D6: Molecular Subtyping Recommendations

Date: September 29, 2017
Contributors: Matthew Brush, Jim Balhoff, Nicole Vasilevsky, Melissa Haendel

Contents

1 Background and Goals

With the emergence of new experimental and analytic technologies, diseases are increasingly characterized and classified by their molecular features. This is particularly relevant for heterogeneous and dynamic diseases such as cancer, where molecular data is driving efforts to define subtypes in a way that will inform detection, diagnosis, and treatment strategies. As the field of cancer informatics matures, once siloed molecular datasets and unstructured published findings become integrated and codified in more standardized community resources. These include data repositories such as the TCGA, curated knowledgebases such as CIViC and ClinVar, and ontological resources such as the NCIt.

Over the past decade, the NCIt has grown to become a widely-used resource for expert-curated knowledge about molecular features of cancers. As an ontology however, the NCIt has evolved independently of other biomedical ontologies and the best practices that have been established in this community. As a result, it lacks a fundamental level of interoperability that would support integrated analysis of data across systems that leverage these different sets of ontological resources.

This report evaluates the representation of knowledge related to molecular subtyping of cancer, and identifies areas for improvement with respect to modeling consistency, content, and alignment with external standards and best practices. We aim to avoid critiques made on purely ontological principles, as the NCIt has been evaluated from this perspective in many reports [1-3] which often devolve into subjective philosophical criticism. We will instead identify opportunities where improved NCIt modeling and infrastructure might accrue real benefits for its curation and maintenance, its ability to interoperate with and leverage external resources, and its utility for bioinformatics and big-data applications.

In the text that follows, **Section 2** provides an objective overview of the scope and representation of information relevant to molecular subtyping in the NCIt. **Section 3** evaluates and makes recommendations concerning identifier strategies. And **Section 4** suggests concrete modeling improvements in three specific areas describing molecular features of cancer.

2 Molecular Subtyping Information in the NCIt

2.1 Molecular Subtyping Conceptual Model

Molecular features relevant to subtyping efforts fall primarily into three categories: (1) the genetic lesions that drive or correlate with disease state; (2) the expression levels and localization of molecular entities in the body; and (3) and the activities of signaling pathways and complexes in the cell. **Figure 1** diagrams the high-level types and design patterns used in the NCIt to describe such molecular features of the cancers it catalogs.

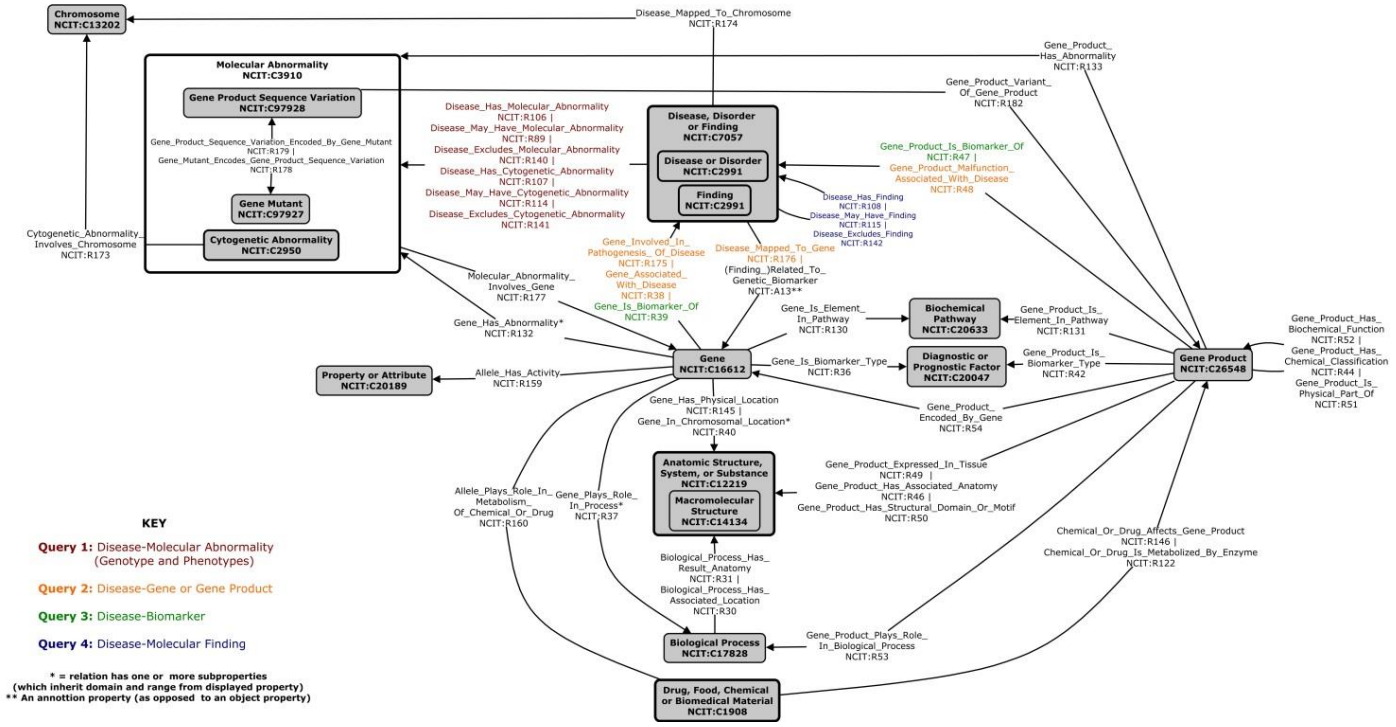

Figure 1. The high-level concepts and relationships used to structure information relevant to molecular subtyping of cancer in the NCIt. *A full resolution version of this image can be found here. This is a subgraph of the complete NCIt diagram here. Key relationships used to link cancers to molecular characteristics are highlighted in colors mapping to each of the four queries summarized in Tables 1-4 below.*

2.2 Molecular Subtyping 'Data' (Query Summaries)

To provide a more detailed view of the types and number of molecular disease associations captured in the NCIt, we executed four queries against a 'flattened' triplestore accessible here, to return four general types of assertions: (1) Disease-Molecular Abnormality; (2) Disease-Gene/Gene Product; (3) Disease-Biomarker; and (4) Disease-Molecular Finding. The complete results from each query can be found in the spreadsheets here. Tables 1-4 below summarize the query results by presenting examples and counts of discrete Association Subcategories that we identified for each of these four types of associations.

Note that results and counts reported here do not include assertions that are part of disjunctive statements that state one of two or more assertions must be true for a given class. For example, axioms on the 'Accelerated Phase Chronic Myelogenous Leukemia, BCR-ABL1 Positive' class (Figure 2) lists several possible molecular and cytogenic abnormalities in two sets of assertions that together comprise a disjunctive statement. As either but not both of these sets are true, none of the atomic assertions here hold universally for this disease, and therefore none are retained in the flattened version of the NCIt that is created to answer the queries reported here.

