

dbPTB

Database for Preterm Birth

dbPTB Guideline version 1.0 2012

About Database for Preterm Birth (dbPTB) and curating guidelines

a) dbPTB

dbPTB is a relational database currently consists of genes and articles which carry information about Preterm Birth (PTB) and links to bioinformatics sources such as UCSC Genome bioinformatics, OMIM (Online Mendelian Inheritance in Man), NCBI Entrez Genes, HGNC (Human Gene Nomenclature Committee).

dbPTB uses SciMiner for extracting the gene and protein information from published articles specific to PTB by creating queries depending on the possible keywords and MeSH terms. Collection of articles and the putative genes for these articles are stored under dbPTB. dbPTB was implemented using a MySQL database running on a Linux server with PERL and PHP scripts used for all data retrieval and output. User interface is built by PHP, HTML and PERL scripts. dbPTB uses SciMiner in regular basis (monthly) to retrieve new articles and their putative genes. Besides computational data mining, also curators should submit newly identified articles into dbPTB. All submitted articles by the curators first pass through SciMiner in order to extract possible genes, then with the gene (if available) information these articles listed in dbPTB.

Curators of dbPTB consist of researchers and students that have knowledge in molecular and cell biology about PTB. Curators' have their own accounts enable to store, edit the curation status of the papers they have chosen or are working on from the list of dbPTB articles. There is no redundancy in picking up the papers. All curators are responsible for the articles they choose. Once the article is assigned to one curator, no curator can pick the same article. Curation status of all the articles and curators are dynamic in dbPTB which means it is up to date. Below, you will find the curating guidelines.

b) Curating guidelines and query process

First step in curation is to read the title and the abstract to see if there is a possibility of any PTB relevant information along with gene/genomics/pathways/genetics in the paper. Sometimes title may not contain direct information then curator should scan through the paper.

Experimental details

- Genes involvement in PTB or any relationship with PTB should be supported by experimental evidence.
- Any experimental results should be noted. Any in vivo or in vitro details should be searched throughout the article.
- Any measured effects, phenotypic changes, genetic associations, and/or biochemical effects should be searched. If they are statistically significant, justifies inclusion of the gene or proteins.
- Acceptance of any gene for the related paper depends on the statistical inference, if it is $p < 0.05$ for a certain gene then it should be accepted into the dbPTB.
- Transgenic animals show important genetic effects. If the genes are associated with a birth phenotype (longer or shorter gestation) the gene(s) should be included. Downstream genes if specified should be included. If not clear or not specified, then it should be saved for future re-evaluation or should be brought to weekly team curation meeting.
- In the proteomics articles, even if the statistics are not significant, the existence of a protein in an abnormal body compartment (e.g. amniotic fluid) or cellular location, then the protein(s) should be included.
- In the array articles, if there is no offline verification, then include all the genes studied with significant fold changes up or down. If there is offline verification such as real-time PCR, then include the ones verified. Genes that failed offline verification should not be included.

Computational results

- Any computational or in silico results should be noted, e.g. imputation. If significant, results should be included.
- The criteria for accepting a gene(s) is again going to be the significance level of findings which is $p < 0.05$.

References, related articles in bibliography for all papers should be searched for inclusion in the database.

The biological systems: Cell, tissue, organism. There are several question may be asked in the curation process such as what kind of cell type was used? In which tissue were the genes expressed? If the study was performed in a species other than human, then which organism was used in this research? These questions are valuable to ask to evaluate the biological systems mentioned in the article and should be included in the notes section. Discrete, non-human data should be included if it directly addresses measured effect,

phenotypic changes, genetic associations, and/or biochemical effects, if statistically significant. There is a tab for common animals/species, e.g. mouse or rat.

References: References should be searched, they may carry additional information. All citations should be checked carefully. Please see the chart in the next slide (See in Figure 2).

Figure 1. Key steps for experimental and in silico details while curating the articles.

