High Frequency Gene Filtering

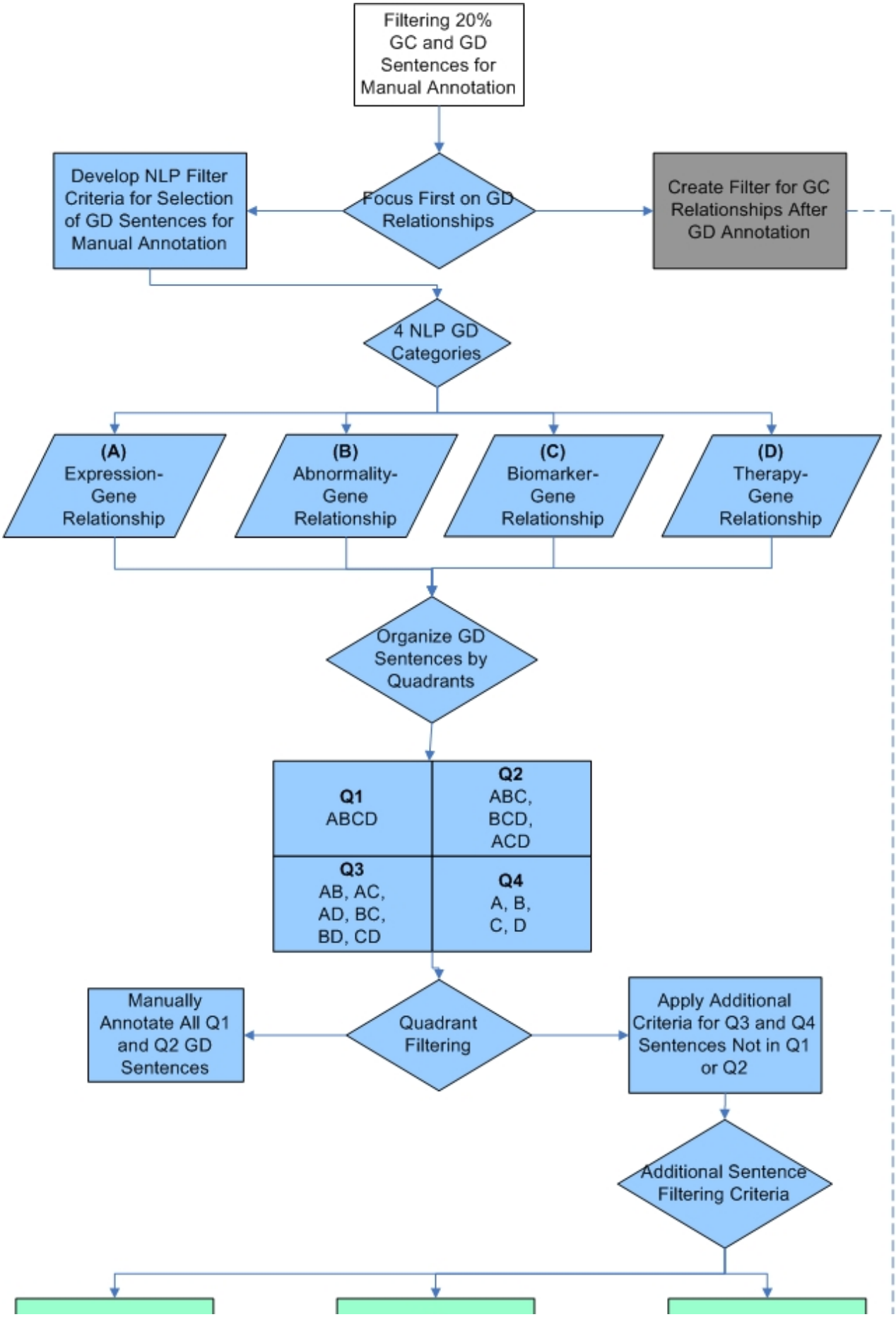
To Print the Guide

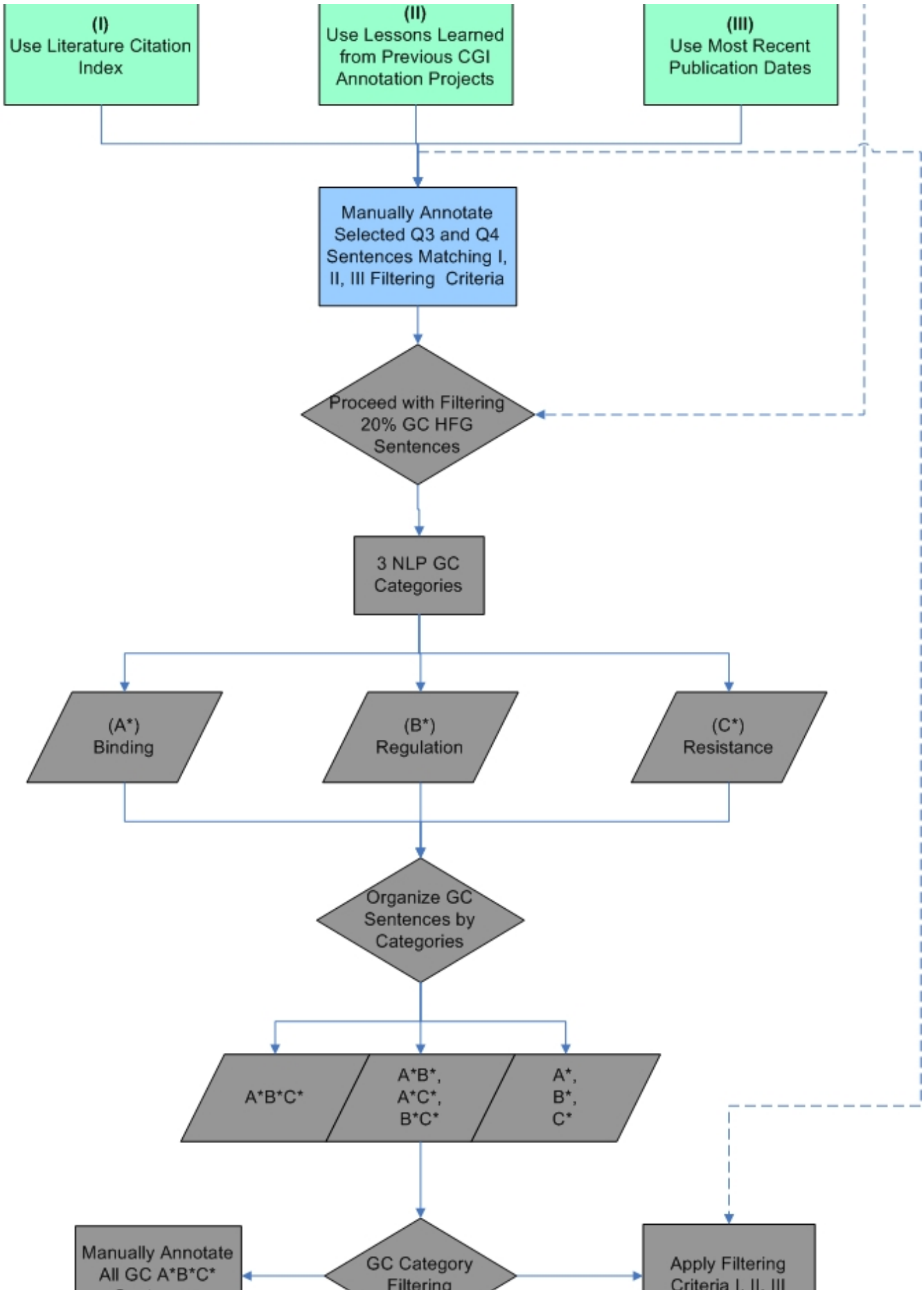
We recommend you print one wiki page of the guide at a time. To do this, click the printer icon at the top right of the page; then from the browser File menu, choose Print. Printing multiple pages at one time is more complex. For instructions, refer to Printing multiple pages.

High Frequency Sentence Count Gene Filtering

Early on, the decision was made to focus first on filtering HFG gene-disease (GD, blue shapes) sentences and then to go back to HFG gene-compound (GC, green shapes and dotted line) sentences. Natural language processing (NLP) filtering found that GD sentences described Expression-Gene Relationships (A), Abnormality-Gene Relationships (B), Biomarker-Gene Relationships (C), and/or Therapy-Gene Relationships (D). Thus, the GD sentences were classified into "quadrants" where Q1 sentences described all four relationship categories, Q2 any three categories, Q3 any two categories, and Q4 only one of the four categories. Q3 and Q4 sentences were all manually curated. Q1 and Q2 sentences were subjected to additional filtering criteria, and the three or four sentences from each of the two categories were selected for manual curation.

A similar approach (green shapes and lower dotted line) was taken for GC sentences, but NLP analysis of these pieces of evidence uncovered three relationship categories: Binding (A*), Regulation (B*), and Resistance (C*). All A*B*C* sentences (i.e., sentences describing all three GC categories) were manually curated. The remaining sentences were subjected to additional filtering steps, as before, to select those sentences that would be manually curated. Here, blue denotes GD flowchart objects, gray GC, and green both GD and GC. Dotted lines represent steps that occurred later in the GD workflow.


Filtering 20% GC and GD Sentences for Manual Annotation
Develop NLP Filter Criteria for Selection of GD Sentences for Manual Annotation
Focus First on GD Relationships
Create Filter for GC Relationships After GD Annotation
4 NLP GD Categories
(A) Expression-Gene Relationship
(B) Abnormality-Gene Relationship
(C) Biomarker-Gene Relationship
(D) Therapy-Gene Relationship
Organize GD Sentences by Quadrants
Q1 ABCD
Q2 ABC, BCD, ACD
Q3 AB, AC, AD, BC, BD, CD
Q4 A, B, C, D
Manually Annotate All Q1 and Q2 GD Sentences
Quadrant Filtering
Apply Additional Criteria for Q3 and Q4 Sentences Not in Q1 or Q2
Additional Sentence Filtering Criteria


(I)
Use Literature Citation Index
(II)
Use Lessons Learned from Previous CGI Annotation Projects
(III)
Use Most Recent Publication Dates
Manually Annotate Selected Q3 and Q4 Sentences Matching I, II, III Filtering Criteria
Proceed with Filtering 20% GC HFG Sentences
3 NLP GC Categories
(A*)
Binding
(B*)
Regulation
(C*)
Resistance
Organize GC Sentences by Categories
A*B*C*
A*B*, A*C*, B*C*
A*, B*, C*
Manually Annotate All GC A*B*C*
GC Category Filtering
Apply Filtering Criteria I, II, III


Sentences
Manually Annotate Selected GC Sentences