Informatics for Integrating Biology and the Bedside

User Guide

Genomic Import Plugin

Document Version: *1.0*

i2b2 Software Version: *1.7*

Table of Contents

DOCUMENT MANAGEMENT

Revision Number	Date	Author	Description of change
1.0	04/18/13	Lori Phillips	Original document

ABSTRACT

This is a User's Guide for the Genomic Import Plugin. This guide will help you import NGS variant data using the Genomic Import Plugin and assist you in querying for genomic variants.

1. BEFORE YOU BEGIN

1.1 Prerequisites

1.1.1 i2b2Workbench

The Genomics import plugin is compatible with i2b2Workbench version 1.7. Download the i2b2Workbench from https://www.i2b2.org/software/ . Follow installation and configuration instructions as given in the *i2b2 Workbench Developers' Guide* which can be found under the Docs tab.

It is important to note that the time necessary to process genomic data files can exceed the default workbench timeout of 30 minutes. Therefore, set the i2b2workbench.properties timeout parameter to something very large (order of hours) prior to starting the workbench. The following sets it for 5 hours, you may need to adjust accordingly.

TimeoutInMilliseconds=18000000

1.1.2 i2b2 Hive

The Genomics import plugin is compatible with 1.7 version of the hive.

1.1.3 PERL_HOME

Download a copy of perl (www.perl.org) onto the same computer as the i2b2 workbench Configure a PERL_HOME environment variable that points to the perl /bin directory.

1.1.4 ANT_HOME

Our database tools rely on a copy of Apache Ant. Download a copy of *Apache Ant* from the following Apache website: http://archive.apache.org/dist/ant/binaries/ and configure an ANT_HOME environment variable.

Our scripts have been tested using ant version 1.6.5. A sample environment variable setting is shown:

ANT_HOME=/opt/apache-ant-1.6.5

2. INSTALLATION AND PREPARATION

2.1 Preparing the Ontology database

This package contains a file called sequenceOntologyData.zip. Locate it now and unzip to a working directory. Scripts are provided for both oracle and sqlserver to create and populate a new metadata table called SEQUENCE_ONTOLOGY.

The following steps outline the process for upgrading the i2b2 metadata tables to support genomics.

1. 'cd data' of your working directory
2. Edit the db.properties file for your database username/password. Do not modify the db.project setting.

If running SQL Server:

db.type=sqlserver

db.username=i2b2metadata

db.password=demouser

db.driver=com.microsoft.sqlserver.jdbc.SQLServerDriver

db.url=jdbc:sqlserver://localhost:1433;database=demo

db.project=demo

- OR -

If running Oracle:

db.type=oracle

db.username=i2b2metadata

db.password=demouser

db.server=localhost:1521:xe

db.driver=oracle.jdbc.driver.OracleDriver

db.url=jdbc:oracle:thin:@127.0.0.1:1521:xe

db.project=demo

- OR -

If running Postgresql:

db.type=postgresql

db.username=i2b2metadata

db.password=demouser

db.driver=org.postgresql.Driver

db.url=jdbc:postgresql://127.0.0.1:5432/i2b2?searchpath=i2b2metadata

db.project=demo

2.1.1 SEQUENCE_ONTOLOGY Table Creation

There are scripts to create and populate a new table called SEQUENCE_ONTOLOGY. This table contains the concept and modifier metadata pertaining to genomics.

```
ant –f data_build.xml create_genomics_metadata_table
```

2.1.2 Update and Load Tables

The next step will load and update several tables:

- Load the SEQUENCE_ONTOLOGY table
- Update the TABLE_ACCESS table to include the SEQUENCE_ONTOLOGY root node.
- Update the SCHEMES table to include the SEQUENCE_ONTOLOGY schemes.
- Update the concept_dimension and modifier_dimension tables with the new genomic-related metadata information.

When running on a database with all ontology and crc tables in one schema run:

```
ant –f data_build.xml load_genomics_metadata
```

When running on database with i2b2demodata and i2b2metadata schemas:

First set up db.properties for the i2b2metadata user and run

```
ant –f data_build.xml load_genomics_metadata2
```

Then set up db.properties for the i2b2demodata user and run

```
ant –f data_build.xml load_genomics_data
```

2.2 Installing the plugin

This package contains a folder called plugins. Locate it now.

Inside this folder is a plugin file edu.harvard.i2b2.eclipse.plugins.genomicsImport_1.7.0.jar. Copy the jar file to the plugins directory of the i2b2Workbench directory.

This package also contains a folder called perlScripts. Locate it now.

Copy the folder and its contents to the dropins directory of the i2b2workbench directory. If the dropins folder does not exist, create one. This folder should exist at the same directory level as the plugins folder.

2.3 Preparing NGS data files for import

The Genomic Import plugin can handle four different types of files:

- Variant Call Format (VCF) files
- ANNOVAR-annotated VCF files
- Genome Variation Format (GVF) files
- I2b2 Observation Fact files

2.3.1 Variant Call Format (VCF) files

This tool is designed to work with version 4.0 or 4.1 VCF files such as those found in the 1000 genomes project. (http://www.1000genomes.org/data#DataAccess) The following files were used to test this project: CEU.exon.2010_03.genotypes.vcf, CEU.trio.2010_07.indel.genotypes.vcf and CEU.low.coverage_2010_07.genotypes.vcf and are mentioned here for comparative purposes. These files are specified as version 4.0.

VCF File Assumptions:

The header line is as follows and lists sample IDs.

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003

In the example above, the sample IDs are NA00001, NA00002 and NA00003.

