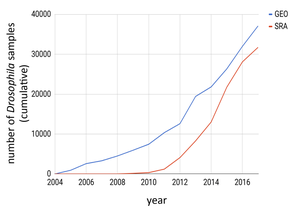
# FlyBase: large-scale datasets and single-cell technologies

Gil dos Santos, Kathleen Falls, David Emmert, Josh Goodman, Chris Tabone, Justin Fear, Gillian Millburn, Marta Costa, Brian Oliver, Norbert Perrimon and the FlyBase Consortium

*Biological Laboratories, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138 USA*

## Introduction

### Bringing "big data" to FlyBase




**Thousands of datasets at GEO & SRA.**

**FlyBase wants to facilitate:**
- Discovery of key results.
- Re-use of data.

**Challenges:**
- Quantity.
- Heterogeneous study descriptions (metadata):
  - different ways to say same thing
  - key details buried

### The limitations of "free-text" metadata

Unstructured experimental descriptions make finding the right datasets difficult.

| search term | GEO hits (*D. melanogaster*) |
|---|---|
| "fat body" | 372 (328 unique to this term) |
| "fatbody" | 107 (73 unique to this term) |
| ("fat body" or "fatbody") | 445 |

**Search term redundancy:**
Small variations affect results.

**Context unclear:**
Was fat body isolated, or removed?

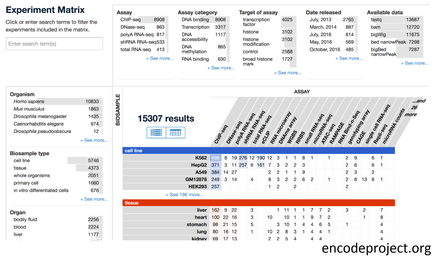
| search term | GEO hits (*D. melanogaster*) |
|---|---|
| ("nej" or "nejire") | 17 (4 unique to this term) |
| "CBP" | 142 (129 unique to this term) |
| ("nej" or "nejire" or CBP") | 146 |

**Search term ambiguity:**
CBP could mean CG15319 or C1435.

**Context unclear:**
CBP a target for ChIP, or RNAi?

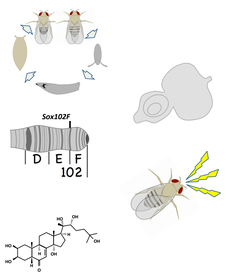
### The advantage of structured metadata

ENCODE project: a model of how structured metadata allows for powerful searching.




encodeproject.org

### FlyBase as a portal for fly datasets



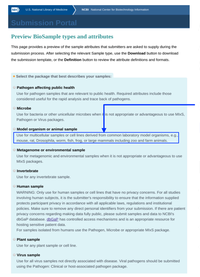

**Goal:**
Catalog datasets using
standardized experimental descriptions.

**In progress:**
Standardize biological sample descriptions.

**Long term:**
Improved search/browse capabilities.

## Drosophila BioSample template for NCBI submission

### Motivation

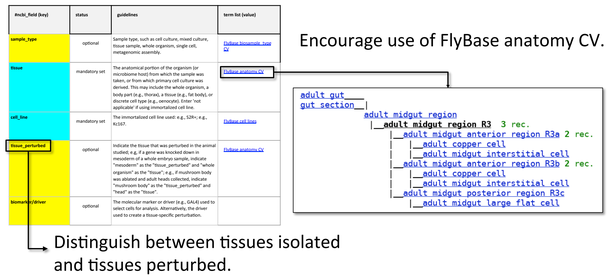



***In progress:***
A *Drosophila*-specific template for biosample description (with Justin Fear and Brian Oliver).

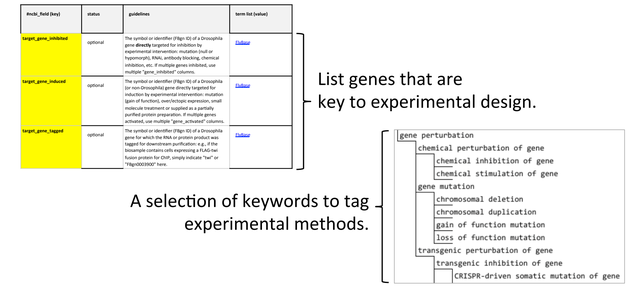
**Goal:**
Make it available at NCBI during submission of *Drosophila* data.

**Features:**
- Improved consistency and **machine** readability.
- Simplified (fewer fields than generic template).
- Covers common aspects of fly studies.
- Fly-specific definitions and examples.

