# bioCADDIE Benchmark Dataset Annotation Guidelines

This document contains guidelines to annotate (judge) datasets resulted from running a set of queries to benchmark the bioCADDIE prototype. To annotate each dataset according to its relevance to the question, three options are used: **Relevant, Partially Relevant, and Not Relevant**.

## *Task Description*

You will be presented with paragraphs from full-text articles. These paragraphs contain the portions of text resulting from a search engine query. The queries are based on three high-level "competency questions" designed to evaluate the quality of the results. The competency questions are listed below:

1. **Search for data types x and y related to the same biological process *z***
2. **Search for data types (genome data) *x* with biological process (mutations) *y* and *z* in species/organism *a* for phenotype *b***
3. **Search for disease *x* data of all types across all databases**

| Abstract Question | Target Answer | Example |
|---|---|---|
| Find datasets describing the role of a gene involved in a given disease | The role of the target gene in the disease. | **Search for gene DRD4 and alcoholism disease:** Look for the role of DRD4 in alcoholism. |
| Find datasets describing the role of a gene in a specific biological process | The **effect** of the target gene on the biological process. | **Search for insulin receptor gene and tumorigenesis:** Look for what effect the insulin receptor gene has on tumorigenesis |
| Find datasets describing interactions (e.g., promote, suppress, inhibit…) between two or more genes in the function of an organ or in a disease | The type of genes' interaction with the functionality of the organ. | Search for **HMG** and **HMGB1** and **hepatitis:** look for interaction between **HMG** and **HMGB1** in **hepatitis**. |
| Find datasets describing one or more mutations of a given gene and its biological impact | How a mutation in a gene **influences** a biological process | **Search for mutation and Ret and thyroid:** How does a mutation in Ret influence thyroid function. |

## *Evaluation Measures*

Before you begin to judge, scan the results and look for frequently appearing terms. Look them up to see how they relate to the concepts in the question. Are they sub-topics or synonyms? There are some questions for which only a few paragraphs are relevant or partially relevant. It is tempting to show leniency when few relevant paragraphs are encountered. Resist the urge to relax criteria for relevance and try to maintain consistent evaluation standards.

Judging consistency is also affected by familiarity with the topic. As you judge results, you will become more familiar with the area of research, and may change your criteria for relevance. For unclear dataset titles and descriptions, check the publication and see if it mentions the query topic. If the dataset/publication has the criteria related to the topic, mark it accordingly.

1. A dataset is **relevant** if it captures all required concepts in the question **AND** it **answers the question** or **there is a relationship between key concepts**. (Even if the dataset does not answer the question, it is still biologically quite important if some other relationship between the concepts exists, e.g., decreased expression if the question asks for increased expression.)
2. If **all key concepts are part of the dataset**, but no relationship exists between them, mark it as **partially relevant**. In other words, a result is **partially relevant** if it contains all of the concepts but has missing elements or doesn't answer the question. Additionally, a result is **partially relevant** if it contains the majority of concepts (2/3, or 3/4).
3. If no related concept exists, or the majority of the concepts are missing, mark the dataset as **not relevant**. E.g. if you found 1 concept out of 3, the dataset will be not relevant although it mentioned a key concept.
4. During the judgement for all three categories, if there is a gene or protein mentioned, identify synonyms for it. For biological processes or diseases, familiarize yourself with more general concepts, as well as sub-topics. E.g. "**mad cow disease**" is formally known as **bovine spongiform encephalopathy (BSE)**.
5. Additionally, check the content of the dataset to see if it is referring to a related disease. For example, "Creutzfeldt-Jakob Disease" (**CJD**) is similar to **BSE** and both belong to a family of diseases known as the **TSEs**, but if the query you are annotating only focused on the **CJD** and not

the other similar diseases in the same family, don't annotate the documents without CJD but with TSE or BSE as relevant.