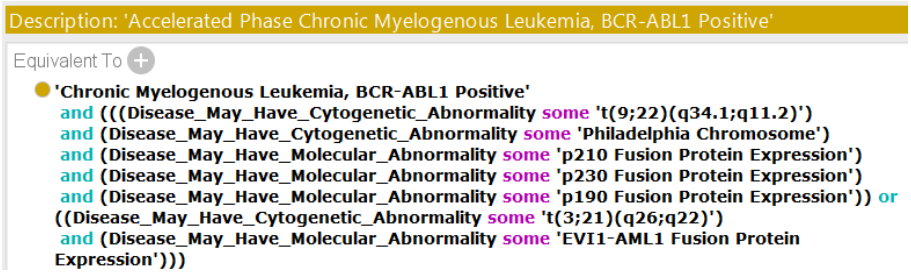

Figure 2: *example of a disjunction on the 'Accelerated Phase Chronic Myelogenous Leukemia, BCR-ABL1 Positive' class*

Query 1: Disease-Molecular Abnormality Associations (count = 2272)

'Molecular Abnormality' is a broad concept in the NCIt that covers content traditionally regarded as 'genotype' information describing genetic variation that causes or contributes to disease, and 'phenotype' information describing aberrations in molecular structure and function that result from such genetic variation. The results of this query across molecular abnormality relations are split into **Table 1A** and **Table 1B** to distinguish disease associations describing genotypes-level concepts (1A) from those describing molecular phenotypes (1B). This distinction will play an important role in recommendations made later in this document.

Table 1A: Disease-Genotype Molecular Abnormality association summaries, with counts for each association subcategory

Disease-Molecular Abnormality Associations (Genotype)			
Association Subcategory	**Relationship(s)**	**Examplar Axiom**	**Count**
gene mutations	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Mucosal Lentiginous Melanoma *Disease_May_Have_Molecular_Abnormality* BRAF Gene Mutation	167
gene deletions	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Anaplastic Astrocytoma *Disease_May_Have_Molecular_Abnormality* CDKN2A Gene Deletion	1
gene amplifications	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Burkitt Leukemia *Disease_Has_Molecular_Abnormality* MYC Gene Amplification	97
gene rearrangements	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Centroblastic Lymphoma *Disease_May_Have_Molecular_Abnormality* Clonal BCL2 Gene Rearrangement	167
gene family mutations	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Diffuse Astrocytoma, IDH-Mutant *Disease_Has_Molecular_Abnormality* IDH Gene Family Mutation	5
cytogenic translocations	*Disease_Has_Cytogenetic_Abnormality* | *Disease_May_Have_Cytogenetic_Abnormality* | *Disease_Excludes_Cytogenetic_Abnormality*	Acute Monocytic Leukemia *Disease_May_Have_Cytogenetic_Abnormality* t(8;16)(p11;p13)	199
cytogenic amplifications or deletions	*Disease_Has_Cytogenetic_Abnormality* | *Disease_May_Have_Cytogenetic_Abnormality* | *Disease_Excludes_Cytogenetic_Abnormality*	B-Cell Prolymphocytic Leukemia *Disease_May_Have_Cytogenetic_Abnormality* del(11q23)	184
chromosomal rearrangements	*Disease_Has_Cytogenetic_Abnormality* | *Disease_May_Have_Cytogenetic_Abnormality* | *Disease_Excludes_Cytogenetic_Abnormality*	Chondromyxoid Fibroma *Disease_May_Have_Cytogenetic_Abnormality* Rearrangement of 6q13	13
specific chromosomal gains/losses	*Disease_Has_Cytogenetic_Abnormality* | *Disease_May_Have_Cytogenetic_Abnormality* | *Disease_Excludes_Cytogenetic_Abnormality*	Classical Glioblastoma *Disease_Has_Cytogenetic_Abnormality* Gain of Chromosome 7	851
aneuploidy	*Disease_Has_Cytogenetic_Abnormality* | *Disease_May_Have_Cytogenetic_Abnormality* | *Disease_Excludes_Cytogenetic_Abnormality*	Class 1a Uveal Melanoma *Disease_Has_Cytogenetic_Abnormality Minimal* Aneuploidy	130

Table 1B: Disease-Phenotype Molecular Abnormality association summaries, with counts for each association subcategory

Disease-Molecular Abnormality Associations (Phenotype)			
Association Subcategory	**Relationship(s)**	**Examplar Axiom**	**Count**
mRNA overexpression	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Activated B-Cell-Like Diffuse Large B-Cell Lymphoma *Disease_Has_Molecular_Abnormality BCL2* Gene mRNA Overexpression	14
gene overexpression	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Mediastinal (Thymic) Large B-Cell Lymphoma Disease_May_Have_Molecular_Abnormality PDCD1LG2 Gene Overexpression	2
protein overexpression	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Ductal Breast Carcinoma *Disease_May_Have_Molecular_Abnormality* ERBB2 Protein Overexpression	46
fusion protein expression	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Epithelioid Hemangioendothelioma *Disease_Has_Molecular_Abnormality* WWTR1-CAMTA1 Fusion Protein Expression	92
loss of protein expression	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Schwannoma *Disease_Has_Molecular_Abnormality* Loss of Merlin Expression	8
loss of gene activity ('gene inactivation')	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Adenosquamous Lung Carcinoma *Disease_May_Have_Molecular_Abnormality* APC Gene Inactivation	225
abnormal molecular modifications	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Hepatocellular Carcinoma *Disease_May_Have_Molecular_Abnormality* Aberrant DNA Methylation	1
abnormal pathway activity	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Philadelphia Chromosome Positive, BCR-ABL1 Positive Chronic Myelogenous Leukemia *Disease_May_Have_Molecular_Abnormality* JAK-STAT Pathway Deregulation	11
genomic instability	*Disease_Has_Molecular_Abnormality* | *Disease_May_Have_Molecular_Abnormality* | *Disease_Excludes_Molecular_Abnormality*	Stage IV Colorectal Cancer AJCC v7 *Disease_May_Have_Molecular_Abnormality* High-Frequency Microsatellite Instability	18

Finally, note that the Molecular Abnormality hierarchy in the NCIt contains additional categories of terms that are never used in axiomatic assertions in the ontology itself (and are therefore absent from the tables above). These include **genotype** concepts based on mutation of a specific gene domain (e.g. 'FLT3 activation loop mutation'), or specific architectural features (e.g. 'EGFR Exon 18 Mutation'), or specific gene variants (e.g. 'EGFR NM_005228.3:c.1474A>C'); and **phenotype** concepts based on things like abnormal protein localization (e.g. 'EGFR Protein Translocation'), increased protein function (e.g. 'Activating PI3K Mutation'), abnormal cellular processes ('Abnormal DNA Repair'), and *resistance to therapeutic treatments* (e.g. 'EGFR-TKI Resistance Mutation').

Query 2: Disease-Gene/Gene Product Associations (count = 2436)

The NCIt uses four relationships to describe disease associations with genes or gene products whose presence or malfunction correlate with disease state. Of these, *Gene_Involved_In_Pathogenesis_Of_Disease* specifically indicates a causative association with disease (i.e. that mutations involving the gene are known to be pathogenic). By contrast, *Gene_Associated_With_Disease*, *Gene_Product_Malfunction_Associated_With_Disease*, and *Disease_Mapped_To_Gene* can be used to describe mere correlations between gene mutation or protein dysfunction, respectively.

