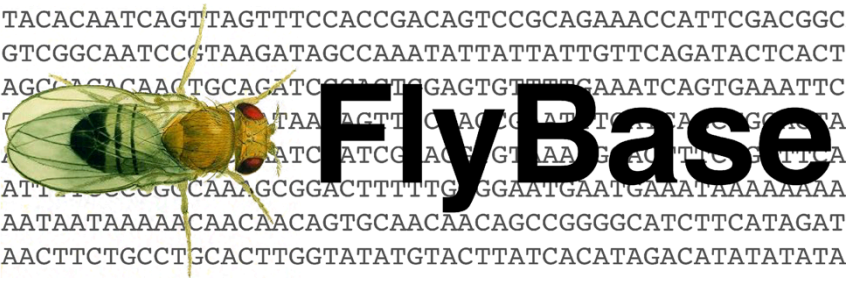


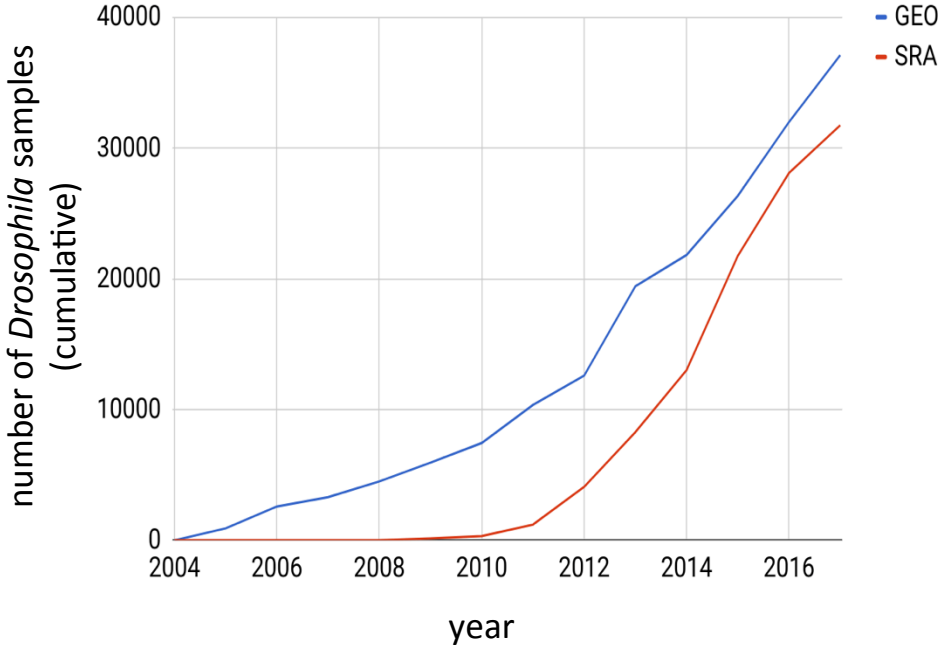
Finding your way around large-scale datasets and single-cell technologies

Gil dos Santos, FlyBase
dossantos@morgan.harvard.edu
helpfb@morgan.harvard.edu

Genetics Society of America

Annual Drosophila Research Conference

"big data": why do we care?



Thousands of samples at GEO & SRA.