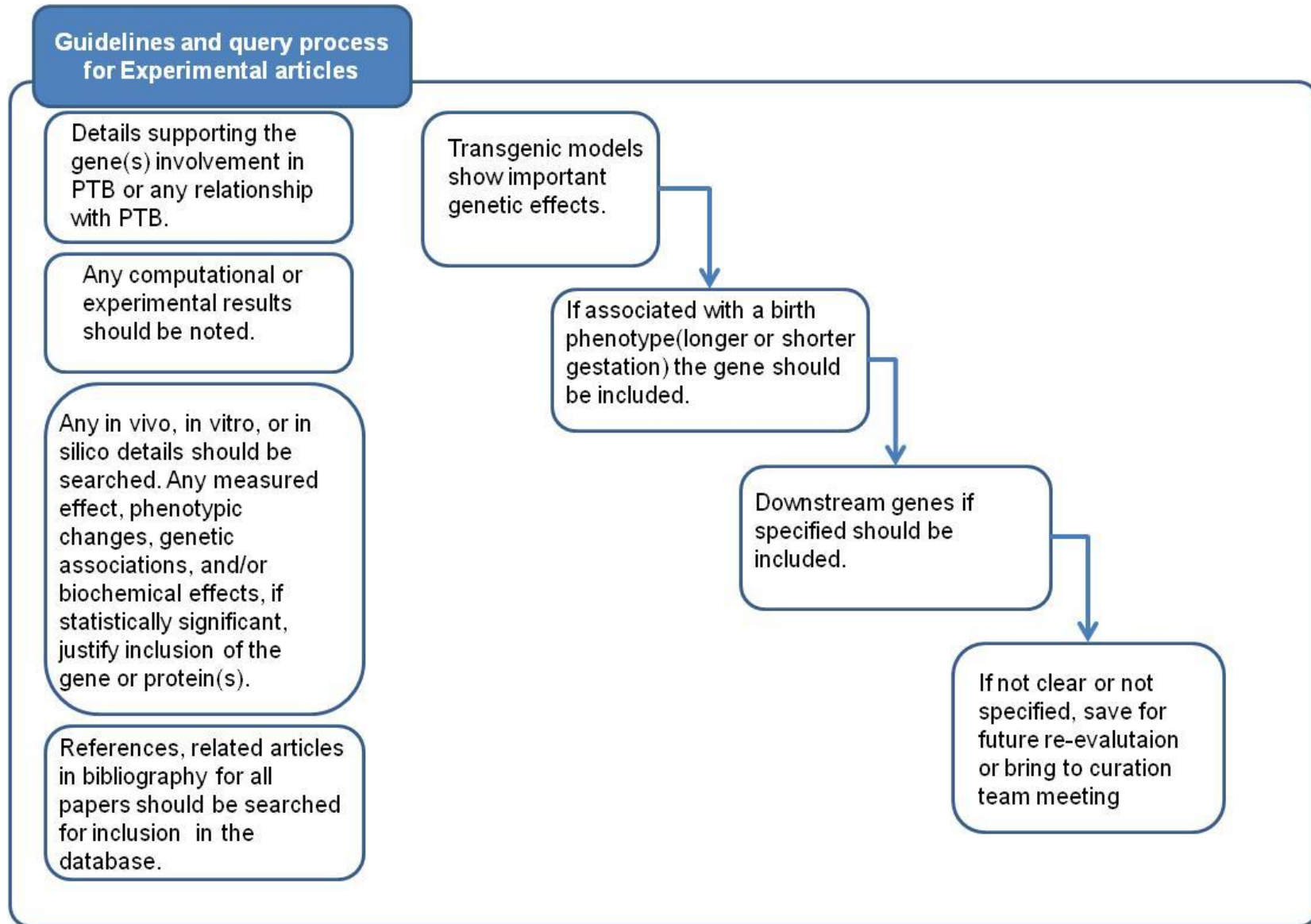


Figure 2. Schematic represents the steps for searching references and submission of the articles.

Flow chart for importing references.