Table 2: Disease-Gene/Product association summaries, with counts for each association subcategory

Disease-Gene/Product Associations			
Association Subcategory	**Relationship(s)**	**Examplar Axiom**	**Count**
genes (whose mutation has clinical significance)	*Gene_Associated_With_Disease*	PAX5 Gene Gene_Associated_With_Disease Small Lymphocytic Lymphoma	1383
genes (with variants known to be pathogenic)	*Gene_Involved_In_Pathogenesis_Of_Disease*	RB1 Gene *Gene_Involved_In_Pathogenesis_Of_Disease* Retinoblastoma	340
gene fusions (known to be pathogenic)	*Gene_Involved_In_Pathogenesis_Of_Disease*	ETV6/NTRK3 Fusion Gene Gene_Involved_In_Pathogenesis_Of_Disease Secretory Breast Carcinoma	226
proteins (whose malfunction has clinical significance)	*Gene_Product_Malfunction_Associated_With_Disease*	BRCD1 Protein *Gene_Product_Malfunction_Associated_With_Disease* *Breast Carcinoma*	487

Query 3: Disease-Biomarker (count = 51)

A distinct set of relations is used to describe associations with genes or gene products whose expression or alteration is used as a proxy or predictor of some clinically relevant state. There are a limited number of such assertions in NCIt at present (51 total). Notably, all asserted biomarker associations are with 'Genes' or 'Gene Products' - but unlike the previous table of disease-gene/product associations, these semantics here imply that the associated entity plays the role of a biomarker for some neoplastic condition.

Table 3: Disease-Biomarker association summaries, with counts for each association subcategory

Disease-Biomarker Associations			
Association Subcategory	**Relationship(s)**	**Examplar Axiom(s)**	**Count**
genes	*Gene_Is_Biomarker_Of*	MCM7 Gene *Gene_Is_Biomarker_Of* Cervical Carcinoma	7
wt gene alleles	*Gene_Is_Biomarker_Of*	KLK15 wt Allele *Gene_Is_Biomarker_Of* Breast Carcinoma	13
proteins	Gene_Product_Is_Biomarker_Of	Mesothelin *Gene_Product_Is_Biomarker_Of* Malignant Ovarian Neoplasm	31

Notably, NCIt contains a Biomarker type hierarchy that allows specification of more specific biomarker roles played by a particular biomarker (e.g. 'Diagnostic Factor', 'Prognostic Marker', 'Proliferation Marker', etc.). However, these are rarely and inconsistently applied, and genes/proteins asserted to be biomarkers for some disease do not classify in the Biomarker hierarchy (discussed in Section 4).

Query 4: Disease-'Molecular' Finding (count = 109*)

A subset of the 'Finding' classes that are linked to diseases in NCIt are molecular in nature - most of which fall under the 'Laboratory Test Result' hierarchy. These findings represent evidence of some molecular state as observed or detected in a patient by a clinician or laboratory test. Ontologically, findings are information level artifacts that are about, but distinct from, the underlying molecular state itself. Some disease-finding assertions are indicative of genotypic states (e.g. 'BRAF V600K Mutation Present', 'Absence of MYCN Gene Amplification') while others are indicative of molecular phenotypes associated with a patient or disease (e.g.'BCL2 Positive', 'Androgen Receptor Positive', 'Increased NFkappaB Pathway Activation'). Notably, there is overlap in scope between what may be represented as a finding and what can be represented more directly as a Disease-Molecular Abnormality or Disease-Gene/Product association (more on this later).

Table 4: Disease-Molecular Finding association summaries, with counts for each association subcategory

Disease-Molecular Finding Associations			
Association Subcategory	**Relationship(s)**	**Examplar Axiom(s)**	**Count**
cytogenic translocations	*Disease_Has_Finding* | *Disease_May_Have_Finding*	B Lymphoblastic Leukemia/Lymphoma, BCR-ABL1-Like *Disease_Has_Finding* Translocations Involving Tyrosine Kinases or Cytokine Receptors Present	2
cytogenic amplifications or deletions	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Embryonal Tumor with Multilayered Rosettes, C19MC-Altered *Disease_Has_Finding* C19MC Amplification	3
gene mutations	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Hairy Cell Leukemia *Disease_Has_Finding* BRAF V600E Mutation Present	10
protein expression (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Triple-Negative Breast Carcinoma *Disease_Has_Finding* Estrogen Receptor Negative	8
cellular protein expression (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Langerhans Cell Histiocytosis *Disease_Has_Finding* CD1a-Positive Neoplastic Cells Present	63
fusion protein expression (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Ph-Like Acute Lymphoblastic Leukemia *Disease_Has_Finding* BCR/ABL1 Fusion Negative	2
peptide/chemical/hormone (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Appendix Goblet Cell Carcinoid *Disease_Has_Finding* Serotonin Positive Neoplastic Cells Present	13
probe-reactive cells (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	Ovarian Insular Carcinoid Tumor *Disease_Has_Finding* Argentaffin Positive Neoplastic Cells Present	7
molecular modifications (absent or present)	*Disease_Has_Finding* | *Disease_May_Have_Finding*	MGMT-Unmethylated Glioblastoma *Disease_Has_Finding* Unmethylated MGMT Promoter	1

** 125 results returned querying 'Genetic Finding', 'Positive Laboratory Test Result', 'Receptor Status', 'Signaling Pathway Status', and 'Negative Test Result' hierarchies – 109 of which are 'molecular' in nature.*

2.3 Molecular Subtyping Logical Definitions

The relationships and patterns illustrated above are used to build logical descriptions of specific subtypes of cancer in the NCIt ontology. **Figure 3** below show an example of a logical definition for the 'Basal-Like Breast Carcinoma' breast cancer subtype. In Section 4 we will use this class as an example in our evaluation and recommendations about modeling design patterns.

```
'Breast Carcinoma by Gene Expression Profile'
 and (Disease_Has_Molecular_Abnormality some 'EGFR Protein Overexpression')
 and (Disease_May_Have_Finding some 'Estrogen Receptor Negative')
 and (Disease_May_Have_Finding some 'Progesterone Receptor Negative')
 and (Disease_May_Have_Finding some 'HER2/Neu Negative')
 and (Disease_May_Have_Finding some 'Aggressive Clinical Course')
 and (Disease_May_Have_Finding some 'Unfavorable Clinical Outcome')
 and (Disease_Mapped_To_Gene some 'KRT5 Gene')
 and (Disease_Mapped_To_Gene some 'KRT17 Gene')
 and (Disease_Mapped_To_Gene some 'EGFR Gene')
```

Figure 3: *Use of Molecular Assertions in the Logical Definitions of the '**Basal-Like Breast Carcinoma'** breast cancer subtype. Color coding as defined in Figure 1.*

3 Identifier Recommendations

The IRIs for NCIt terms are all defined in an NCIt namespace, with no re-use of terms from other community ontologies or vocabularies. However, the NCIt does provide mappings for many of its terms to identifiers from external systems. These are implemented as OWL annotation property axioms, where a unique annotation property is defined for each external identifier system. **Table 5** shows the label and IRI for all such properties, along with a usage count indicating the number of mappings of each type.

Table 5: Existing Mappings implemented using owl annotation property axioms in NCIt.