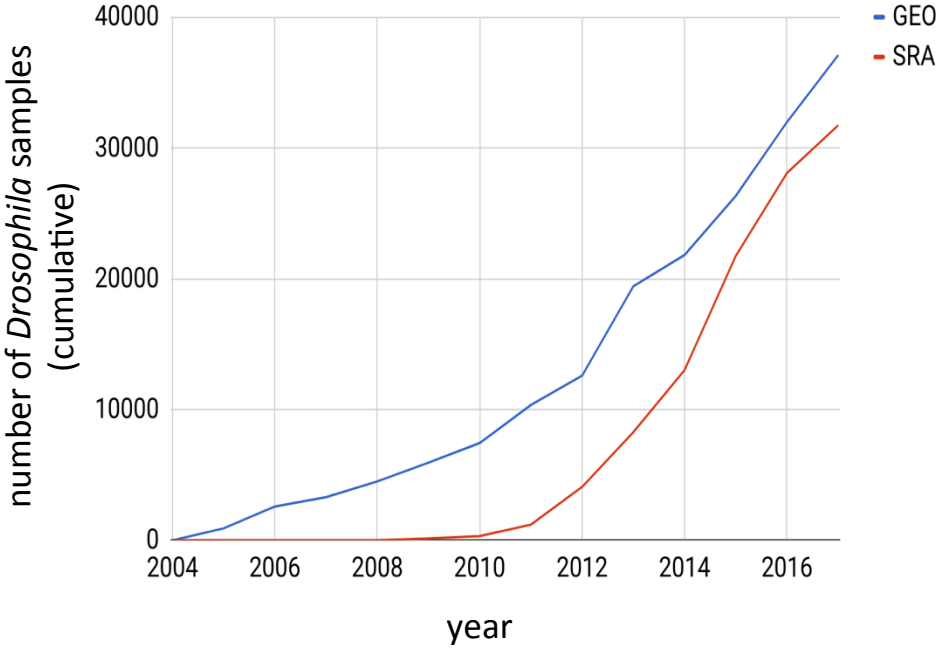
What's in there?

- Gene expression in "**normal**" tissues.
- Gene expression in "**perturbed**" tissues.
- Genome binding profiles.
- Nucleotide variation.

Why do we care?

- Insights into various genes and processes.
- Opportunity for meta-analysis discovery.

Bringing "big data" to FlyBase



FlyBase wants to facilitate:

- Discovery of key results.
- Re-use of data.

Challenges:

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

The limitations of "free-text" metadata

search term	GEO hits (*D. melanogaster*)
"fat body"	372 (338 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, perturbed?

The limitations of "free-text" metadata

search term	GEO hits (*D. melanogaster*)
"fat body"	372 (338 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, perturbed?

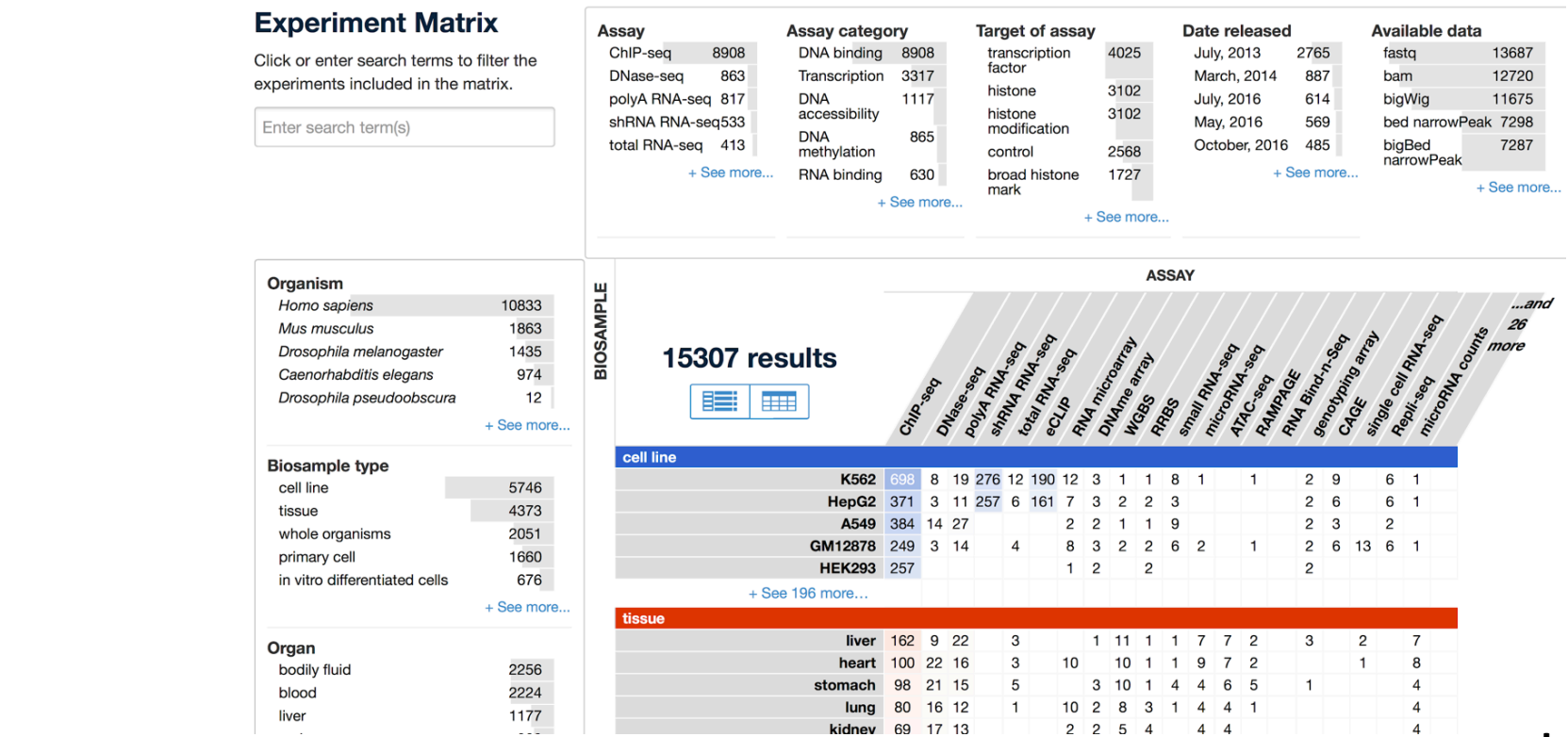
search term	GEO hits (*D. melanogaster*)
"nej" or "nejire"	17 (4 unique to this term)
"CBP"	142 (129 unique to this term)
"nej" or "nejire" or CBP"	146

