FlyBase

Improving sample descriptions at large dataset repositories
(Drosophila template for NCBI BioSample submissions)

Survey Questions December 2017

Improving sample descriptions at large dataset repositories (Drosophila template for NCBI BioSample submissions)

Introduction

Inconsistent sample descriptions make it hard to find relevant datasets at repositories like SRA (Sequence Read Archive). With this in mind, FlyBase is working with NCBI to create a Drosophila-specific BioSample submission template that provides fields more relevant to Drosophila research, and guidelines to standardize experimental descriptions. This survey asks for feedback in developing this template.

Improving sample descriptions at large dataset repositories (Drosophila template for NCBI BioSample submissions)

General Background

The following questions establish your general familiarity with and interest in next-generation sequencing data.

1. Please indicate which of the following apply to you (choose one or more).

□ I have generated or processed a sample used for next-generation sequencing (e.g., RNA-Seq).

□ I have processed and analyzed raw data output from a next-generation sequencing experiment (i.e., bioinformatics).

□ I have searched NCBI GEO, SRA or ArrayExpress for next-generation sequencing studies relevant to my research.

□ I have used FlyBase tools to assess next-generation sequencing data (e.g., RNA-Seq coverage plots, RPKM gene expression).

□ I have used tools at other (not FlyBase) websites to assess next-generation sequencing data.

□ None of the above.

□ Not sure.

2. Please indicate the data repositories to which you have made a direct submission(s) (choose one or more).

□ NCBI BioSample

□ NCBI Sequence Read Archive (SRA)

□ NCBI GEO

□ EMBL-EBI ArrayExpress

□ EMBL-EBI European Nucleotide Archive (ENA)

□ DDBJ BioSample

□ DDBJ Sequence Read Archive (DRA)

□ Other data repository.

□ None of the above - I have never submitted to a data repository.

□ Not sure.

□ Other (please indicate the data repository):

3. FlyBase wants to catalog Drosophila datasets and develop tools to facilitate the identification of relevant datasets by researchers. How would you rate the importance of this effort?

○ Very important

○ Somewhat important

○ Not sure

○ Somewhat unimportant

○ Not at all important

Improving sample descriptions at large dataset repositories (Drosophila template for NCBI BioSample submissions)

Prioritizing aspects of biosample description:

The NCBI Biosample describes how a tissue specimen was obtained (up to the point of cell lysis). Certain biosample attributes are crucial (and mandatory at NCBI) for an informative description: organism, sex, age/developmental stage, tissue (or cell line). FlyBase is considering **additional** attributes that should be emphasized in a Drosophila-specific NCBI BioSample submission template.

*** 4. From the list of 10 attributes under consideration by FlyBase (listed below), please select the five that are the most informative in an experimental description.**

□ biomarker/driver: The molecular biomarker/driver used to select cells for analysis, or the driver used to create a tissue-specific perturbation: e.g., GFP-neur; e.g., ey-GAL4; e.g., en-lacZ.

□ chemical_studied: The chemicals that are used to treat the organism, and for which a biological response is studied: e.g., ecdysone; e.g., cadmium.

□ culture_medium: In general terms, the fly or cell culture medium used (including cases where the medium is the same for control and treatment samples): e.g., M3+BPYE medium; e.g., cornmeal-yeast-molasses medium.

□ gene_manipulated: The gene(s) that is directly manipulated by some experimental technique: e.g., mutation, overexpression, RNAi, antibody blocking, chemical inhibition, epitope tagging, etc.: e.g., engrailed.

□ genotype: The genotype of the biosample (e.g., fz3(J29)/fz3(G10)), as well as details of the genetic cross used to generated the genotype.

□ methods: In general terms, the methods used to perturb the animal/cells: e.g., null mutation, RNAi, bacterial exposure, caloric restriction, etc.

□ sample_role: A simple classification of the sample as either a control or a treatment in the study.

□ sample_type: A simple classification based on the biological material: i.e., whole organism, tissue, isolated cells, single cell, primary cell line, immortalized cell line, metagenomic collection, synthetic molecules, etc.

□ temperature_regimen: The temperature at which flies were raised, whether it be at a constant temperature (e.g., 25oC), or a complex regimen used to tune, for example, inducible transgene expression.

□ tissue_perturbed: The tissue manipulated in a study (not necessarily the tissue that was harvested): e.g., a biosample of whole embryos (tissue) in which a gene was knocked down only in the mesoderm (tissue_perturbed).

5. What additional attributes, not listed above, do you value?

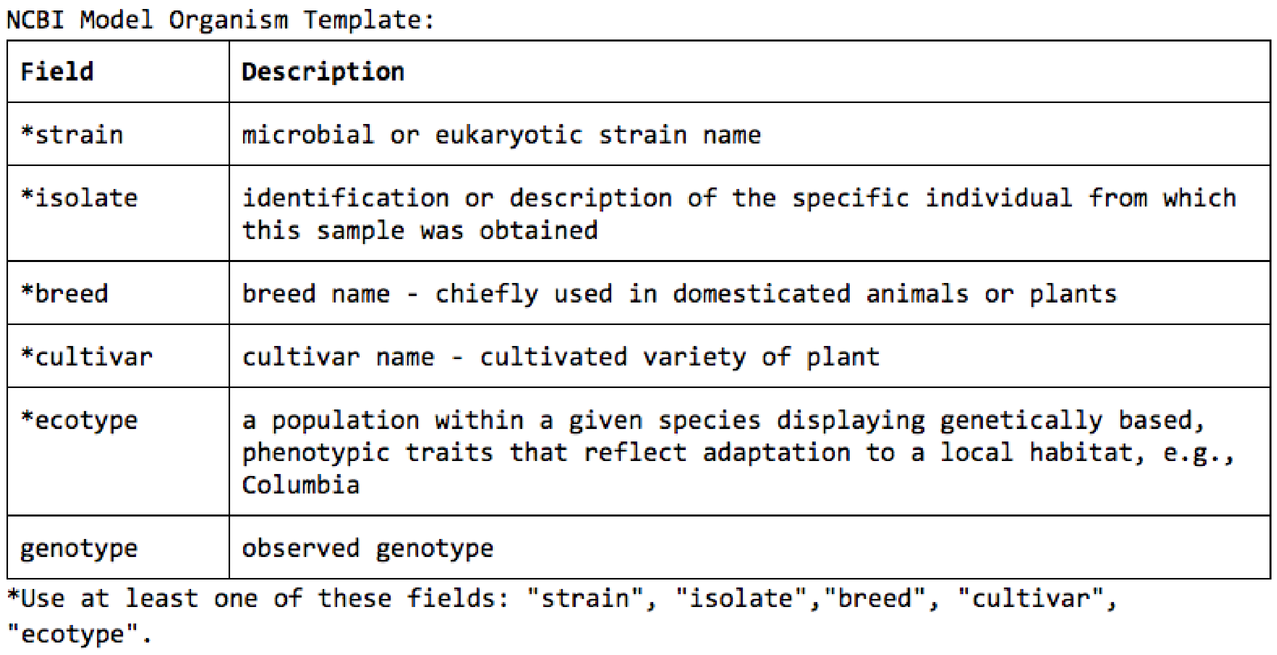
6. Do you want to answer additional questions on specific aspects of the Drosophila template being prepared for NCBI? (optional)