Mapping Property Label	Mapping Property IRI	Usage Count
General Mappings		
code	NCIT:NHC0	127970
UMLS_CUI	NCIT:P207	110307
NCI_META_CUI	NCIT:P208	7828
PubMedID_Primary_Reference	NCIT:P171	601
Gene/Feature Mappings		
EntrezGene_ID	NCIT:P321	4621
GenBank_Accession_Number	NCIT:P102	4807
HGNC_ID	NCIT:P102	4605
miRBase_ID	NCIT:P362	176
OMIM_Number	NCIT:P100	13457
Swiss_Prot	NCIT:P93	4505
SNP_ID	NCIT:P315	137
Drug/Food/Chemical Substance Mappings		
FDA_UNII_Code	NCIT:P319	13022
CAS_Registry	NCIT:P210	12352
ChEBI_ID	NCIT:P386	3531
NSC_Code	NCIT:P175	2367
USDA_ID	NCIT:P354	134
Pathway Mappings		
BioCarta_ID	NCIT:P216	335
KEGG_ID	NCIT:P215	237
PID_ID	NCIT:P367	169
Disease Mappings		
ICD-O-3_Code	NCIT:P334	1223
Organism/Taxon Mappings		
MGI_Accession_ID	NCIT:P332	154
NCBI_Taxon_ID	NCIT:P331	2203

Evolving the NCIT into a more interoperable resource will require a deeper integration with identifiers from community systems and standards. This can be achieved through more comprehensive mappings to select external systems, and also by direct re-use of external IRIs in the NCIt. Such efforts will facilitate data integration and federated query and analysis operations, enabling the NCIt to leverage knowledge in external systems, and external systems to leverage the knowledge in the NCIt.

Recommendations:

1. **Gene identifier mapping:** Most existing mappings seem fairly consistent and robust. One exception is gene identifier mappings, where HGNC gene identifiers are consistently mapped to named gene classes (e.g. 'BRCA2 Gene'), EntrezGene identifiers are consistently mapped to the 'wt allele' child of gene classes (e.g. 'BRCA2 wt allele'), and OMIM gene identifiers are typically mapped to both the parent gene and wt allele child. The rationale for this approach is not clear from inspection; we would recommend using a single authoritative gene resource, such as EntrezGene.
2. **Variant mappings to dbSNP vs ClinVar:** The fact that no single, comprehensive identifier system for genetic alleles has been adopted poses a significant barrier to efforts to integrate variant data. While many efforts use the HGVS syntax in lieu of an identifier, the two most common sources for variant

identifiers are dbSNP and ClinVar. Here it is important to recognize that there is not a 1:1 mapping of identifiers between these systems, as dbSNP ids identify a ***location*** where a particular kind of variation exists, and ClinVar ids more precisely identify a ***specific variation*** that occurs at this location. For example, the dbSNP identifier rs121913238 is linked to both C>G and C>A alleles that affect the locus at NM_004985.4:c.181. Each of these SNVs has its own ClinVar identifier that uniquely identifies the specific variant at this location (ClinVar:177777 for C>G and ClinVar:376324 for C>A). For this reason, we recommend mapping to ClinVar in addition to or instead of dbSNP. Longer term, keep an eye on efforts such as the ClinGen Allele Registry [4] which we hope will emerge as the long awaited single centralized authority for variant identifiers.

3. **Direct re-use of OBO ontology terms in NCIt axioms:** Ultimately, direct re-use of terms from established vocabularies promotes more efficient, maintainable, and effective integration between resources. This directly connects ontologies and annotated data across the linked data landscape, dramatically lowering barriers to data integration and cross-resource queries and analyses that can leverage ontological semantics. Participation in community re-use efforts, and alignment with the principles espoused by groups such as the OBO Foundry, will also raise the profile of the NCIt and engender re-use of its terms by other ontology and data developers.

   In pursuit of these goals, we would advocate expanding on our initial work to integrate Uberon anatomy and CL cell types into the NCIt (see D10 deliverable report), to explore additional integrations and re-use experiments. Here, we should target ontologies and identifier systems that are broadly and deeply integrated into the data landscape. Our recommendations for high-value mapping and integration targets is presented in Table 6, which suggests identifier systems for several high-level molecular types in the NCIt.

Table 6: Proposed high-value targets for mapping or reuse in NCIt.

NCIT Molecular Type	Mapping/Integration Target
Molecular Abnormalities (Genotypes)	Genotype Ontology (GENO), Sequence Ontology (SO)
Molecular Abnormalities (Phenotypes)	Human Phenotype Ontology (HPO)
Genes	NCBI Gene database, Ensembl
Proteins	Uniprot database, Protein Ontology (PRO)
Pathways	Reactome database
Drugs/Chemicals	ChEBI Ontology, UNII chemical substance registry, NCATS Global Ingredient Archival System (GINAS)
Biological Processes	Gene Ontology (GO)

These integrations will vary in difficulty and in the benefits they afford. For example, swapping in NCBI gene identifiers for NCIt gene classes will be relatively straightforward, given that numerous gene ID mappings are already encoded in annotation axioms. But most integrations will require de novo equivalency and sub/superclass relationships to be established before refactoring can occur. For example, replacement of gene products with Uniprot or Protein ontology terms, or Pathway terms with Pathway Ontology or Reactome identifiers may leverage lexical mapping tools such as Ontobio [5] to make initial mappings that can be validated by manual curation. Integrations of Molecular Abnormality types with ontologies such as the Genotype Ontology (GENO) and Human Phenotype Ontology (HPO) would require the additional steps of refactoring classification and logic implemented in the NCIt to align with these target ontologies. Finally,

integration of more complex data types such as diseases with the MONDO disease ontology may require the use of probabilistic algorithms such as the Monarch kBOOM [6] tool, to seed initial equivalence or subclass recommendations for manual validation. Note here that determination of the best integration strategy (e.g. mapping vs direct re-use), and which external resources may be the best fit for the NCIt, will require deeper requirements analysis for each data type.

Despite barriers to doing so, we highly recommend continued work towards integration of the above nature. Specifically, as cancer moves towards precision medicine, it will be critically important to be able to leverage multiple datasets in the interpretation of any given patient's cancer phenotypes. This would mean improved pathogenicity determinations by being able to determine degree of corroboration of evidence relating to the pathogenicity, improved inference between treatment selections and their potential therapeutic value and the patient's molecular lesions, and identification of new drug targets. The utility of an NCIt that includes these molecular interoperability improvements will impact the NCIt's ability to integrate data, query across disparate datasets, provision improved data validation tools for data submission to the GDC and other sources, analytical capabilities, and development of new and improved algorithms. Further, the ongoing benefits of inclusion and harmonization of knowledge being represented in other ontological and biological contexts could be extremely impactful in that it would necessarily increase the reach of the NCIt across a much wider arena of both clinical and biomedical informatics, providing increased community contributions and usage.

4 Modeling Recommendations

We have evaluated aspects of the NCIt concerning molecular features of cancer to identify areas where modeling patterns could be improved in their consistency, richness, and alignment with standards and practices in the broader biomedical ontology community. Below we present recommendations in three areas that we believe to represent pragmatic and potentially high-impact improvements to the NCIt:

(1) Refactoring Molecular Abnormality Classification
(2) Refactoring Biomarker Modeling
(3) Harmonizing Molecular Subtype Representation

4.1 Refactoring Molecular Abnormality Classification

Molecular Abnormalities in the NCIt describe abnormalities in macromolecular sequences, or the abundance, structure, or activity of molecular entities in a cell. Traditionally, biologists bucket such anomalies into two high level categories: (1) **Genotypes** describing aberrations in genomic sequence molecules, or (2) **Phenotypes** describing downstream manifestations of this genotypic variation (here at the molecular level). Below we outline a recommendation to refactor the Molecular Abnormality hierarchy to explicitly distinguish genotype from phenotype level abnormalities, and define logical design patterns that can be applied to enable consistent, automated, and maintainable classification of these terms. We believe this is a tenable and valuable exercise for the following reasons:

1. All Molecular Abnormality terms describe either a genetic variation or molecular phenotype, and thus would naturally partition into one of these two categories.
2. From a biological perspective this distinction is fundamental, as "Genotype + Environment = Phenotype" is a universal paradigm guiding research and educational pursuits.
3. From a technical perspective, this approach is aligned with models implemented in existing biomedical ontologies and knowledgebases.

