



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

Gil dos Santos, FlyBase

[dossantos@morgan.harvard.edu](mailto:dossantos@morgan.harvard.edu)

[helpfb@morgan.harvard.edu](mailto:helpfb@morgan.harvard.edu)



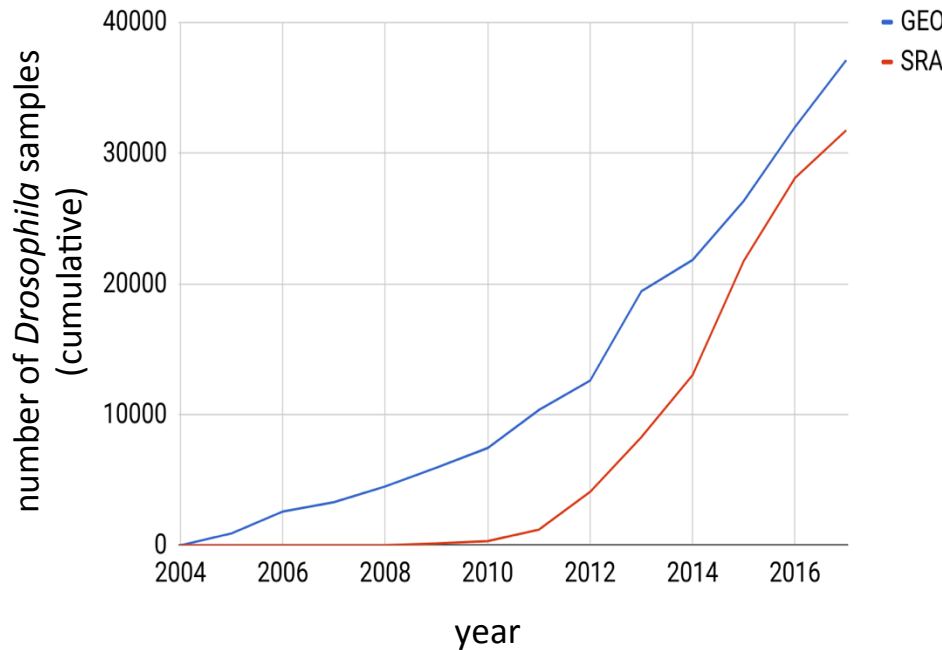
Genetics Society of America



Annual Drosophila  
Research Conference



# "big data": why do we care?



Thousands of samples at GEO & SRA.

## What's in there?

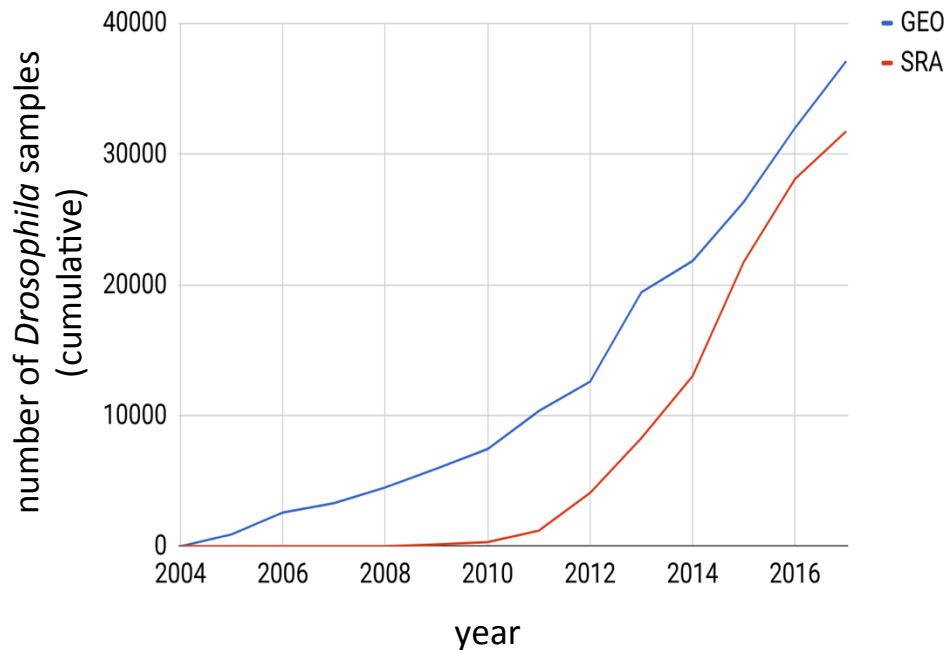
- Gene expression in "normal" tissues.
- Gene expression in "perturbed" tissues.
- Genome binding profiles.
- Nucleotide variation.

