BDK12 Data annotation and curation – ANSWER KEY

Unit 3: Annotating and curating data

Exercise:

Data annotation and curation can be performed at each stage of the research lifecycle:

* Experiment
* Consult
* Publish
* Deposit
* Fund

**Experiment:**

In this scenario, you are a researcher in a lab, and you want to study the effects of consuming red meat and cancer progression. Certain inflammatory cytokines are elevated in colorectal cancer patients, and you hypothesize that these markers may be elevated in patients who consume red meat compared to people who do not consume red meat. You want to analyze the serum levels of the proteins listed below, using an immunoassay. An immunoassay is a biochemical test that measures the presence or concentration of a molecule using antibodies, which are specialized reagents that recognize target molecules. Antibodies are widely used in biomedical research; however, they are finicky and difficult to use, so having all the relevant metadata about the antibody reagent is important for the success of your experiment.

* What kind of metadata should you collect about these antibodies according to the guidelines published here: <https://www.force11.org/node/4433>? (Hint, see Recommendation 2 – Reporting of Antibodies).

Provide at least one of the following:

1. An identifier resolving to a stable public identifier, such as:
   1. Antibody Registry ID ([www.antibodyregistry.org](http://www.antibodyregistry.org/)), OR
   2. eagle-i repository identifier  ([http://www.eagle-i.net](http://www.eagle-i.net/)) OR
   3. vendor name and catalog number for resolving to a single offering
2. For antibodies lacking identifiers, that are not publicly available, provide sufficient protocol details on production of the antibody so as to allow reproduction. (Note: An antibody registry ID can be obtained for such antibodies.) Minimally:
   1. the host organism (NCBI taxonomy ID and/or strain ID from a MOD or vendor)
   2. the identity of the immunogen used (as per sequence molecule identification recommendations above)

* What kind of metadata is included in the Antibody Registry ([www.antibodyregistry.org](http://www.antibodyregistry.org)) for antibodies? (Hint, click the search icon (the magnifying glass) to see the metadata for the results.)

Antibody ID

Antibody Name

Target Antigen

Vendor

Cat Num

Proper Citation

Reference

Clonality

Clone ID

Host Organism

Comments

* Supposed you used the antibodies listed below in your study. Search for the antibodies in the Antibody Registry and note the Antibody ID and manufacturer (vendor) for each antibody:
  + Interleukin-6 (catalog number 11-7061-41) AB\_1633408, eBioscience
  + Interleukin-1a (catalog number 13-7111-81) AB\_466922, eBioscience
  + VEGF (catalog number LS-C18446-25) AB\_867070, LifeSpan

The Antibody Registry IDs are persistent identifiers for each antibody. So if the manufacturer discontinues the catalog number, the metadata about the antibody will still be accessible via the Antibody Registry ID.

When reporting antibodies in the your lab notebook or in the published literature, it is important to capture the appropriate metadata. This will enable you or others to find the reagents that you used.

**Consult:**

In your hypothetical experiment, you next want to collect information about different genes that are implicated in susceptibility to colorectal cancer. You are given this list of **human** genes below. Go to NCBI gene (<http://www.ncbi.nlm.nih.gov/gene/>) to curate this list of genes, with the official gene name, NCBI gene ID, and synonyms (if applicable).

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene symbol** | **Gene ID** | **Gene name** | **Synonyms** |
| TLR2 | 7097 | Toll-Like Receptor 2 | TIL4; CD282 |
| TLR4 | 7097 | Toll-Like Receptor 4 | TIL4; CD282 |
| PLA2G2A | 5320 | phospholipase A2, group IIA (platelets, synovial fluid) | MOM1; PLA2; PLA2B; PLA2L; PLA2S; PLAS1; sPLA2 |
| CCND1 | 595 | Cyclin D1 | BCL1; PRAD1; U21B31; D11S287E |
| ODC1 | 4953 | Ornithine Decarboxylase 1 | ODC |
| APC | 324 | adenomatous polyposis coli | GS; DP2; DP3; BTPS2; DP2.5; PPP1R46 |
| TP53 | 7157 | tumor protein p53 | P53; BCC7; LFS1; TRP53 |
| CTNNB1 | 1499 | catenin (cadherin-associated protein), beta 1, 88kDa | CTNNB; MRD19; armadillo |
| PIK3CA | 5290 | phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha | MCM; CWS5; MCAP; PI3K; CLOVE; MCMTC; p110-alpha |

Read the following excerpt from a methods section of a paper, then answer the following questions.

