

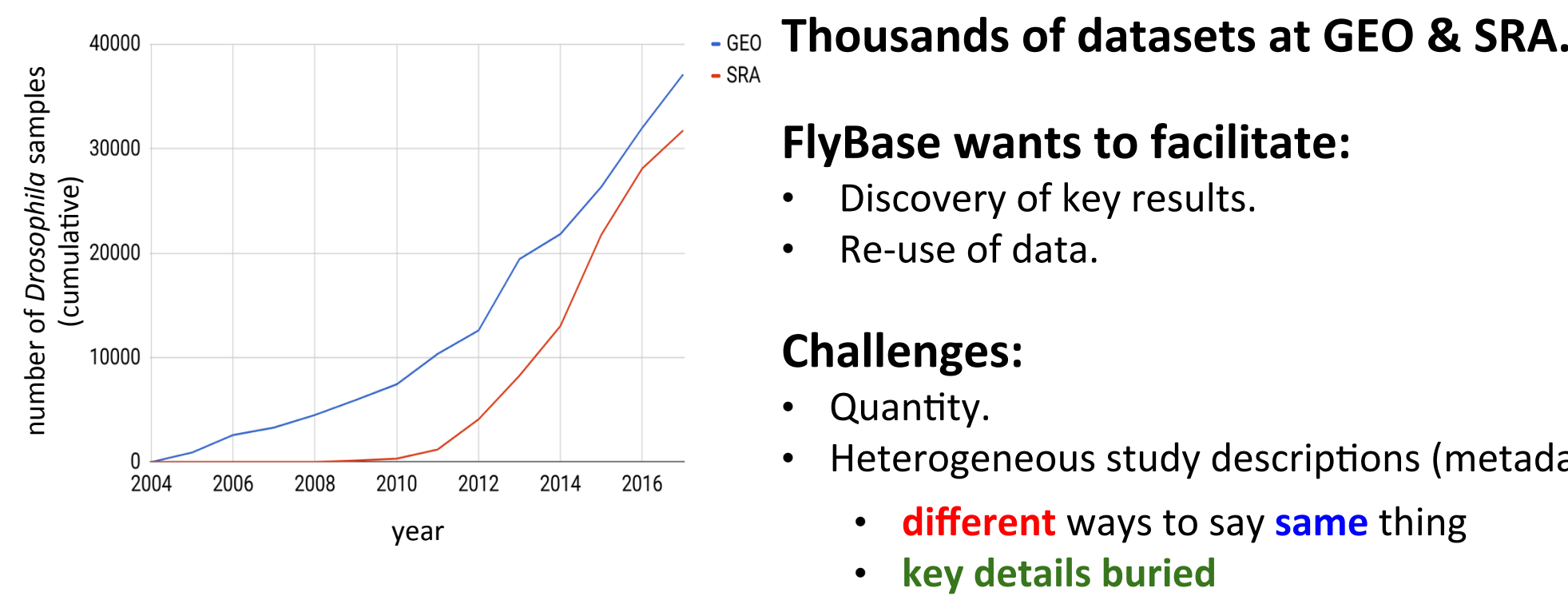
FlyBase: large-scale datasets and single-cell technologies

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

GTCGGCAATCCGTAAGAT
 ACCGCAATGTCAGAT
 ATATTCCCGC
 CAAAGCG
 AATAATAAAAACAACAAC

Introduction

Bringing "big data" to FlyBase



The limitations of "free-text" metadata

Unstructured experimental descriptions make finding the right datasets difficult.

search term	GEO hits (<i>D. melanogaster</i>)
"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 (<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 C1435.