It is recommended that you download the file CEU.exon.2010_03.genotypes.vcf as we will be referring to it in the examples that follow.

2.3.1.1 VCF MAP FILES

Map files that identify the file date, reference genome and sample to patient number, encounter number mappings are required in the following format:

```
##genome-build hg18
##file-date 2010-03-03
#sample|patient_num|encounter_num
NA00001|1000000001|1880001001
NA00002|1000000002|1880001002
NA00003|1000000003|1880001003
```

Each sample represents a unique patient and encounter.

The date format must be as shown.

```
##genome-build hg18
##file-date 2010-03-03
#sample|patient_num|encounter_num
NA00001|1000000001|1880001001
NA00002|1000000001|1880001002
NA00003|1000000001|1880001003
```

Each sample represents a unique encounter for a single patient.

The date format must be as shown.

A sample map file that corresponds to CEU.exon.2010_03.genotypes.vcf can be found in the data/mapFile folder that was extracted in section 2.1. Locate CEU.exon.2010_03.map.txt now.

2.3.2 ANNOVAR-annotated VCF (VCF-ANNOVAR) files

It is possible to also import VCF files that have been annotated by ANNOVAR. The assumptions for the source VCF file is as above in section 2.3.1.

Download ANNOVAR main package from here:

http://www.openbioinformatics.org/annovar/annovar_download.html

Using the 1000 genomes CEU exon VCF file as an example run the following two annovar steps. Be sure to include the –includeinfo and –comment settings as shown.

- perl convert2annovar.pl –format vcf4 –includeinfo –comment CEU.exon.2010_03.genotypes.vcf > /ANNOVAR/CEU.exon.2010_03.genotypes.vcf
- perl annotate_variation.pl –comment /ANNOVAR/CEU.exon.2010_03.genotypes.vcf /ANNOVAR/annovar/humandb/

One output is /ANNOVAR/CEU.exon.2010_03.genotypes.vcf.variant_function which I refer to as the VCF-ANNOVAR file.

2.3.2.1 VCF MAP FILES

As with the regular VCF files, a VCF map file is required. A sample map file that corresponds to CEU.exon.2010_03.genotypes.vcf.variant_function can be found in folder that was extracted in section 2.1. Locate CEU.exon.2010_03.map.txt now. See section 2.3.1.1 above for details.

2.3.3 Genome Variation Format (GVF) files

GVF version 1.05 or 1.06 files are recognized by this tool.

```
chr1    VCF    SNV    3537996    3537996    .    +    .
    ID=6;Reference_seq=T;Variant_seq=C;Variant_feature=exonic;Ge
ne=WRAP73;Genotype=heterozygous
```

The first eight columns are as per the GVF standard and represent:

Chromosome

Source (unused)

*Sequence Ontology type

*The following Sequence Ontology types are recognized by this tool: SNV, MNP, complex_substitution, insertion, inversion, deletion, indel, copy_number_variation, gap

Start location

End location

Score (unused)

Strand

Phase (unused)

The last column is a series of value/type pairs. The following are currently recognized by the tool and itemized for i2b2: Variant_feature, Gene and Genotype.

GVF files are entered individually; as such the tool requires the user to enter the corresponding file date, patient number, encounter number and reference genome per submission.

2.3.4 i2b2 files

This tool may also be used as a means to provide | -delimited observation_fact i2b2 files to a bulk loader or any import tool. Any type of observation fact data can be submitted as long each row is formatted as follows:

ENCOUNTER_NUM|PATIENT_NUM|CONCEPT_CD|PROVIDER_ID|START_DATE|MODIFIER_CD|INSTANCE_NUM|VALTYPE_CD|TVAL_CHAR|NVAL_NUM|VALUEFLAG_CD|QUANTITY_NUM|UNITS_CD|END_DATE|LOCATION_CD|CONFIDENCE_NUM|UPDATE_DATE|DOWNLOAD_DATE|IMPORT_DATE|UPLOAD_ID|SOURCESYSTEM_CD|OBSERVATION_BLOB

All text fields must be in “quotes”. Null values should appear as empty fields. The first row of the file must be the header row shown above.

2.4 Data Storage Requirements

2.4.1 Input/Output files

VCF files can be very large (order of 5 GB). Our experience suggests that the final i2b2 file is just as large for VCF conversions and 60% larger for VCF-ANNOVAR files. Considering that one VCF file results in multiple i2b2 files (one per sample), the storage requirements add up quickly.

The tool is designed such that that i2b2 output file is co-located in the same folder as the input file.

2.4.2 File Shares

Another important storage requirement to consider is the use of file shares. Ultimately the output file location will be provided as input to a bulk loader. Creating a file share that both the import tool and the bulk loader have access to simplifies storage requirements.

2.4.3 Database

You can expect that the final i2b2 file for a single sample will result in approximately 20 million rows in the observation_fact table. Your database storage requirements must account for this as well as additional space for indexing.