○ Yes

○ No

Strain and Genetic Background:

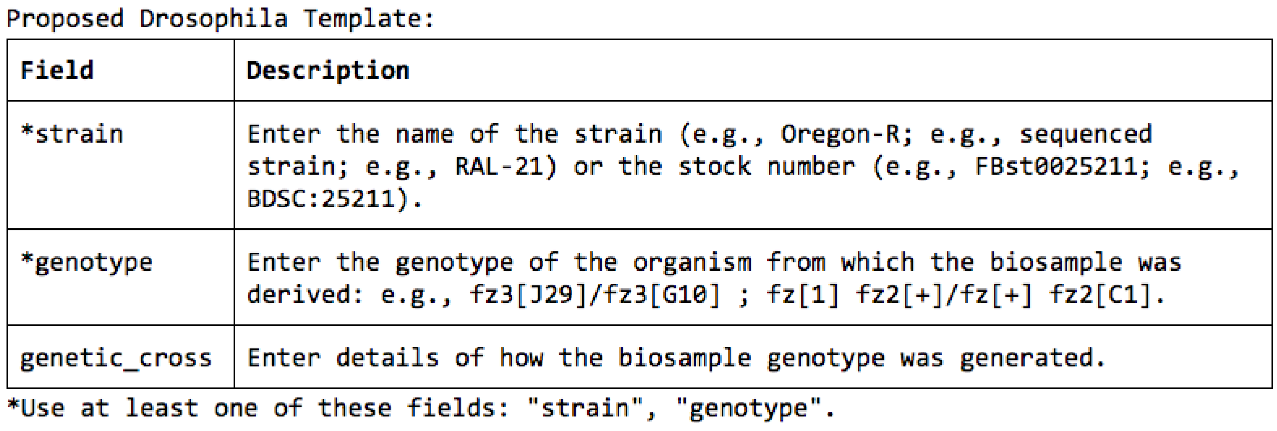
The BioSample submission template provides various fields that allow for a description of the biosample's species sub-type and/or genotype. For the following questions, please review the fields provided for this purpose in the current NCBI "Model Organism" and proposed "Drosophila" templates, shown below.



NCBI Model Organism Template:

Field	Description
*strain	microbial or eukaryotic strain name
*isolate	identification or description of the specific individual from which this sample was obtained
*breed	breed name - chiefly used in domesticated animals or plants
*cultivar	cultivar name - cultivated variety of plant
*ecotype	a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat, e.g., Columbia
genotype	observed genotype

*Use at least one of these fields: "strain", "isolate","breed", "cultivar", "ecotype".



Proposed Drosophila Template:

Field	Description
*strain	Enter the name of the strain (e.g., Oregon-R; e.g., sequenced strain; e.g., RAL-21) or the stock number (e.g., FBst0025211; e.g., BDSC:25211).
*genotype	Enter the genotype of the organism from which the biosample was derived: e.g., fz3[J29]/fz3[G10] ; fz[1] fz2[+]/fz[+] fz2[C1].
genetic_cross	Enter details of how the biosample genotype was generated.

*Use at least one of these fields: "strain", "genotype".

7. In the proposed Drosophila template, the "strain" field is retained, but other fields considered redundant or rarely applicable to Drosophila have been removed ("isolate", "breed", "cultivar" and "ecotype"). Compared to the current NCBI template, how would you rate this proposed change?

○ Great

○ A good idea, but needs improvement

○ Unhelpful

○ Not sure

Please expand on your answer (optional).

8. In the proposed Drosophila template, "strain" is no longer mandatory, and one can instead report a "genotype" for the biosample. Compared to the current NCBI template, how would you rate this proposed change?