Search term ambiguity:
CBP could mean CG15319 or CG1435.

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

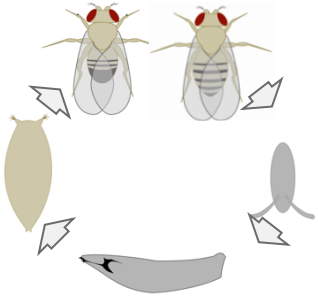
The advantage of structured metadata

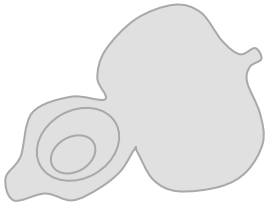
ENCODE project: a model of how structured metadata allows for powerful searching.

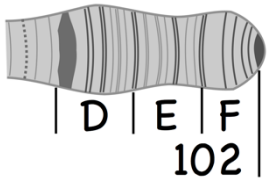


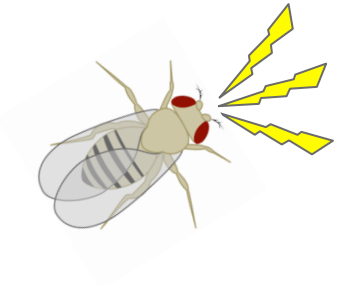
encodeproject.org

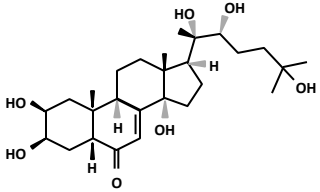
FlyBase as a portal for fly datasets









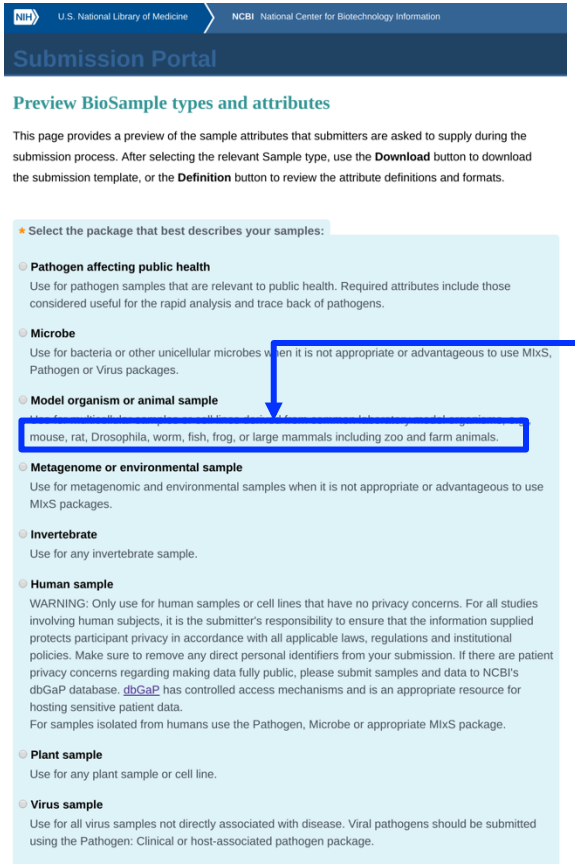


Goal:
Create an indexed catalog of *Drosophila* datasets.

