# Building a comprehensive catalog of *Drosophila* datasets at FlyBase

136

Gilberto dos Santos, Kathleen Falls, Chris Tabone, David Emmert, Gillian Millburn, Marta Costa, Madeline Crosby, Norbert Perrimon and the FlyBase Consortium

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

## ABSTRACT

FlyBase (**http://flybase.org**) is an essential resource of genetic and molecular data for the Drosophila research community. In the past decade, the exponential growth in the number of large datasets has presented FlyBase (and similar databases) with the challenge of incorporating big data alongside more traditional data types. With the goal of increasing the research community's accessibility to big data, we propose a system of dataset tracking intended to provide the Drosophila researcher with a unified, comprehensive and well-indexed catalog of large datasets that provides links to data repository submissions and their related published results. Datasets are curated according to a four-level classification system (project, biosample, assay and result) to permit more sophisticated tracking of data provenance and facilitate metadata presentation to the public. Metadata for these datasets are described using a variety of controlled vocabularies to allow for improved searchability. The key genes related to datasets, both those that are experimentally manipulated and those that come up as hits in analyses, are captured to provide a listing of the most relevant datasets for a given gene. We anticipate that this dataset tracking system will not only increase accessibility to researchers, but also serve as a useful foundation for future incorporation of genomic data into FlyBase.

## INTRODUCTION

**MOTIVATION:**

1. Help researchers find datasets of interest at data repositories.
2. Facilitate large scale dataset processing.

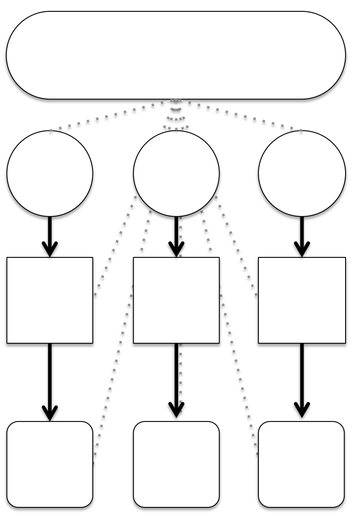
**GOALS:**

1. Create a single unified list of datasets at various data repositories.
2. Associate repository accessions with published “high-level” analyses.
3. Standardize metadata to provide a useful index of datasets.
4. Associate datasets with FlyBase entities (e.g., genes) to increase visibility.

## SCHEME

**TRACK DISTINCT COMPONENTS OF DATASETS**

Flexible, compatible with various data repositories, NCBI GEO (Barrett *et al.*, 2013) and ENCODE (Hong *et al.*, 2016) in particular.



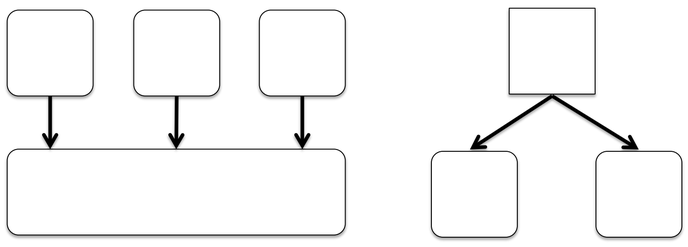
**PROJECT**
Groups objects from a given study.
*NCBI BioProject, GEO series*

