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:
• Overhead/missing:
  • 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 IDs (D. melanogeter)
• “fat body” 372 (398 unique to this term)
• “fatbody” 107 (73 unique to this term)
Search term redundancy: Small variations affect results.
Context unclear: Was fat body isolated, or removed?

Search term GEO IDs (D. melanogeter)
• “CSP” or “ngene” 17 (7 unique to this term)
• “CSP” or “ngene” or “CSP” 142 (125 unique to this term)
Search term ambiguity: CSP could mean CG125320 or C1435.
Context unclear: CSP a target for ChIP, or RNAi?

The advantage of structured metadata

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

FlyBase as a portal for fly datasets

Goal: Catalog datasets using standardized experimental descriptions.
In progress: Standardize biological sample descriptions.
Long term: Improved search/browsing capabilities.

Dataset section (gene report)

A list of datasets for which a gene is a key player:
• Gene’s experimental role
• Project type.

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

Encourage use of FlyBase anatomy CV.

Clear listing of genes and perturbations

List genes that are key to experimental design.

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

Markers/signatures
Observed cell types
Cell cluster generation
Clustering method
Consensus gene expression profiles
Relationships between clusters (similarity, lineage)
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

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

Submit single-cell data? Send it to EBI scAtlas:

• Submit directly to EBI ArrayExpress
• Share preprints with FCA Slack/Flybase/Virtual Fly Brain:
  • flycellatlas.slack.com#data-submission
  • FCA Slack/Flybase/Virtual Fly Brain

References


Funding

FlyBase is supported grant #R41 HG007039 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. NO301177) and the Indiana Genomics Initiative. Hosting of this site is supported in part by the National Science Foundation (ROG-1553375) through XSEDE resources via Indiana University.