Sophic™
Integrated Knowledge Environment
Software and Services

A White Paper on SCan-MarK™ Explorer

The Sophic Cancer Biomarker Knowledge Environment

I. Abstract: The three-year SCan-MarK™ Explorer Phase I and II NCI Small Business Innovation Research (SBIR) Project entailed mining and manually curating select full text clinical research papers to extract Critical Data Elements (CDEs) compiled into Sophic Cancer Biomarker Objects (SCBOs) of breast, ovarian, colorectal, non-Hodgkin's lymphoma and melanoma biomarker targets. SCBOs are computational objects that include valuable, hard-to-find experimental data elements that support scientists *across* the cancer community. SCBO CDEs include related elements that support and connect basic research, diagnostics, drug discovery and, ultimately, clinical patient care. SCan-MarK Explorer is intended to reduce time wasted and delays in finding cures for cancer. SCBOs can be individual biomarkers or panels that have been integrated by Sophic into the Biomax BioXM Knowledge Management System (developed by Biomax AG, Munich, Germany). The SCan-MarK Explorer configuration of BioXM allows non-technical researchers to mine, search, discover and visualize complex networks of valid scientific or semantic relationships, connecting biomarkers to diseases to critical data elements and enrichment databases. During the SBIR project, Sophic collaborated with a scientific advisory board (SAB) that included senior cancer researchers at NCI, cancer hospitals, research centers and pharmaceutical companies. SAB members provided recommendations and feedback during the project in quarterly review meetings and were critical to the success of the project. The aim of this ongoing project is to support research, diagnostics, and drug discovery, improve the accuracy of disease diagnosis, increase the effectiveness of treatments and accelerate the discovery of cancer cures.

II. Sophic Cancer Biomarker Objects (SCBOs). A Sophic Cancer Biomarker Object (**SCBO**) may consist of individual or panels of genes, proteins or chemical elements that include all or most of the Critical Data Elements (**CDEs**) detailed below. Currently, there are up to 33 CDEs that are mined, manually curated by Sophic scientists, and compiled into disease-specific SCBOs. Each CDE has validated scientific relationships with other CDEs, forming a knowledge network of "building blocks" that represent complex biomarker information. While SCan-MarK SCBOs target cancer diseases, SCBOs can be used to create similar "building blocks" for other complex diseases such as cardio vascular, Alzheimer's, Autism, etc.

III. Sophic's Automated and Manual Curation Process. The Sophic mining and manual curation process is a highly structured, quality-driven, "factory-like" process conducted by experienced, US-based, PhD scientists . SCan-MarK information is continually mined from a range of high-quality scientific and clinical research databases rich with CDEs, HIPAA-compliant clinical information from de-identified patient-related sources (none of PHI) and

peer reviewed papers from basic, translational and clinical publications with data that are associated with clinical outcomes.

Sophic's "factory-like" manual curation process includes 6 steps:

1. Sophic scientists use a range of powerful linguistic tools to mine Medline abstracts to identify ONLY papers focused on human clinical studies for 5 cancer disease types (breast, ovarian, colorectal, melanoma and Non-Hodgkin's lymphoma) that are rich in documented, detailed experimental results.
2. Quality assurance reviews are conducted on each candidate paper identified in the automated mining process to confirm that selected papers are rich in bench test results prior to manual curation.
3. Manual curation entails PhD scientists reading, studying and manually curating the papers to identify up to 33 critical data elements (CDEs) in the text and supplemental data.
4. Detailed curation results are carefully recorded on annotation forms that allow a second review of the paper, and on a final annotation form to insure quality, accuracy and completeness.
5. The CDEs in the annotation forms are organized and integrated into a SCBO (individual and panels).
6. Quality-assured SCBOs are finally uploaded and integrated into the SCan-MarK data model as building blocks that are the primary reference and information backbone for SCan-MarK Explorer

IV. SCBOs and CDEs. Common Data Elements including Experimental Evidence, Sensitivity, Specificity, Number of Patients, Survival p Values, Risk factors and Clinical Trials, are used to rank our biomarkers. The CDEs provide a structured metric with detailed information about the quality, research and confidence stage of the target biomarkers.

Evidence codes (Karp et al., 2004) and evidence count metrics used in the Cancer Gene Project (5 year, Sophic NCI-funded mining and manual curation project) are used to rank the quality of evidence for each SCBO (target biomarker). Evidence codes are divided into 3 tiers based on the rigor of the authors' experiments. The result is a confidence ranking in three tiers: Experimental Evidence (EXP), Computational Evidence (COMP), and Author Statement and Assertion (AS). Both evidence type and evidence count are used to establish biomarker ranking.

