Oracle® Health Sciences Translational Research Center

User's Guide

Release 3.1.0.3

E66623-05

September 2016

ORACLE®

Oracle Health Sciences Translational Research Center User's Guide, Release 3.1.0.3

E66623-05

Contents

4 Cohort Query

5 Cohort Viewer

6 Single Patient or Subject Viewer

Preface

Oracle Health Sciences Translational Research Center (TRC) comprises of two products - Oracle Health Sciences Omics Data Bank (ODB) and Oracle Health Sciences Cohort Explorer (OHSCE). This guide provides information about how to use TRC.

Audience

This document is intended for the following job classifications:

- Life and health sciences investigators
- Researchers within a group led by a principal investigator
- Clinicians
- Biostatisticians
- Bioinformaticians

Documentation Accessibility

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc.

Access to Oracle Support

Oracle customers that have purchased support have access to electronic support through My Oracle Support. For information, visit http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info or visit http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs if you are hearing impaired.

Finding Information and Patches on My Oracle Support

Your source for the latest information about Oracle Health Sciences Translational Research Center is Oracle Support's self-service Web site, My Oracle Support (formerly MetaLink).

Before you install and use an Oracle software release, always visit the My Oracle Support Web site for the latest information, including alerts, release notes, and patches. Visit the Oracle Technology Network for the latest documentation.

Creating a My Oracle Support Account

You must register at My Oracle Support to obtain a user name and password account before you can enter the Web site.

To register for My Oracle Support:

1. Open a Web browser to http://support.oracle.com.
2. Click the **Register here** link to create a My Oracle Support account. The registration page opens.
3. Follow the instructions on the registration page.

Signing In to My Oracle Support

To sign in to My Oracle Support:

1. Open a Web browser to http://support.oracle.com.
2. Click **Sign In**.
3. Enter your user name and password.
4. Click **Go** to open the My Oracle Support home page.

Searching for Knowledge Articles by ID Number or Text String

The fastest way to search for product documentation, release notes, and white papers is by the article ID number.

To search by the article ID number:

1. Sign in to My Oracle Support at http://support.oracle.com.
2. Locate the Search box in the upper right corner of the My Oracle Support page.
3. Click the sources icon to the left of the search box, and then select Article ID from the list.
4. Enter the article ID number in the text box.
5. Click the magnifying glass icon to the right of the search box (or press the Enter key) to execute your search.

   The Knowledge page displays the results of your search. If the article is found, click the link to view the abstract, text, attachments, and related products.

In addition to searching by article ID, you can use the following My Oracle Support tools to browse and search the knowledge base:

- Product Focus — On the Knowledge page, you can drill into a product area through the Browse Knowledge menu on the left side of the page. In the Browse any Product, By Name field, type in part of the product name, and then select the product from the list. Alternatively, you can click the arrow icon to view the complete list of Oracle products and then select your product. This option lets you focus your browsing and searching on a specific product or set of products.
- Refine Search — Once you have results from a search, use the Refine Search options on the right side of the Knowledge page to narrow your search and make the results more relevant.
- Advanced Search — You can specify one or more search criteria, such as source, exact phrase, and related product, to find knowledge articles and documentation.

Finding Patches on My Oracle Support

Be sure to check My Oracle Support for the latest patches, if any, for your product. You can search for patches by patch ID or number, or by product or family.

To locate and download a patch:

1. Sign in to My Oracle Support at http://support.oracle.com.
2. Click the **Patches & Updates** tab.

   The Patches & Updates page opens and displays the Patch Search region. You have the following options:

   - In the Patch ID or Number is field, enter the primary bug number of the patch you want. This option is useful if you already know the patch number.
   - To find a patch by product name, release, and platform, click the Product or Family link to enter one or more search criteria.
3. Click **Search** to execute your query. The Patch Search Results page opens.
4. Click the patch ID number. The system displays details about the patch. In addition, you can view the Read Me file before downloading the patch.
5. Click **Download**. Follow the instructions on the screen to download, save, and install the patch files.

Finding Documentation on Oracle Technology Network

The Oracle Technology Network Web site contains links to all the latest Oracle user and reference documentation. To find the latest user documentation for Oracle products:

1. Go to the Oracle Technology Network at

   http://www.oracle.com/technetwork/index.html and log in.
2. Mouse over the Support tab, then click the **Documentation** hyperlink.