In progress:
Standardize experimental descriptions,
focus on biological sample descriptions.

Long term:
Improved search/browse capabilities.
Support large-scale data re-analysis.

Drosophila BioSample template



In progress:

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

The Goal:

Make it available at NCBI during submission of fly data.

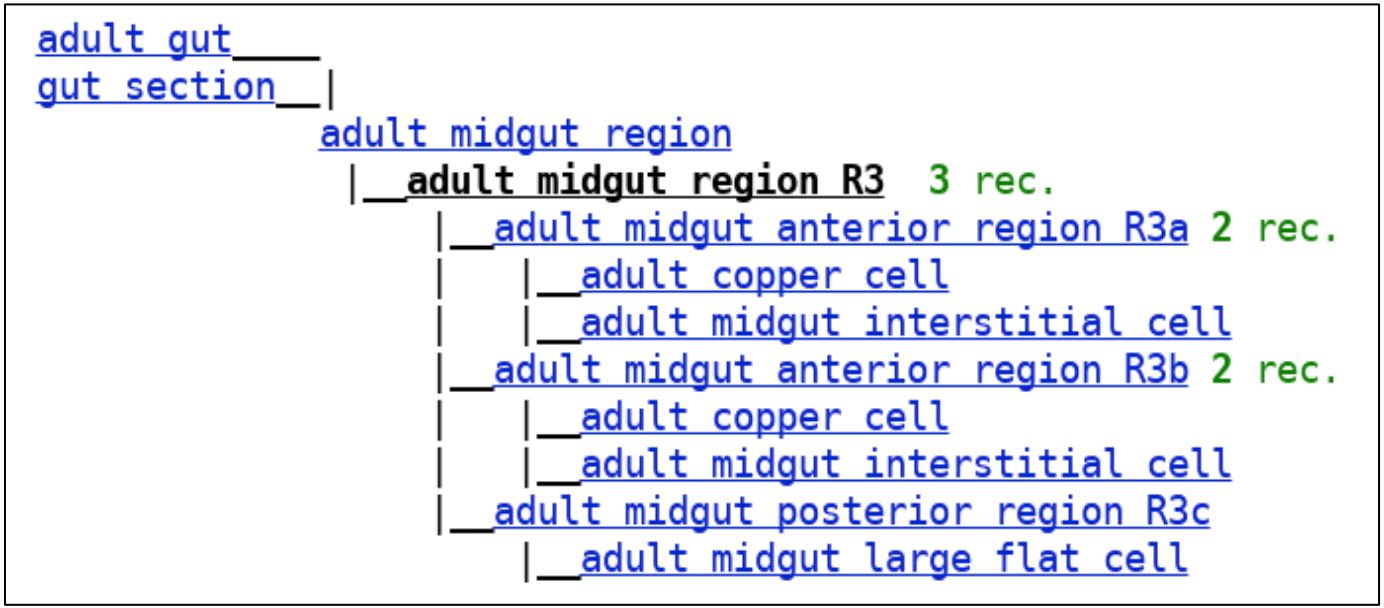
Features:

- Simplified (fewer fields than generic template).
- Covers common aspects of fly studies.
- Improved consistency for **machine** readability.