○ Great

○ A good idea, but needs improvement

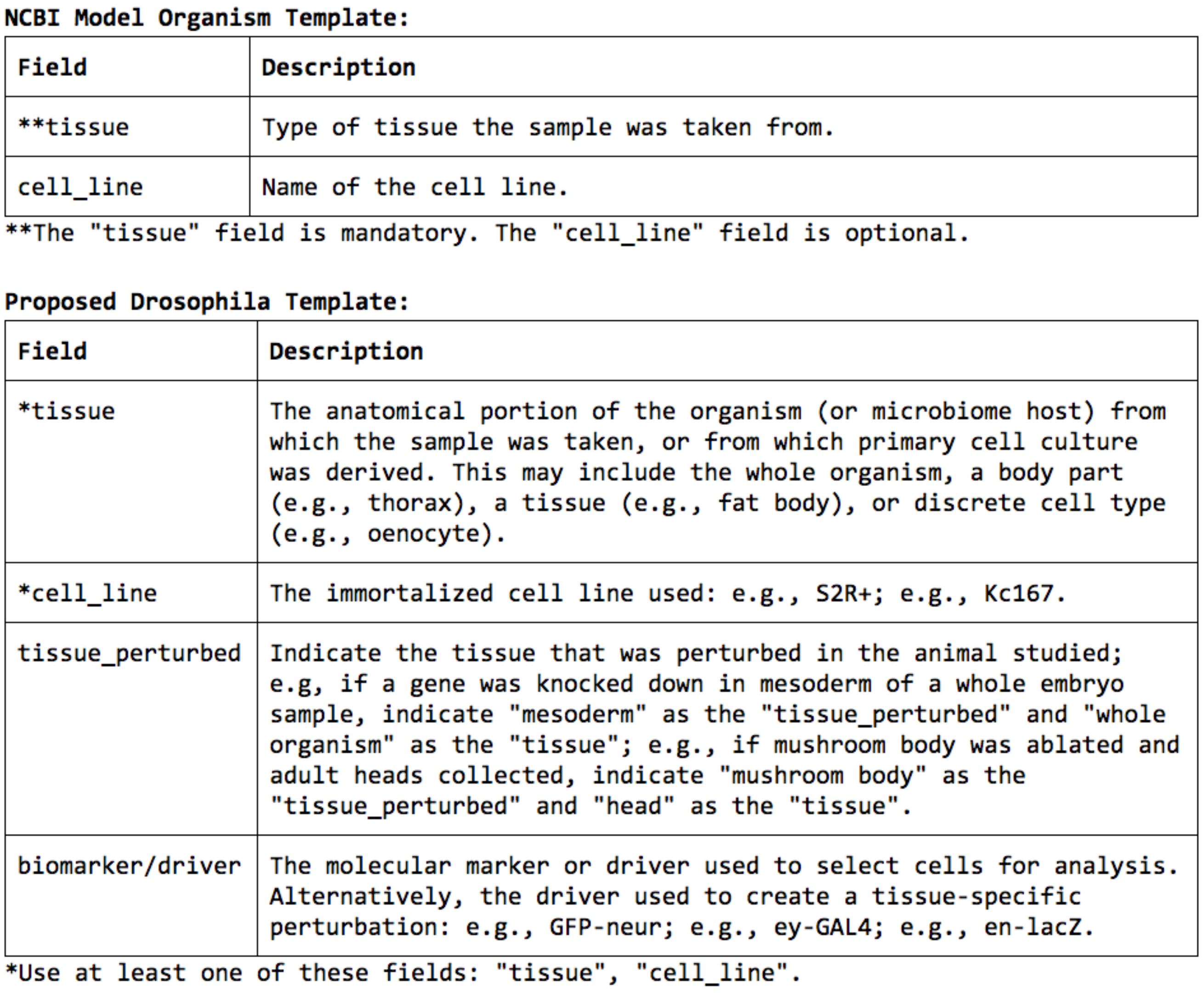
○ Unhelpful

○ Not sure

Please expand on your answer (optional).

Tissue:

One section of the BioSample submission template provides various fields that allow for a description of the biosample's tissue or cell line of origin. For the following questions, please review the set of fields provided for this section in the current NCBI "Model Organism" and proposed "Drosophila" templates, shown below.



NCBI Model Organism Template:

Field	Description
**tissue	Type of tissue the sample was taken from.
cell_line	Name of the cell line.

**The "tissue" field is mandatory. The "cell_line" field is optional.

Proposed Drosophila Template:

Field	Description
*tissue	The anatomical portion of the organism (or microbiome host) from which the sample was taken, or from which primary cell culture was derived. This may include the whole organism, a body part (e.g., thorax), a tissue (e.g., fat body), or discrete cell type (e.g., oenocyte).
*cell_line	The immortalized cell line used: e.g., S2R+; e.g., Kc167.
tissue_perturbed	Indicate the tissue that was perturbed in the animal studied; e.g, if a gene was knocked down in mesoderm of a whole embryo sample, indicate "mesoderm" as the "tissue_perturbed" and "whole organism" as the "tissue"; e.g., if mushroom body was ablated and adult heads collected, indicate "mushroom body" as the "tissue_perturbed" and "head" as the "tissue".
biomarker/driver	The molecular marker or driver used to select cells for analysis. Alternatively, the driver used to create a tissue-specific perturbation: e.g., GFP-neur; e.g., ey-GAL4; e.g., en-lacZ.

*Use at least one of these fields: "tissue", "cell_line".

9. In the proposed Drosophila template, "tissue" is no longer mandatory, and one can instead report a "cell_line" for the biosample. Compared to the current NCBI template, how would you rate this proposed change?

○ Great

○ A good idea, but needs improvement

○ Unhelpful

○ Not sure

Please expand on your answer (optional).

10. In the proposed Drosophila template, a new "tissue_perturbed" field is provided, distinct from the "tissue" field. From the descriptions provided above, is the distinction between these two fields clear?

○ Yes

○ Needs improvement

○ No

○ Not sure

Please expand on your answer (optional).

11. How would you rate this newly proposed "tissue_perturbed" field?

○ Great

○ A good idea, but needs improvement

○ Unhelpful

○ Not sure

Please expand on your answer (optional).

12. In the proposed Drosophila template, a new "biomarker/driver" field is provided. From the descriptions provided above, is the definition of this field clear?

○ Yes

○ Needs improvement

○ No

○ Not sure

Please expand on your answer (optional).

13. How would you rate this newly proposed "biomarker/driver" field?

○ Great

○ A good idea, but needs improvement

○ Unhelpful

○ Not sure

Please expand on your answer (optional).

Improving sample descriptions at large dataset repositories (Drosophila template for NCBI BioSample submissions)

Genes

Key genes in a study are not always immediately obvious (and difficult to extract automatically) from the genotype or methods provided for a biosample. FlyBase is proposing a dedicated field that clearly flags key genes that are subjected to direct experimental intervention, as defined below.

Field	**Description**
gene_manipulated	The symbol and identifier (FBgn ID) of a Drosophila gene directly subjected to experimental intervention: mutation, overexpression, RNAi, antibody blocking, chemical inhibition, epitope tagging, etc: e.g., wg (FBgn0284084).

14. To assess the clarity of the "gene_manipulated" definition, please consider a hypothetical study in which progeny of females homozygous for a gastrulation-defective null mutation (*gd*^{7}) are collected as embryos and processed for RNA-Seq. In these embryos, the expression of hundreds of genes changes significantly, including the *twist (twi)* and *eiger (egr)* genes. According to the definition (above), which gene(s) should be reported in the "gene_manipulated" field?

○ *gd* (FBgn0000808)

○ *twi* (FBgn0003900) AND *egr* (FBgn0033483)

○ *gd* (FBgn0000808) AND *twi* (FBgn0003900) AND *egr* (FBgn0033483)

○ None of the above

○ Not sure

15. How would you rate this newly proposed "gene_manipulated" field?

○ Great

○ A good idea, but needs improvement

○ Unhelpful

○ Not sure

Please expand on your answer (optional).

