Pathways Download, Extraction, and Data Loading Process Preprocessing - Manual and/or One-time processing Download and process Create interaction Extract Pathway and Gene **Create Interaction** Download and process NCI Pathways search terms Predictive Networks Types config file and Subpathways web page to information from interactions file and save to config file NCI-Nature Web API "interaction_types.txt" "NCI pathways and subpathways.txt "interaction_search_terms.txt" "Extracted Predictive Networks.txt" caBioPathwayGenes.txt Extracted_NCI_pathways_and_subpathways.txt Quackenbush extract edited.tsv Download, Extract, and FilterData Download HGNC data NCBI Gene To PubMed NCBI GeneRIFS NCI Nature SIF text file to Download .gz file Download .zip file Download .gz file "HGNC dataset.txt" Decompress file Remove records for Decompress file Decompress file "withrawn" symbols and "homo-sapiens-9606.sif" to "gene2pubmed" to "interactions" to "ncbi_gene2pubmed.txt" "NCBI_generifs_interactions.txt" those missing symbol or name "nci-nature.sif" Drop some fields Read CSV name synonym Drop many fields Drop many fields Extracted Pathway Commons NCI nature sif.txt data from column "Syonyms" Remove records not related Remove records not related Extract name synonym data to humans to humans Read CSV symbol synonym data from column Extracted_NCBI_gene2pubmed.txt Extracted_NCBI_generifs_interactions.txt "Previous Symbols" Extract symbol synonym data Extract List of Gene Symbol-Entrez Id pairs Extracted_HGNC_dataset.txt Extracted_Gene_Symbol_Entrez_List.txt GeneNameSynonyms.txt GeneSymbolSynonyms.txt Quackenbush_extract_edited.tsv Extracted_Pathway_Commons_NCI_nature_sif.txt Extracted_NCBI_generifs_interactions.txt Remove records where gene symbols Remove records that have neither Remove records where gene symbols are empty are empty or null and gene symbol is null predicate of "functional interaction" Assign interactionDetailId and Assign interactionDetailId and nor source of "Pathway Commons" interactionDirection as appropriate. interactionDirection as appropriate. Assign interactionId, interactionDetailId and interactionDirection as appropriate. Extract assigned interaction ids and related documentation ids for PubMed Replace non-approved symbol synonyms with HGNC approved symbols where possible. Extract assigned interaction Form superset of ids and related documentation interaction details objects ids for PubMed and Pubmed Central from all three interaction sources Modify superset of interactions and interaction details objects Calculate number of interaction to arrange geneA and geneB Replace non-approved symbol detail records in each direction alphabetically and then combine Extracted_Gene_Symbol_Entrez_List.txt synonyms with HGNC approved and assign interaction direction all interaction details objects and symbols where possible. for overall interaction create interaction objects and assign interaction ids as appropriate Interactions.txt Extract assigned interaction detail ids and related documentation ids for PubMed Interactions_Details.txt Interactions_Documentation.txt Interactions_Predicates.txt Reactome Reactome Pathways Reactome SQL Database Pathways and Genes Hierarchy XML file Download .zip file Download Download .gz file "ReactomePathwaysHierarchy.xml" Decompress file to Decompress file to "ReactomePathways.gmt" "reactome.sql" Extract multiple tables from MySQL database dump Combine MySQL database table extracts with gmt and xml file and output multiple files Correct some know gene symbol issues with Reactome data in pathways ReactomePathwaysAndGenes.txt Extracted_ReactomePathwaysHierarchy.tsv ReactomeDatabaseObject.txt ReactomeStableIdentifier.txt GeneOntology Search for online file name of format go_.+-assocdb-data.gz file GeneOntology: Download .gz file containing MySQL dump Decompress .gz to file "geneOntology.sql" Extract multiple tables from MySQL database dump Write gene_product table to Write species table to Write gene_product table to Write term definition table to Write term2term table to Write association.txt table to "GO base species.txt" "GO_base_gene_product.txt" "GO_base_term.txt" "GO_base_term_definition.txt" "GO_base_term2term.txt" "GO_base_association.txt" Process term relationships Remove all but Remove records with invalid Strip out records with identify root term Remove all but unique Human and Mouse terms or relationship types drop obsolete terms negative, "is_not" relationship, Human and Mouse records records Remove duplicate records drop duplicated terms records will empty gene symbols or terms and separate terms used for associating terms duplicate symbol-term pairs Write to multiple files GO_species.txt GO_gene_products.txt GO_term_types.txt GO terms.txt GO_term_relationship_types.txt GO_gene_products_to_go_terms.txt GO_term_relationships.txt **Correct and Extend Data and Add to Database** Connect to database Set database to skip unusable indexes for session Retrieve version id of latest current data load version Write new version record to database



