Sophic
Integrated Life Science Software and Services

The Cancer Genome Atlas Project Overview

In September 2006, The Cancer Genome Atlas (TCGA) program announced plans to map genomic changes of lung, brain and ovarian cancer tumors using large scale sequencing technologies. Seven institutions in five states were selected as Cancer Genome Characterization Centers (CGCCs). CGCCs included the Broad Institute of MIT and Harvard; Harvard Medical School and Brigham and Women's Hospital, Lawrence Berkeley National Laboratory, Memorial Sloan-Kettering Cancer Center, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Stanford University School of Medicine, and the University of North Carolina Lineberger Comprehensive Cancer Center.

The analysis of sequence information generated by the CGCCs would entail trying to identify genes that were involved in the mutation or metastasis process for the lung, brain or ovarian tumors. To support the identification process, in April 2007, the Sophic-Biomax teamed with NCI to annotate the 3,168 Brain, Lung and Ovarian cancer genes previously identified in the Cancer Gene Index Project. This project entailed a deeper review of both abstracts and full text papers to find only those genes involved in mutation and metastasis processes. Evidence for the abnormality and mutation process would be annotated by PhD scientists by reviewing both Medline abstracts and full text papers. Below is a breakdown of the 3,168 genes by disease type to be reviewed in the project.

Brain – 1,143
Lung – 1,182
Ovarian – 843
Total- 3,168

The Medline abstract corpus was updated to include all publications through January 1, 2007 and NCI made full text papers available for review. Linguistic analysis of the new material was performed to update and refresh the gene name, cancer disease terms and new categories were added for abnormality and metastasis processes. PhD scientists followed the roadmap below to identify Brain (B) , Lung (L) and Ovarian (O) cancer genes involved in the abnormality and metastasis process.

2275 Research Blvd., #500, Rockville, MD 20850
Office (301) 296-4554 Fax (301) 330-0740
www.sophicalliance.com

The annotators organized the target annotated genes into three categories:

- **Category A:** Genes which relate to all three types of cancer disease (B&L&O)
- **Category B:** Genes which relate to two of the three types of cancer diseases - (B&L, B&O, L&O).
- **Category C:** Genes which relate to only a single type of cancer disease (B,O,L).

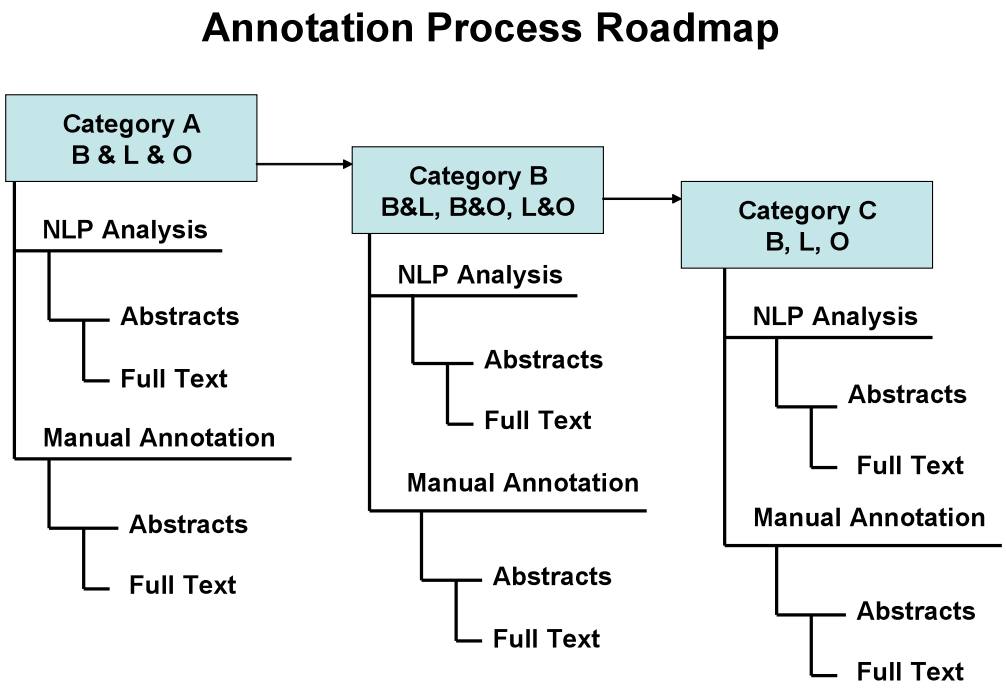

Figure 1: The categorization process was designed to structure the targeted genes set and identify the molecular abnormality abstracts and full text papers. NLP analysis was performed on abstracts and full text papers using Biomax BioLT Linguistic tool for each category gene to detect molecular abnormality information. Custom dictionaries with expanded abnormality and metastasis terms were used in the NLP analysis.

