# Use of Genomic Variants in Informatics for Integrating Biology and the Bedside (i2b2)

**Lori C. Phillips MS[1], Simon Minovitsky[2], Igor Ratnere[2], Inna Dubchak Ph.D.[2,3], Isaac Kohane MD Ph.D.[4], Shawn N. Murphy MD Ph.D.[5]**
**[1]Partners Healthcare Systems, Charlestown, MA, [2]DOE Joint Genome Institute, Walnut Creek, CA, [3]Lawrence Berkeley National Laboratory, Berkeley, CA, [4]Children's Hospital, Boston, MA, [5]Massachusetts General Hospital, Boston, MA**

## Abstract

*An electronic Clinical Research Chart (CRC) has been developed under the NIH Roadmap National Centers for Biomedical Computing (NCBC) Informatics for Integrating Biology and the Bedside (i2b2) effort to organize and integrate clinical data, trials data and genomic data with knowledge annotations. This paper describes a method to classify genomic variants within the CRC. A set of new tools to explore and annotate these variants have been developed and are demonstrated. Finally, the inherent architecture of the CRC permits translational querying and computational analysis of genomic and clinical data.*

## Introduction

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap NCBC (http://www.ncbc.org). A primary goal of i2b2 is to provide clinical investigators with a cohesive set of software tools necessary to collect and manage clinical research data in the genomics age—a software suite to construct and manage the modern clinical research chart.

The i2b2 Hive is a collection of interoperable software objects, or "cells" that communicate through scalable web services. The i2b2 Hive consists of a number of core cells that provide basic services to access the data in the CRC clinical repository and present it in a form for consumption by the researcher. The software architecture allows additional cells to be developed for specific forms of analysis, and then integrated into a software "hive" to form a cohesive whole.

## Exploration of genomic variants

Genomic variant exploration has two contexts within the scope of i2b2. First, the variants have to be named and classified within the Ontology cell such that the user may easily discover the variants they wish to query against the CRC. Secondly, it would be helpful if the workbench could provide some useful information and annotations about the variant.

When designing a method to represent genomic variant data for the CRC, we analyzed properties available to us from both the available literature and our genomic lab reporting system (Figure 1).

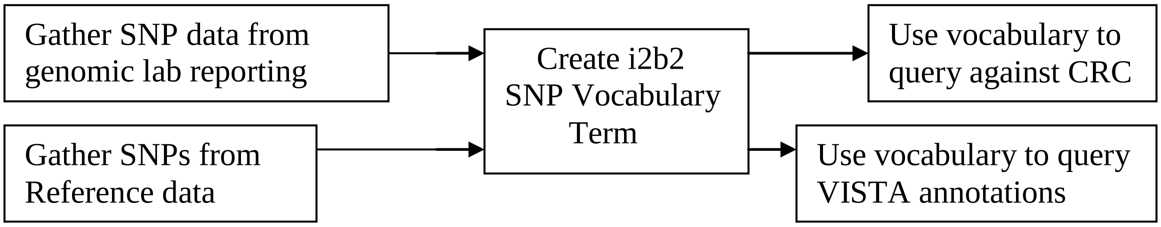
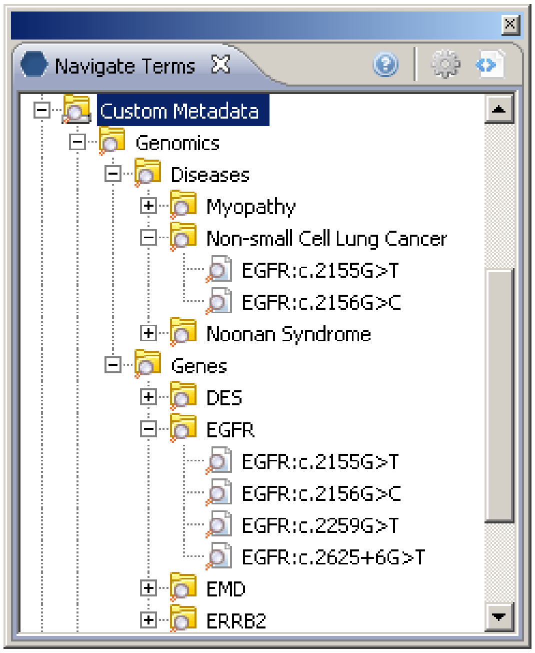