**BIOSAMPLE**
The biological material analyzed.
*NCBI BioSample*

**ASSAY**
‘Raw’ or minimally ‘processed’ data generated from biosample.
*NCBI GEO sample*

**RESULT**
Data derived from analysis of ‘raw’ or other ‘processed’ data.
*Supplementary files*

**VARIOUS RELATIONSHIPS PERMITTED BETWEEN ENTITIES**



**one result from the analysis of many results**



**many results from one assay**

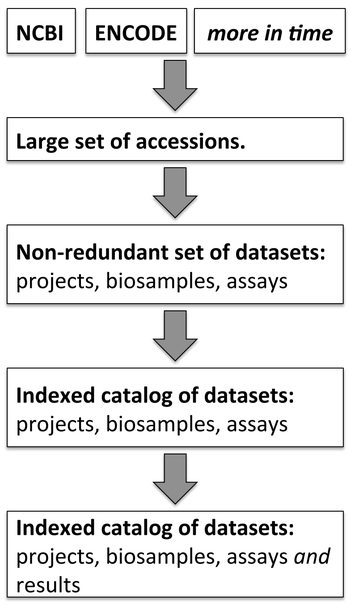
## CURATION PIPELINE

**COLLECT:**
Gather accessions from repositories.
(automated)

**RECONCILE:**
Find accessions representing same data.
(automated)

**STANDARDIZE:**
Map metadata to FlyBase CVs.
(manual curation)

**CURATE RESULTS:**
Identify high-level analyses in literature, typically from supplementary files.
(publication triage, manual curation)




## METADATA STANDARDIZATION

**CATEGORIZATION:**
Dataset entities will be categorized by type. (FlyBase CVs, modeled after ENCODE and NCBI GEO types)

**ATTRIBUTES:**
Additional terms will be associated to further describe the dataset component. (FlyBase CVs)

| PROJECT | BIOSAMPLE | ASSAY | RESULT |
|---|---|---|---|
| **Type**<br>genome variation<br>genome binding<br>transcriptome<br>… | **Type**<br>whole organism<br>tissue<br>immortalized cell line<br>… | **Type**<br>RNA-Seq<br>ChIP-Seq<br>two hybrid screen<br>… | **Type**<br>binding site identification<br>expression clustering<br>RNA-Seq profile<br>… |
| **Study design**<br>development stage study<br>tissue type study<br>cell cycle study<br>… | **Sample attributes**<br>developmental stage<br>anatomy<br>cell line<br>… | **Assay material**<br>chromatin<br>genomic DNA<br>poly(A) RNA<br>… | **Result attributes**<br>stranded profile<br>uniquely mapping alignment<br>RPKM calculation<br>… |
| | **Treatments**<br>biotic treatment<br>chemical treatment<br>… | **Assay platform**<br>Affymetrix gene array<br>Illumina sequencing<br>… | |

**ASSOCIATED ENTITIES:**
Genes, alleles, transgenes, etc. will be associated with datasets.
Genes will be further distinguished by their role: bait_protein, ectopic_factor, depletion_target, experimental result, etc.

## DATASET REPORTS

**DISTINCT REPORTS FOR PROJECTS, BIOSAMPLES, ASSAYS AND RESULTS:**

**Sample Assay Report**

General Information about the assay

Information imported from related biosample

Detailed information about the assay

Related dataset entities

**General Information**

| | | | |
|---|---|---|---|
| Name | Composite_example_assay | Species | *D. melanogaster* |
| Assay type | RNA-Seq | FlyBase ID | FBlc0001317 |
| Project | modENCODE_mRNA-Seq_transcriptome | Data Provider | *modENCODE* |
| Title | RNA-Seq of D. melanogaster, iso-1 strain, larva, nuclear transcripts, where ftz is knocked down in the ring gland. | | |
| Accessions | modENCODE_4439, SRX0078111, SRX008227, SRX008258 | | |

Identifier
Provenance
Brief description
Links to data repositories

**[-] Biosample Source**

**Overview**

| | |
|---|---|
| Strain | iso-1 |
| Stage | larval stage |
| Sex | female |
| Tissue isolated | organism |
| Other tissues st | ring gland |
| Cell component | nucleus |
| Cell line | |
| Key genes | RNAi target: ftz |
| Methods | gene perturbation |
| Sample prepara | ftz was knocked down in the ring gland. Wandering third instar larvae were homogenized and fractionated to isolate the nucleus, which was immediately resuspended in Trizol reagent for RNA isolation. |

**Biosamples used**

| Biosample | Type | Title |
|---|---|---|
| BS_mE_em0-2hr | whole organism | D. melanogaster, iso-1 strain, embryo (0-2hr AEL), source for RNA. |