### Consistent and nuanced reporting of tissues

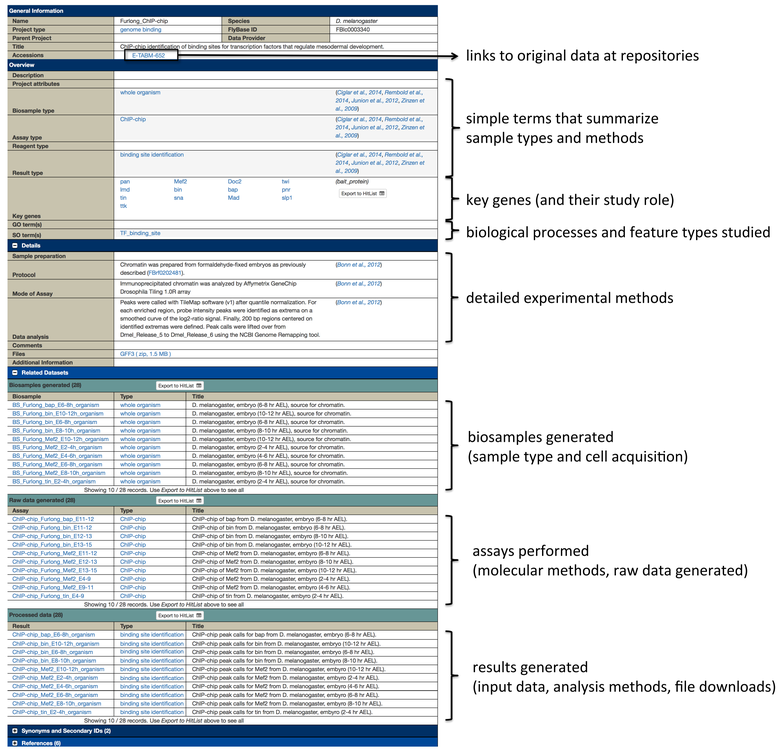



### Clear listing of genes and perturbations

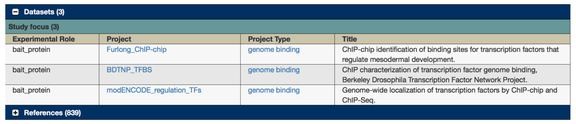



## FlyBase Dataset Reports

### Dataset report



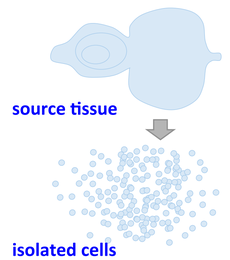

### Dataset section (gene report)



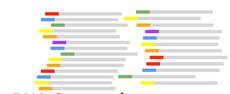

**A list of datasets for which a gene is a key player:**
- Gene's experimental role
- Project type.

## Single-cell technologies

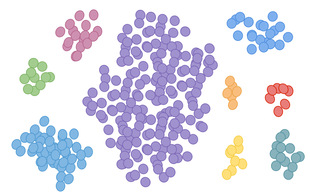
### Facilitating access to single-cell data, at FlyBase and beyond




**Ensure standardized descriptions of source tissue:**
- Anatomy
- Developmental stage
- Cell type enrichment
- Method of cell isolation and barcoding
- Strain, genotype, perturbations (diet, chemical, etc.)




**Ensure proper data formatting at data repositories:**
- Facilitate data re-use.
- Harness existing data analysis pipelines (EBI scAtlas):
  - Gene expression across individual cells.
  - Find similar cell types across experiments.
- Linkouts to data from FlyBase and Virtual Fly Brain.




**Cell clusters as the key result output:**

**Methods used to identify cell clusters:**
- RNA-Seq mapping (annotation set, method, depth)
- Gene expression measurement
- Clustering method

**Cell cluster characteristics:**
- Observed cell types
- Novel cell types
- Markers/signatures
- Consensus gene expression profiles
- Relationships between clusters (similarity, lineage)
- Genes involved in cell type specification

### The Fly Cell Atlas community

**Fly Cell Atlas**
flycellatlas.org

**FlyBase**
flybase.org

**EBI scAtlas/ArrayExpress**
ebi.ac.uk/arrayexpress/

**Virtual Fly Brain**
virtualflybrain.org

Submitting single-cell data? Send it to **EBI scAtlas:**

**Submit directly to EBI ArrayExpress**
**OR**
**Share preprints with FCA Slack/FlyBase/Virtual Fly Brain/:**
flycellatlas.slack.com#data-submission
dossantos@morgan.harvard.edu
virtualflybrain@googlegroups.com

## References

*Barrett, et al.* NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.
*Hong, et al.* Principles of metadata organization at the ENCODE data coordination center. Database (Oxford). 2016 Mar 15.

## Funding

FlyBase is supported grant **#U41 HG000739** from the National Human Genome Research Institute at the U.S. National Institutes of Health.
*Support is also provided by the British Medical Research Council (**#MR / N030117 / 1**) and the Indiana Genomics Initiative.*
*Hosting of this site is supported in part by the National Science Foundation (**#OCI-1053575**) through XSEDE resources via Indiana University.*

flybase.org
@FlyBaseDotOrg
FlyBase TV
alliancegenome.org