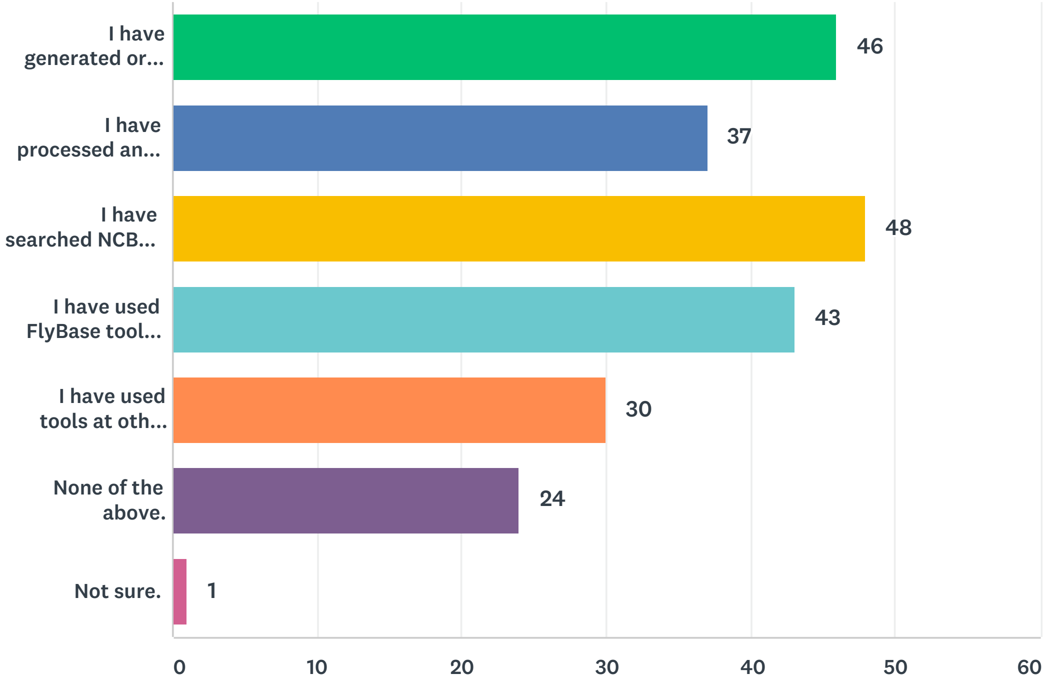
FlyBase

Improving sample descriptions at large dataset repositories
(Drosophila template for NCBI BioSample submissions)

Survey Answers
December 2017

Q1 Please indicate which of the following apply to you (choose one or more).

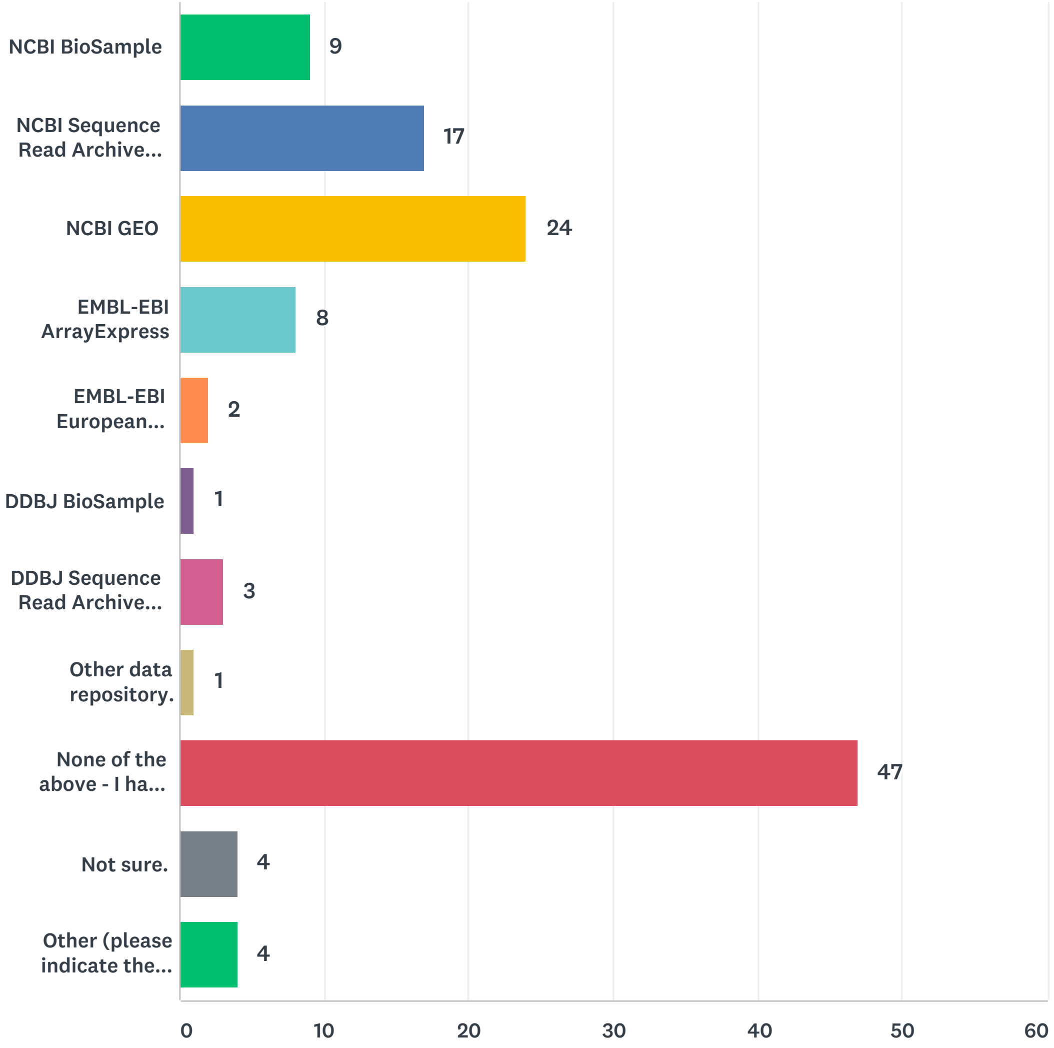
Answered: 94 Skipped: 3



ANSWER CHOICES	RESPONSES	
I have generated or processed a sample used for next-generation sequencing (e.g., RNA-Seq).	48.94%	46
I have processed and analyzed raw data output from a next-generation sequencing experiment (i.e., bioinformatics).	39.36%	37
I have searched NCBI GEO, SRA or ArrayExpress for next-generation sequencing studies relevant to my research.	51.06%	48
I have used FlyBase tools to assess next-generation sequencing data (e.g., RNA-Seq coverage plots, RPKM gene expression).	45.74%	43
I have used tools at other (not FlyBase) websites to assess next-generation sequencing data.	31.91%	30
None of the above.	25.53%	24
Not sure.	1.06%	1
Total Respondents: 94		

Q2 Please indicate the data repositories to which you have made a direct submission(s) (choose one or more).