## 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 ( <i>D. melanogaster</i> )
"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 ( <i>D. melanogaster</i> )
"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 ( <i>D. melanogaster</i> )
"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.

## Experiment Matrix

Click or enter search terms to filter the experiments included in the matrix.

### Organism

<i>Homo sapiens</i>	10833
<i>Mus musculus</i>	1863
<i>Drosophila melanogaster</i>	1435
<i>Caenorhabditis elegans</i>	974
<i>Drosophila pseudoobscura</i>	12

[+ See more...](#)

### Biosample type

cell line	5746
tissue	4373
whole organisms	2051
primary cell	1660
in vitro differentiated cells	676

[+ See more...](#)

### Organ

bodily fluid	2256
blood	2224
liver	1177

Assay	Assay category	Target of assay	Date released	Available data
ChIP-seq 8908	DNA binding 8908	transcription factor 4025	July, 2013 2765	fastq 13687
DNase-seq 863	Transcription 3317	histone 3102	March, 2014 887	bam 12720
polyA RNA-seq 817	DNA accessibility 1117	histone modification 3102	July, 2016 614	bigWig 11675
shRNA RNA-seq533	DNA methylation 865	control 2568	May, 2016 569	bed narrowPeak 7298
total RNA-seq 413	RNA binding 630	broad histone mark 1727	October, 2016 485	bigBed narrowPeak 7287

[+ See more...](#)

[+ See more...](#)

[+ See more...](#)

[+ See more...](#)

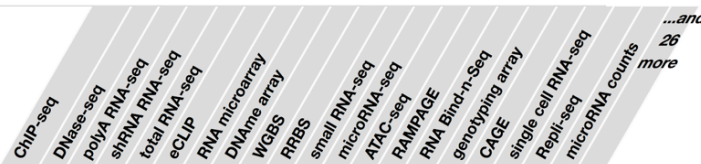
[+ See more...](#)

BIOSAMPLE

15307 results



ASSAY



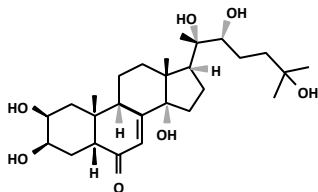
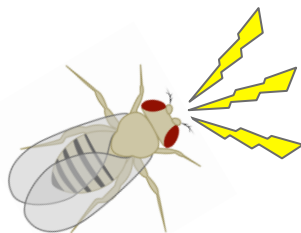
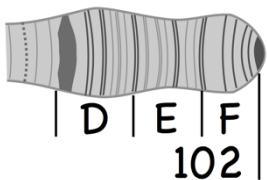
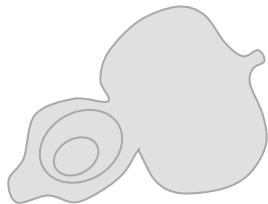
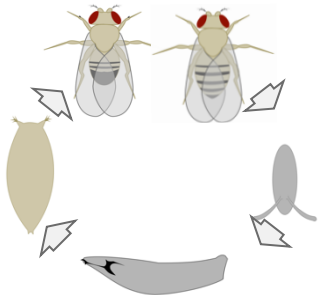
cell line		K562	698	8	19	276	12	190	12	3	1	1	8	1	1	2	9	6	1
HepG2	371	3	11	257	6	161	7	3	2	2	3					2	6	6	1
A549	384	14	27				2	2	1	1	9					2	3	2	
GM12878	249	3	14	4	8	3	2	2	6	2	1				2	6	13	6	1
HEK293	257						1	2	2						2				