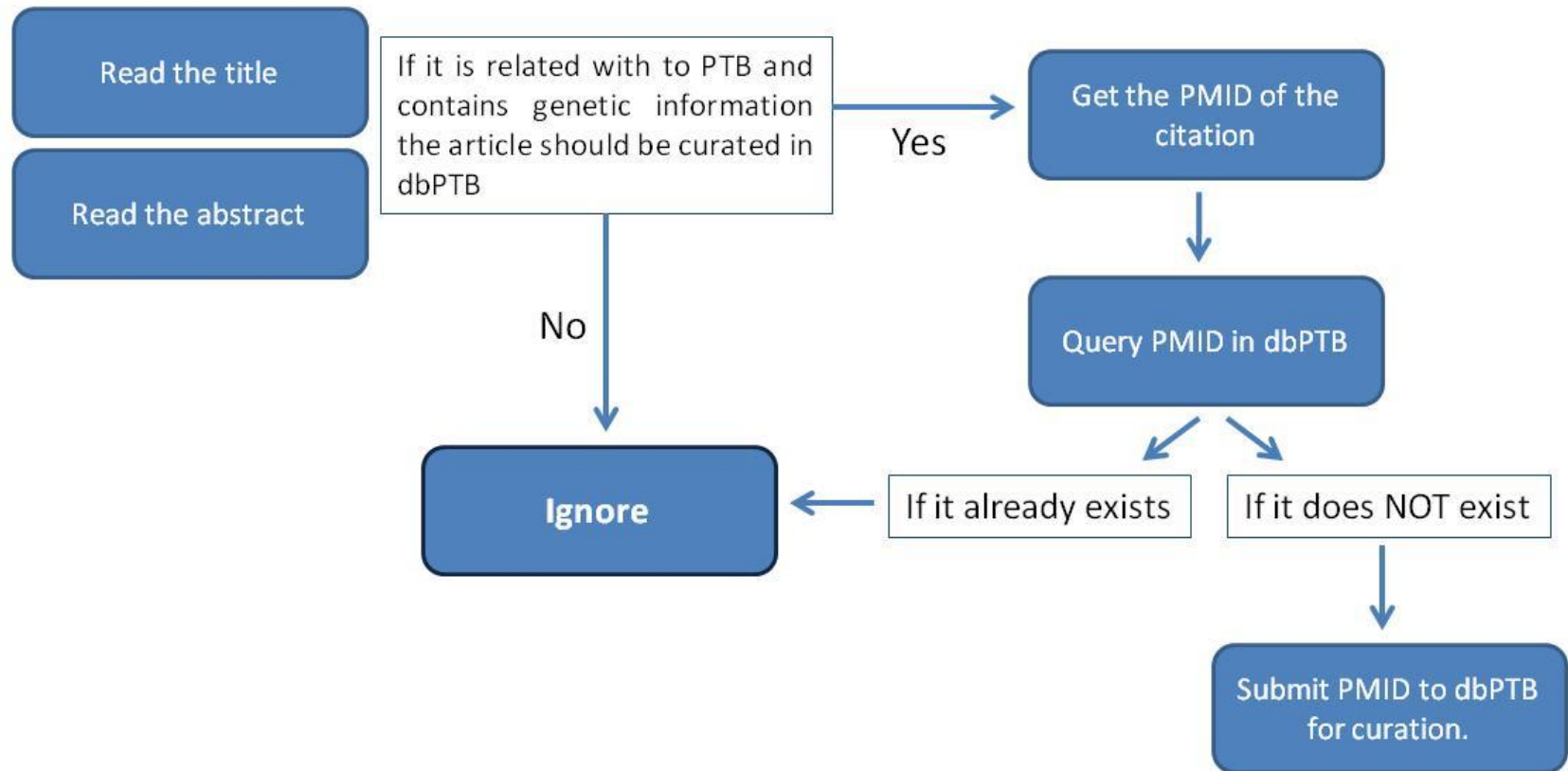


Figure 3. Schematic for keeping curation notes

How to keep a curation log notes