Gather SNP data from genomic lab reporting
Gather SNPs from Reference data
Create i2b2 SNP Vocabulary Term
Use vocabulary to query against CRC
Use vocabulary to query VISTA annotations

**Figure 1.** An i2b2 SNP vocabulary term can be created after review of reference data or data collected from the lab as a result of a post-generation sequencer. The variants are named and classified for insertion into the CRC. Existing i2b2 cells permit querying and analysis of the variants within the i2b2 framework. A new i2b2 cell has been created to annotate and compare these variants.

Exploration of a genomic variant requires, at a minimum, information to both reliably locate the variant within the genome and to describe the variant using an easily understood nomenclature. To this means, we found that the Human Genome Variation Society (HGVS) nomenclature[2] for the description of sequence variants works very well for both exonic and intergenic variants and adopted it as our variant naming convention.

Within the Navigate Terms view, we organized the variants both by associated disease and by gene name. As shown in Figure 2, the two sets may not be identical as some variants are not necessarily disease associated. We chose the aforementioned HGVS names for the variants over simple dbSNP[1] rs numbers in that they are more helpful and descriptive to users uninitiated in genomic variants. The dbSNP rs number is also underspecified as it does not specify the specific nucleotides exchanged in the polymorphism.



- Custom Metadata
  - Genomics
    - Diseases
      - Myopathy
      - Non-small Cell Lung Cancer
        - EGFR:c.2155G>T
        - EGFR:c.2156G>C
      - Noonan Syndrome
    - Genes
      - DES
      - EGFR
        - EGFR:c.2155G>T
        - EGFR:c.2156G>C
        - EGFR:c.2259G>T
        - EGFR:c.2625+6G>T
      - EMD
      - ERRB2

**Figure 2.** Navigate Terms view with variants classified by disease and gene

Since i2b2 is primarily a clinical repository, we have chosen to omit intergenic variants from the scope of this paper. However the HGVS naming convention does extend to intergenic variants and the tools described in this paper can be extended to include them.

**Gathering Genomic Metadata**

We investigated collections of genomic metadata such as the National Center for Biomedical Ontology (NCBO) SNP Ontology[3], dbSNP[8] and the HUGO Mutation Database Initiative recommendations for variant submissions[7] and data sharing[9] to understand what data is typically collected for genomic variants. The NCBO SNP Ontology alone lists 23 classes to fully describe a genomic variant ranging from amino acid classifications, genomic region, haplotype, sequencing data, and variant type. In a perfect world, we would have access to all of this metadata for all the variants collected in our system. Since this was unlikely, we focused on a subset of this metadata that we felt we could reliably obtain from our genomic lab reporting system (Figure 3).

```
GenomicMetadata
   Version  1.0
   ReferenceGenomeVersion hg18
   SequenceVariant
      HGVSName  NM_0005228.3:c.2155G>T
      SystematicName c.2155G>T
      SystematicNameProtein  p.Glu719Cys
      AaChange missense
      DnaChange substitution
   SequenceVariantLocation
      GeneName EGFR
      FlankingSeq_5  GAATTCAAAAAGATC
AAAGTGCTG
      FlankingSeq_3  GCTCCGGTGCGTTCG
GCACGGTGT
      RegionType exon
      RegionName Exon 18
        Accessions
           Accession
              Name NM_005228
              Type mrna (NCBI)
           Accession
              Name NP_005219
              Type protein (NCBI)
          Accession
              Name NT_004487
              Type contig (NCBI)
   ChromosomeLocation
      Chromosome chr7
      Region 7p12
      Orientation +
```

**Figure 3.** Amalgamated i2b2 genomic variant metadata.

This metadata is divided into a Reference Genome Version associated with the variant; Sequence Variant data to describe the name and variant type; Sequence Variant Location data to describe the associated gene name and id, flanking sequences, region type, region name, and associated accessions; and Chromosome Location data to list chromosome region and orientation.