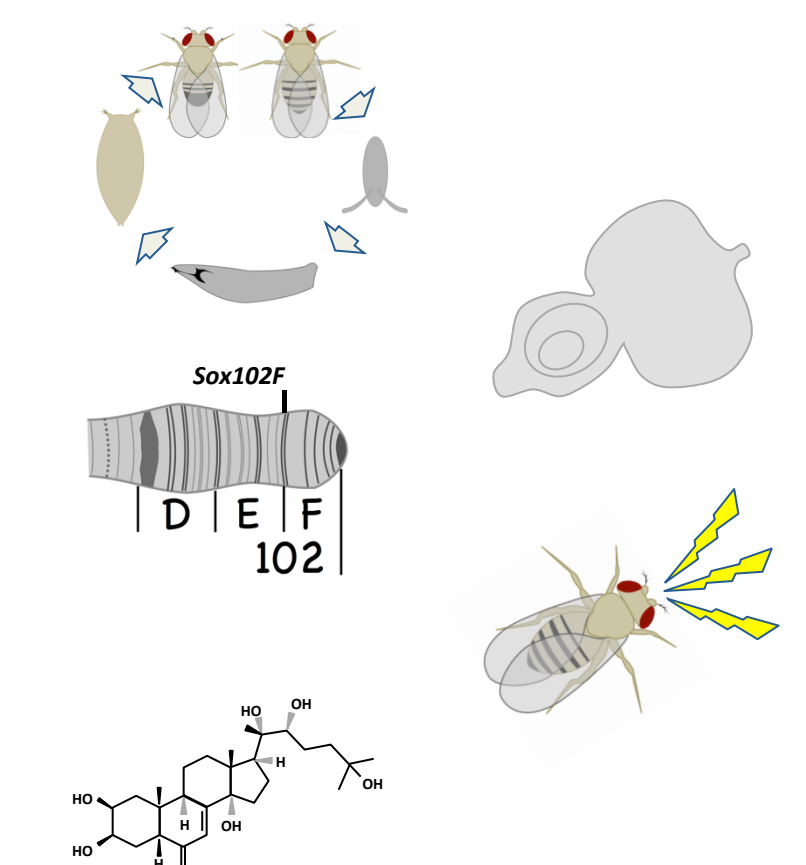
Context unclear:
CBP a target for ChIP, or RNAi?

The advantage of structured metadata

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

experimentproject.org

FlyBase as a portal for fly datasets



Goal:
Catalog datasets using standardized experimental descriptions.

In progress:
Standardize biological sample descriptions.

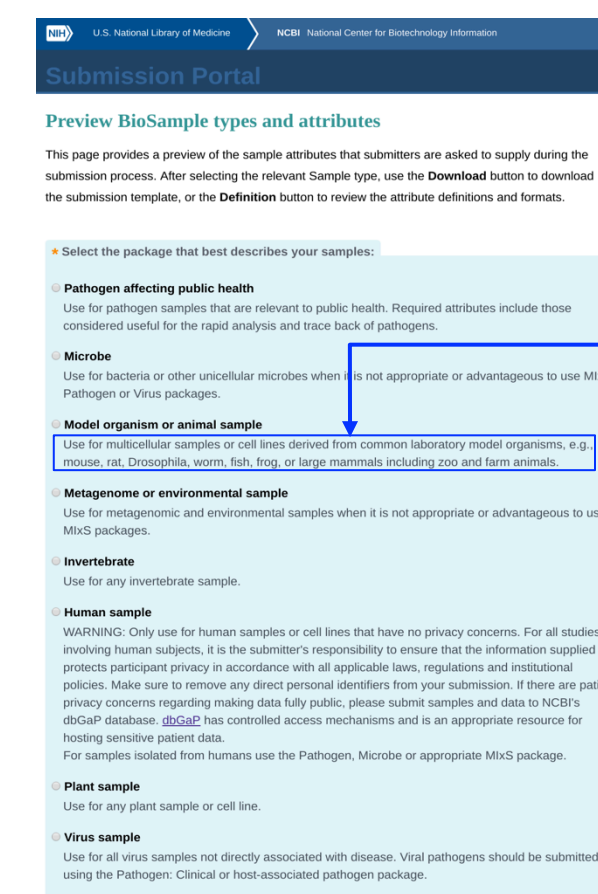
Long term:
Improved search/browse capabilities.

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.