   Alternatively, go to Oracle Documentation page at

   http://www.oracle.com/technology/documentation/index.html
3. Navigate to the product you need and click the link.

   For example, scroll down to the Applications section and click Oracle Health Sciences Applications.
4. Click the link for the documentation you need.

Related Documents

For more information, see the following documents in the *Oracle Business Intelligence Suite Enterprise Edition 11g Release 1 (11.1.1)* documentation set, and the *Oracle Healthcare Data Warehouse Foundation Release 6.1* documentation set:

Oracle Business Intelligence Enterprise Edition Documentation

The *Oracle Business Intelligence Suite Enterprise Edition Online Documentation Library* (Part E21764) documentation set includes:

- *Oracle® Fusion Middleware User's Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E10544)

- *Oracle® Fusion Middleware Metadata Repository Builder's Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E10540)
- *Oracle® Fusion Middleware System Administrator's Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E10541)
- *Oracle® Fusion Middleware Scheduling Jobs Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E18562)
- *Oracle® Fusion Middleware Security Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E10543)
- *Oracle® Fusion Middleware Developer's Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E10545)
- *Oracle® Fusion Middleware Integrator's Guide for Oracle Business Intelligence Enterprise Edition 11g Release 1 (11.1.1)* (Part E16364)

Oracle Healthcare Data Warehouse Foundation Documentation

The *Oracle Healthcare Data Warehouse Foundation* documentation set includes:

- *Oracle Healthcare Data Warehouse Foundation Release Notes*
- *Oracle Healthcare Data Warehouse Foundation Secure Installation and Configuration Guide*
- *Oracle Healthcare Data Warehouse Foundation Programmer's Guide*
- *Oracle Healthcare Data Warehouse Foundation Interface Table Programmer's Guide*

Oracle Data Integrator Documentation

The *Oracle Data Integrator* documentation is a part of the *Oracle Fusion Middleware 11.1.1.6* documentation. Oracle Data Integrator documents in the Fusion Middleware Documentation Library are as follows:

- *Oracle® Fusion Middleware Getting Started with Oracle Data Integrator 11g Release 1 (11.1.1)*
- *Oracle® Fusion Middleware Developer's Guide for Oracle Data Integrator 11g Release 1 (11.1.1)*
- *Oracle® Fusion Middleware Installation Guide for Oracle Data Integrator 11g Release 1 (11.1.1)*
- *Oracle® Fusion Middleware Application Adapters Guide for Oracle Data Integrator 11g Release 1 (11.1.1)*
- *Oracle® Fusion Middleware Knowledge Module Developer's Guide for Oracle Data Integrator 11g Release 1 (11.1.1)*
- *Oracle® Fusion Middleware Connectivity and Knowledge Modules Guide for Oracle Data Integrator 11g Release 1 (11.1.1)*

Conventions

The following text conventions are used in this document:

Convention	Meaning
boldface	Boldface type indicates graphical user interface elements associated with an action, or terms defined in text or the glossary.

Convention	Meaning
italic	Italic type indicates book titles, emphasis, or placeholder variables for which you supply particular values.
`monospace`	Monospace type indicates commands within a paragraph, URLs, code in examples, text that appears on the screen, or text that you enter.

1 Getting Started

This section contains the following topics:

1.1 Overview

Oracle Health Sciences Translational Research Center (TRC) suite comprises of Oracle Health Sciences Cohort Explorer (CE) 3.1 and Oracle Health Sciences Omics Data Bank (ODB) 3.1. TRC enables storing, integrating, controlling, and providing means to analyze clinical and omics data required to support the complete biomarker lifecycle. This includes the data acquisition, discovery, and research as well as clinical use of patient and specimen information.