**Location, Location, Location**

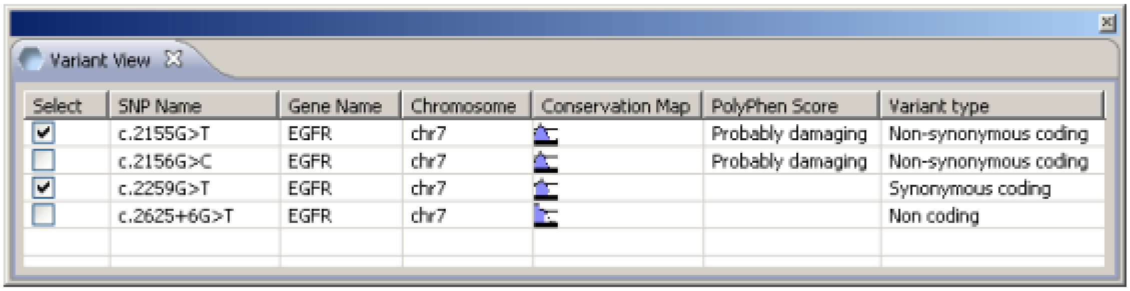
Even within our own lab reporting system, we knew we may not always have all of this information, and certainly our i2b2 users would not always have such data, so we then focused on what data we felt we needed as an absolute minimum. Data to derive variant position on the chromosome is very important, since chromosomal position is a gateway to various external genomic databases and tools such as: VISTA, UCSC Genome Browser, PolyPhen, dbSNP, HapMap, 1000Genomes, and the NCBI Map Viewer. Once genomic position is known, the variants can be explored for comparative conservation data against other organisms, predicted functional effects (PolyPhen[4]), and genomic type (non-coding, synonymous/non-synonymous coding).

An i2b2 VISTA cell was created to provide variant chromosomal position when it is not known or is ambiguous. The cell is flexible enough to accept the following formats as input: dbSNP rs number; HGVS name; or gene name/flanking sequences. In addition, the desired output reference genome version must be specified. For the variant identified in figure 3, the location chr7:55209201 – 55209201 on reference genome hg18 is returned. In doing so, (identifying variants by genomic position across a common reference point) we have effectively provided a way to normalize our data across experiments.

With a method in place to reliably ascertain variant location on a common reference genome, we can use the i2b2 framework to leverage this knowledge and data from the comprehensive set of VISTA programs and databases for comparative analysis of genomic sequences suite hosted at Lawrence Berkeley National Laboratory (LBNL). The VISTA software suite provides biomedical investigators with a unified framework for alignment of long genomic sequences and interactive visual analysis of alignments along with functional annotation such as variant type and predicted functional effects for each variant. An i2b2 comparative genomics cell was developed at LBNL to provide this information and is shown in figure 4.

Genomic variants may be dragged and dropped from the Navigate Terms view to the Variant View. This table summarizes features of a variant that may be of interest to a user: associated gene name, chromosomal location, a comparative analysis summary against the mouse genome, predicted functional effects and variant type. Clicking on the summary conservation image causes a full conservation map to appear for a region around selected variants. Figure 5 shows the alignment results of six organisms to human. Each point on a curve shows the percent identity of a particular base pair in the 100bp window surrounding it. The blue color represents exons and the pink color represents conserved non-coding sequences (CNSes).