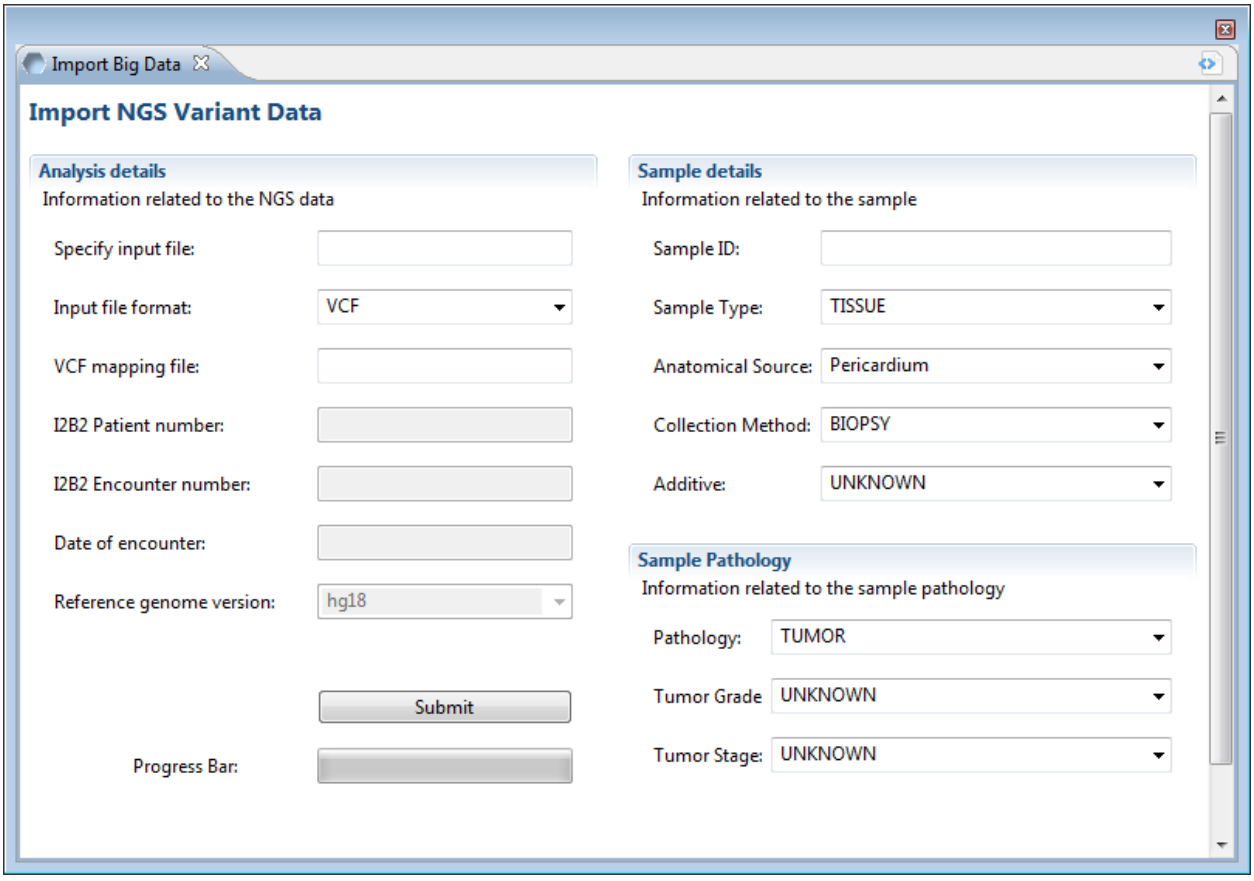
3. IMPORTING GENOMICS DATA

3.1 Import NGS Observation_fact Data

3.1.1 Genomics Import View

Start up the workbench. To bring the new Import Big Data View into focus, select Window -> Show View -> Other -> Genomics Import Category -> Genomics Import. Double click on the view tab to display in full screen view.

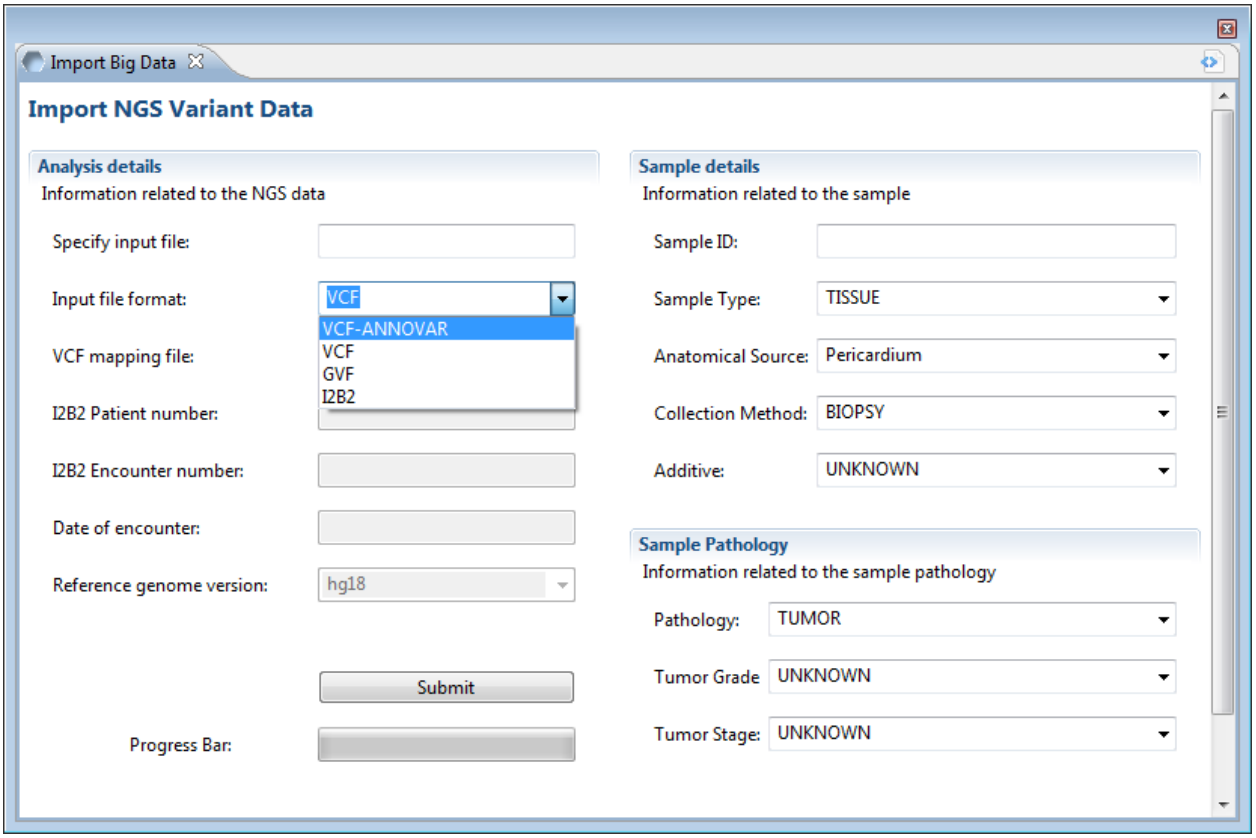
This view allows you to import the VCF, VCF-ANNOVAR, GVF and I2B2 file types described in section 2.3 to the observation_fact table.