Answered: 94 Skipped: 3

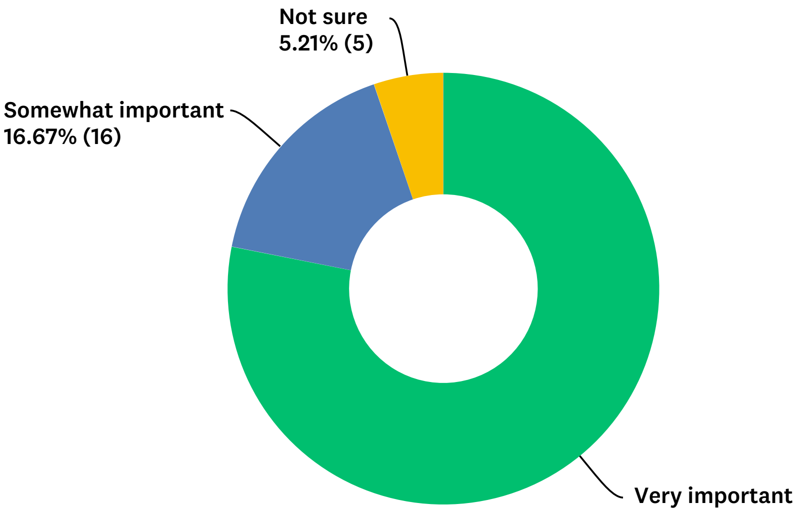


ANSWER CHOICES	RESPONSES	
NCBI BioSample	9.57%	9
NCBI Sequence Read Archive (SRA)	18.09%	17
NCBI GEO	25.53%	24
EMBL-EBI ArrayExpress	8.51%	8
EMBL-EBI European Nucleotide Archive (ENA)	2.13%	2
DDBJ BioSample	1.06%	1
DDBJ Sequence Read Archive (DRA)	3.19%	3
Other data repository.	1.06%	1
None of the above - I have never submitted to a data repository.	50.00%	47

Not sure.	4.26%	4
Other (please indicate the data repository):	4.26%	4
Total Respondents: 94		

Q3 FlyBase wants to catalog Drosophila datasets and develop tools to facilitate the identification of relevant datasets by researchers. How would you rate the importance of this effort?

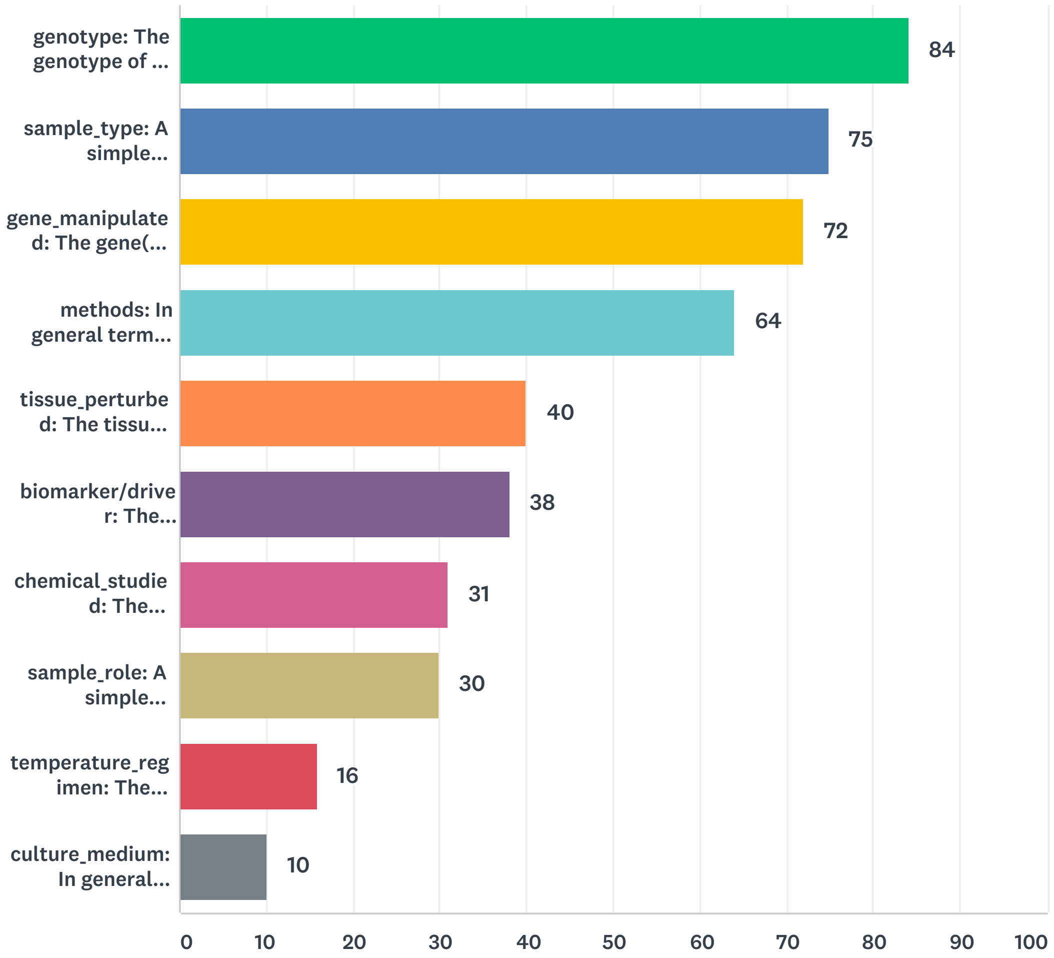
Answered: 96   Skipped: 1



ANSWER CHOICES	RESPONSES	
Very important	78.13%	75
Somewhat important	16.67%	16
Not sure	5.21%	5
Somewhat unimportant	0.00%	0
Not at all important	0.00%	0
TOTAL		96

Q4 From the list of 10 attributes under consideration by FlyBase (listed below), please select the five that are the most informative in an experimental description.