1.1.1 What You Can Do Using Oracle Health Sciences Translational Research Center

Translational Research Center v3.1 features the following:

- Cohort-driven built-in reports:
  - Demographic statistics reports including age, gender, ethnicity
  - Clinical statistics reports
  - Genomic reports including mutations, copy number variation, drill-into single and dual channel gene expression
  - Genomic reports also include mutation reports as gene vs sample matrix and also specific variants vs sample matrix
  - Structural Variation histograms based on occurrence frequency in genes and (or) gene pairs
  - Genomic report to view the percentage of patients or subjects in a cohort containing genomic data
- Web infrastructure to support collaboration
  - UIs and workflows to enable sharing cohort queries and lists
  - User group creation and maintenance through custom-built UIs

- Usability enhancements
  - Autocomplete and type-in enabled for all searchable concepts
  - Improvements to layouts and interface flexibility
  - Workflow improvements to simplify user experience
  - New icons, images and so on to improve the look and feel
- More than twofold increase in the Cohort Data Model schema
  - Subject study tables and unidirectional link between subject and patient
  - Attributes have been added across multiple concept areas including Observation, Clinical Encounter, Patient and Subject Family, Allergy and so on Histories, Familial Relationship
  - Personally Identifiable attributes such as First Name, Last Name, Contact Information are now supported in the schema with obfuscation governed by individual customer's requirements
- RNA-sequence based querying for cohorts
- Genome Viewers for Variants and Copy Number Variation
- Drill in hierarchy viewer for Diagnosis and Anatomical Site
- Job Scheduling
- Clinical concepts for search: Encounters, Observation types, Sources of data, Coding systems, Familial History and so on
- Gene Set generation through cut-and-paste or import from files
- Genomic assembly version selection for searching reference and result data with the option to select preferred ensembl annotation for the selected assembly version. Supports multiple assembly version selection for most usecases.
- Genomic Data export now has the option to export data from the last-loaded VCF files, when there is duplicate data loaded for a specimen.
- Single patient viewer now supports Dalliance genomic browser from gene and variant
- Variant search can now accept a list of variant IDs as input for a search in various screens
- Whole genome and whole chromosome VCF export is much faster
- Search medication codes using hierarchical drill down
- Define observation query using result reference range
- Context-based values in *Result String* and *Result UoM* search popups in the test and observation search
- Enhancement in Cohort timelines with single line mode option to display same events on the same line with visual separation, and a table to display additional information on the selected event.
- Single patient or subject viewer now has the ability to navigate in cohort through Previous or Next buttons.

1.2 Architecture

TRC consists of the following three tiers:

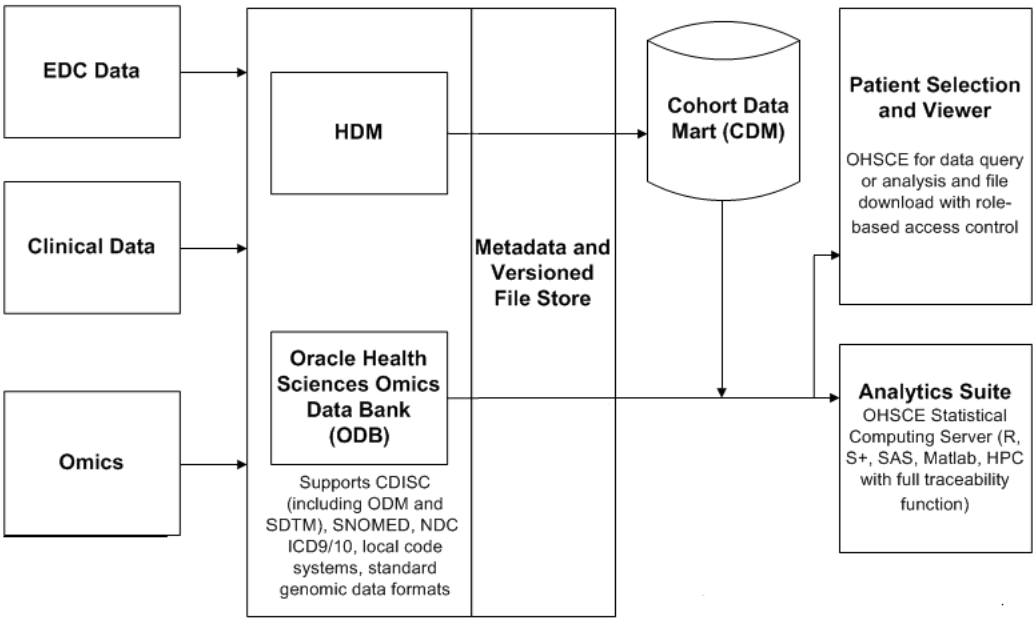
1. Database tier that includes tables and views (to simplify creating certain patterns of queries), indexes, sequences, and PL/SQL packages. PL/SQL packages are of two types:
   - **Utility:** For example, supporting integration between Cohort Data Mart (CDM) and ODB.
   - **Data Movement:** Processes data loaded into staging tables.
2. Client tier comprising of the following functions:
   - Java loaders packaged in jar file for loading of reference data.
   - Shell or batch files facilitating execution of Java loaders.
   - Shell or batch files (one per file type) that facilitate executing PL/SQL based loaders.
3. Middle tier consists of a set of ADF-based UIs deployed into WebLogic Server.