[+ See 196 more...](#)

tissue		liver	162	9	22	3		1	11	1	1	7	7	2	3	2	7
heart	100	22	16	3	10	10	1	1	9	7	2				1		8
stomach	98	21	15	5		3	10	1	4	4	6	5		1			4
lung	80	16	12	1	10	2	8	3	1	4	4	1					4
kidney	69	17	13			2	2	5	4	4	4						4

TACCAATCAGTTAGTTTC  
GTCCGCAATCCSTAAGATF  
AGCAGCAATGCGAGATC  
TAF  
ATTC  
ATTTCCGGCAAGCGC  
AATAATAAAAACACACAC  
AACTTCTGCCTGCACCTTC

# FlyBase as a portal for fly datasets



## Goal:

Create an indexed catalog of *Drosophila* datasets.

## In progress:

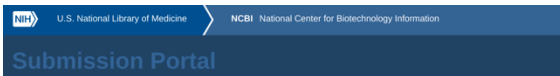
Standardize experimental descriptions,  
focus on biological sample descriptions.

## Long term:

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

TGCACATCAGTGTAGTTTC  
GTCCGGCAATCCSTAAGATF  
AGCCAGCAATGTCAGATC  
TAA  
ATTTCCGGCCAAAGCGG  
AATAATAAAAACAACAC  
AACTTCTGCCTGCATTGC

# Drosophila BioSample template



## Preview BioSample types and attributes

This page provides a preview of the sample attributes that submitters are asked to supply during the submission process. After selecting the relevant Sample type, use the **Download** button to download the submission template, or the **Definition** button to review the attribute definitions and formats.

★ Select the package that best describes your samples:

### Pathogen affecting public health

Use for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

### Microbe

Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use MixS, Pathogen or Virus packages.

### Model organism or animal sample

mouse, rat, Drosophila, worm, fish, frog, or large mammals including zoo and farm animals.

### Metagenome or environmental sample

Use for metagenomic and environmental samples when it is not appropriate or advantageous to use MixS packages.

### Invertebrate

Use for any invertebrate sample.

### Human sample

WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is the submitter's responsibility to ensure that the information supplied protects participant privacy in accordance with all applicable laws, regulations and institutional policies. Make sure to remove any direct personal identifiers from your submission. If there are patient privacy concerns regarding making data fully public, please submit samples and data to NCBI's dbGaP database. [dbGaP](#) has controlled access mechanisms and is an appropriate resource for hosting sensitive patient data.

For samples isolated from humans use the Pathogen, Microbe or appropriate MixS package.

### Plant sample

Use for any plant sample or cell line.

### Virus sample

Use for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen: Clinical or host-associated pathogen package.