Genomic variant exploration via the Variant View is intended to help users discover which variants contained within their i2b2 system may be of interest to study further. Now that we have a method to classify, identify and explore variants, we are in a position to query the CRC for patients with a known variant.



| Select | SNP Name | Gene Name | Chromosome | Conservation Map | PolyPhen Score | Variant type |
|---|---|---|---|---|---|---|
| ☑ | c.2155G>T | EGFR | chr7 | | Probably damaging | Non-synonymous coding |
| □ | c.2156G>C | EGFR | chr7 | | Probably damaging | Non-synonymous coding |
| ☑ | c.2259G>T | EGFR | chr7 | | | Synonymous coding |
| □ | c.2625+6G>T | EGFR | chr7 | | | Non coding |

**Figure 4.** i2b2 Variant View containing variants of interest.

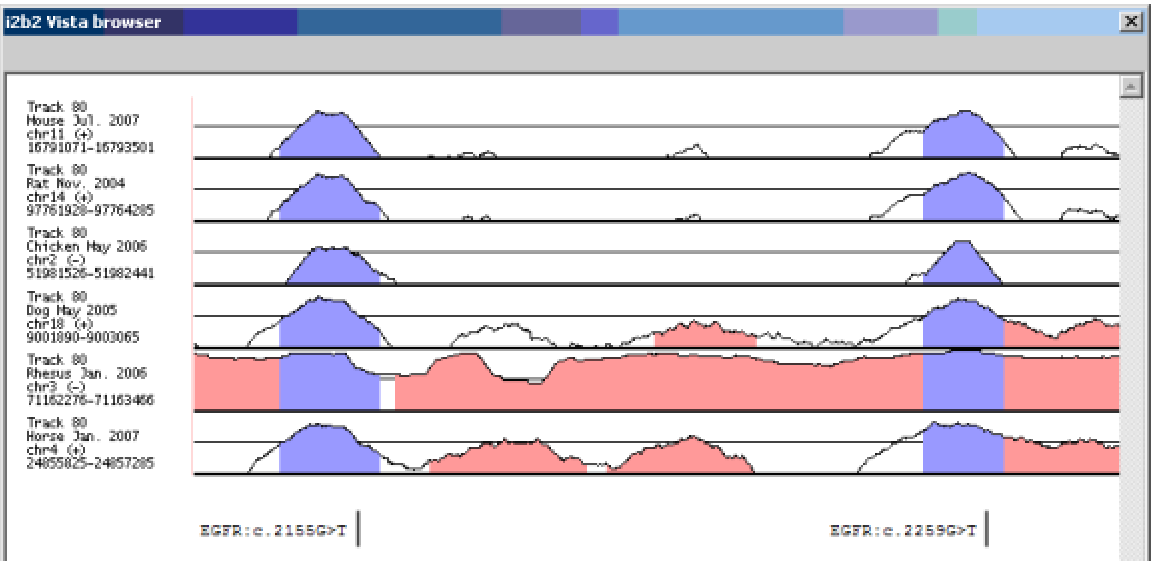


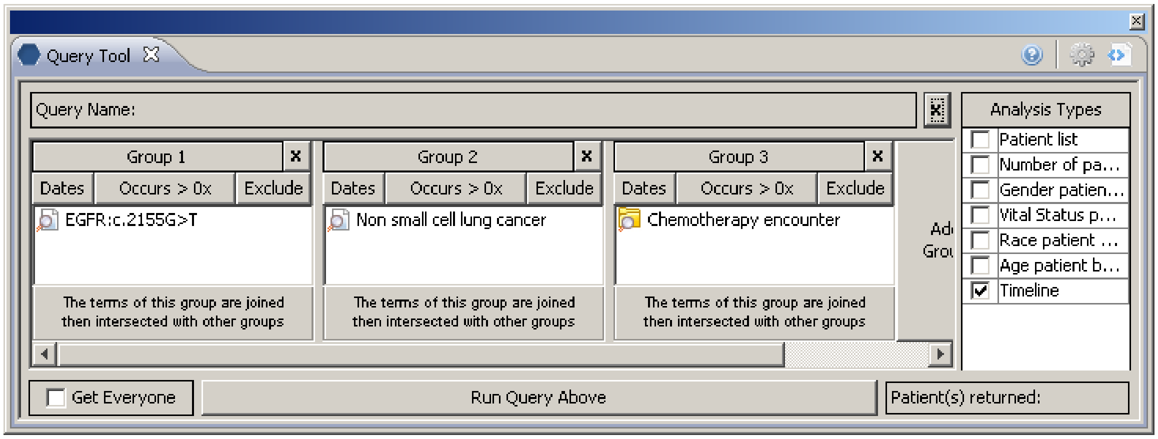
**Figure 5.** i2b2 Vista browser showing conservation map of selected variants. The blue color represents exons and the pink color represents conserved non-coding sequences (CNSes). The picture clearly shows that two variants are within exons.