Drosophila BioSample template for NCBI submission

Motivation



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

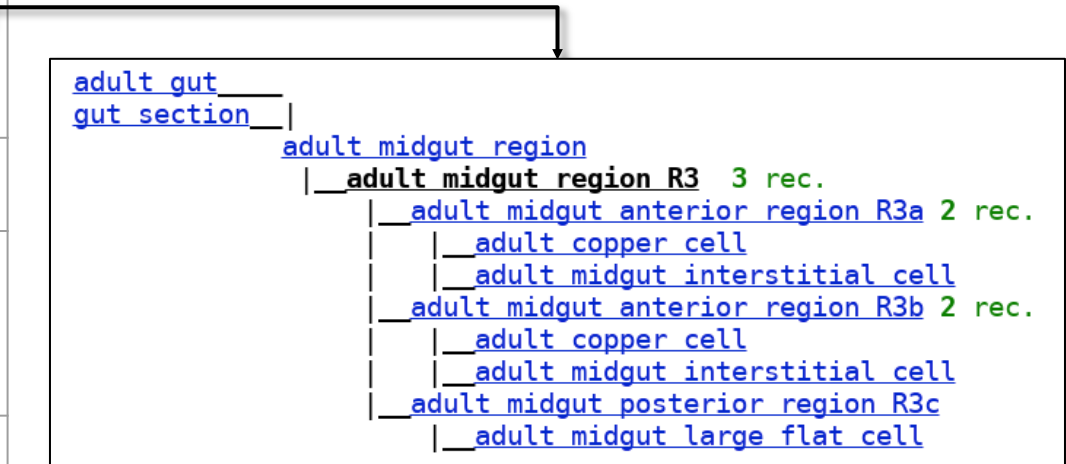
Goal:
Make it available at NCBI during submission of *Drosophila* data.

- 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

task_list (key)	status	guidelines	term list (value)
adult_type	optional	Sample types such as cell cultures, mixed cultures, tissue samples, whole organisms, single cell, developmental stage.	FlyBase anatomy CV
adult_tissue	mandatory set	This field contains the names of the regions or tissues from which the sample was taken. This field includes the entire organism, or body part (e.g. "head", "eye", "heart", "gut", "wing", "leg", "antenna", "testis", "ovary", "embryo", "larva", "pupa", "adult", "post-embryo", etc.) or a subset of these (e.g., "embryo + heart" or "antenna + wing").	FlyBase anatomy CV
adult_age	mandatory set	This field contains the age of the sample (e.g., "embryo", "larva", "pupa", "adult", "post-embryo", etc.).	FlyBase anatomy CV
tissue_perturbation	optional	Indicate the tissue that was perturbed in the animal (adult, e.g. if a gene was knocked down in a specific tissue).	FlyBase anatomy CV
tissue_isolation	optional	Indicate the tissue that was isolated (adult, e.g. if a gene was knocked down in a specific tissue).	FlyBase anatomy CV

Encourage use of FlyBase anatomy CV.



Distinguish between tissues isolated and tissues perturbed.

Clear listing of genes and perturbations

task_list (key)	status	guidelines	term list (value)
target_gene_listed	optional	The number or identifier (Fly ID) of a <i>Drosophila</i> gene. Genes are listed to indicate the gene(s) represented in the experiment. Genes not listed here are not represented in the experiment. Genes listed here are not necessarily the only genes represented in the experiment.	FlyBase
target_gene_perturbed	optional	The number or identifier (Fly ID) of a <i>Drosophila</i> gene. Genes are listed to indicate the gene(s) represented in the experiment. Genes not listed here are not represented in the experiment. Genes listed here are not necessarily the only genes represented in the experiment.	FlyBase
target_gene_reagent	optional	The number or identifier (Fly ID) of a <i>Drosophila</i> gene. Genes are listed to indicate the gene(s) represented in the experiment. Genes not listed here are not represented in the experiment. Genes listed here are not necessarily the only genes represented in the experiment.	FlyBase

List genes that are key to experimental design.

A selection of keywords to tag experimental methods.