This view is designed to transform VCF, VCF-ANNOVAR and GVF data files to an i2b2 formatted file that can either be loaded manually or, as described later in this document, submitted to an automated bulk loader process.

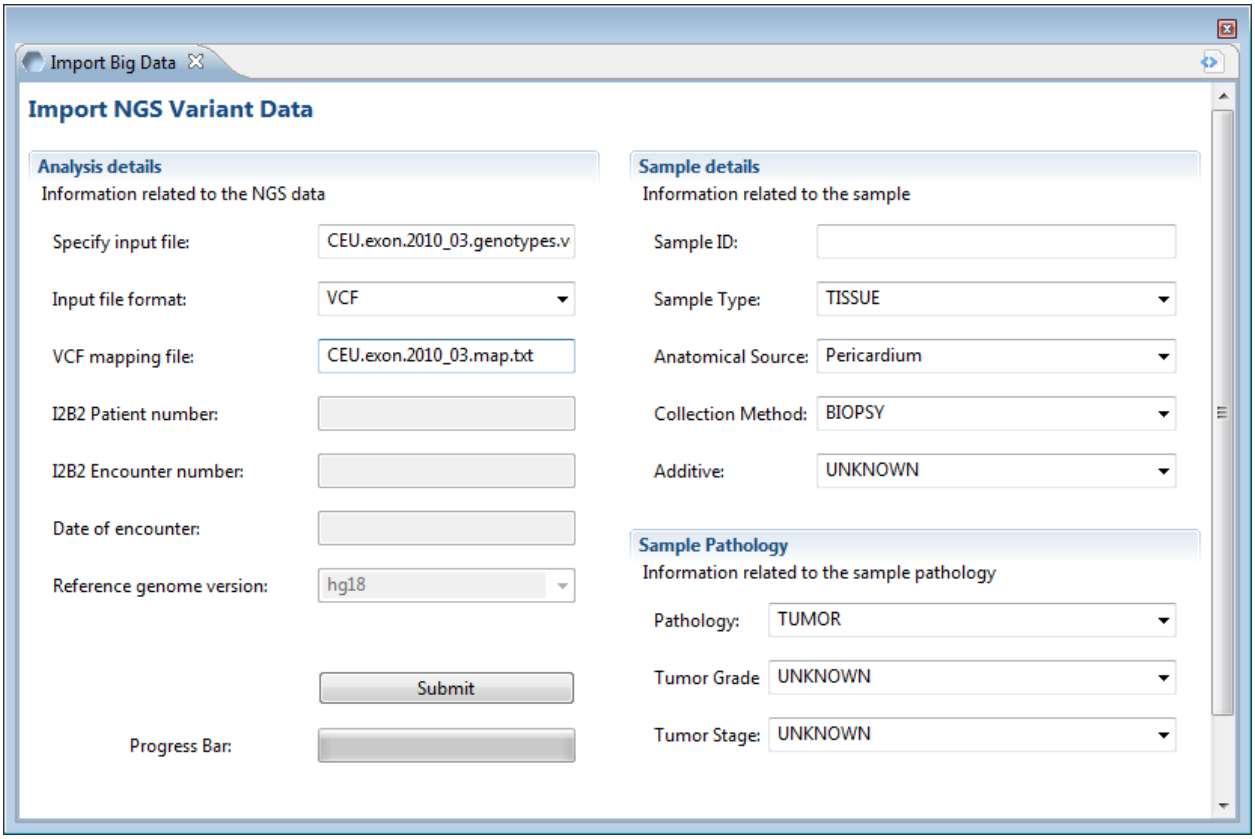
3.1.1.1 IMPORTING VCF OR VCF-ANNOVAR FILES

Choose VCF or VCF-ANNOVAR from the Input file format drop down menu.



When importing large VCF or VCF-ANNOVAR files, set the i2b2workbench.properties timeout parameter to something very large (order of hours) prior to starting the workbench.