### Enabling Translational Queries

Figure 6 shows a translational query across three domains. The CRC query seeks all patients that have a certain variant [EGFR:c.2155G>T], a certain disease diagnosis [Non small cell lung cancer] and are receiving a certain treatment [Chemotherapy].

### Computational analysis

The CRC also provides a platform to correlate the presence of genomic variants against clinical findings. The Correlation Analysis cell uses model-less data mining and clustering techniques based on direct pair-wise correlation calculations. It is capable of calculating correlations between heterogeneous concepts such as diagnoses and genomic variants across a set of patients.



Query Name:

Group 1 — Dates | Occurs > 0x | Exclude
EGFR:c.2155G>T
The terms of this group are joined then intersected with other groups

Group 2 — Dates | Occurs > 0x | Exclude
Non small cell lung cancer
The terms of this group are joined then intersected with other groups

Group 3 — Dates | Occurs > 0x | Exclude
Chemotherapy encounter
The terms of this group are joined then intersected with other groups

Analysis Types
- □ Patient list
- □ Number of pa...
- □ Gender patien...
- □ Vital Status p...
- □ Race patient ...
- □ Age patient b...
- ☑ Timeline

□ Get Everyone

Patient(s) returned:

**Figure 6.** A sample translational query across three domains.

Lab results
1. Sequencer
2 Mic

## Conclusion

Designing a genomic variant ontology for common use within the open i2b2 community poses a unique challenge. As described in this paper, variants may be defined across several domains: dbSNP rs number, HGVS name, or gene/flanking sequence pairs. It is often difficult to know if variants from different domains are in fact equivalent. We have shown that we can form a common basis for these variants by ultimately defining them by chromosomal location and nucleotide substitution on a reference genome. The two parameters that are necessary to unequivocally define a variant in these terms are the HGVS name (provides nucleotide substitution) and flanking sequences (along with HGVS name provides chromosomal location).

This strategy allows us to map variants from systems that may use different naming conventions. Regardless of the name assigned by the system, if the variants' HGVS name and flanking sequences are identical, then we know that the differently named variants are equal. A common organizational strategy is to gather known equivalent terms into a common parent folder; this folder may then be used to collectively query for all the equivalent terms.

Through careful consideration, a genomic variant ontology can be organized that lends itself to further exploration, annotation and analysis of the variant within i2b2.

## Acknowledgments

This work was funded by the NIH Roadmap for Medical Research, Grant U54LM00874.

## References

1. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11.
2. Dunnen JT den, Antonarakis E. Nomenclature for the description of human sequence variations. Hum Genet 2000; 109:121-124.
3. Noy, NF., Shah, NH, Whetzel, PL, Dai, B, Dorf, M, Griffith, N, Jonquet, C, Rubin, DL, Storey, MA Chute, CG, Musen, MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009; Jul 1:37.
4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 2010; 7(4):248-249.
5. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32 (Web Server issue):W273-9
6. Dubchak I, Poliakov A, Kislyuk A, Brudno M. Multiple whole-genome alignments without a reference organism. Genome Res. 2009 Apr; 19(4):682-9.
7. Auerbach AD, Horaitis O, Cotton RGH, et al, Allele variant entry form, http://www.hgvs.org/entry.html
8. dbSNP Entity Relationship Diagram, Build 125, 12/6/05
9. Dunnen JT den, Sijmons RH, et al, Sharing data between LSDBs and central repositories, Hum. Mut. 2009 Apr;30(4):493-5.