## 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).

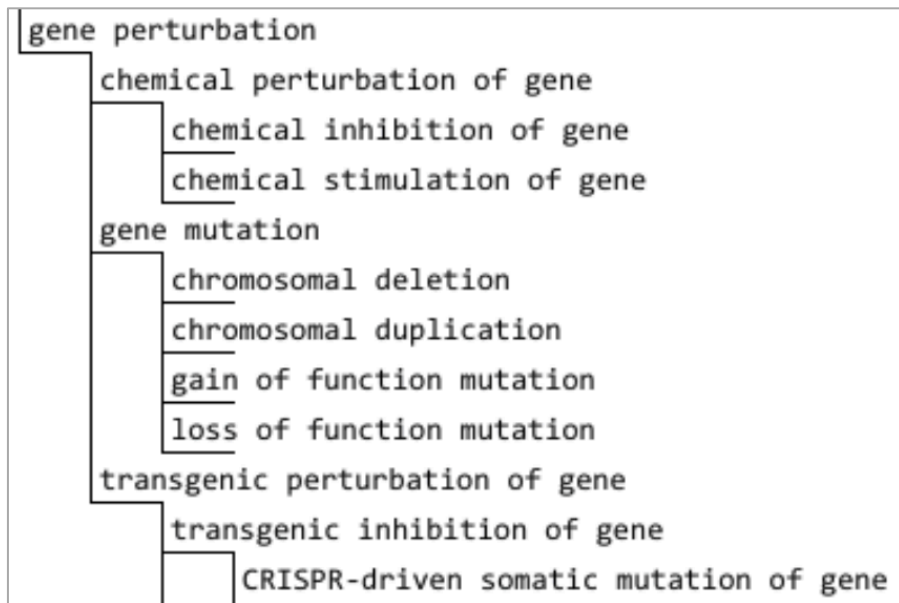
```
adult gut _____  
gut section _____ |  
    adult midgut region  
    |__ adult midgut region R3 3 rec.  
    |   |__ adult midgut anterior region R3a 2 rec.  
    |   |   |__ adult copper cell  
    |   |   |__ adult midgut interstitial cell  
    |   |__ adult midgut anterior region R3b 2 rec.  
    |   |   |__ adult copper cell  
    |   |   |__ adult midgut interstitial cell  
    |   |__ adult midgut posterior region R3c  
    |   |__ adult midgut large flat cell
```



# 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	<i>D. melanogaster</i>
Project type	genome binding	FlyBase ID	FBic0003340
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	<i>(Ciglar et al., 2014, Rembold et al., 2014, Junion et al., 2012, Zinzen et al., 2009)</i>	
Assay type	ChIP-chip	<i>(Ciglar et al., 2014, Rembold et al., 2014, Junion et al., 2012, Zinzen et al., 2009)</i>	
Reagent type			
Result type	binding site identification	<i>(Ciglar et al., 2014, Rembold et al., 2014, Junion et al., 2012, Zinzen et al., 2009)</i>	
Key genes	pan lmd tin ttk	Mef2 bin sna	Doc2 bap Mad
GO term(s)			
SO term(s)	TF_binding_site		
Details			
Sample preparation			
Protocol	Chromatin was prepared from formaldehyde-fixed embryos as previously described <i>(FBrf0202481)</i> . <i>(Bonn et al., 2012)</i>		
Mode of Assay	Immunoprecipitated chromatin was analyzed by Affymetrix GeneChip Drosophila Tiling 1.0R array <i>(Bonn et al., 2012)</i>		
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. <i>(Bonn et al., 2012)</i>		
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) <span>Export to HitList</span>		
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, embryo (8-10 hr AEL), source for chromatin.
BS_Furlong_Mef2_E10-12h_organism	whole organism	D. melanogaster, embryo (10-12 hr AEL), source for chromatin.
BS_Furlong_Mef2_E2-4h_organism	whole organism	D. melanogaster, embryo (2-4 hr AEL), source for chromatin.
BS_Furlong_Mef2_E4-6h_organism	whole organism	D. melanogaster, embryo (4-6 hr AEL), source for chromatin.
BS_Furlong_Mef2_E6-8h_organism	whole organism	D. melanogaster, embryo (6-8 hr AEL), source for chromatin.
BS_Furlong_Mef2_E8-10h_organism	whole organism	D. melanogaster, embryo (8-10 hr AEL), source for chromatin.
BS_Furlong_tin_E2-4h_organism	whole organism	D. melanogaster, embryo (2-4 hr AEL), source for chromatin.
Showing 10 / 28 records. Use <a href="#">Export to HitList</a> above to see all		
Raw data generated (28) <span>Export to HitList</span>		
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, embryo (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, embryo (6-8 hr AEL).
ChIP-chip_Furlong_Mef2_E12-13	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embryo (8-10 hr AEL).
ChIP-chip_Furlong_Mef2_E13-15	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embryo (10-12 hr AEL).
ChIP-chip_Furlong_Mef2_E4-9	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embryo (2-4 hr AEL).
ChIP-chip_Furlong_Mef2_E9-11	ChIP-chip	ChIP-chip of Mef2 from D. melanogaster, embryo (4-6 hr AEL).
ChIP-chip_Furlong_tin_E4-9	ChIP-chip	ChIP-chip of tin from D. melanogaster, embryo (2-4 hr AEL).
Showing 10 / 28 records. Use <a href="#">Export to HitList</a> above to see all		
Processed data (28) <span>Export to HitList</span>		
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, embryo (8-10 hr AEL).
ChIP-chip_Mef2_E10-12h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embryo (10-12 hr AEL).
ChIP-chip_Mef2_E2-4h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embryo (2-4 hr AEL).
ChIP-chip_Mef2_E4-6h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embryo (4-6 hr AEL).
ChIP-chip_Mef2_E6-8h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embryo (6-8 hr AEL).
ChIP-chip_Mef2_E8-10h_organism	binding site identification	ChIP-chip peak calls for Mef2 from D. melanogaster, embryo (8-10 hr AEL).
ChIP-chip_tin_E2-4h_organism	binding site identification	ChIP-chip peak calls for tin from D. melanogaster, embryo (2-4 hr AEL).
Showing 10 / 28 records. Use <a href="#">Export to HitList</a> 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	<a href="#">Furlong_ChIP-chip</a>	<a href="#">genome binding</a>	ChIP-chip identification of binding sites for transcription factors that regulate mesodermal development.
bait_protein	<a href="#">BDTNP_TFBS</a>	<a href="#">genome binding</a>	ChIP characterization of transcription factor genome binding, Berkeley Drosophila Transcription Factor Network Project.
bait_protein	<a href="#">modENCODE_regulation_TFs</a>	<a href="#">genome binding</a>	Genome-wide localization of transcription factors by ChIP-chip and ChIP-Seq.