Drosophila BioSample template: tissues

A focus on tissues:

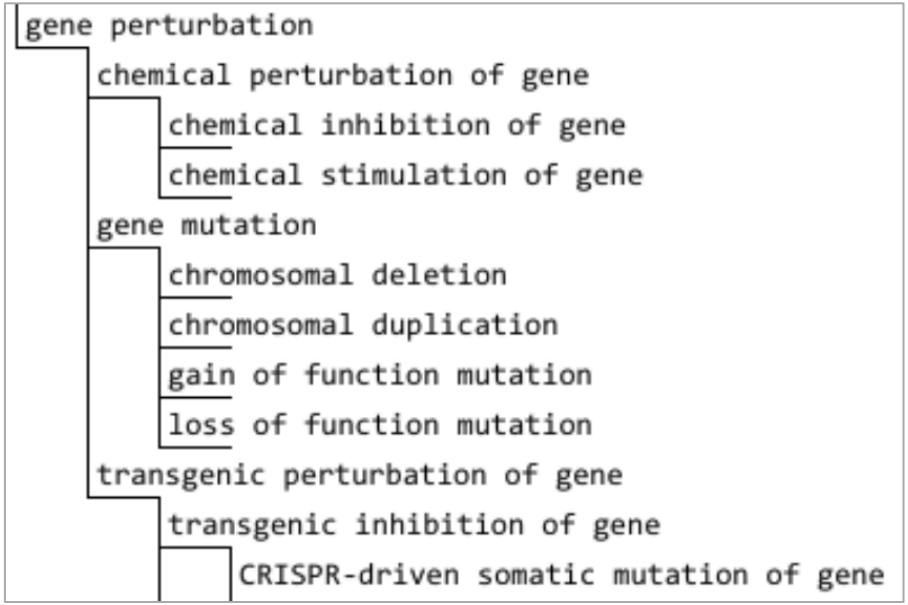
- Encourage submitters to use FlyBase's controlled anatomy terms.
- Distinguish the tissue **isolated** from the tissue **perturbed** (these often differ).



Drosophila BioSample template: genes

A focus on genes:

- Specify genes that are key experimental variables.
- Specify **how** these genes are manipulated, using keywords (under development).



Dataset report: project overview

General Information			
Name	Furlong_ChIP-chip	**Species**	*D. melanogaster*
Project type	genome binding	**FlyBase ID**	FBlc0003340
Parent Project		**Data Provider**	
Title	ChIP-chip identification of binding sites for transcription factors that regulate mesodermal development.		
Accessions	E-TABM-652		
Overview			
Description			
Project attributes			
Biosample type	whole organism		(*Ciglar et al., 2014*, *Rembold et al., 2014*, *Junion et al., 2012*, *Zinzen et al., 2009*)
Assay type	ChIP-chip		(*Ciglar et al., 2014*, *Rembold et al., 2014*, *Junion et al., 2012*, *Zinzen et al., 2009*)
Reagent type			
Result type	binding site identification		(*Ciglar et al., 2014*, *Rembold et al., 2014*, *Junion et al., 2012*, *Zinzen et al., 2009*)
Key genes	pan, lmd, tin, ttk, Mef2, bin, sna, Doc2, bap, Mad, twi, pnr, slp1		(*bait_protein*) Export to HitList
GO term(s)			
SO term(s)	TF_binding_site		
Details			
Sample preparation			
Protocol	Chromatin was prepared from formaldehyde-fixed embryos as previously described (FBrf0202481).		(*Bonn et al., 2012*)
Mode of Assay	Immunoprecipitated chromatin was analyzed by Affymetrix GeneChip Drosophila Tiling 1.0R array		(*Bonn et al., 2012*)
Data analysis	Peaks were called with TileMap software (v1) after quantile normalization. For each enriched region, probe intensity peaks were identified as extrema on a smoothed curve of the log2-ratio signal. Finally, 200 bp regions centered on identified extremas were defined. Peak calls were lifted over from Dmel_Release_5 to Dmel_Release_6 using the NCBI Genome Remapping tool.		(*Bonn et al., 2012*)
Comments			
Files	GFF3 (zip, 1.5 MB)		
Additional Information			

data repository record

sample types and methods

key genes (and their study role)