TimeoutInMilliseconds=18000000

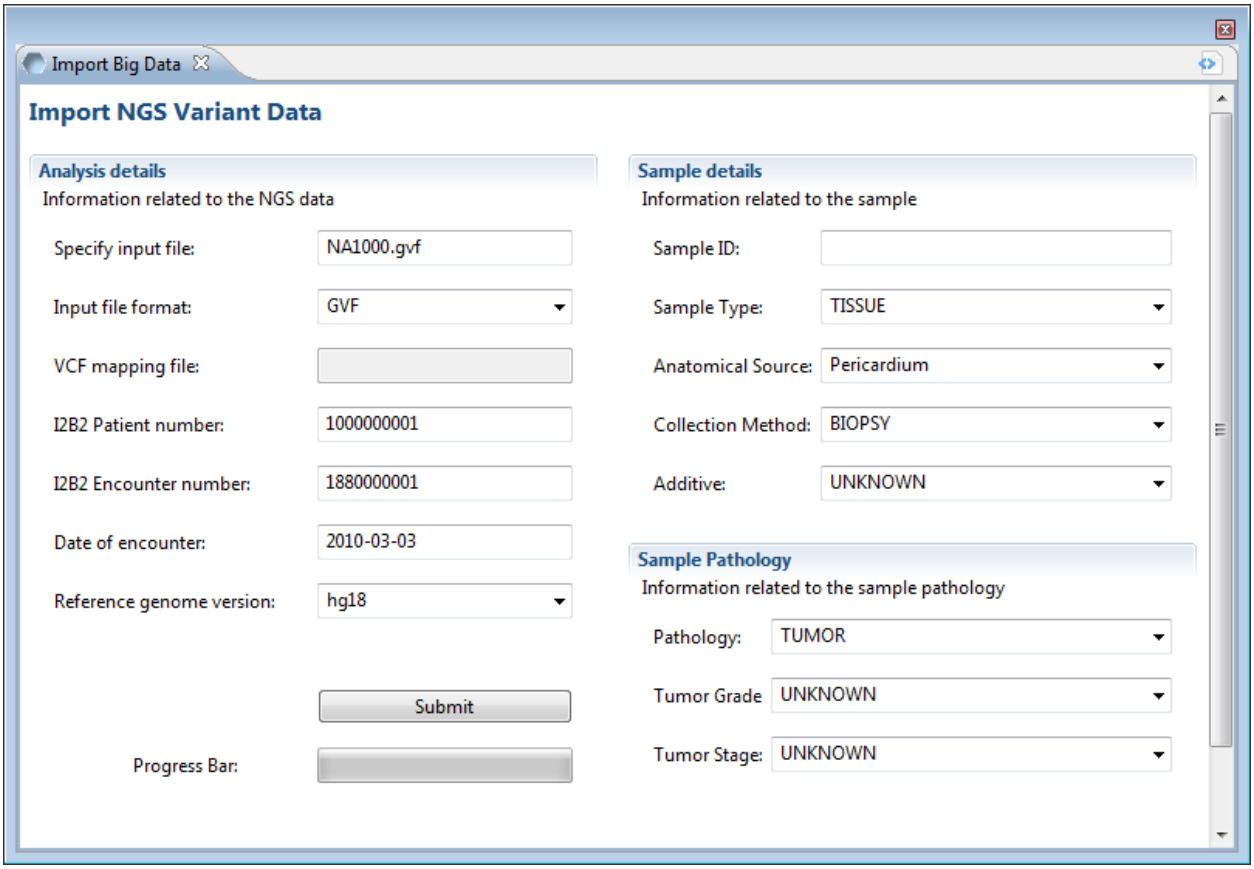


Specify the VCF or VCF-ANNOVAR file and its associated mapping file*. In the example above we show the 1000genomes CEU exon VCF file. Click on Submit; status messages will appear below the progress bar indicating first a VCF (VCF-ANNOVAR) to GVF conversion process, followed by a GVF to I2B2 conversion process.

At the end of this process you will have one I2B2 file for each subject/sample listed in the VCF file and as specified by the mapping file. The resulting files will be named "SAMPLE.encounter.i2b2", where SAMPLE is the VCF file sample ID (e.g NA06984), and encounter is the encounter_num associated with that sample as dictated by the mapping file.

*Always list the full path of the input file, including the location of the share://share.i2b2.org/Project/Genomics/ CEU.exon.2010_03.genotypes.vcf or ://share.i2b2.org/Project/Genomics/ ANNOVAR/CEU.exon.2010_03.genotypes.vcf.variant_function

3.1.1.2 IMPORTING GVF FILES



GVF files must be imported one at a time as shown above*. For each file, specify the associated I2B2 patient number and encounter number, date of encounter in YYYY-MM-DD format and reference genome version.