+ References (839)

## Available now:

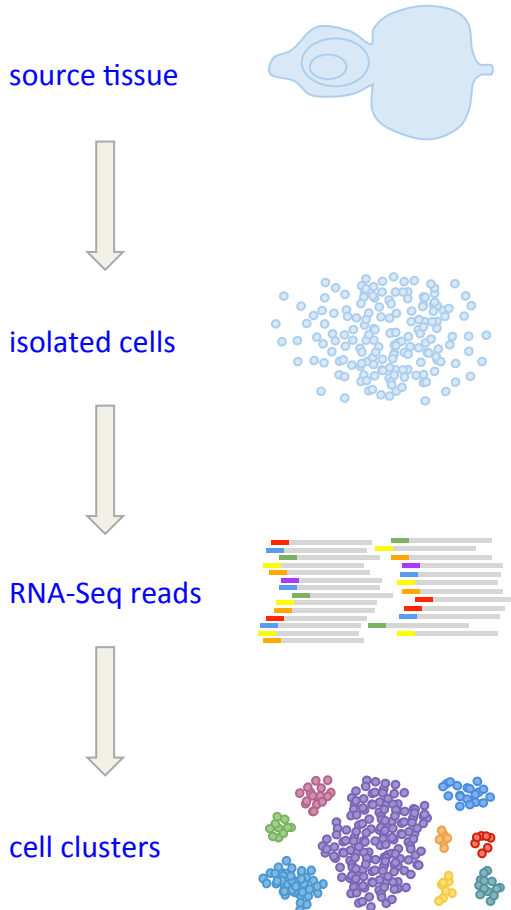
Capture genes that are key variables in an experiment: for example, targets of RNAi.