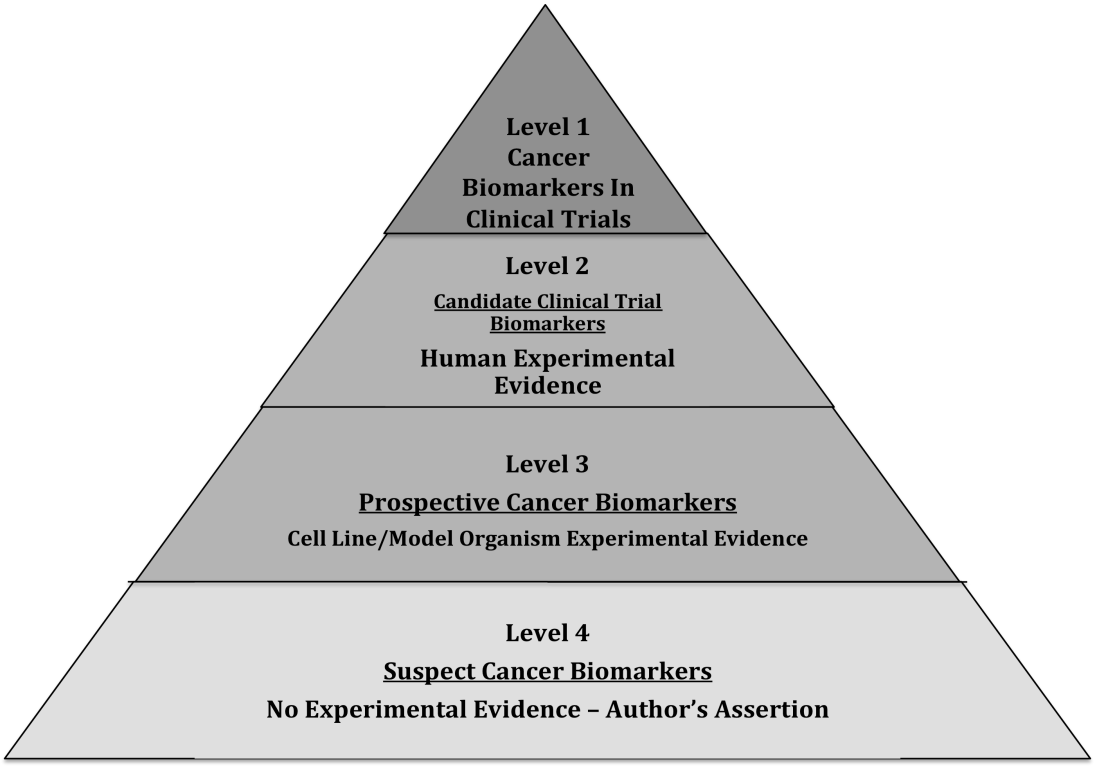
The Sophic curators collect the following critical data elements (CDEs) on each potential individual biomarker or panel of potential biomarkers that can be genes, cellular proteins, microRNAs, DNA-sequence changes (i.e. SNP or mutations) or serum proteins. CDEs include:

1. **Gene Name:** HUGO gene name
2. **Pubmed ID:** The PubMed ID of the paper with research and/or clinical biomarker information that is validated with patient samples.
3. **Other Source:** If the source is other than a publication, list the name of the source, the URL, etc., that provides a link to the source.
4. **Source Organism(s):** Human, animal models or cell line studies in vitro
5. **Communication Author:** Full name of the communication author
6. **Email:** Communication author's email address
7. **Evidence code:** Evidence codes are divided into 3 tiers based on the rigor of the authors' experiments. The result is a confidence ranking in three tiers: Experimental Evidence (EXP), Computational Evidence (COMP), and Author Statement and Assertion (AS). Sophic scientists start from the Evidence codes cited in the Cancer Gene Index detailed reports and modify them to reflect the full text paper and the data provided.
8. **Experimental evidence**: genomics (i.e. Microarray), proteomics, Enzyme assays, in vitro reconstitution (e.g. transcription), Immunofluorescence, Cell fractionation, etc. that are used to validate potential biomarkers.
9. **Compound or treatment:** compounds or treatments that are used during the experiments and may be associated with the biomarker.
10. **Biomarker Usage**: Screening markers (diagnostic), Prognostic markers (Prediction of course of disease), Stratification markers (Predict drug response), Efficacy markers (Monitor treatments), Toxicity markers (demonstrate adverse reaction to therapy) will be documented based on author's claim.
11. **Biomarker Type**: **Wild-type Gene**: Increased expression, decreased expression, copy number change; **Mutant Gene**: insertion, deletion, point mutation, repeat; **Modified Gene** – methylated; Splice Variants; **Wild-type Protein**: Increased expression, decreased expression; **Mutant Protein**: insertion, deletion, point mutation, repeat; **Modified protein**: phosphorylated, acetylated, methylated, ubiquitylated, sumoylated, neddylated, carbohydrate, lipid etc.
12. **Cancer Type:** Generic name. e.g. Ovarian Cancer.
13. **Cancer Subtype**: subtype of the cancer, e.g. serous carcinoma**.**
14. **Cell Line:** cell line(s) used for the analysis.
15. **Patient Tissue:** surgical specimen, blood, frozen, paraffin embedded: fresh tissues used cases get the highest scores, followed by frozen and then paraffin embedded.
16. **Clinical Trial** (yes, no, and Phase)
17. **Number of patients:** number of patients involved in the study
18. **Number of patients, subtype:** if the data is broken down by cancer subtype, use the highest number here. Use the format "cancer subtype: patient number"
19. **Number of controls:** number of normal controls involved in the study.
20. **Number of controls, subtype:** If the data is broken down by cancer subtype, use the highest number here. Use the format "control subtype: control number".
21. **Sensitivity:** if the data is broken down by cancer subtype, use the highest number here.

22. **Sensitivity Subtype:** if the data is broken down by cancer subtype, use the following format for data entry: *subtype1:xx%; subtype2:yy%*
23. **Specificity:** if the data is broken down by cancer subtype, use the highest number here.
24. **Specificity Subtype:** if the data is broken down by cancer subtype, use the following format for data entry: *subtype1:xx%; subtype2:yy%*
25. **Correlation: P value:** experimentally determined association, where each correlated factor is listed and the p-value for the association is reported (only for significant – p-value < 0.05 – correlations)
26. **Specificity P Value**: The P value associated with the biomarker and the clinical element described. Data is in the form *Clinical_Element: P value*
27. **Disease Risk Factor:** odds ratio for epidemiology, familial mutations, and prognostic studies to show whether the patient develops the disease.
28. **Disease-specific survival P Value**: The P value for the biomarker and the ability to predict a change in disease specific survival (DSS). Data is in the form *Element: P value*
similar scoring system as in #21.
29. **Relapse-free survival P Value**: The P value for the biomarker and the ability to predict a change in disease free survival (RFS). Data is in the form *Element: P value*
30. **Disease-specific survival Risk Factor**: Risk factor for the biomarker in terms of disease specific survival (DSS). Can be described as Odds Ratio (OR), Hazard Ratio (HR), or Relative Risk (RR). Odds ratio is the ratio of probabilities between 2 conditions. OR >1 is more likely, OR <1 is less likely. Hazard ratio is the risk of an event between 2 populations (often used with Kaplan-Meier curves). Relative Risk is similar to odds ratio but is a simple ratio of events occurring.
31. **Relapse-free survival Risk factor**: Risk factor for the biomarker in terms of relapse free survival (RFS). Can be described as Odds Ratio (OR), Hazard Ratio (HR), or Relative Risk (RR). Odds ratio is the ratio of probabilities between 2 conditions. OR >1 is more likely, OR <1 is less likely. Hazard ratio is the risk of an event between 2 populations (often used with Kaplan-Meier curves). Relative Risk is similar to odds ratio but is a simple ratio of events occurring.
32. **Gene Description**: A brief description of this gene or protein. Link to gene information. Form: http://www.ncbi.nlm.nih.gov/gene/ <<insert gene ID here>>
33. **Paper Summary:** Link to the PubMed abstract. Form: http://www.ncbi.nlm.nih.gov/pubmed/ <<insert PubMed id number here>>. Based on the number of criteria counts and quantitative/statistical approaches we will classify potential biomarkers into four evidence based categories displayed in a four level pyramid. FDA approved biomarkers are not included in this "pipeline."
34. **Charts and Images:** Kaplan-Meier curves or other relevant information in graphical format will be extracted from the publications for selected SCBOs.