Our recommendations are made with the goals of making the NCIt more internally consistent and user-friendly, but more importantly to facilitate integration with extant ontological frameworks (e.g. GENO, SO, HPO) and data resources (e.g. the Monarch Initiative and the NCATS Biomedical Data Translator) - so that knowledge can be seamlessly exchanged and re-used to drive novel insights into cancer. Below we outline the tasks that would be performed and tools that could be used to achieve these goals.

Task 1: Evaluation and Partitioning of Genotype and Phenotype Concepts

We must initially evaluate which of the more than 1600 Molecular Abnormality terms represent genotype-level concepts (variations in genomic sequence) versus phenotype-level concepts (the 'post-genomic' abnormalities that can result from genomic (or environmental) factors.

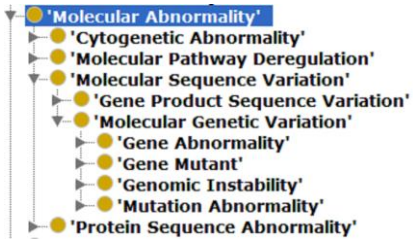

Figure 4: *Top-level organizational classes in the Molecular Abnormality hierarchy.*

The Molecular Abnormality hierarchy contains four direct subclasses (Figure 4). We found all terms in the 'Cytogenic Abnormality' hierarchy to represent genotype concepts, and all terms in the 'Molecular Pathway Deregulation' and 'Protein Sequence Abnormality' hierarchies to represent phenotype level concepts. The most variable and abundant set of terms is found under 'Molecular Sequence Variation', specifically in the 'Molecular Genetic Variation' subhierarchy. The labels and definitions of the four organizing subclasses under 'Molecular Genetic Variation' gave no clear indication of how they were different and how classes would be partitioned here. Across these hierarchies, we found a mix of genotype level terms (e.g.' BRAF Gene Mutation', 'EGFR NM_005228.3:c.1474A>C') and phenotype level terms (e.g. 'ERBB2 Protein Overexpression', ' Aberrant DNA Methylation'). Many terms were ambiguous - e.g. 'EGFR Activating Mutation' seems to describe a genetic mutation that is defined by a specific phenotypic outcome - so it is not clear if axioms using this term are intended to link a disease to a genetic lesion or an increase in EGFR activity. Finally several terms seemed inconsistent with their classification, e.g. 'Abnormal DNA Repair' asserted as a child of 'Gene Abnormality'.

Overall, we found the Molecular Abnormality hierarchy to provide no clear distinction between genotype and phenotype concepts, and inconsistent classification of terms under ambiguous organizational concepts.

Task 2: Identification of Classification Axes

Our next step was to group genotype and phenotype terms into subcategories based on common underlying classification criteria. For example, a category grouping all molecular phenotype classes based on increased expression of a particular protein. Here we focused on classes that were used in axioms, and identified an initial set of genotype and phenotype class categories that covered the majority of classes used in axioms in the NCIt. These categories are reflected in the categories of Molecular Abnormality Associations in **Table 1B**, and the molecular phenotype categories are further detailed in **Table 7** below.

Task 3: Definition of Logical Design Patterns

For each category identified above, we can then define a logical design pattern that will support automated and consistent classification of genotype and phenotype hierarchies. To promote interoperability and re-use, these patterns should be aligned with established design patterns used in genotype and phenotype ontologies and data models. Here, the OBO Phenotype consortium [7] defines a standard that uses clearly defined “EQ” patterns [8] to describe phenotypes as qualities (Q) that inhere in an physical or processual entity (E). Examples can be found in the “EQ Design Pattern” column of **Table 7**, which present proposed design patterns for each molecular phenotype category.

Table 7: Proposed design patterns for logical definitions of classes in each molecular phenotype category.

Molecular Phenotype Category	Example Phenotype Class	Phenotype EQ Design Pattern
gene overexpression	PDCD1LG2 Gene Overexpression	has_part' some ('increased amount' and ('inheres_in' some ('gene_product_of' some 'PDCD1LG2 Gene')) and ('qualifier' some 'abnormal'))
mRNA overexpression	BCL2 Gene mRNA Overexpression	has_part' some ('increased amount' and ('inheres_in' some ('mRNA' and 'transcribed_from' some 'BCL2 Gene')) and ('qualifier' some 'abnormal'))
protein overexpression	ERBB2 Protein Overexpression	has_part' some ('increased amount' and ('inheres_in' some 'ERBB2 Protein') and ('qualifier' some 'abnormal'))
fusion protein expression	WWTR1-CAMTA1 Fusion Protein Expression	has_part' some ('present' and ('inheres_in' some 'WWTR1-CAMTA1 Fusion Protein') and ('qualifier' some 'abnormal'))
loss of protein expression	Loss of Merlin Expression	has_part' some ('absent' and ('inheres_in' some ('gene_product_of' some 'Merlin Gene')) and ('qualifier' some 'abnormal'))
loss of gene activity ('gene inactivation')	APC Gene Inactivation	has_part' some ('inactive' and ('inheres_in' some ('gene_product_of' some 'APC Gene')) and ('qualifier' some 'abnormal'))
aberrant molecular modifications	Aberrant DNA Methylation	'has_part' some ('amount' and ('inheres_in' some ('methyl group' and 'part_of' some 'DNA')) and ('has_modifier' some 'abnormal'))
aberrant pathway activity	JAK-STAT Pathway Deregulation	has_part' some ('quality' and ('inheres_in' some 'regulation of JAK-STAT cascade') and ('has_modifier' some 'abnormal'))
genomic instability	High-Frequency Microsatellite Instability	

Creating logical definitions for each Molecular Abnormality term in the NCIt will allow a reasoner to automatically classify these terms into distinct genotype and phenotype hierarchies. Furthermore, the task of creating logical definitions forces us to be explicit and precise about the ontological type of a given concept. For example, we would have to decide if 'gene overexpression' is or is not different from 'mRNA overexpression' in order to craft a logical definition for classes in these categories. This exercise improves the rigor and consistency of the ontology, and facilitates user understanding and re-use and integration with external sources.

Notably, molecular level phenotypes are poorly represented across the OBO phenotype corpus, and we are leading efforts to address this by defining design patterns and supporting vocabularies in this space. The molecular phenotype terms in the NCIt fit nicely into the initial framework that we have established, and can provide useful requirements for evolving and validating this work.

Task 4: Implementing Logical Design Patterns

Creating logical definitions for each 'Molecular Abnormality' term in the NCIt will allow a reasoner to automatically classify these terms into distinct genotype and phenotype hierarchies. But manual addition of these axioms would be time consuming and error prone. Here, we can leverage the Dead Simple OWL Design Patterns (DOS-DPs)[9] tools to semi-automate this process. DOS-DPs provide a light-weight standard for specifying these design patterns that can then be used for generating documentation, generating new terms and retrofitting old ones. Design patterns such as those in **Table 7** are first captured in a structured YAML format that defines axiom template and variables to populate them. Variable values that will

populate this template (in this case the NCIt terms are specified in a separate file - typically a tabular format. Then DOS-DP scripts operate on these two inputs to generate the OWL axioms that logically define NCIt molecular phenotype classes. **Figure 5** below provides an example DOS-DP YAML specification of the 'mRNA overexpression' pattern defined in **Table 7**.

