Disease Portal Curation

Outline

1. **Target disease gene list**
2. **Curation process:**

   A. **Disease (rat, mouse, human)**

      1. **Disease-gene association criteria**

         a. **How is gene related to disease?**
         - **Biomarker – gene product is associated with disease by expression level**
         - **Molecular mechanism – gene product is known to be functionally involved in the disease process**
         - **Therapeutic target – gene product is drug target for drug which is therapeutic for the disease in question**
         - **Variant association (usually polymorphism or mutation) – gene sequence is being distinguished from others that may be involved in the disease**

         b. **What is the molecular level of the alteration?**
         - **DNA**
         - **mRNA**
         - **Protein**

         c. **What type of alteration is involved?**
         - **See Excel spreadsheet of modifier terms ("Disease ontology modifiers") for free text notes field of curation tool**

      2. **Annotation of species in reference**
      3. **Assignment of annotation to other species (currently automated)**
         - **Use ISO as evidence code**
         - **Put RGD ID of referenced species in "With Info" box**
         - **Use same qualifier and free text note, but if a sequence-specific identifier is used in the free text note (ex:DNA:polymorphism:exon:1257G>A), follow note with original species in parentheses: (rat, mouse, or human)**

   B. **Gene Ontology (rat)**

   C. **MP/HPO (rat, human)**

   D. **Pathway (rat, mouse, human)**

Gene list assembly procedure for the target disease (performed by curation team leader):

1. Look up disease category in RDO and assemble approximately top ten high level terms. Take note of all the adjectives, synonyms, sub-types, etc. specific for the disease terms in question.
2. Select at least three disease/gene databases which can be queried with your terms. **Phenopedia** (https://phgkb.cdc.gov/HuGENavigator/startPagePhenoPedia.do), **GeneCards** (http://www.genecards.org), and **Genatlas** (www.genatlas.org) are good sites for returning general disease queries with lists of genes in tables that can be easily copied and pasted or downloaded to Excel (Phenopedia, GeneCards v3). For databases dedicated to the specific disease category in question, search the internet for any possible sites.
3. Copy and paste the gene lists to an Excel spreadsheet. Keep the gene symbol, name, species, source, and hit number (if available). Assemble all the lists from different term searches, different databases, and the RGD genes into one large list. Alphabetize the entire combined list based on gene symbol. Rank each gene for the number of times it appears in the list. If any database gave a number of "hits"

based on relevance (ex: GeneCards), add that number in for the ranking. Order the list based on the numerical ranking. Separate the list in thirds or fourths based on rank order.

Searching the literature:

1. The master list will be divided among curators “on the fly”. Each curator will be given a list of 10-20 genes at a time from the top of the master list. Starting with the top of the list find the RGD gene that matches the name and symbol on the list. Check the RGD gene page to see what ontology terms and evidence codes have previously been curated for the gene in question. Enter the gene name/symbol into the curation tool and load the appropriate ortholog group. Start the disease curation with OntoMate to search with the parent term or terms given to you with the gene list. If there are papers to curate for the disease in question, curate all of those papers to the point of completeness, but not redundancy. Redundancy would be annotations that have the same term, same species, same evidence code, same qualifier, and same information in “notes”. Also, curate all rat papers for the chosen gene for gene ontology (GO), mammalian phenotype ontology (MP), and pathway ontology (PW), if applicable. Any disease paper with human RDO annotations should also be curated with HPO, if possible. If it is determined that the target gene has no literature for the disease in question, make a note of this, don’t bother with any curation for that gene and move on to the next gene on the list.
2. For GO, repeat the literature search with OntoMate, but search with gene only, no ontology term. In the OntoMate results page choose “rat” from the “Organisms” list on the left side of the page. This will filter the list accordingly. If manual GO annotations are already on the gene report page for the targeted gene, determine when the newest annotations were made and filter out older dates using one of the options on the upper left side of the OntoMate results page.
3. To begin curating, first read the titles to determine if there may be disease-related and/or gene-related information in the journal article.
4. If an article looks interesting from the title, read or scan the abstract, looking for disease terms related to your target disease and your target gene if you are doing RDO. For GO, look for papers that describe what a gene product does, what processes it is involved in, or where it is located in the cell. The information about what is done to a gene/gene product it would be GO data for another gene. If you can tell from the title or first couple of lines that there is curatable information in the article and if there is a “conclusions” section in the abstract, check the summary sentence(s) for information suitable for annotation. If you need more details, finish reading the abstract. RDO annotations can often be found in the abstract. For GO annotations you often need to check the full article, scanning for the specific details you seek (see below). Be aware of the possibility that authors may use synonyms and symbols different from the official gene name and symbol. You may need to check the RGD gene report pages for rat, mouse and human to confirm synonyms and alternate or previously used symbols. If you find symbols or synonyms that are not in RGD, you should add those to the rat gene report page via the “Edit Me!” tab found at the top left corner of the page. On the edit page select “Add alias” at the upper right corner of the “Aliases” section. Select “old_gene_name” or “old_gene_symbol” from the drop down menu on the left side of the empty text box. Enter the symbol or synonym in the text box and click “Update” above the “Add Alias” link.
5. For GO, don’t look at expression papers unless they look like they might show subcellular localization as opposed to tissue distribution or developmental expression.
6. “Full paper” curation doesn’t mean you need to read every word: Scan/read applicable methods to determine what species is being studied. Look at figures and methods to find details of experiments. Read only the parts of the results that appear curatable—section headers are often very helpful in scanning for information. You don’t usually need to read the discussion (unless you need it to understand what the results mean).
7. To save time, we do not usually do multiple annotations to the same term. If the annotation has already been made from one reference it is not necessary to make the same annotation from other papers. When one reference gives experimental evidence and another only makes a statement, use the reference with

the experimental evidence. This is not a hard and fast rule. If the experimental evidence is different in two different papers such that annotations can be made using two different evidence codes, both annotations can be made. Generally, it is preferred to have a stronger evidence code where IDA, IMP, or IPI > IEP.

8. MP/HPO curation is closely related to RDO curation, in that most of the phenotype data in the literature usually relates to disease and the evidence codes of the annotations are shared by both types of curation (see "Ontology evidence code charts (Appendix 12A) SJL 4-4-17" document). The curation of MP/HPO terms can often be done simultaneously with RDO terms, because the terms will be similar, and many MP/HPO terms will show up simultaneously with RDO terms when searching for terms in the curation tool. See the document "Mammalian Phenotype curation supplement SJL 5-18-16"
9. PW curation is closely related to GO curation. Many of the interactions described by pathway terms can also be described with GO terms. As among MP/HPO/RDO, the annotation evidence codes are shared between PW and GO. While curating GO, if pathway-specific data is found, the annotation may be made to PW or GO. There are more pathway-related terms in GO than in PW. If an appropriate pathway term can be found in GO, annotate with that term. If an appropriate pathway term is not found in GO, try PW. If neither ontology has an appropriate term, a new term can be requested for either, or both ontologies. If an appropriate term is found in both ontologies, two annotations may be made, although one is sufficient.

Making the annotation:

1. **Select target gene and orthologs as core objects in the curation tool**. Choose "Select Objects" under the "Core Objects" bucket and enter the gene name or symbol in the "Search for an object" search field in the "Select Core Objects" frame. "Rat Gene & Orthologs" is the default choice. From the returned list click on the bucket icon in the "Select" column that corresponds to the appropriate gene and the gene name will be put into the "Core Objects" bucket. If curating disease or pathway, choose "Add the ortholog group" link at the top or bottom of the appropriate ortholog group in the results list and rat, mouse, and human gene members of the ortholog group will be added to the "Core Objects" bucket. (**Note**: The core objects search has a pull-down menu with the options "Contains", "Equals", "Begins with", and "Ends with" for tweaking the search).
2. **Select applicable ontology term in the curation tool**. When using OntoMate, look at the terms listed under the abstract to help choose an ontology term. Click the bucket icon to the left of the term to place the term into the term bucket in the curation tool. To see the term in the ontology browser embedded in the curation tool, click on the ontology term in the OntoMate list. The browser may then be used to find the most appropriate term. Clicking the "select" button to the left of any term in the browser will put that term into the ontology term bucket.
3. **Import the PubMed reference into curation tool**. When using OntoMate, click the bucket icon to the left of the PubMed ID for any entry in the results list. This action will import the PubMed ID and abstract into RGD, assign an RGD ID to that reference, and put the RGD ID into the reference bucket of the curation tool. If OntoMate is not available, click "Select References" and enter the PubMed ID into the "Import a Reference Directly from NCBI" text box in the "Select / Create References" frame in the curation tool. Clicking "Import Directly" will also import the PMID and abstract from PubMed.