Answered: 95 Skipped: 2



ANSWER CHOICES	RESPONSES	
genotype: The genotype of the biosample (e.g., fz3(J29)/fz3(G10)), as well as details of the genetic cross used to generated the genotype.	88.42%	84
sample_type: A simple classification based on the biological material: i.e., whole organism, tissue, isolated cells, single cell, primary cell line, immortalized cell line, metagenomic collection, synthetic molecules, etc.	78.95%	75
gene_manipulated: The gene(s) that is directly manipulated by some experimental technique: e.g., mutation, overexpression, RNAi, antibody blocking, chemical inhibition, epitope tagging, etc.: e.g., engrailed.	75.79%	72
methods: In general terms, the methods used to perturb the animal/cells: e.g., null mutation, RNAi, bacterial exposure, caloric restriction, etc.	67.37%	64
tissue_perturbed: The tissue manipulated in a study (not necessarily the tissue that was harvested): e.g., a biosample of whole embryos (tissue) in which a gene was knocked down only in the mesoderm (tissue_perturbed).	42.11%	40
biomarker/driver: The molecular biomarker/driver used to select cells for analysis, or the driver used to create a tissue-specific perturbation: e.g., GFP-neur; e.g., ey-GAL4; e.g., en-lacZ.	40.00%	38

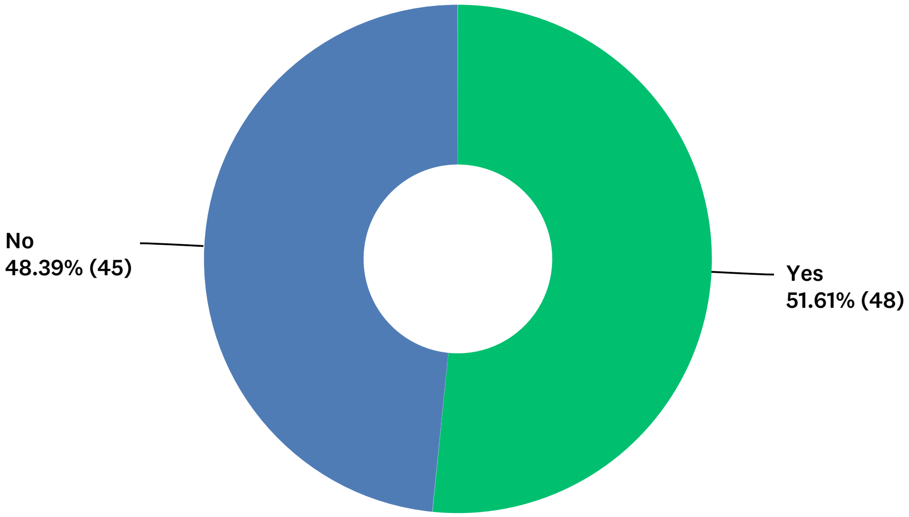
chemical_studied: The chemicals that are used to treat the organism, and for which a biological response is studied: e.g., ecdysone; e.g., cadmium.	32.63%	31
sample_role: A simple classification of the sample as either a control or a treatment in the study.	31.58%	30
temperature_regimen: The temperature at which flies were raised, whether it be at a constant temperature (e.g., 25oC), or a complex regimen used to tune, for example, inducible transgene expression.	16.84%	16
culture_medium: In general terms, the fly or cell culture medium used (including cases where the medium is the same for control and treatment samples): e.g., M3+BPYE medium; e.g., cornmeal-yeast-molasses medium.	10.53%	10
Total Respondents: 95		

Q5 What additional attributes, not listed above, do you value?

Answered: 16 Skipped: 81

Q6 Do you want to answer additional questions on specific aspects of the Drosophila template being prepared for NCBI? (optional)

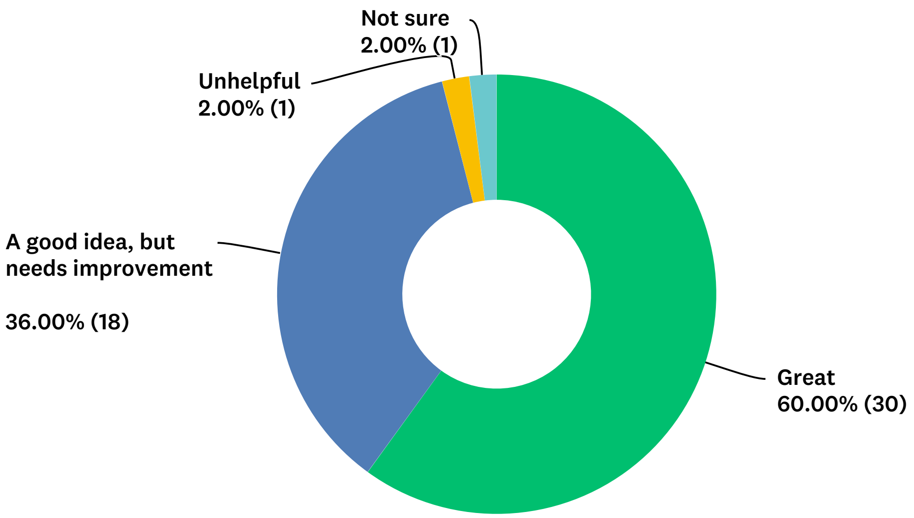
Answered: 93 Skipped: 4



ANSWER CHOICES	RESPONSES	
Yes	51.61%	48
No	48.39%	45
TOTAL		93

Q7 In the proposed Drosophila template, the "strain" field is retained, but other fields considered redundant or rarely applicable to Drosophila have been removed ("isolate", "breed", "cultivar" and "ecotype"). Compared to the current NCBI template, how would you rate this proposed change?

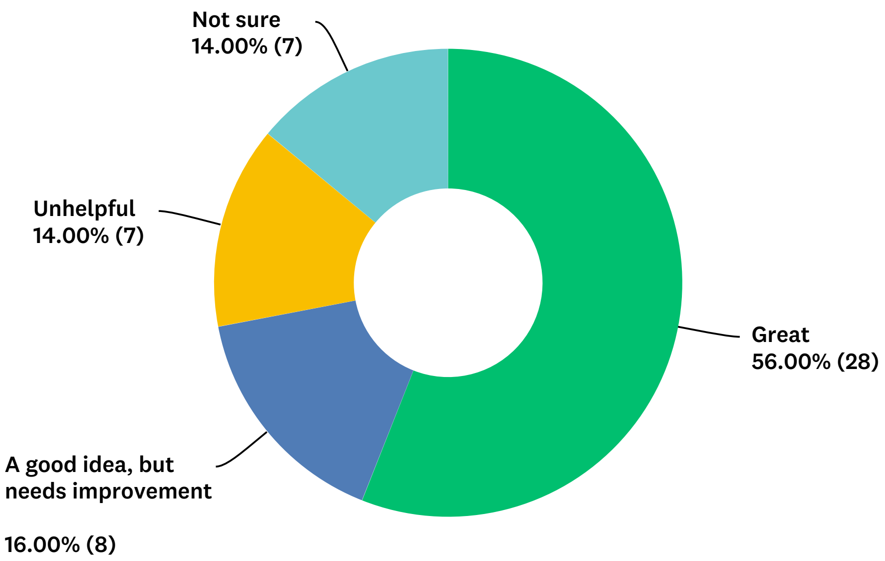
Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	60.00%	30
A good idea, but needs improvement	36.00%	18
Unhelpful	2.00%	1
Not sure	2.00%	1
TOTAL		50

Q8 In the proposed Drosophila template, "strain" is no longer mandatory, and one can instead report a "genotype" for the biosample. Compared to the current NCBI template, how would you rate this proposed change?