Excerpt #1:

#### Semi-quantitative PCR (Polymerase Chain Reaction)

To check the relative representation of clones in the cDNA library, semi-quantitative PCR was performed on Pax6, Homer3, Dncl1, Trim11and the constitutively expressed genes GapdhandAtp5a1[[29](http://www.biomedcentral.com/1471-2156/6/43#B29),[30](http://www.biomedcentral.com/1471-2156/6/43#B30)]. Primers were designed to cross at least one intron, so that only correctly spliced clones were amplified. Primer sequences were: Pax6-F CAG CCA AAA TAG ATC TAC CTG; Pax6-R CGA TCA CAT GCT CTC TCC TT; Homer3-F CCC AGG TGG CTG TAG AGC; Homer3-R CTC TAC ACA GTG CAA AGC TCA G; Trim11-F GTG CAG GAT GTG AAG CTG; Trim11-R GCC TGC AGA TAG TCA TAG GG; Dncl1-F CAA AAA TGC AGA CAT GTC G; Dncl1-R CTA AGG GAG AAA AAA ATG GGG; Gapdh-F: CAT CAC CAT CTT CCA GGA GC; Gapdh-R: ATG ACC TTG CCC ACA GCC TT; Atp5a1-F: CAC ACG TGA GAT GTC CTC CA; Atp5a1-R: CAC AGA GAT TCG GGG ATA A. 10 ng library cDNA were amplified in a reaction containing 1xAmpliTaq polymerase buffer (Perkin Elmer), 1.5 mM MgCl2, 200μM each primer and 2.5 units of AmpliTaq polymerase (Perkin Elmer). PCR conditions were (95°C for 30 sec) × 1, (94°C for 30 sec, 55°C for 30 sec, 72°C for 30 sec) × 32 and (72°C for 2 min) × 1. Products were resolved on a 2.5% agarose gel with ΦX174/HaeIII size markers (Promega).

(Reference: http://www.biomedcentral.com/1471-2156/6/43)

Excerpt #2:

* For each study:  
  A) Which genes were used?

B) How can you tell (how were the genes described/identified)?

C) Do you think there is enough information present to replicate the study?

D) Based on this, what information should you include in your hypothetical paper?

Excerpt #1:

A. Pax6, Homer3, Dncl1, Trim11, Gapdh, Atp5a1

B. The genes were described in the methods section describing PCR. If you go to the original paper, you can see they are studying human Pax6, so it’s likely they are screening only human genes, but they don’t explicitly say that, so you can’t be totally sure.

C. They did give the primer sequences, so that could be enough information if the human sequences are different enough from the other species, but there is probably not enough information here to replicate the study exactly.

D. If they provided the Entrez Gene ID for each gene, then that would help.

**Deposit:**

In your hypothetical scientific study, you cloned a new gene and you want to deposit your new sequence in GenBank, so you can obtain an accession number for that gene. Go to GenBank’s website (<http://www.ncbi.nlm.nih.gov/genbank/submit>) and describe below the metadata would you include? (Hint, navigate to the submissions requirements page)

* sequence data
* date for public release
* basic information (authors and a working title) for a corresponding reference paper
* name(s) of the organism(s) from which the sequence data were isolated and any other related [descriptive data](http://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers)
* sequence features (for example: CDS, gene, rRNA, tRNA, with nucleotide intervals and product names)

**Fund:**

Many researchers are required by their funding source to make their data publicly available. What are some ways you could share your data?

* Making it available upon request
* Sharing it in your institutional repository or third-party repository
* Supplementary data in your journal submission