IV. Classification of SCan-MarK Biomarkers. Sophic scientists developed a structured method to classify potential cancer biomarkers in a "pipeline" used to identify stages of research progress, development and confidence levels starting with "biomarker suspects" (low evidence and confidence) through to biomarkers that are "in clinical trials". FDA-approved biomarkers are used as positive controls to test Scan-MarK algorithms to validate metrics and individual biomarker scores. Below is an image of the Sophic Biomarker Classification Pyramid and the category classification metrics.

Sophic has focused almost entirely on mining and manually curating Level 2, high evidence and confidence Candidate Clinical Trial Biomarkers for SCan-MarK Explorer. These are the most promising biomarker candidates that can accelerate basic research, diagnostics and drug development and eventually inform clinicians on the best therapies for patients.



Sophic's metrics for classifying SCBOs into one of the four levels in the pyramid are:

Level 1 - Cancer Biomarkers In Clinical Trials. Cancer biomarkers that have successfully been tested in clinical trials or are currently being evaluated in one or more clinical trials.

Level 2 - Candidate Clinical Trial Cancer Biomarkers. Potential biomarkers with evidence from human tissue experiments. Most SCBOs with manually curated CDEs fall into this category.
Level 3 - Prospective Cancer Biomarkers. Potential biomarkers with evidence from cell or animal experiments.
Level 4 - Suspect Cancer Biomarkers. Potential biomarkers with no experimental evidence. Associations with known biomarkers and author's assertion of biomarker potential are included in this level.

V. SCan-MarK Explorer Enrichment Databases: *Scientific information viewed in isolation is rarely as enlightening as information viewed in a network of valid, scientific or semantic relationships.* Visualization of "knowledge networks" provide a broad, connected view of complex scientific and research maps. SCan-MarK uses SCBOs as anchors for connecting related information found in over 27 enrichment databases. SCan-MarK Explorer Enrichment Databases include**:**

1. **Biosystems** currently contains records from several source database: KEGG, BioCyc, Reactone, and the National Cancer Institute's Pathway Interaction Database. The BioSystems database includes several types of records such as pathways, structural complexes, and functional sets. It is designed to accommodate other record types, such as diseases, as data become available. Through these collaborations, the BioSystems database facilitates access to, and provides the ability to compute on, a wide range of biosystems data. Detailed diagrams and annotations for individual biosystems are then available on the web sites of the source databases.

2. **Cancer Gene Index** is a NCI founded and Sophic and Biomax carried 5 year project (CaBIG initiative) to manually identify, curate and annotate all cancer genes found in PubMed abstracts. Over 18 million PubMed abstracts and 94 million sentences are mined, and 1.3 million abstract sentences are manually curated to generate the 6,695 "true" cancer genes. NCI Thesaurus, Role Codes (Carp et al) are used to classify each relationship, and Evidence Codes are used to qualify the methods used by authors to generate data and conclusions.

3. **ChEBI ontology** is an ontology for biologically interesting chemistry. It consists of three sub-ontologies, namely: **Molecular Structure**, in which molecular entities or parts thereof are classified according to their structure; **Role**, in which entities are classified on the basis of their role within a biological context, e.g. as antibiotics, antiviral agents,

coenzymes, enzyme inhibitors, or on the basis of their intended use by humans, e.g. as pesticides, detergents, healthcare products, fuel; and **Subatomic Particle**, in which are classified particles which are smaller than atoms.

4. **ChEMBL** is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) The data is abstracted and curated from the primary scientific literature, and cover a significant fraction of the SAR and discovery of modern drugs

5. **ChEMBL Assays** is the abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data) from the ChEMBL database. EMBL attempts to normalize the bioactivities into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels

6. **ClinicalTrials** is a registry and results database of federally and privately supported clinical trials conducted in the United States and around the world. It gives you information about a trial's purpose, who may participate, locations, and phone numbers for more details.

7. **dbSNP** is the single nucleotide polymorphism database from the NCBI. The database is meant to be a central repository for single base substitutions as well as short deletions and insertions. SNPs that occur within genes are annotated with the gene accession number and can be easily mapped back to the gene. Data is provided for the location on the contig sequence from a variety of genome builds so the mutations can be correlated with the sequence.

8. **DrugBank** is offered to the public as a freely available resource. Use and re-distribution of the data, in whole or in part, for commercial purposes requires explicit permission of the authors and explicit acknowledgment of the source material (DrugBank) and the original publication. We ask that users who download significant portions of the database cite the DrugBank paper in any resulting publications.