## Coming soon:

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

TACCAATCAGTAGTTTC  
GTCGGCAATCCSTAAGATF  
AGCAGCAAGTGCAGATC  
TAF  
ATC  
ATTTCCGGCCAAAGCGC  
AATAATAAAAACAACAC/  
AACTTCTGCCTGCACTTGC

# Single-cell technologies



## 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?

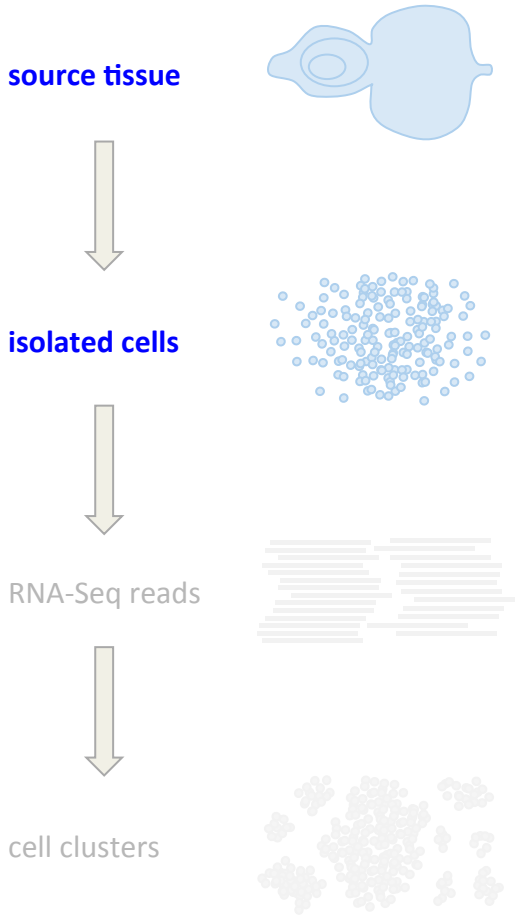
We can make it easier to find and use single-cell data, at FlyBase and beyond.

## Partners:

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

TACCAATCAGTAGTTTC  
GTCGGCAATCC TAAGATF  
AGCAGCAAGTGCAGATC  
TAF  
ATTTCCGGCAAAGCGC  
AATAATAAAA CAACAACT  
AACTTCTGCCCTGCATTGC