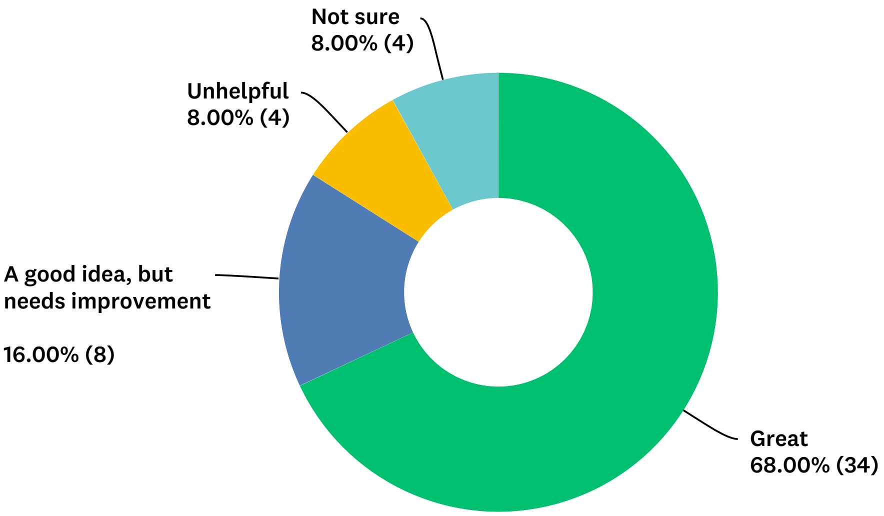
Answered: 50   Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	56.00%	28
A good idea, but needs improvement	16.00%	8
Unhelpful	14.00%	7
Not sure	14.00%	7
TOTAL		50

Q9 In the proposed Drosophila template, "tissue" is no longer mandatory, and one can instead report a "cell_line" for the biosample. Compared to the current NCBI template, how would you rate this proposed change?

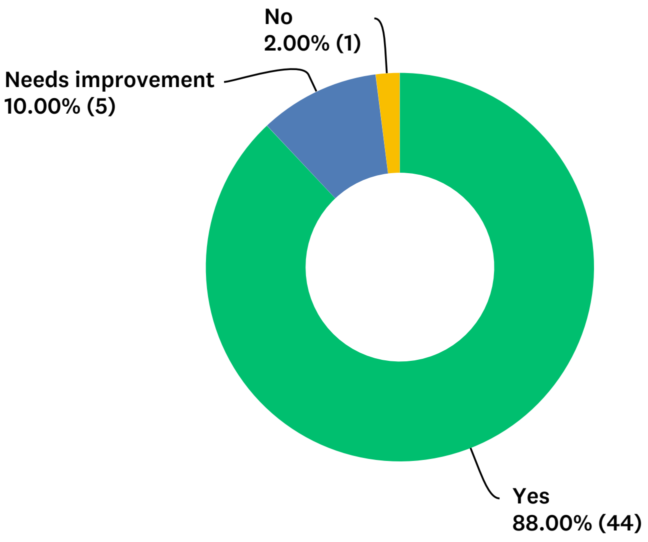
Answered: 50   Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	68.00%	34
A good idea, but needs improvement	16.00%	8
Unhelpful	8.00%	4
Not sure	8.00%	4
TOTAL		50

Q10 In the proposed Drosophila template, a new "tissue_perturbed" field is provided, distinct from the "tissue" field. From the descriptions provided above, is the distinction between these two fields clear?

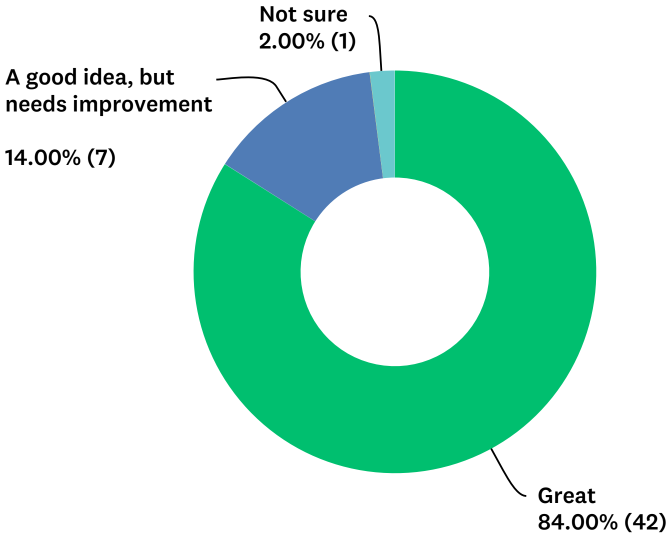
Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
Yes	88.00%	44
Needs improvement	10.00%	5
No	2.00%	1
Not sure	0.00%	0
TOTAL		50

Q11 How would you rate this newly proposed "tissue_perturbed" field?

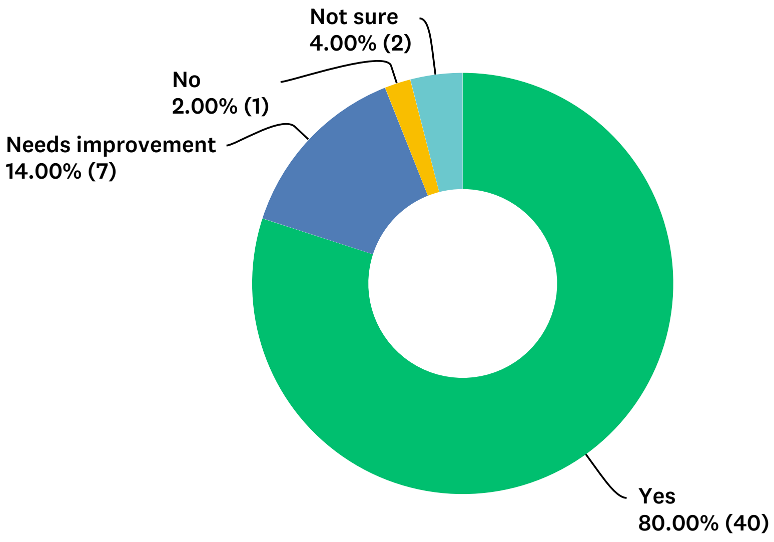
Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	84.00%	42
A good idea, but needs improvement	14.00%	7
Unhelpful	0.00%	0
Not sure	2.00%	1
TOTAL		50

Q12 In the proposed Drosophila template, a new "biomarker/driver" field is provided. From the descriptions provided above, is the definition of this field clear?