9. **ENSEMBL** is a joint project of the European Bioinformatics Institute (EBI) and the Sanger Centre. The Ensembl database project provides a bioinformatics framework to organize biology around the sequences of large genomes. It is a comprehensive source of stable

automatic annotation of the human genome sequence, with confirmed gene predictions that have been integrated with external data sources, and is available as either an interactive web site or as flat files.

10. **EntrezGene** integrates information from a wide range of species. The record may include nomenclature, Reference Sequences, maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

11. **Enzyme Nomenclature Database, ENZYME**, is a repository of information relative to the nomenclature of enzymes provided by the Swiss Institute of Bioinformatics. ENZYME is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Each type of characterized enzyme for which an Enzyme Commission (EC) number has been provided is described.

12. **FDA drug labels** and other drug-specific information represent the most recent drug listing information companies have submitted to the Food and Drug Administration (FDA).

13. **GO (Gene Ontology) Database**, maintained by the Gene Ontology Consortium, provides controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products. The terms are used as attributes of gene products by collaborating databases, facilitating uniform queries across them. The controlled vocabularies of terms are structured to allow both attribution and querying to be at different levels of granularity. Databases external to GO collaborate with GO in three ways: by making database cross-links between GO terms and objects in their database (typically, gene products, or their surrogates, genes), and then providing tables of these links to GO (and hence the community), second by supporting queries that use these terms in their database, and third by contributing to the development of the GO database itself expanding the vocabularies and refining the terms.

14. **Homologene** is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

15. **Interpro** (The Integrated Resource of Protein Families, Domains and Sites) is collaborative effort funded by the European Commission to provide an integrated interface for the

commonly used signature databases for text- and sequence-based searches. Collaborating databases include PROSITE, PRINTS, ProDom and Pfam, SMART and TIGRFAMs.

16. ***KEGG** is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information).

17. **Medline** is compiled by the U.S. National Library of Medicine (NLM, Bethesda, USA) and published on the Web by Community of Science, MEDLINE is the world's most comprehensive source of life sciences and biomedical bibliographic information. It contains nearly eleven million records from over 7,300 different publications from 1965 to present. MEDLINE is updated weekly.

18. **miRBase** is a searchable database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database), with information on the location and sequence of the mature miRNA sequence (termed miR). Both hairpin and mature sequences are available.

19. ***OMIM** is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

20. **PDB** archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

21. **PDQ** is NCI's comprehensive cancer database. It contains summaries of 8,000+ open and 19,000+ closed cancer clinical trials from around the world.

22. **Pfam (PfamA and PfamB)** is a database of protein domain families produced by the Sanger Centre (Cambridge, UK). Pfam contains curated multiple sequence alignments for each family, as well as profile Hidden Markov Models for finding these domains in new sequences. Pfam contains functional annotation, literature references and database links for each family.

23. **Pubchem** provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative.

24. **Reactome Pathway** annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many bioinformatics databases. The rationale behind Reactome is to convey the rich information in the visual representations of biological pathways familiar from textbooks and articles in a detailed, computationally accessible format. The core unit of the Reactome data model is the reaction. Entities (nucleic acids, proteins, complexes and small molecules) participating in reactions form a network of biological interactions and are grouped into pathways.

25. **Refseq** (NCBI Reference Sequence) project provides reference sequence standards for the naturally occurring biological molecules, from chromosomes to proteins. RefSeq standards provide a foundation for the functional annotation of the human genome. In addition, these standards provide a stable reference point for mutation analysis, gene expression studies and polymorphism discovery.

26. **Sanger's COSMIC Database and Sophic's Non-redundant COSMIC Database. Sanger's COSMIC Database** is designed to store and display somatic mutation information and related details and contains information relating to human cancers. The mutation data and associated information is extracted from the primary literature and entered into the COSMIC database. In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene. The data can be queried by tissue, histology or gene and displayed as a graph, as a table or exported in various formats. **Sophic's Non-redundant COSMIC Database –** Sophic removed the redundant entries and added statistics including number of mutations reported for the same mutation, and the total occurrence of mutations found for a particular gene.

27. **Taxonomy Database** is a curated classification and nomenclature for all the organisms in the public sequence database.

28. **The Cancer Genome Atlas (TCGA)** is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. It is a project that seeks to improve our ability to diagnose, treat, and prevent cancer through a better understanding of the molecular basis of this disease.