6. When annotating a query pertaining to a disease, make sure that you understand all the related concepts and synonyms and they are logically describe the disease. Some specific diseases are related to more general disease and these general diseases are not always related to the query. E.g. "**bone marrow edema**" implies **Bone Marrow Diseases** and **Edema**. So, if the concept represented in the dataset does not imply the disease which is mentioned in the query, mark it as **not relevant**. Rule of thumb is that we are not only extracting keywords, but those keywords must be related to the targeted query.
7. In another example, a question asks about **liver function**. The definition of "liver development" in the Gene Ontology states that the ***liver secretes bile***, ***synthesizes blood clotting factors*** and ***vitamin A***, and ***stores glycogen***. Those functions that are unique to the liver and are supplied as an answer are relevant. References to functions that occur in other organs in addition to the liver are partially relevant.

## *Examples and Reasoning for Relevance Judgments*

| # | Question | Excerpt | Judgment | Reason |
|---|---|---|---|---|
| 1 | Find Data on **T-cell homeostasis** related to **multiple sclerosis (MS)** | Polymorphisms within the locus encoding a transcription factor **BACH2 are associated with diverse immune-mediated diseases including asthma, multiple sclerosis, Crohn's disease, coeliac disease, vitiligo and type 1 diabetes.** A role for Bach in maintaining immune homeostasis, however, has not been established. **Bach2 was required for efficient formation of regulatory (Treg) cells and consequently for suppression of lethal inflammation in a manner that was Treg cell dependent.** | Relevant | It doesn't directly mention anything about T-cell homeostasis but Bach 2 is involved in regulation of level of Treg (Which is regulatory T-cells). Also, it mentions the role of Bach 2 in multiple diseases as highlighted. |
| 2 | | The purpose of this study was to characterize the transcriptional effects induced by subcutaneous IFN-beta-1b treatment (Betaferon, 250 ?g every other day) in patients with relapsing-remitting form of multiple sclerosis (MS). | Partially Relevant | It talks about Multiple Sclerosis but doesn't have anything related to T-cell homeostasis. |
| 3 | | We investigated the regulation of the size of Interleukin-2-producing CD4+ T-cell (IL-2p) pool using different IL-2-reporter mice. | Not Relevant | It hasn't mentioned MS and therefore, the majority of concepts do not exist. |

| | | | | |
|---|---|---|---|---|
| 4 | | Suppression of **T-Cell** Activation and Collagen Accumulation by an Anti-IFNAR1 mAb, Anifrolumab, in Adult Patients with **Systemic Sclerosis** | Not Relevant | Although it mentions **Systemic Sclerosis,** there is no mention of **MS**. |
| 5 | Search for gene **expression and genetic deletion data** that mention **CD69** in **memory augmentation** | Follicular helper T cell signature in Type 1 Diabetes. This study set out to examine CD4 T cell differentiation in a mouse model of diabetes based on transgenic expression of ovalbumin under the control of the rat insulin promoter T cells were sorted based on expression of CD4, DO11.10 TCR (KJ-126), CD25 and **CD69**. | Relevant | The publication (external resource) indicates that the study considered the impact of the gene on memory augmentation. It also evaluated similar genes, including the one in the question. |
| 6 | | Human **CD4+ memory T cells** are preferential targets for bystander activation and apoptosis. There is much evidence that T cells may be activated via mechanisms which act independently of direct TCR ligation. We have investigated the genetic, phenotypic, and functional characteristics of bystander activated T cells and show that bystander T cell activation is observed during a specific immune response, and that it occurs preferentially amongst CD4+ memory T cells. | Partially Relevant | The dataset talks about the CD4+ memory cell and focused on memory augmentation as well. But it does not focus on CD69 directly or indirectly. |
| 7 | | The high-resolution structure of the extracellular domain of human CD69 using a novel polymer. The structure of the extracellular domain of human CD69 has been determined by single-crystal X-ray diffraction. The structure refined to 1.37 A resolution provides further details of the overall structure and the asymmetric interface between the monomers in the native dimer. | Not Relevant | Although the publication mentions CD69, it is not related to expression and genetic deletion data and it doesn't mention memory augmentation. |
| 8 | Search for **proteomic data** related to **regulation of calcium** in **blind D. melanogaster** | Analysis of genes regulated by Akirin and Brahma upon IMD pathway activation in drosophila S2 cells. This micro-array helps establishing the function of Akirin, a nuclear protein with unknown domains, a putative interacting partner, in the transcriptional **regulation** of the targets of Relish, a NF-kB factor required to fight Gram(-) bacteria infection in **Drosophila melanogaster**. S2 cells were knocked down for Relish, Akirin and immune-challenged by **Calcium**-phosphate transient transfection of dsRNA and over-expressing PGRP-LC vector. Positively transfected cells were sorted by co-expressed Tomato and RNA was purified and analysed by a micro-array | Relevant | The dataset clearly is related to regulation of the calcium in Drosophila melanogaster. |
| 9 | | Data from: 'Escaping' the X chromosome leads to increased gene expression in the male germline of **Drosophila melanogaster**. Genomic analyses of | Partially Relevant | The topic is about D. melanogaster (one element), but it doesn't |