Rules for the Manual Annotation Process:

1. Abnormalities related to expression
    a. For abnormalities related to gene or protein expression it is sufficient to give information about the type of expression change (expression up-regulated,

expression down-regulated, expression changed). It is not required to give detailed numerical information about the respective change in expression.
   b. Usually the information about the expression change can be found in the abstract. Otherwise the annotator has to check the full text.
2. Abnormalities related to epigenetic changes
   a. For abnormalities related to epigenetic changes it is sufficient to give information about the type of change (hyper-methylation, hypo-methylation). It is not required to give information about the respective nucleotide(s) affected.
   b. Usually the information about the epigenetic change can be found in the abstract. Otherwise the annotator has to check the full text.
3. Abnormalities related to posttranslational modifications
   a. Posttranslational modifications include changes of phosphorylation, glycosylation (not in Version 1 of the abnormality dictionary, will be included in Version 2)
   b. For abnormalities related to posttranslational modifications it is sufficient to give information about the type of change (hyper-phosphorylation, hypo-phosphorylation). It is not required to give detailed information on the amino-acid(s) that is affected.
   c. Usually the information about the posttranslational modification can be found in the abstract. Otherwise the annotator has to check the full text.
4. Abnormalities related to mutations
   a. Mutations include: gene mutations, chromosomal changes, SNPs, polymorphisms, rearrangements
   b. For abnormalities related to mutations a detailed description of the respective mutation is to be given, e.g. detailed information about the respective base-pair change, amino-acid change.
   c. Usually this detailed information can not be deduced from the abstract. Thus the full text has to be used to get the detailed information.

Atlas Project Results

In September 2007, the manually annotated genes were delivered to NCI. Of the 3,168 cancer brain, lung and ovarian genes reviewed, 577 were identified as genes related to mutation and metastasis process.

Sophic
Integrated Life Science Software and Services

Insilco Research using the Cancer Genome Atlas Genes

The Atlas Cancer Genes have been integrated into various use cases in the BioXM Knowledge Management System to support Insilco research. The system allows scientists to explore brain, ovarian and lung cancer genes involved in the abnormality process and find relationships with other scientific elements such as pathways, chemical compounds and information contained in clinical trials databases.

The screen image below (figure ?) shows 11 brain cancer genes with increases in gene expression experiments which contributed to the abnormality process. Figure ? shows a sample report the CNTF cancer gene with the sentence, NCI thesaurus role code and evidence code in the specific sentence from the PubMed abstract annotated by PhD curators.

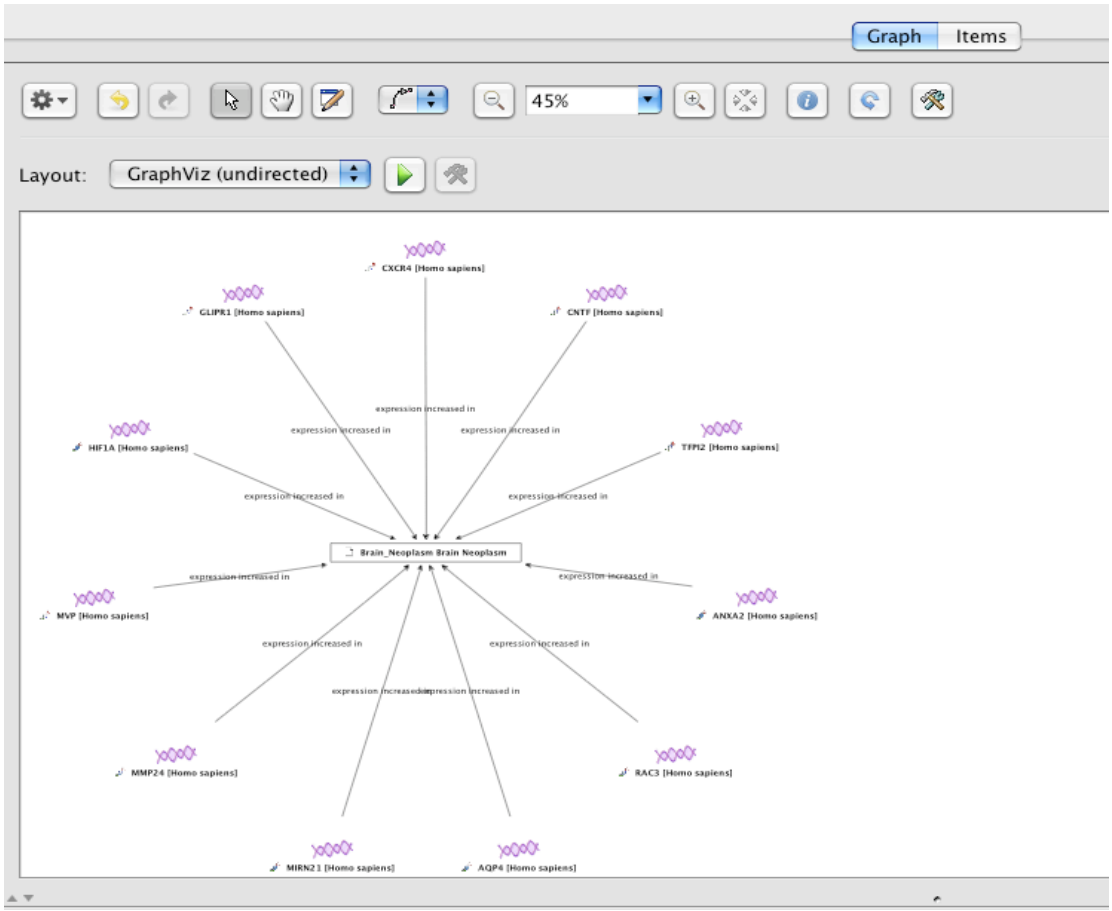

Figure 2: 11 Brain Neoplasm genes with increased gene expression involved in mutation and abnormality process.