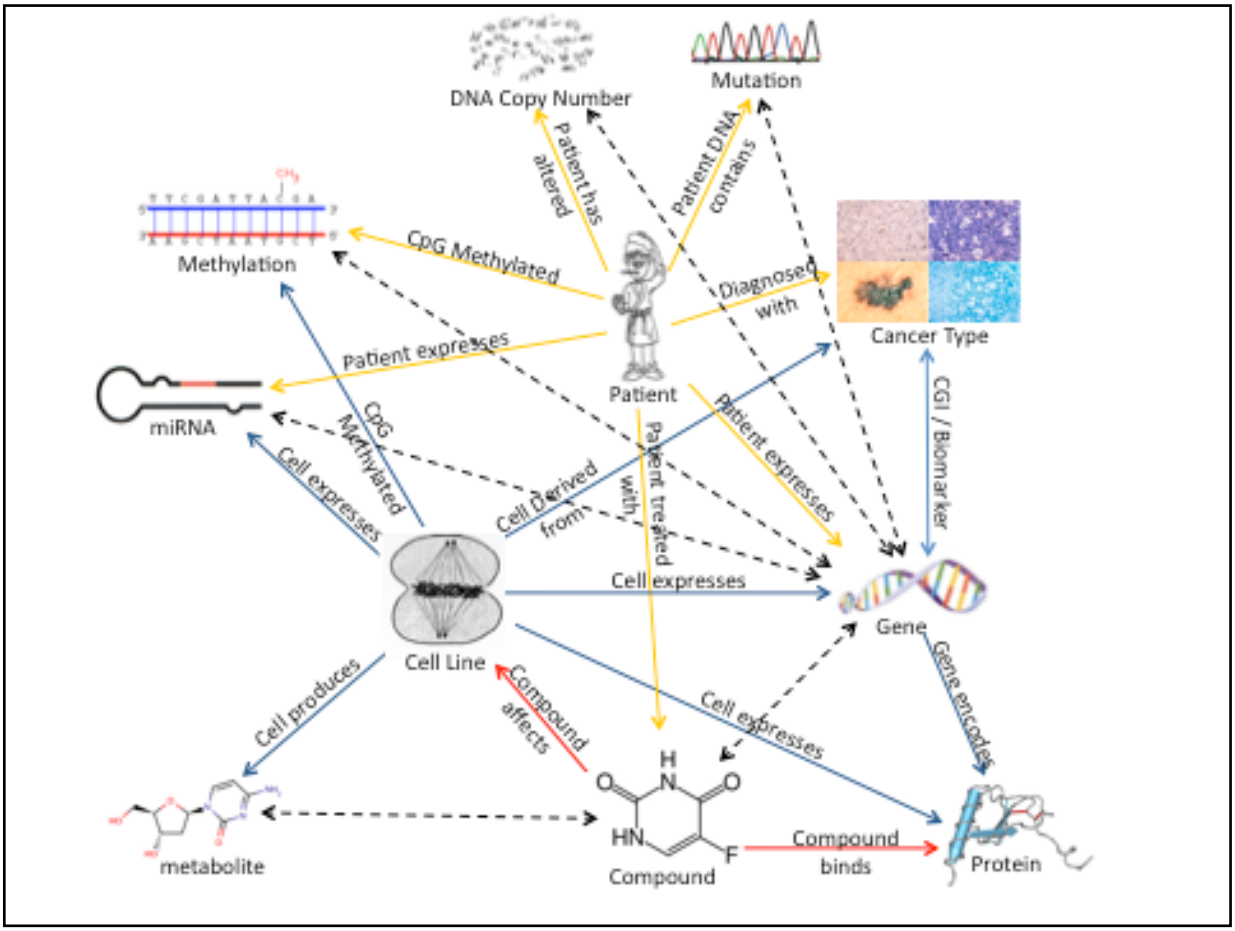
29. **Unigene** is a NCBI experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

30. **Uniprot** (Universal Protein Resource) is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components, each optimized for different uses. The UniProt Knowledgebase (UniProt) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

.

∗ **indicate that these databases require customers to pay a license fee to the developers of these databases.**

VI SCan-MarK Explorer Data Model. The SCan-MarK BioXM data model is the GPS that organizes, connects and gives direction to queries launched in SCan-MarK Explorer. The data model evolves as SCan-MarK expands and is configured to support the specific requirements of each customer.



VI. SCan-MarK WIKI End User Console. Many non-technical scientists and clinicians are often confused and frustrated with the complexity of integrated software. SCan-MarK Explorer is designed to "look and feel familiar" to the users due the close resemblance of today's cell phone interface-like console. Tool tips make using SCan-MarK simple and reinforce the intuitive design.