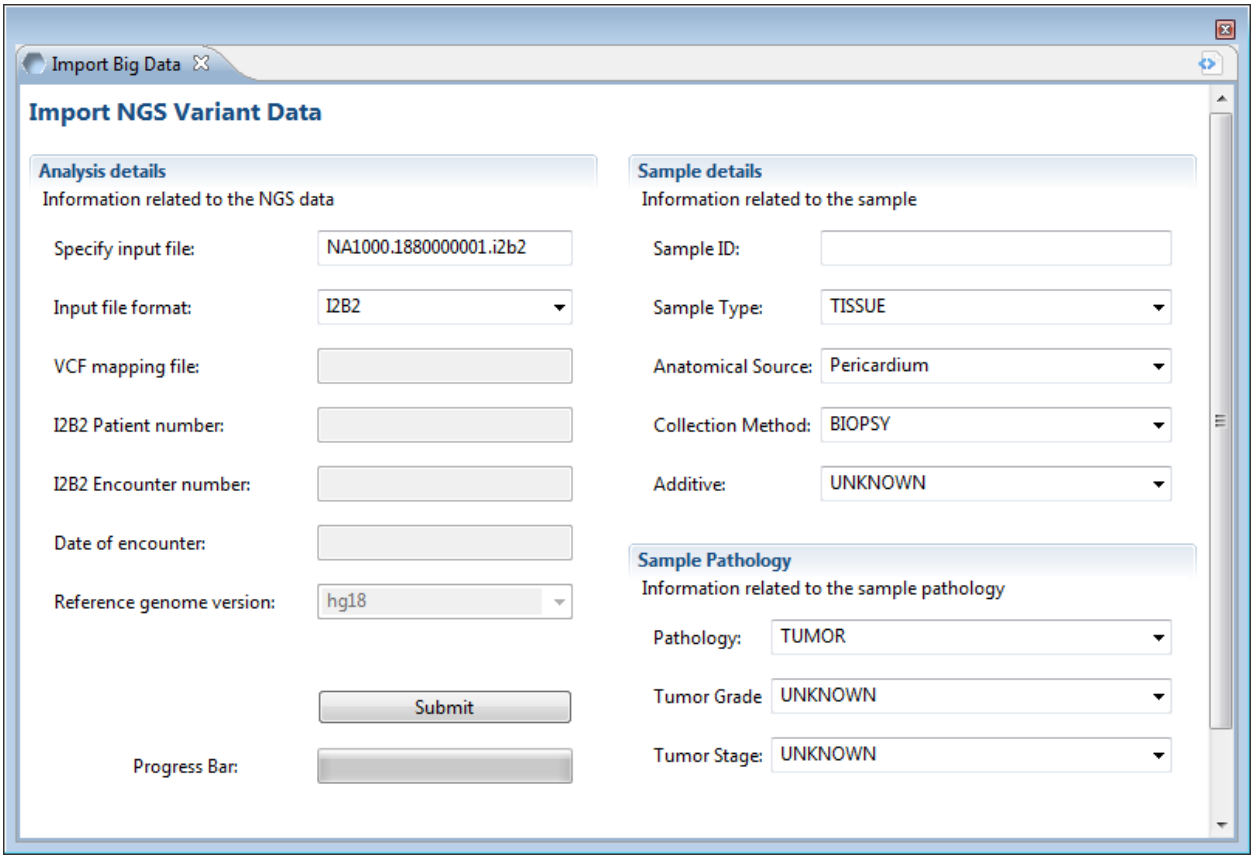
Click on Submit; status messages will appear below the progress bar communicating the GVF to I2B2 conversion process. At the end of this process you will have an I2B2 file that corresponds to your input GVF file. For the example shown above, the input file NA1000.gvf would result in an i2b2 file called NA1000.1880000001.i2b2, where "1880000001" is the encounter number specified in the form.

*Always list the full path of the input file, including the location of the share: //share.i2b2.org/Project/Genomics/NA1000.gvf

3.1.1.3 IMPORTING I2B2 FILES

Large observation_fact data files of any nature (not just genomics) may also be staged by this tool for bulk loading. Simply enter the file name*, and select an input file format of 'I2B2'.

.



*Always list the full path of the input file, including the location of the share: //share.i2b2.org/Project/Genomics/NA1000.180000001.i2b2

4. BULK LOADING BIG DATA

4.1 UPLOAD_STATUS Table

Each time an i2b2 file is created by this tool, an entry is made to the UPLOAD_STATUS table via a publishDataRequest message to the CRC.

UPLOAD_STATUS		
PK	**UPLOAD_ID**	**int**
	UPLOAD_LABEL	**varchar(100)**
	USER_ID	**varchar(50)**
	SOURCE_CD	**varchar(50)**
	NO_OF_RECORD	bigint
	LOADED_RECORD	bigint
	DELETED_RECORD	bigint
	LOAD_DATE	**Datetime**
	END_DATE	Datetime
	LOAD_STATUS	varchar(100)
	MESSAGE	varchar(4000)
	INPUT_FILE_NAME	varchar(500)
	LOG_FILE_NAME	varchar(500)
	TRANSFORM_NAME	varchar(500)

A typical entry resulting from the import tool is as follows:

1|BULK_LOAD_OBS_FACT|demo|GENOMIC_CLIENT|0|0|0|2013-04-22||QUEUED|
|\\share.i2b2.org\Project\Genomics\NA1000.180000001.i2b2||GENOMIC_IMPORT

4.2 Bulk Loader

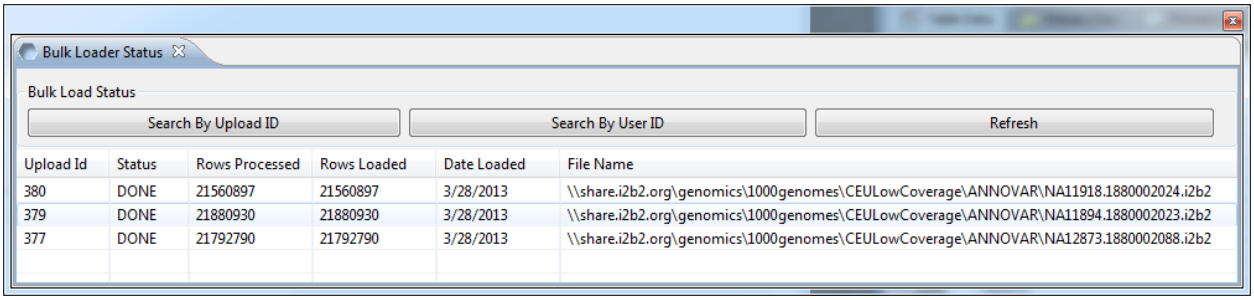
A bulk loader can be written that queries the UPLOAD_STATUS table and then loads any file with a load_status of 'QUEUED'. It is important to note that the input_file_name must have a full path name to the file whose location is understood by the bulk loader. In our example above we show a file that exists on a theoretical file share: \\share.i2b2.org\Project\Genomics\NA1000.180000001.i2b2

In our lab, we have created a package that may be run as a SQLSERVER Agent job. This package queries the UPLOAD_STATUS table for files with a load_status of 'QUEUED'. The file is then bulk loaded to the observation_fact table, while the load_status is set to 'LOADING'. Upon completion the load_status is set to 'DONE' and the no_of_record (number of records in the sample file) and loaded_record (number of records written to the table) fields are updated accordingly.