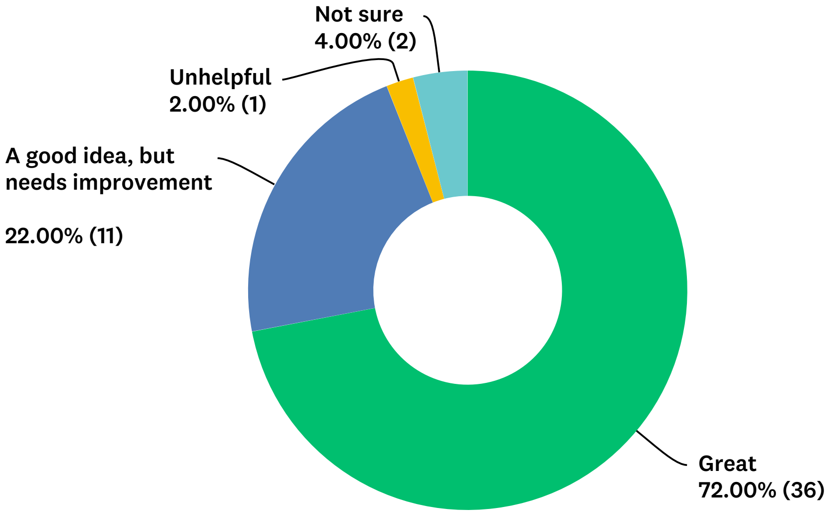
Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
Yes	80.00%	40
Needs improvement	14.00%	7
No	2.00%	1
Not sure	4.00%	2
TOTAL		50

Q13 How would you rate this newly proposed "biomarker/driver" field?

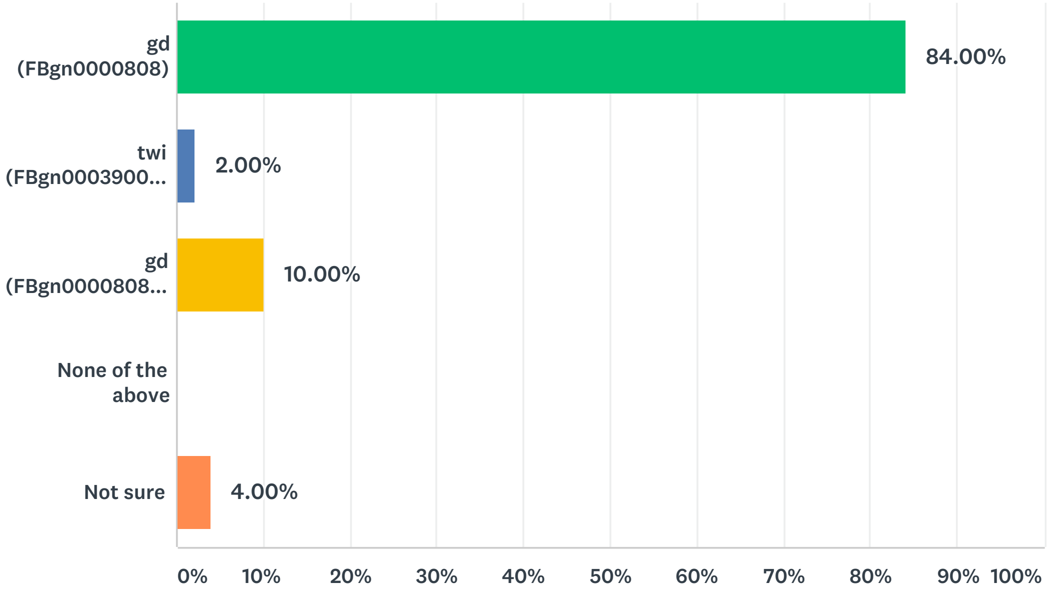
Answered: 50   Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	72.00%	36
A good idea, but needs improvement	22.00%	11
Unhelpful	2.00%	1
Not sure	4.00%	2
TOTAL		50

Q14 To assess the clarity of the "gene_manipulated" definition, please consider a hypothetical study in which progeny of females homozygous for a gastrulation-defective null mutation (gd7) are collected as embryos and processed for RNA-Seq. In these embryos, the expression of hundreds of genes changes significantly, including the twist (twi) and eiger (egr) genes. According to the definition (above), which gene(s) should be reported in the "gene_manipulated" field?

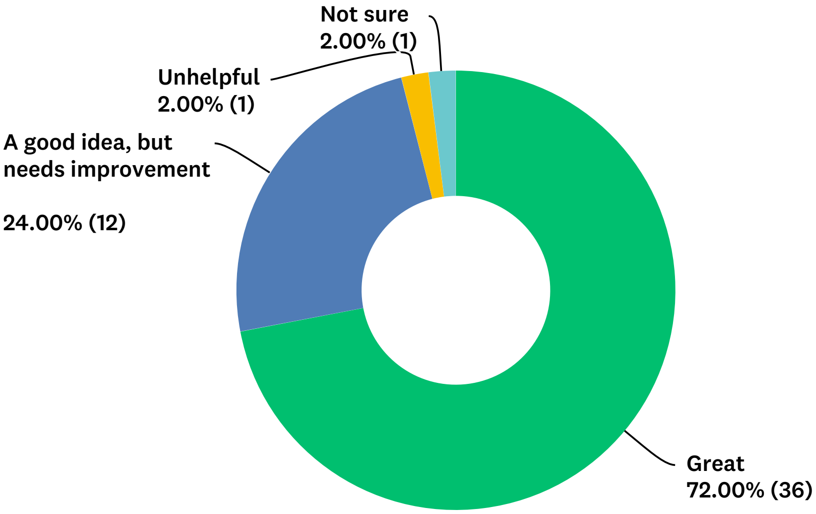
Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
gd (FBgn0000808)	84.00%	42
twi (FBgn0003900) AND egr (FBgn0033483)	2.00%	1
gd (FBgn0000808) AND twi (FBgn0003900) AND egr (FBgn0033483)	10.00%	5
None of the above	0.00%	0
Not sure	4.00%	2
TOTAL		50

Q15 How would you rate this newly proposed "gene_manipulated" field?

Answered: 50 Skipped: 47



ANSWER CHOICES	RESPONSES	
Great	72.00%	36
A good idea, but needs improvement	24.00%	12
Unhelpful	2.00%	1
Not sure	2.00%	1
TOTAL		50