   If you are curating to a reference that already has an RGD ID, enter that number in the "Keywords" box of the "Select / Create References" frame. Click "Find References" and the corresponding reference entry will be returned in a new frame of the curation tool and the reference ID will simultaneously be put into the reference bucket.
4. **Generate annotation in the curation tool**. First select "Associations" from the left frame of the curation tool page, and then select "Make an Object/Term/Reference Annotation" in the right frame to open the "Make an Annotation" frame. Select the appropriate combination of primary Core Object(s), Ontology Term(s), Reference(s), Qualifier, and Evidence Code (for the species of the gene(s) used in the journal article). The tool automatically generates annotations (with ISO evidence

code) for other members of the ortholog group for disease and pathway ontology annotations. For associations with the ISO evidence code, the RGD ID for the gene/homolog having experimental evidence is put into the "With Info" box. The "With Info" box also must be used for IPI evidence of "binding" (to a protein or other molecule) or IGI evidence, usually of biological process. The database ID of the "other" gene, protein, or molecule is entered. (Note: The correct format for the "With Info" box is "RGD:######". If more than one ID is necessary use a pipe "|" to separate the two entries. "|" means "or" in the With Info field, while a "," means "and".) MP/HPO annotations are only made for rat/human and only for the species in the journal article.

For RDO annotations the "Notes" box in the curation tool can be used to add additional, optional "modifier" information to the annotation. This information relates to what change(s) in the gene or gene product is/are associated with the disease in the annotation. The format is "molecular level of alteration (DNA, mRNA, or protein):modifier term:location of alteration:specific polymorphism or mutation". Modifier terms can be found in the document "Disease Ontology Modifiers". Terms for Organ/tissue (Uberon Anatomy Ontology) and cell type (Cell Type Ontology) for the "location of alteration" field can be found in the RGD ontology term browser. Cellular Component terms from the Gene Ontology may also be used in the "location of alteration" field. If more than one term is needed to describe the location, use the following format and order: "organ, cell type, cellular component". The notes field may also be used to add modifier information to MP or HPO annotations.

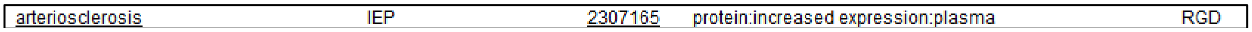
If the disease in the annotation is being studied in the context of another disease, the other disease is also put in the notes field as "associated with (specific RDO disease term);" in front of any modifier information. If any sequence-specific information is included in the "modifier" information, add the species ("rat", "mouse", or "human") in parentheses at the end of the note.

Click "Generate List" to display the list of associations created. The boxes in the "Select" column are checked by default. If you do not want to make a particular annotation, uncheck the appropriate box for associations you do not want to record as annotations. Click "Create these Annotations in DataBase" to record annotations.

If any errors are made in the annotation process within the curation, the annotation will be blocked and an error message will appear at the top of the "Make an Annotation" frame. Correct the error and proceed.