A sample copy of the SSIS package we created is included in the software package.(ObsFactBulkLoader_2005.dtsx) It is designed to run with SQLSERVER 2005 and has been scrubbed of any database connection information. It is presented as representative sample only; users may use this file as a starting point and would need to configure it for their system accordingly.

4.3 Bulk Loader Status

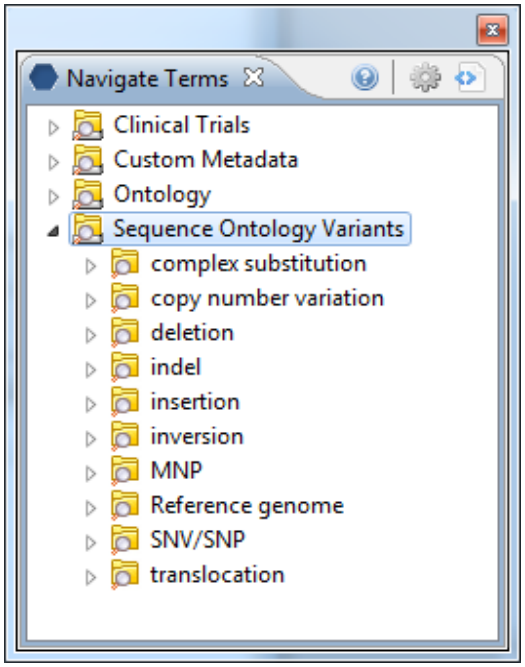
The Bulk Loader Status view displays the status of UPLOAD_STATUS entries by upload_id or user_id. To obtain status for a single job, search by Upload ID for that job. Shown below are the results of Search by User ID for a single user.



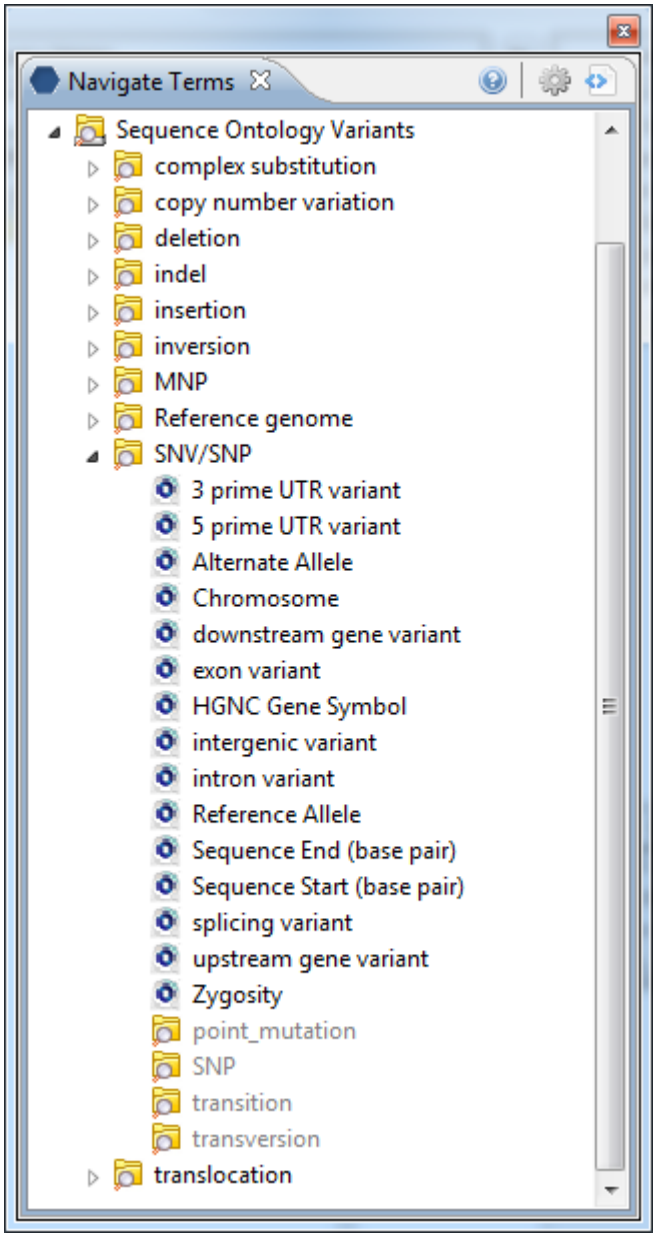
5. QUERYING FOR GENOMICS DATA

5.1 Navigate Terms View

Start up the i2b2Workbench as you normally do and select the project that has been configured with Sequence Ontology (SO) metadata. The Sequence Ontology Variants container should appear in the Navigate Terms view. When expanded, you should see the SO concepts as shown below.