| | | | | |
|---|---|---|---|---|
| | | Drosophila species suggest that the X chromosome presents an unfavourable environment for the expression of genes in the male germline. A previous study in **D. melanogaster** used a reporter gene driven by a testis-specific promoter to show that expression was greatly reduced when the gene was inserted onto the X chromosome as compared with the autosomes. | | directly talk about regulation of the calcium. |
| 10 | | Genome-wide binding of the homeodomain transcription factor Sine oculis (So) in the Drosophila eye imaginal disc. We report genome-wide binding of the highly conserved TF Sine oculis (So), which is necessary for **Drosophila** eye development and has few previously known direct transcriptional targets. biological replicates of ChIP-seq with anti-So antibody on chromatin from **D. melanogaster** third instar eye-antennal imaginal discs; negative control - same sample and ChIP-seq protocol without anti-So antibody | Not Relevant | Although it has the keywords "Drosophila", or "melanogaster", the dataset is not about proteomic data related to regulation of calcium. |
| 11 | Search for data of all types related to the **ob gene** in **obese M. musculus** | Lean **Mouse** 1 Gut Metagenome. Comparisons of the distal gut microbiota of genetically **obese mice** and their lean littermates, as well as those of obese and lean human volunteers have revealed that obesity is associated with changes in the relative abundance of the two dominant bacterial divisions, the Bacteroidetes and the Firmicutes. DNA was isolated from the distal gut (ceca) of eight-week old C57BL/6J obese (**ob/ob**) and lean (ob/+ and +/+) mice using a bead beater. | Relevant | The dataset is clearly related to the topic. It's not only includes the keywords, it is relevant to the question. |
| 12 | | Expression data from bone marrow macrophage derived from **lean and obese mice** after 1,4, and 24h exposition to P.gingivalis. Since M1 polarization is crucial during acute infectious diseases, we hypothesized that diet-induced obesity inhibits M1 polarization of macrophages in the response to bacterial infection. Bone marrow macrophages (BMM?) from lean and obese mice were exposed to live Porphyromonas gingivalis for three incubation times (1 h, 4 h and 24 h). | Partially Relevant | The dataset is related to the obese mice and their genes. But it doesn't directly mention the ob gene, which has been asked in the question. |
| 13 | | High Fat Diet Reduces the Expression of Glutathione Peroxidase 3 in Mouse Prostate. We investigated the effect of high fat diet on mouse prostate gene expression. C57BL/6J mice were fed either a control or high fat diet for 12 weeks. | Not Relevant | Rather than focusing on the obese mouse, the study is related to a high fat diet in mouse prostate. It is not related to ob gene as well. |

## *Resources for Judging*

**IHoP – Information Hyperlinked over Proteins**
http://www.ihop-net.org/UniPub/iHOP/
This database lists synonyms for proteins and provides excerpts from the literature, allowing you to familiarize yourself with the biology of the protein.

**PubMed Books**
http://www.ncbi.nlm.nih.gov/entrez/query/Books.live/Help/bookhelp.html#search
Good for general overviews of biological processes and diseases. Link takes you to instructions for searching books.

**AmiGO, the Gene Ontology browser**
http://www.godatabase.org/cgi-bin/amigo/go.cgi
Good for brief definitions of biological processes. No disease information. Use MeSH or PubMed books.

**MeSH – Medical Subject Headings**
http://www.nlm.nih.gov/mesh/MBrowser.html
For biological processes and diseases, provides synonyms or constituent processes that are part of the indicated concept.