameter</td>
<td>decreased anterior-posterior diameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phenotype of M100566</td>
<td>M100566</td>
<td>cranial</td>
<td>decreased diameter</td>
<td>decreased anterior-posterior diameter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Automatically connected “data network” around mouse strain

Chromosomal position
Chr2:155766759-155771103 bp, - strand

Mouse allele
has allele

Mouse gene
has gene

Mouse Strain
has allele

Mouse Strain
Stock
T(2;11)3OH/+ x AEJ-a Gdf5bp-H,J or A/J-a Gdf5bp-J,J

Institute
depositor

RIKEN GSC

doctoror

Another mouse strain with Gdf5 gene mutation

expression of Gdf5 gene

Gene expression data of Gdf5 gene

Other strains developed by RIKEN GSC

Other alleles showing brachypodism

M100451

Mouse Strain
C57BL6/J DBA2/J

Mouse allele
has allele

Mouse gene
has gene

brachypodism

phenotype

MGI

CDT-DB

has gene

phenotype

phenotype

has allele

Mou se allele
Gdf5

Mou se allele
Gdf5

Gdf5

Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5

Mou se allele
Gdf5
Representation of complicated data

At first minimum data link is represented
Clincking Shows detailed data
You can order a resource
Recommended mice with “Amazon” like representation
What is "Mus musculus"?

This strain is derived from C57BL

The other bio-resources in Mus musculus

Other strains derived from C57BL

This strain has "Kit<W-sh>" allele

References of this strain

What phenotypes are induced by Kit<W-sh>?

Kit<W-sh> is a variant of Kit gene
Collection of mice show similar phenotype

Collection of related information with links

(Ranking of strains with phenotypic similarity)
Promotion of “Linked Open Data (LOD)”

Data integration on the World Wide Web

LOD Cloud
Linked together
Open to use
Using URI and RDF

Data dissemination
Making common use

Data collection
Data analysis

Access to bioresource will be drastically improved!!
Releasing whole data via BioLOD

http://biolod.org/

Open data repository

Whole data is available in various data formats
Common-resource will promote progress of science

Common Use of Data and Bioresources

- Biological experiments
- Medicine and drug discoveries
- Simulations
- System biology
- Environment