Figure 5: DOS-DP YAML template for the 'mRNA overexpression' design pattern. Additional templates for other design patterns can be found here.

```
pattern_name: mrna_overexpression

classes:
  increased amount: PATO:0000470
  abnormal: PATO:0000460
  mRNA: SO:0000234

relations:
  transcribed_from: RO:0002510
  has_modifier: RO:0002573
  has_part: BFO:0000051
  inheres_in: RO:0000052

vars:
  gene: "'gene'"

name:
  text: "%s mRNA overexpression"
  vars:
   - gene

def:
  text: "Overexpression of %s mRNA."
  vars:
    - gene

equivalentTo:
  text: "'has_part' some ('increased amount' and
  ('inheres_in' some ('mRNA' and 'transcribed_from' some %s)) and
  ('has_modifier' some 'abnormal'))"
  vars:
    - gene
```

Conclusions

Implementing the tasks above will result in a Molecular Abnormality hierarchy that is split at the top-level into genotype and molecular phenotype sub-hierarchies. An analogous approach to that described for molecular phenotypes could be taken to logically define and re-classify the genotype terms. Importantly, all Molecular Abnormality classes currently defined in the NCIt will remain, but their hierarchical organization will be different, and each will have a logical definition that will allow it to automatically classify at one or more locations in the ontology. Furthermore, the labels for all classes should be reviewed and a consistent naming scheme applied that reflects the updated perspective and organization in this part of the NCIt. This can be achieved with some straightforward reporting and standardization scripts. Alignment with OBO Phenotype design patterns would allow for all NCIt molecular phenotype classes to be implemented in the HPO namespace (or some molecular phenotype spin-off), and re-used directly in the NCIt. Ultimately, we

believe that the work proposed here will provide a characterization of molecular abnormalities that is better integrated into the broader landscape of biomedical ontologies and data sources. This will enable the NCIt to leverage these resources toward improved molecular characterization and subtyping of cancer.

4.2 Refactoring Biomarker Modeling

Current Implementation

Disease Biomarkers are defined in the NCIt as "*a specific molecular signature of disease, physiological measurement, genotype structural or functional characteristic, metabolic changes, or other determinant that may simplify the diagnostic process, make diagnoses more accurate, distinguish different causes of disease, or enable physicians to make diagnoses before symptoms appear and to track disease progression.*" The NCIt presently contains 51 Biomarker assertions that link a gene or gene product to a disease using the *Gene_Is_Biomarker_Of* or *Gene_Product_Is_Biomarker_Of* properties, respectively. **Gene** biomarker assertions are made when *"expression or alteration of a gene is correlated with a particular disease or disease state or is predictive of the disease or disease state".* **Gene product** biomarker assertions are made when *"the presence of a gene product may contribute to clinical diagnosis, treatment selection, or prediction of clinical outcome".* In addition, two separate properties, *Gene_Is_Biomarker_Type* and *Gene_Product_Is_Biomarker_Type*, are used to link genes or gene product to one of a classes representing 'types' of biomarkers (e.g. 'Tumor Marker', 'Prognostic Marker', Proliferation Marker').

Evaluation

1. **Data Richness:** There are a very limited number of biomarker assertions (51), and biomarkers limited to genes or gene products. Expanding properties and design patterns to accommodate a broader representation of biomarker relationships would support a much richer set of knowledge in this space to be captured in the NCIt. Furthermore, alignment and re-use efforts described elsewhere will facilitate integrations with external sources of biomarker associations that could be leveraged to pull more comprehensive data into the NCIt framework.
2. **Modeling Consistency and Precision**: In some gene-biomarker associations a specific wt allele of a gene is asserted as the biomarker (e.g. 'SLPI wt Allele' *Gene_Is_Biomarker_Of 'Malignant* Ovarian Neoplasm'), and in others a parent gene is asserted as the biomarker ('ELF3 Gene' *Gene_Is_Biomarker_Of* ' Malignant Ovarian Neoplasm'). The significance of these two patterns and how to choose between then when making assertions should be made more apparent. Likewise, the significance of asserting gene vs gene product level biomarker associations is unclear. From the definitions above, one would assume that a **gene biomarker assertion** is made based on gene expression data or sequencing data that reveal expression of the gene or an alteration in the gene correlates with some disease feature state, and a **gene product biomarker assertion** is made based on detection of a protein that correlates with a disease feature or state. It is currently not clear if and how these distinctions are captured in the model. The NCIt should support more nuanced and explicit representation of what the biomarker is, and what evidence exists to support these assertions. It should also aspire to support other types of biomarkers besides genes and gene products - e.g. metabolic molecules, or expression signatures/profiles that may include multiple genes, etc.
3. **Biomarker Classification**: We would expect specific genes or gene products asserted as biomarkers for some disease to classify as subtypes of one of the NCIt Biomarker type classes. But such classifications are not made - of the 51 genes or gene products asserted to be biomarkers for some disease, none are placed in the Biomarker hierarchy. For example, 'Melanoma-Associated Antigen 3' is asserted as a

biomarker_of 'Melanoma', and thus we might expect it to classify as a subtype of the NCIt 'Melanoma Biomarker' class. But this relationship is not asserted, nor does the logic exist in the NCIt to infer it.

Notably, there are many asserted Biomarker subclasses that do represent specific molecular entities. But in all such cases, there is no corresponding assertion indicating what it is a *Biomarker_Of.* For example, 'Phosphorylated Epidermal Growth Factor Receptor' is asserted as a subclass of 'Prognostic Marker'. But it is not classified as a Gene or Protein alongside other genes and proteins, nor is there any assertion about what disease this may be prognostic for. So a disconnect exists whereby we know this entity is a prognostic biomarker, but not the types of diseases where this is relevant. As a general rule, such molecular entities should be asserted according to their biochemical type, and inferred as subtypes of 'defined classes' in the 'Biomarker' hierarchy. The modeling recommendations below present a possible approach to take here.

Modeling Recommendations

We recommend the following as a possible approach to enable automated classification of entities that serve as biomarkers into an informative and consistent Disease Biomarker Type hierarchy.

First, logically define the 'Biomarker' class with the following axiom:

```
Biomarker =
(Gene_Is_Biomarker_Of some 'Disease or Disorder') or
(Gene_Product_Is_Biomarker_Of some 'Disease or Disorder')
```

This axiom will enable a reasoner to classify any entity (here a gene or gene product) asserted to be a biomarker for a disease in the Biomarker hierarchy. Notably, this definition is based on the limited perspective on Biomarkers currently implemented in the NCIt. A broader definition would be needed if NCIt decides to create patterns to assert entities besides genes and proteins to be biomarker for entities besides diseases.

Second, additional axioms may then be required to achieve more fine-grained classification under specific subtypes in the Biomarker hierarchy (e.g. classification of 'Melanoma-Associated Antigen 3' as a subclass of 'Melanoma Biomarker'). Here, we see that the majority of Biomarker subclasses seem to be defined based on three axes:

(Axis 1) The entity it is a biomarker for (e.g. 'Melanoma Biomarker', 'Proliferation Biomarker')
(Axis 2) The natural kind of thing the biomarker is (e.g. Genetic Biomarker', 'Proteomic Profile')
(Axis 3) The clinical task it informs (e.g. 'Prognostic Marker', 'Diagnostic Factor').