CDM consists of the following two tiers:

1. Database tier that includes tables, views (to simplify creation of certain patterns of queries), indexes, sequences and PL/SQL packages
2. Middle tier consist of set of ADF-based User Interfaces deployed into WebLogic Server. It also contains ETL that brings data from Oracle Healthcare Data Warehouse Foundation (HDWF) into CDM.

Figure 1–1 illustrates the overall data flow in TRC.

Figure 1–1 Data Flow in TRC



The TRC suite of products enables the following functions:

- querying for patient count based on a combination of genomics and clinical attributes. The clinical attributes are available through querying clinical data model, namely CDM as part of OHSCE. The omics attributes are available through

the same interface yet the data will come from Omics Data Bank model, namely ODB

- listing the actual patients that correspond to the obtained count and their clinical profiles
- looking at the timeline view of the clinical history of such patients
- comprehensive reporting on the cohort including cohorts demographics, clinical, genomic features and further drilling into the details on the statistics
- looking at the individual patient's clinical and genomic data together. Also, integration of such APIs as Visquick enables display of omics data in a user-friendly way for bioinformaticians
- exporting selected ODB omics data for patient cohort into standard file formats such as VCF, SEG, RES, which can then be loaded into genome viewers that support such data formats
- downloading genomic files to the desktop through links on the application. This is available only for selected user groups
- searching for genes and variants using a simple search interface that lets the end user perform live queries to find any genomic results that are present in Omics Data Bank model.
- loading data into the Omics Data Bank model from several public reference sources using provided autoloader scripts
- loading your own result data into the ODB using provided autoloader scripts

1.3 Regulatory Compliance

Cohort Explorer is developed with HIPAA regulations in mind. The software enables the customer to easily implement obfuscation rules to protect any patient identifiable medical information. Cohort Explorer development also follows the software development guidelines and requirements for FDA 21 CFR Part 11 compliant software.

1.3.1 Tracking Data

The origin of any data stored in CDM must be traceable to its source, and all transformations applied to the data must be accessible. Data sourced from HDWF is traced by the following criteria:

- **ETL Load**: When data was loaded from HDWF into CDM.
- **ODI Interfaces**: The version of ODI interface used to transform the data from HDWF to CDM, and when the data was executed.
- **Informatica Interfaces**: The version of Informatica interface used to transform the data from HDWF to CDM, and when the data was executed.
- **Configuration Seed Data**: There are two tables which contain seed data. Based on these tables, data is loaded in CDM and another configuration table, which is automatically seeded during ETL load.

Data in Oracle HDWF also keeps audit trails of all modifications.

1.3.2 Managing ETL Versions

You can use a third-party versioning tool or the in-built functionality of ODI and Informatica versioning to manage ETL versions. Currently, all ODI and Informatica objects are in the default version.

1.3.3 Security

Data within the data warehouse is secure from updates by unauthorized personnel and can only be updated through controlled execution of ETL mappings. You can define custom standards in ODI and Informatica to modify and execute ETL routines.