- gene perturbation
- chemical perturbation of gene
- chemical inhibition of gene
- gain of function mutation
- loss of function mutation
- transgenic perturbation of gene
- transgenic inhibition of gene
- CRISPR-driven somatic mutation of gene
- chromosomal deletion
- chromosomal duplication

FlyBase Dataset Reports

Dataset report

- links to original data at repositories
- simple terms that summarize sample types and methods
- key genes (and their study role)
- biological processes and feature types studied
- detailed experimental methods
- biosamples generated (sample type and cell acquisition)
- assays performed (molecular methods, raw data generated)
- results generated (input data, analysis methods, file downloads)

Dataset section (gene report)

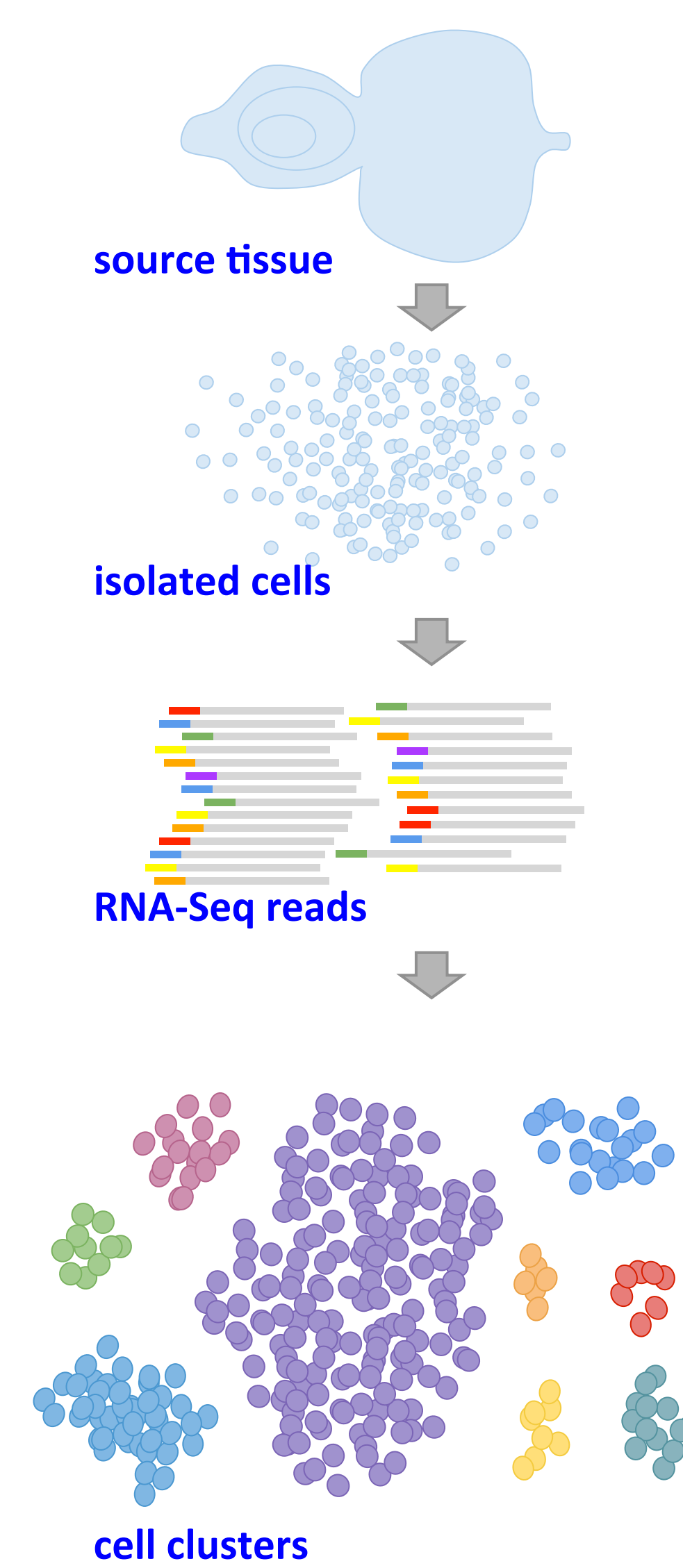
Study focus ID	Project	Project Type	Title
bait.protein	Furlong_CHIP-chip	genome binding	ChIP-chip identification of binding sites for transcription factors that regulate mesodermal development.
bait.protein	BOT1P_TBSB	genome binding	ChIP characterization of transcription factor genome binding. Berkeley Drosophila Transcription Factor Network Project.
bait.protein	modENCODE_regulation.TFs	genome binding	Genome-wide localization of transcription factors by ChIP-chip and ChIP-seq

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

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

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

Virtual Fly Brain
virtualflybrain.org

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.