For any given concept you should see the following modifiers:



5.2 Querying for Genomic Data

Genomic variant data can be queried by:

Chromosomal location (chromosome number, start, end position, reference genome)

Relative position or type (integenic, intronic, exonic, downstream, upstream, UTR, splicing)

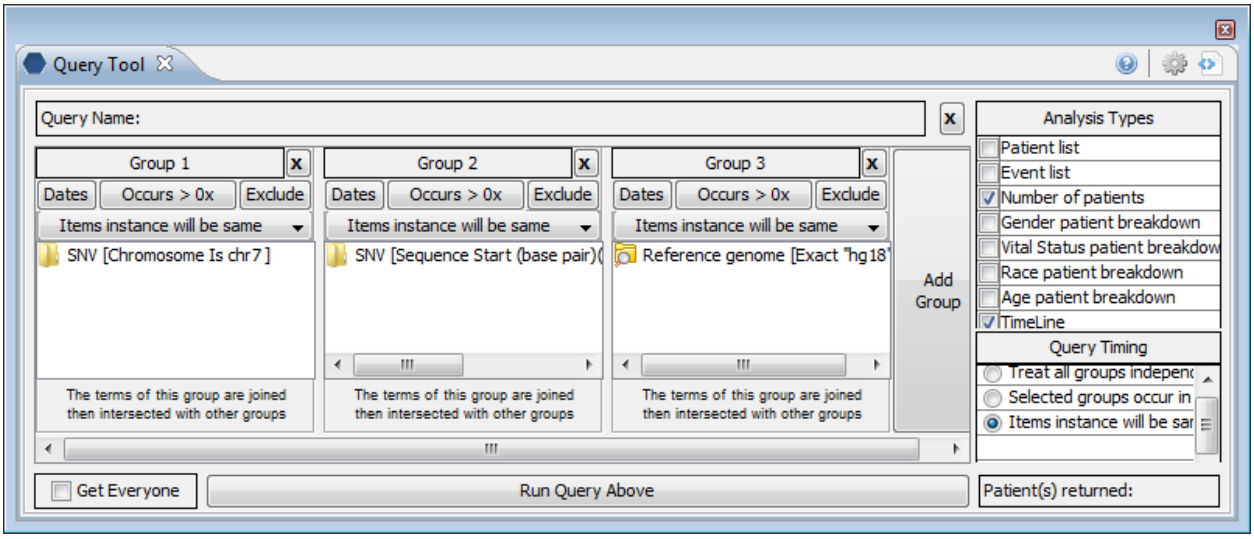
Gene location

Zygosity

Consider variant rs121434568 or NM_005228.3:c.2573T>G. It is a heterozygous exonic variant located on gene EGFR at location 55227009 (hg18) of chromosome 7. This information will be used in the examples that follow.

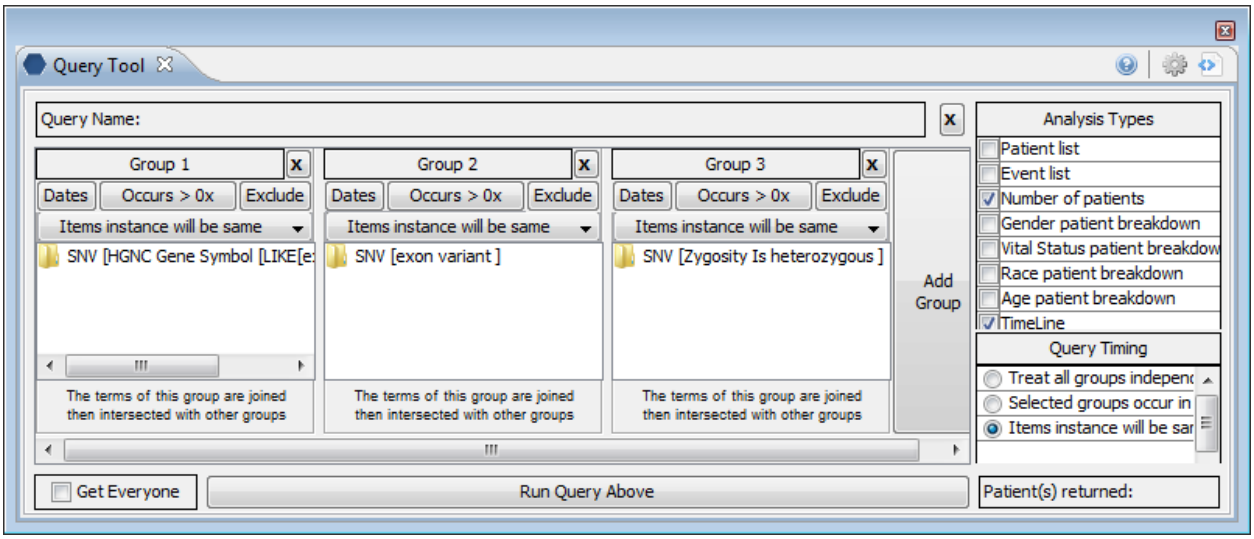
5.2.1 Querying by Chromosomal Location

This is particularly useful when you know the chromosomal location of the variant(s) of interest within a certain reference genome. In this example we query for a patients with an SNV variant on CHR7 at start location between 5522700 and 5522720 on reference genome hg18. Whenever start or end locations are entered, the reference genome must be specified. Be sure to select “Items instance will be the same”.



5.2.2 Querying by Associated Gene Name

This is particularly useful when you are searching for variants associated with a gene. In this example we query for a patients with a heterozygous SNV variant located in an exonic region of gene EGFR. Be sure to select “Items instance will be the same”.



6. REFERENCES

1. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data, Nucleic Acids Research, 38:e164, 2010).

2. VCF2GVF.pl Kong, SekWon, Lee, Joon, Boston Childrens Hospital, email correspondence Nov 2012.