1.4 Known Issues

- If you change the context from patient to subject when no patient ID is present, the demographic text is struck out. You must delete the criteria and add it again.
- When searching for genes in the gene search popup, if no value is provided for a search, no results are displayed.
- When patients (or subjects) are marked as deleted, the function indexes in ODB have to be recreated. When specimens are marked as deleted, the downstream linked tables have to be marked as deleted.
- When you search for a gene (hugo_name), you may also see results for a different hugo_name. This is because of synonyms or aliases present in the gene cross-reference in the ODB data model.
- In the variant search tab, when the context is changed from patient to subject or vice versa, the DNA reference version stays blank and is not loaded with default.
- Specimen Number value (obfuscated and non-obfuscated) is not consistent.
- Description column length in UI for Single Patient page is restricted to 2000 characters.
- Export functionality of Cohort List shows incorrect Start Date and End date.
- The existing query cannot handle genes that have multiple identifiers of the same type.
- The current ETL package does not include ETL for populating the table W_EHA_STUDY_PATIENT_H, which stores the association of Patients and Studies.To associate Patients to Studies, you can develop custom code to populate the above table. Any existing data in W_EHA_STUDY_PATIENT_H is unaffected during an upgrade.
- The drop-down DNA Reference Version is blank or empty while switching context or navigating to Genomic query with Variant Search from other tabs or pages. You can reset the value by changing, or selecting All, from the Assembly version.
- When you search for a gene, pathway or geneset, which does not exist in the database, the message "<name> is not found." is displayed. Immediately after, if you search for a gene, pathway, or geneset which is present in database, the earlier message does not disappear. Ignore the message and proceed with you search, it will not affect your search.
- When you add multiple codes in a single item in Cohort Query (for example, Diagnosis) you may see following error while saving the query:

"An error has occurred. Please contact support with reference 'LOG-1082033954.'"

- There is an extra empty column in Genomic Query > Variant Search report in Google Chrome v 50.0.2661.94 m (64 bit).
- Searching for a whole chromosome (for example, chr1) in Genomic Query can take more than 5min in some cases. Use a specific region for your search (for example, chr1:1-100000) to get a faster response.

1.5 Disclaimer Regarding Third Party Data or Software

1.5.1 Public Domain Data

Oracle makes no express or implied warranty, including but not limited to warranties regarding the accuracy, completeness, merchantability, or fitness for a particular purpose, with respect to third party data loaded into this application or the results of any functions of the application using such data. It may be used for information purposes only, and no medical, clinical or other health related decisions may be based upon such results. You are solely responsible for your use of the third party data, including your right to use the data for your purposes.

1.5.2 Software

Visquick 1.0.2

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the Institute for Systems Biology nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Protovis 3.3.1

Dalliance 0.13.0

ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

2

Logging in to Oracle Health Sciences Translational Research Center

This section contains the following topics:

2.1 User Account

In the current release of Oracle Health Sciences Cohort Explorer (CE), a user's account has permissions to make minimal changes for basic accessibility options, but the user cannot customize the viewing screens.

If Oracle Access Manager (OAM) is configured, the user has the following functions:

1. User logs into the CE application by using Single Sign-On interface which can be shared among multiple applications. For example, if the user also purchased Oracle Business Intelligence Enterprise Edition Plus (OBIEE) full license the same credentials can be used for generating OBIEE reports.
2. Password is created by the user and not visible to an administrator.
3. User is locked out after a configurable number of unsuccessful login attempts.
4. Login session times out after a configurable time of inactivity.
5. Roles are automatically setup as listed in the Roles and Permissions section on page 2.

If OAM is not configured, the identified roles must be manually set up in a Weblogic instance. For more information, refer to Create Users and Add Users to Groups in the *Oracle® Fusion Middleware Oracle WebLogic Server Administration Console Online Help*. In OBIEE, you may need to perform Policy Migration to see the roles. Policy Migration documentation can be found in the *Oracle Health Sciences Translational Research Center Installation and Configuration Guide*.

Note: Due to the caching approach leveraged and in compliance with the licensing agreement, multiple users should not log into the system with the same user name and password at the same time. Doing so may result in inconsistent data.