biological processes and sequence types

detailed experimental methods

Dataset report: biosamples, assays, and results

Related Datasets

Biosamples generated (28) Export to HitList

Biosample	Type	Title
BS_Furlong_bap_E6-8h_organism	whole organism	D. melanogaster, embryo (6-8 hr AEL), source for chromatin.
BS_Furlong_bin_E10-12h_organism	whole organism	D. melanogaster, embryo (10-12 hr AEL), source for chromatin.
BS_Furlong_bin_E6-8h_organism	whole organism	D. melanogaster, embryo (6-8 hr AEL), source for chromatin.
BS_Furlong_bin_E8-10h_organism	whole organism	D. melanogaster, embyro (8-10 hr AEL), source for chromatin.
BS_Furlong_Mef2_E10-12h_organism	whole organism	D. melanogaster, embyro (10-12 hr AEL), source for chromatin.
BS_Furlong_Mef2_E2-4h_organism	whole organism	D. melanogaster, embyro (2-4 hr AEL), source for chromatin.
BS_Furlong_Mef2_E4-6h_organism	whole organism	D. melanogaster, embyro (4-6 hr AEL), source for chromatin.
BS_Furlong_Mef2_E6-8h_organism	whole organism	D. melanogaster, embyro (6-8 hr AEL), source for chromatin.
BS_Furlong_Mef2_E8-10h_organism	whole organism	D. melanogaster, embyro (8-10 hr AEL), source for chromatin.
BS_Furlong_tin_E2-4h_organism	whole organism	D. melanogaster, embyro (2-4 hr AEL), source for chromatin.

Showing 10 / 28 records. Use *Export to HitList* above to see all

Raw data generated (28) Export to HitList

Assay	Type	Title
ChIP-chip_Furlong_bap_E11-12	ChIP-chip	ChIP-chip of bap from D. melanogaster, embryo (6-8 hr AEL).
ChIP-chip_Furlong_bin_E11-12	ChIP-chip	ChIP-chip of bin from D. melanogaster, embryo (6-8 hr AEL).
ChIP-chip_Furlong_bin_E12-13	ChIP-chip	ChIP-chip of bin from D. melanogaster, embyro (8-10 hr AEL).
ChIP-chip_Furlong_bin_E13-15	ChIP-chip	ChIP-chip of bin from D. melanogaster, embryo (10-12 hr AEL).
ChIP-chip_Furlong_Mef2_E11-12	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embyro (6-8 hr AEL).
ChIP-chip_Furlong_Mef2_E12-13	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embyro (8-10 hr AEL).
ChIP-chip_Furlong_Mef2_E13-15	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embyro (10-12 hr AEL).
ChIP-chip_Furlong_Mef2_E4-9	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embyro (2-4 hr AEL).
ChIP-chip_Furlong_Mef2_E9-11	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embyro (4-6 hr AEL).
ChIP-chip_Furlong_tin_E4-9	ChIP-chip	ChIP-chip of tin from D. melanogaster, embyro (2-4 hr AEL).

Showing 10 / 28 records. Use *Export to HitList* above to see all

Processed data (28) Export to HitList

Result	Type	Title
ChIP-chip_bap_E6-8h_organism	binding site identification	ChIP-chip peak calls for bap from D. melanogaster, embryo (6-8 hr AEL).
ChIP-chip_bin_E10-12h_organism	binding site identification	ChIP-chip peak calls for bin from D. melanogaster, embryo (10-12 hr AEL).
ChIP-chip_bin_E6-8h_organism	binding site identification	ChIP-chip peak calls for bin from D. melanogaster, embryo (6-8 hr AEL).
ChIP-chip_bin_E8-10h_organism	binding site identification	ChIP-chip peak calls for bin from D. melanogaster, embyro (8-10 hr AEL).
ChIP-chip_Mef2_E10-12h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embyro (10-12 hr AEL).
ChIP-chip_Mef2_E2-4h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embyro (2-4 hr AEL).
ChIP-chip_Mef2_E4-6h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embyro (4-6 hr AEL).
ChIP-chip_Mef2_E6-8h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embyro (6-8 hr AEL).
ChIP-chip_Mef2_E8-10h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embyro (8-10 hr AEL).
ChIP-chip_tin_E2-4h_organism	binding site identification	ChIP-chip peak calls for tin from D. melanogaster, embyro (2-4 hr AEL).