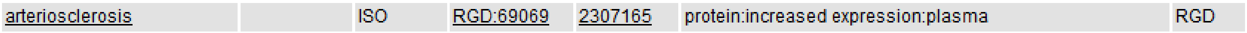
Example of a disease annotation from one paper curated to rat, mouse, and human with the experimental evidence coming from rat:

Rat:



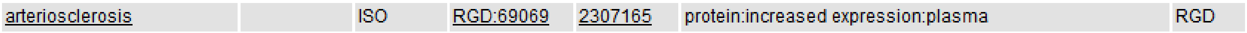
arteriosclerosis		IEP		2307165	protein:increased expression:plasma	RGD

Mouse:



arteriosclerosis		ISO	RGD:69069	2307165	protein:increased expression:plasma	RGD

Human:



arteriosclerosis		ISO	RGD:69069	2307165	protein:increased expression:plasma	RGD

5. If you notice a curation error after the annotation is committed to the database, scroll down the "Make an Annotation" frame to see all of the current annotations to all the genes in the "Core Objects" bucket. Click the "Edit" icon to the left of the bogus annotation. On the edit page change the field(s) that contain the error. Click the "Enter" key or the "Update and forward to curation tool" button to save the correction and return to the curation tool. Alternately, in the "Make an Association" frame you can click "Show/Delete My Annotations" if you prefer to delete your annotation and start over. On the "Your Annotations" page you can delete any of your annotations by clicking the "Annotation with Mapping" link in the "Delete" column on the left side of the appropriate annotation line. Note that this will delete the annotation and the association of the gene object with the reference, while the "Annotation" link on the right side of the line will only delete the annotation.

Additional Notes:

1. Manual evidence codes for GO and PW consist of IDA, IEP, IGI, IMP, IPI, ND, and IC. IDA, IGI, IPI, IMP and IEP are experimental evidence codes. Use of the experimental codes is strongly encouraged and use of IC is strongly discouraged. IC can be used for GO but only when the inference is from an annotation to another GO term. In this case, the originating GO term ID must be entered in the "With Info" field.

   ND is used when GO curation has found no rat data for any one of the three branches: molecular function, biological process, or cellular component. The reference code used in the curation tool is RGD ID "1598407". The terms used would be "molecular_function", "biological_process", or "cellular_component". In addition it is necessary to put "MM/YYYY: no relevant rat data" in the "notes" box when making the annotation. "MM/YYYY" is the month and year of the annotation, which lets users and other curators know when the last curation was done for that gene.

   Evidence codes for RDO and MP/HPO consist of IAGP, IDA, IEP, IMP, plus ISO (automated assignment to orthologs) for RDO only.
2. Annotations with evidence codes of IPI, IGI and ISS must have a valid database ID (RGD, UniProt, MGI, etc) in the "With Info" field. The list of allowed database abbreviations is available on the GO Consortium website (http://www.geneontology.org/doc/GO.xrf_abbs)
3. When an annotation to a term such as "protein binding" is made, the reciprocal annotation for the gene in the "With Info" field in the first annotation should be made if that second gene is rat and the annotation does not already exist. For example, if a paper says that rat protein X binds to rat protein Y, both gene X and gene Y should receive the "protein binding" annotation with the ID for the appropriate binding partner in the "With Info" field.
4. Qualifiers are used sparingly and only when they add information to the term itself (not to the evidence code, for instance). The qualifiers approved by the GO Consortium for use with GO terms are "colocalizes with", "contributes to" and "not" (see "Using the Qualifier column" at http://geneontology.org/page/go-annotation-conventions) for more information on their use. The qualifiers that are used with RDO are "no association", "resistance", "susceptibility", "disease progression", "severity", "model" (in reference to rat strains), "induced" (in reference to rat strains), "spontaneous" (in reference to rat strains), and "onset". Those for use with MP are "induced" (in reference to rat strains) and "spontaneous" (in reference to rat strains). There are currently no qualifiers that are used with PW.
5. When an annotation is made to a gene using a reference, the link between the gene and the reference is automatically made. If you want to link a gene to a reference without making an annotation this can be done using the "Make Association"→ "Reference to Core Object" option. If you look at a paper which describes a rat gene, but find it is not appropriate for current curation purposes, the link can still be made. An example would be an expression paper that would be valuable if RGD decides to curate expression data in the future.
6. When adding information in the "Notes" text box, any missing information can be designated by a space in that particular field. For example "DNA:missense mutation: :p.D307H (human)" has just a space in Field 3 because the location (exon, intron, etc) of the mutation is not specified.
7. For cellular component annotations put an MMO ID (MMO:#######) in the "Notes" text box in the curation tool to indicate the type of assay used to support the annotation.