2275 Research Blvd., #500, Rockville, MD 20850
Office (301) 296-4554 Fax (301) 330-0740
www.sophicalliance.com

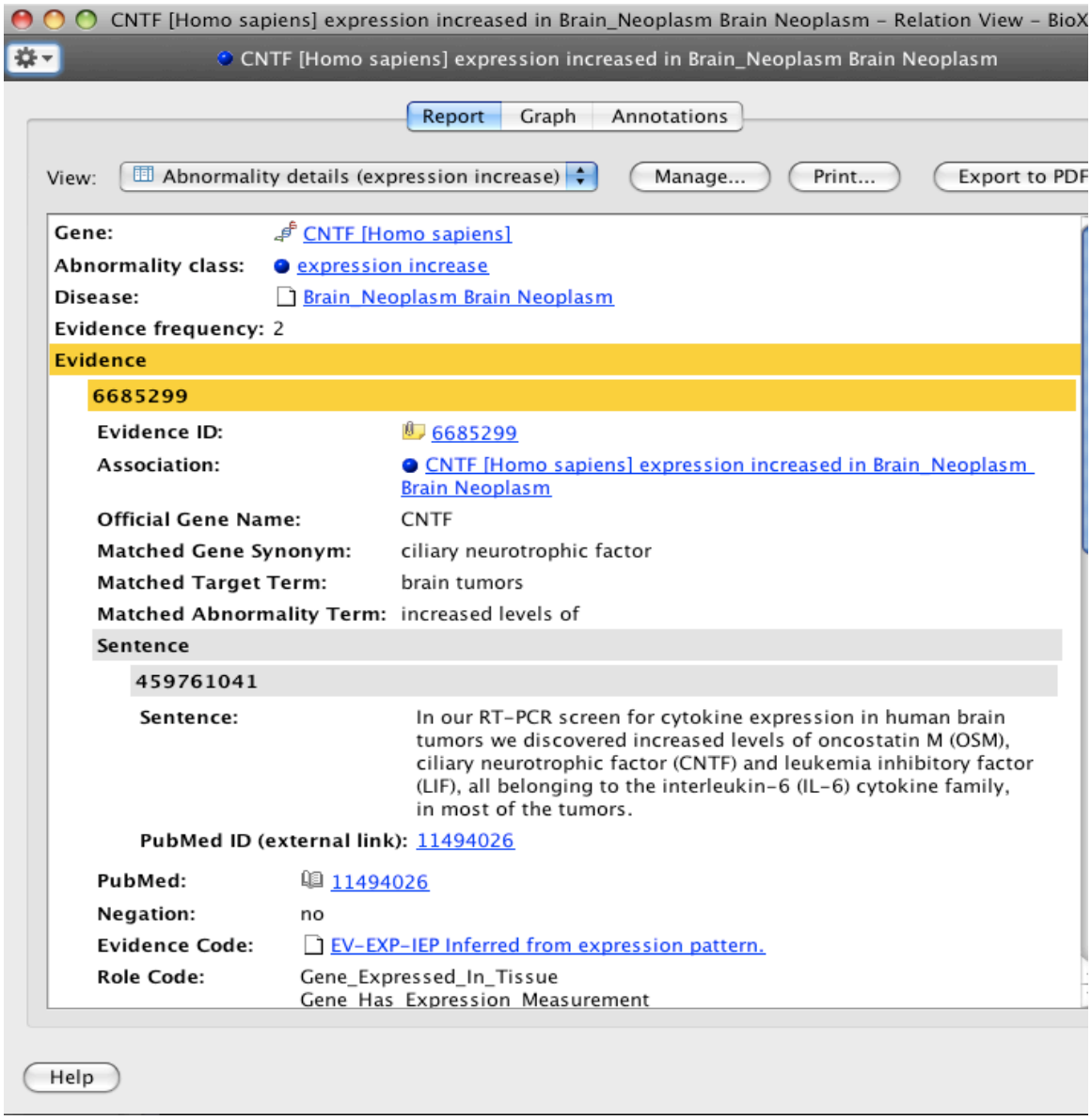

Figure 3: A gene report on CNTF brain cancer gene provides the PubMed sentence, a link to the abstract and NCI Role Codes and Evidence Codes for the abnormality process. All Atlas Genes have similar levels of information manually annotated by PhD curators.