**Assay Details**

| | |
|---|---|
| Methods | Illumina sequencing, paired-end layout, poly(A) RNA isolation |
| Key genes | bait_protein: Ubx |
| Protocol | Total RNA was extracted using the TRIzol reagent protocol (Invitrogen). RNA was purified on an RNeasy spin column (Qiagen), and DNase treated. Polyadenylated RNAs were purified from total RNA extracts via oligo(dT) binding, using standard Illumina protocol. The poly(A)+ RNA was fragmented using divalent cations under elevated temperature, following by first and second strand cDNA synthesis primed with random hexamers. The cDNA fragments were end-repaired using T4 DNA polymerase and Klenow DNA polymerase, and phosphorylated at their 5' ends with T4 polynucleotide kinase. |
| Mode of assay | Read length (bases):76 |
| Comments | |

Assay attributes (CV)
Related genes

**Associated Data**

| | |
|---|---|
| Size | 86,980,459 Uniquely aligned reads; as reported. |
| Associated feat | 3945 exon junctions [Download] [HitList] |
| Files | *fastq* (raw reads, 3 MB) |

Lists of related objects
File downloads

**[-] Related Datasets**

**Parent projects**

| Project | Type | Title |
|---|---|---|
| modENCODE_mRNA-Se | transcriptome | Transcriptional profile of D. melanogaster developmental stages, unstranded RNA-Seq, modENCODE. |

**Processed data generated from this assay**

| Result | Type | Title |
|---|---|---|
| mE_mRNA_ER_1 | RNA-seq profile | RNA-Seq coverage profile of D. melanogaster, iso-1 strain, larva, nuclear transcripts, where ftz is knocked down in the ring gland. |
| mE_mRNA_junctions | exon junction identification | RNA-Seq exon junctions in D. melanogaster, iso-1 strain, larva, nuclear transcripts, where ftz is knocked down in the ring gland. |
| co-regulated_gene_set | expression clustering | Clustering of genes based on similar expression dynamics throughout development, as determined from modENCODE RNA-Seq data. |

**Other assays (via shared biosamples)**

| Result | Type | Title |
|---|---|---|
| CAGE-Seq_mE5338_E1 | CAGE-Seq | CAGE-Seq of D. melanogaster, iso-1 strain, larva, nuclear transcripts, where ftz is knocked down in the ring gland. |

[+] Update History
[+] Synonyms and Secondary IDs
[+] References

## INCREASE VISIBILITY – GENE REPORTS

**HIGHLIGHT DATASETS RELEVANT TO A GIVEN GENE:**
FlyBase gene reports are the most visited sites.

**[-] Datasets**

**Study focus**

Gene is a **variable**

| Experimental Role | Project | Project Type | Title |
|---|---|---|---|
| allele_used | modENCODE_replication_factor_ChIP | genome binding | Genome-wide localization of essential replication factors characterized by ChIP-Seq. |
| bait_protein | FBGSE22069 | genome binding | Protein profiling reveals five principal chromatin types in Drosophila cells. |

**Study result**

Gene is **hit** in analysis

| Result | Result Type | Title |
|---|---|---|
| mE1_20_mRNA_expression_cluster_05 | expression cluster | Genes expressed at moderate expression levels, enriched in early-to-mid embryogenesis and adult females. |
| L3_disc_N[12]_down | expression cluster | Genes downregulated in L3 N[12] discs, > 1.5-fold, p < 0.05. |
| L3_proventriculus_top100 | expression cluster | The most highly expressed genes in larval proventriculus (top 100). |
| Ciblin | gene list | Identification of genes conserved in bilaterians but lost in nematodes. |

## 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 by a grant from the National Human Genome Research Institute at the U.S. National Institutes of Health **#U41 HG000739**. Support is also provided by the British Medical Research Council (**#G1000968**) and the Indiana Genomics Initiative. Hosting of this site is supported in part by the National Science Foundation (**#OCI-1053575**) through XSEDE resources provided by Indiana University.
Please contact Gil dos Santos for more information: **dossantos@morgan.harvard.edu** or visit **flybase.org**