Axis 1: Classification along the first axis would leverage existing *Gene_Is_Biomarker_Of* assertions to achieve finer-grained classification. For example:

```
'Melanoma-Associated Antigen 3' Gene_Product_Is_Biomarker_Of some Melanoma
```

The only additional axioms required would be logical definition on the more specific Biomarker subtypes that are defined using this axis. For our Melanoma example:

```
'Melanoma Biomarker =
(Gene_Is_Biomarker_Of some 'Melanoma') or
(Gene_Product_Is_Biomarker_Of some 'Melanoma')
```

This would allow automated classification of 'Melanoma-Associated Antigen 3' as a subtype of ‘Melanoma Biomarker’.

Axis 2: Classification along the second axis could leverage the parent type of the entity asserted as a biomarker to achieve finer grained classification. For example, a 'Genetic Biomarker' is defined as "*any alteration in DNA that may indicate an increased risk of developing a specific disease or disorder".* A logical definition for this class could therefore be:

```
'Genetic Biomarker =
('Molecular Genetic Variation' or 'Cytogenetic Abnormality') and
(Gene_Is_Biomarker_Of some 'Disease or Disorder')
```

As noted in section 3 above, the 'Molecular Genetic Variation' hierarchy is currently littered with things that are not 'alterations in DNA' (e.g. 'DNA Damage Response Deficiency') - so this is an area where the implemented the recommended explicit distinction between genotype and phenotype level abnormalities would be useful.

Axis 3: Finally, classification along the third axis could leverage additional axioms made on the entity asserted to be a biomarker, possibly using the existing *Gene_Is_Biomarker_Type* relation. For example, consider a biomarker type such as 'Melanoma Prognostic Biomarker' (currently not in the NCIt):

```
'Melanoma Prognostic Biomarker' =
((Gene_Is_Biomarker_Of some 'Melanoma') or
(Gene_Product_Is_Biomarker_Of some 'Melanoma')) and
((Gene_Is_Biomarker_Type some 'Prognostic Marker') or
(Gene_Product_Is_Biomarker_Type some 'Prognostic Marker'))
```

NCIt curators/editors would have to consider for every entity they make a biomarker assertion on, if there is a specific clinical task where the biomarker is used (e.g. assessing a diagnosis, prognosis, risk, treatment response, etc), and axiomatize accordingly.

Conclusions

Addressing the issues outlined above with the proposed recommendations (or other approaches) would ultimately afford numerous benefits for creating, interrogating, and re-using biomarker assertions in the NCIt to define molecular subtypes of cancer. The result would be a structure where all molecular entities are asserted according to their natural, biochemical kind (e.g. proteins in the 'Protein' hierarchy, genes in the 'Gene' hierarchy, molecular entities such as 'Histone H3 Lysine 27' in the 'Macromolecular Structure' hierarchy), and given axioms that enable their inferred classification in the 'Biomarker' hierarchy based on the logical definitions of the classes there. Ideally, a richer model with new relations and would be defined to allow assertion of additional types of molecular entities as biomarkers (i.e. not just genes and proteins). Notably, implementing these recommendations could also leverage the DOS-DP framework to semi-automate the creation of axioms on existing classes, and the creation of new biomarker classes.

4.3 Harmonizing Molecular Subtype Representations

A primary goal of this work is to better enable semantic technologies to leverage the knowledge encoded in the NCIt to inform molecular subtyping of disease. This is highly dependent on a rich, consistent, and interconnected representation of molecular phenotype and finding information in the NCIt. Consider for example, the well-established 'Basal' Breast Cancer subtype, which we will use the as a test case to illustrate issues and recommendations for improving molecular subtype descriptions.

The minimal molecular features defining this subtype are described and compared to other subtypes in the following chart adapted from [10]. As described in [10], these tumors are ER, PR, and HER2 negative, and typically express basal markers such as keratins 5, 6, 14, 17, EGFR, and proliferation related genes [10, 11,12]. They are also marked by basal cytokeratin expression often exhibit low BRCA1 expression [13] and TP53 mutation [14,15].

Intrinsic subtype	IHC status	Grade	Outcome	Prevalence[Δ]
Luminal A[*]	[ER+|PR+] HER2-KI67-	1|2	Good	23.7% [p1] [10]
Luminal B[*]	[ER+|PR+] HER2-KI67+	2|3	Intermediate	38.8% [p1] [10]
	[ER+|PR+] HER2+KI67+		|Poor	14% [p1] [10]
HER2 over-expression[*]	[ER-PR-] HER2+	2|3	Poor	11.2% [p1] [10]
Basal[*]	[ER-PR-] HER2-, basal marker+	3	Poor	12.3% [p1] [10]
Normal-like	[ER+|PR+] HER2-KI67-	1|2|3	Intermediate	7.8% [p2] [15]

Using the 'Basal-Like Breast Carcinoma' class as a test case, we will evaluate the **completeness** of its representation, the **consistency** with which molecular features are represented, and the **connectedness** of this knowledge to related biomedical concepts (e.g. related genes, pathways, and functions they perform). The axioms used to describe this class in the NCIt shown in **Figure 6.**

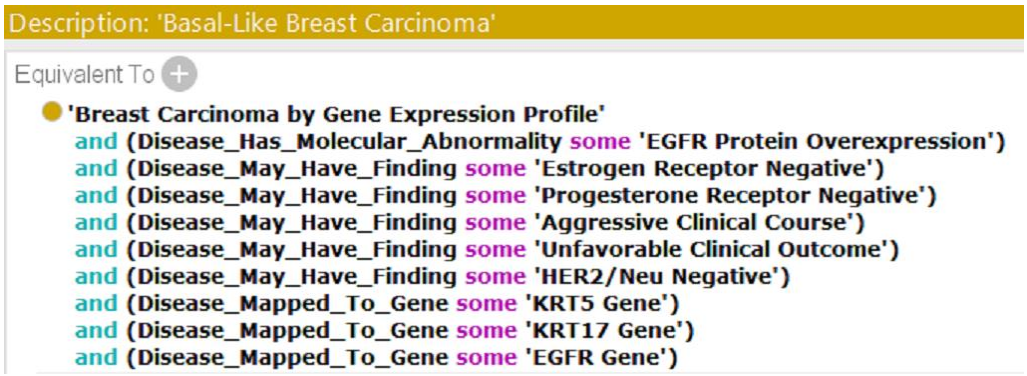

Figure 6: *Protégé screen shot of axioms describing the 'Basal-Like Breast Carcinoma' subtype.*

1. Completeness:

The NCIt representation of 'Basal-Like Breast Carcinoma' is relatively complete – describing all core IHC status markers including three of the five basal markers listed above (keratins 5, 17, EGFR, but not keratins 6 and 14). Also missing are descriptions of BRCA1 and TP53 related markers. These omissions may of course be for good reason – curators may not believe the evidence for these assertions is strong enough to include, or may have not curated the literature reporting them yet. However, it would be beneficial to include purposeful omissions in a comment field.

2. Consistency

Within this single class, the molecular features that characterize this disease are described using three relations that capture different levels of nuance and precision. Some features are described simply as ***mapped/associated genes*** (KRT5 and KRT17), with no representation of what it is about these genes that is interpreted as definitive for the subtype. For other features, the relation used captures how the gene/protein serves as a marker for the disease – e.g. it is the ***abnormal overexpression*** of the EGFR protein that is indicative of this subtype. And yet other features are described as ***findings*** that represents the evidence or observation made by a clinician as opposed to the molecular entity itself (e.g. ER, PR and HER2 receptors). These choices may of course reflect a variable level of knowledge or evidence about how each gene or product characterizes this subtype of breast cancer – in which case this may be the best possible representation. But it introduces a lack of precision and internal inconsistency that will pose barriers to use and discovery of knowledge in the NCIt.