Showing 10 / 28 records. Use *Export to HitList* above to see all

Synonyms and Secondary IDs (2)

References (6)

Biosamples
(animal → tissue sample)

Assays
(tissue → raw data)

Results
(input data → processed output data)

Dataset section (gene report)

Datasets (3)

Study focus (3)

Experimental Role	Project	Project Type	Title
bait_protein	Furlong_ChIP-chip	genome binding	ChIP-chip identification of binding sites for transcription factors that regulate mesodermal development.
bait_protein	BDTNP_TFBS	genome binding	ChIP characterization of transcription factor genome binding, Berkeley Drosophila Transcription Factor Network Project.
bait_protein	modENCODE_regulation_TFs	genome binding	Genome-wide localization of transcription factors by ChIP-chip and ChIP-Seq.

References (839)

Available now:
Capture genes that are key variables in an experiment: for example, targets of RNAi.

Coming soon:
Capturing genes that are “hits” of screens and differential expression studies.

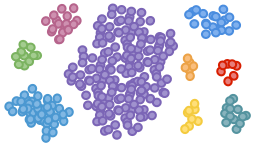
Single-cell technologies

source tissue

isolated cells

RNA-Seq reads

cell clusters



What is it?

- Sequence information is obtained for single cells:
- Deeper insights into expression and regulation.
- Sometimes, novel cell types are discovered.

Where does FlyBase fit in?

We can make it easier to find and use single-cell data, at FlyBase and beyond.

Partners:

- Fly Cell Atlas group
- Virtual Fly Brain
- EBI scAtlas

FlyBase goal: catalog single-cell datasets

source tissue

isolated cells

RNA-Seq reads

cell clusters

We want to ensure standardized descriptions of source tissue:

- Anatomy
- Developmental stage
- Cell type enrichment
- Method of cell isolation and barcoding
- Strain, genotype, perturbations (diet, chemical, etc.)

FlyBase goal: help data into repositories

source tissue

isolated cells

RNA-Seq reads

cell clusters

We want to ensure standardized descriptions of source tissue:

- Anatomy
- Developmental stage
- Cell type enrichment
- Method of cell isolation and barcoding
- Strain, genotype, perturbations (diet, chemical, etc.)

We want to 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 EBI data from FlyBase and Virtual Fly Brain.

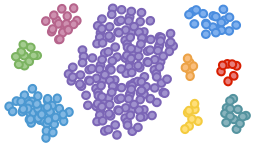
FlyBase goal: curate cell clusters

source tissue

isolated cells

RNA-Seq reads

cell clusters



We want to curate cell clusters as the key output.

Cell cluster characteristics:

- Observed cell types (in a given tissue).
- Novel cell types.
- Markers/signatures.
- Consensus gene expression profiles.
- Relationships between clusters (similarity, lineage).
- Genes involved in cell type specification.

Methods used to identify cell clusters:

- RNA-Seq mapping (annotation set, method, depth).
- Gene expression measurement.
- Clustering method.

The Fly Cell Atlas community

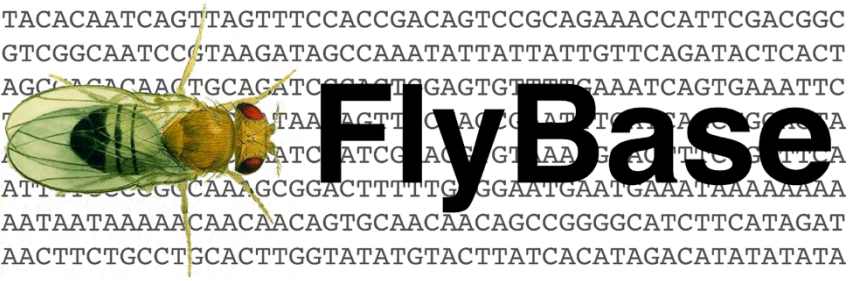
Join the Fly Cell Atlas community (Slack channel):
flycellatlas.org

Watch for the single-cell workshop at ADRC 2019.

Producing scRNA-Seq data? Please submit to **EBI scAtlas** (via ArrayExpress).
ebi.ac.uk/arrayexpress/

Please share your preprints with us:

Slack channel:
flycellatlas.slack.com#data-submission
FlyBase contact:
dossantos@morgan.harvard.edu
Virtual Fly Brain contact:
virtualflybrain@googlegroups.com



@FlyBaseDotOrg

Thank you for your support.

FlyBase TV

alliancegenome.org