# FlyBase goal: catalog single-cell datasets

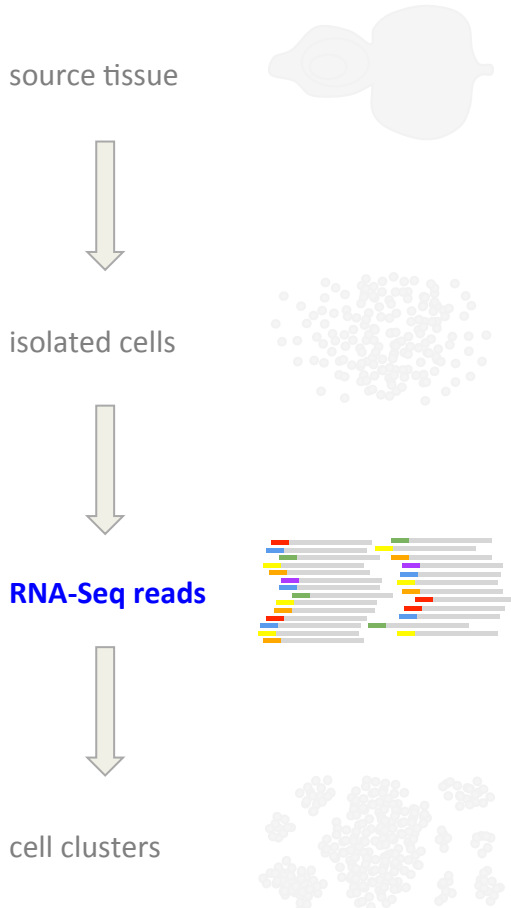


**We want to ensure standardized descriptions of source tissue:**

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



# FlyBase goal: help data into repositories



**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.)

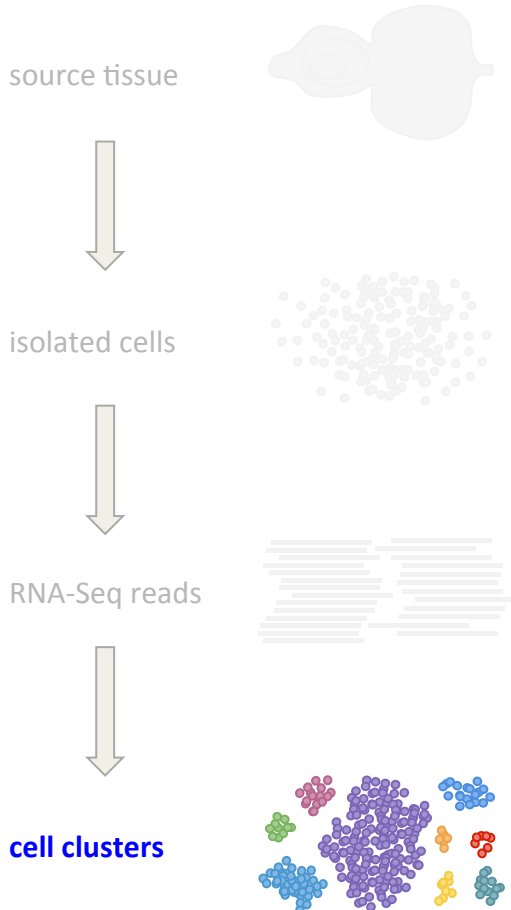
**We want to ensure proper data formatting at data repositories:**

- Facilitate data re-use.
- **Harmonize 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.



TACCAATCAGT TAGTTTC  
GTCGGCAATCC TAAGATF  
AGCAGCAATGCCAGATC  
TAA  
ATTC  
ATTTCCGGCCAAAGCGC  
AATAATAAAAA CAACAACT  
AACTTCTGCCTGCCTTC

# FlyBase goal: curate 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.

TACCAATCAGT TAGTTTC  
GTCGGCAATCC TAAGATF  
AGCAGCAATGCCAGATC  
ATAF  
ATC  
ATTTCCGGCAAAGCGC  
AATAATAAAAA CAACAACT  
AACTTCTGCCTCACATTGC

# The Fly Cell Atlas community

Join the Fly Cell Atlas community (Slack channel):

[flycellatlas.org](https://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/](https://ebi.ac.uk/arrayexpress/)

Please share your preprints with us:

Slack channel:

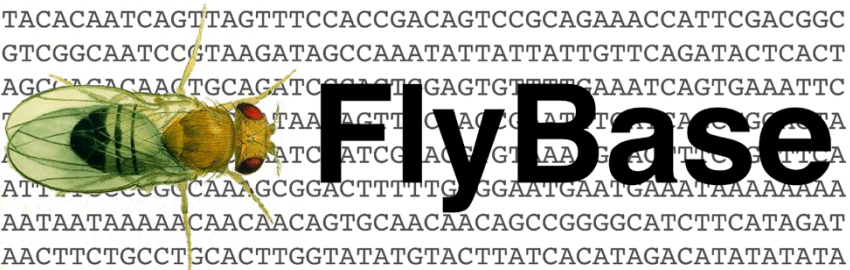
[flycellatlas.slack.com#data-submission](https://flycellatlas.slack.com/#data-submission)

FlyBase contact:

[dossantos@morgan.harvard.edu](mailto:dossantos@morgan.harvard.edu)

Virtual Fly Brain contact:

[virtualflybrain@googlegroups.com](mailto:virtualflybrain@googlegroups.com)



@FlyBaseDotOrg

Thank you for your support.



FlyBase TV



[alliancegenome.org](http://alliancegenome.org)