We also noted inconsistency in representation of the same molecular feature in different ways ***across*** breast cancer subtype classes. Compare for example the representation of HER2 receptor status for 'Basal-Like Breast Carcinoma' above as a finding, with that for the 'HER2 Positive Breast Carcinoma' class below as a molecular abnormality (note erbB2 is another name for HER2) (**Figure 7**).

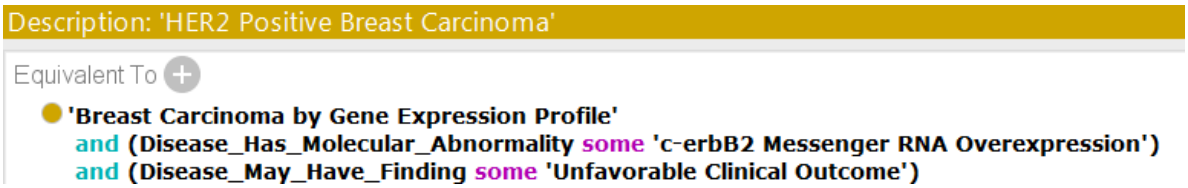

Figure 7: *Protégé screen shot of axioms describing the 'HER2 Positive Breast Carcinoma' subtype.*

Note also that HER2 Positive Breast Carcinoma has much sparser logical description – lacking even core estrogen and progesterone receptor status. This is true of most other breast carcinoma subtypes in the NCIt, and re-iterates need for enhancing content of NCIt - ideally through semi-automated curation approaches that leverage computational and manual efforts.

3. Connectedness

Issues associated with the variable ways of describing molecular traits and abnormalities could be tempered to some degree if there were logical connections between classes that describe related biological concepts. For the most part, the NCIt does a good job of this – linking molecular abnormalities to affected genes and proteins, and genes and proteins to each other, chromosomes, molecular functions, and signaling pathways. NCIt also uses the annotation property '*Related_to_Genetic_Biomarker*' to link Findings to Genes that they describe some aspect of. So in our evaluation above, the 'HER2/Neu Negative Finding' class does have a link to the ERBB2 gene - but because it uses an annotation property, it is invisible to a reasoner.

Recommendations

Below we offer recommendations regarding both evaluation and ***curation*** practices, and ***modeling*** approaches.

Curation

1. NCIt curators may want to evaluate the semantics of these relations so as to be clear about how they are different and when they should be used. They may also wish to assess the curation processes/context that led to their selection to represent molecular features of this cancer subtype in different ways.

2. NCIt curators should be as specific as possible when describing how a molecular feature relates to a disease. Here, can we be more specific about the say in which KR5 and KR17 characterize this disease? It is likely their expression, overexpression, or mutation, and more precise relations exist in the NCIt to capture this nuance.

3. Curators may want to consider re-visiting associations to genes that are represented only using *Disease_Mapped_To_Gene* to see if evidence can be found to make a more precise assertion. For 'Basal-Like Breast Carcinoma' this would be KRT5 and KRT17 – but note that EGFR is linked both with this generic relation and a more specific axiom describing its overexpression.

Modeling

1. The ability to describe molecular features as findings or molecular phenotypes lead to inconsistent representation and missed connections in the data. We would recommend whenever possible to

describe such features as close to biology as possible – using molecular trait and phenotype terms instead of findings. So for the 'HER2 Positive Breast Carcinoma' example, use *Disease_Has_Molecular_Abnormality* or *Disease_Mapped_To_Gene* over *Disese_Has_Finding* to describe the HER2/Neu receptor status associated with this disease. Note here that a new property/pattern might be needed to describe expression that is normal for a cell (i.e. it is a ***trait*** rather than an ***abnormal phenotype***). *Disease_Has_Molecular_Abnormality* describes abnormally low or high expression phenotypes. Finally, if a curator wants to capture information about the evidence or provenance for an assertion, a Finding object can be used here, and linked to the molecular abnormality it describes. This would allow NCIt to capture things like methods used to make the finding, who did it, and when.

Note that our recommendations to re-cast molecular findings as trait or phenotype concepts parallels our earlier recommendation from the first contract to refactor findings generally as phenotypes.

2. The *Related_to_Genetic_Biomarker* annotation property used to link Findings to Genes that they describe should be made an object property if possible so axioms using it are not invisible to the reasoner. Also, there are cases where this property could more link a Finding to terms that more precisely define it. For example, the 'BRAF V600E Mutation Present' finding is currently linked to the 'BRAF Gene' class, but would be better described with a link to the 'BRAF NP_004324.2:p.V600E' class representing this specific mutation.

Conclusions

Standardizing descriptions of molecular traits and phenotypes associated with cancer subtypes receptor status would improve consistency, connectedness, and query-ability of the data. An approach that directly describes the biology as opposed to findings about the biology is preferred, as it is aligned with modeling in most biomedical ontologies and knowledgebases. As before, DOS-DP tools can help facilitate the process of generating molecular trait and phenotype classes for any findings that don't yet have them.

5 Supplementary Materials

1. NCIt_Molecular_Query_Results.xlsx – spreadsheets containing full results of queries sued to explore and evaluate molecular associations asserted in the NCIt, and generate tables in Section 2 above.
2. NCIT_OBO_Full_cmap.jpg – diagram of high level structure of the NCIt, generated from domain and range axioms on Object properties
3. NCIT_OBO_Molecular_cmap.jpg – diagram of high level structure of the 'molecular subset' of the full NCIt cmap above.

6 References

1. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of information in medicine. 2005 Jan 1;44(4):498
2. Schulz S, Schober D, Tudose I, Stenzhorn H. The pitfalls of thesaurus ontologization–the case of the NCI Thesaurus. In AMIA Annual Symposium Proceedings 2010 (Vol. 2010, p. 727). American Medical Informatics Association.
3. de Coronado S, Tuttle MS, Solbrig HR. Using the UMLS Semantic Network to validate NCI Thesaurus structure and analyze its alignment with the OBO relations ontology. In AMIA Annual Symposium Proceedings 2007 (Vol. 2007, p. 165). American Medical Informatics Association.
4. ClinGen Allelel Registry web page, http://reg.clinicalgenome.org/redmine/projects/registry/genboree_registry/landing
5. Ontobio Github repo web page, https://github.com/biolink/ontobio
6. kBOOM Github repo web page, https://github.com/monarch-initiative/kboom
7. OBO Phenotype Organization Github web page, https://github.com/obophenotype
8. Mungall C, Gkoutos G, Washington N, Lewis S. Representing Phenotypes in OWL. In OWLED 2007 Jun 6.
9. DOS-DP Github web page, https://github.com/dosumis/dead_simple_owl_design_patterns
10. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, Shi B. Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research. 2015;5(10):2929.
11. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature. 2000;406:747–752
12. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci U S A. 2003;100:10393–10398.
13. Abd El-Rehim DM, Ball G, Pinder SE, Rakha E, Paish C, Robertson JF, Macmillan D, Blamey RW, Ellis IO. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. Int J Cancer. 2005;116:340–350.
14. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lonning PE, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98:10869–10874.
15. O'Brien KM, Cole SR, Tse CK, Perou CM, Carey LA, Foulkes WD, Dressler LG, Geradts J, Millikan RC. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. Clin Cancer Res. 2010;16:6100–6110.