2.2 Roles and Permissions

A user is assigned one or more of the role groups in the following list:

1. trc-basic-user-group
   - Able to view and execute OBIEE-based reports in the Dashboard page but is unable to modify individual reports.
   - Can access the following screens:
     - Dashboard: Standard Reports
     - Dashboard: Custom Reports
     - Dashboard (top level tab)
2. trc-bioinformatician-group
   - Can download files from the web interface. Files must have a link stored in ODB and can be located in the middle tier in an accessible location.
   - Can access the following screens:
     - Dashboard: Standard Reports
     - Dashboard: Custom Reports
     - Dashboard (top level tab)
     - Cohort Query Tab
     - Cohort Query Tab: Genomic Data (tab in accordion)
     - Query Tab: Relative Time Events: gene variant (radio button)
     - Cohort Viewer (top tab)
     - Cohort Viewer: Cohort List
     - Cohort Viewer: Cohort Timeline
     - Cohort Reports
     - Cohort Viewer: Genomic Data Export
     - Single Patient Viewer: View Record
     - Single Patient Viewer: View Record: Genomic Data Collected
     - Circular Genomic Viewer (Visquick)
     - Genomic Query
     - Genomic Query: Columns after searching for gene/variant coming from CDM (Patient Count, Specimen Anatomical Site columns)
     - Genomic Query - Gene/Variant Search-Result summary Export VIA GENE/VARIANT SEARCH
     - My Workspace: My Recent Queries
     - Short Cuts
     - My Cohort Lists
     - Queries or Lists shared with me
     - Queries or Lists shared with All
     - My Queries

  - Gene Sets
  - Manage Queries
  - Manage Gene Sets
  - Manage Cohort List
  - Jobs

3. trc-cohort-group
   - Can query any data from CDM by using any of the CE User Interfaces. This group is unable to query ODB directly.
   - Can export the clinical data in any supported format, and is able to view the Dashboard.
   - Can access the following screens:
     - Dashboard: Standard Reports
     - Dashboard: Custom Reports
     - Dashboard (top level tab)
     - Cohort Query Tab
     - Cohort Viewer (top tab)
     - Cohort Viewer: Cohort List
     - Cohort Viewer: Cohort Timeline
     - Cohort Reports (Only Clinical)
     - Single Patient Viewer: View Record
     - Circular Genomic Viewer (Visquick)
     - My Workspace: My Recent Queries
     - Short Cuts
     - My Cohort Lists
     - Queries or Lists shared with me
     - Queries or Lists shared with All
     - My Queries
     - Manage Queries
     - Manage Cohort List
4. trc-limited-user-group
   - Can only see the Query Patients page (that is the count of patients) but is unable to view other detailed data.
   - Can access the following screens:
     - Dashboard: Standard Reports
     - Dashboard: Custom Reports
     - Dashboard (top level tab)
     - Cohort Query Tab (**Save Query** button should not be accessible)
5. trc-omics-group

- Can query and read data from ODB by using User Interfaces and components marked as licensed with ODB only.
- Can export omics data in file formats readable by genome viewers. For example, VCF, SEG, RES formats for IGV.
- Although unlikely, if the user has only this role as standalone, you can only view from My Workspace page the Recent Gene Sets report and Omics Explorer and Gene Sets User Interfaces. All other CE bound components of the User Interface is grayed out.
- Can access the following screens:
  - Cohort Viewer: Genomic Data Export
  - Genomic Query
  - Short Cuts
  - Gene Sets
  - Manage Gene Sets
  - Jobs

6. trc-comics-limited-user-group
   - Can query data from CDM and ODB using Query Patients interface, however cannot access Patient Viewer User Interfaces except Patient Genomic Data export
   - Can access MyWorkspace, view queries and gene sets
   - Can create his own Gene Sets
   - Can access the following screens:
     - Cohort Query Tab (**Save Query** button should not be accessible)
     - Cohort Query Tab: Genomic Data (tab in accordion)
     - Query Tab: Relative Time Events: gene variant (radio button)
     - Cohort Viewer (top tab)
     - Cohort Viewer: Genomic Data Export
     - My Workspace: My Recent Queries
     - Short Cuts
     - My Queries
     - Gene Sets
     - Manage Queries
     - Manage Gene Sets
     - Jobs
7. trc-pi-user-group
   - Has specific privileges allowing access to identifiable information on patients or subjects.
   - Can view personally identifiable (PI) data in Cohort List page. Together with user's VPD profile, this role enables the user to have access to PI attributes from the UI

- Can view personally identifiable data in Single Patient or Subject Viewer > View Record page.
- When in Subject context, View Record page allows this user to see merged version of both Patient and Subject clinical data all in one page.
- Can access the following screens:
  - Dashboard: Standard Reports
  - Dashboard: Custom Reports
  - Dashboard (top level tab)
  - Cohort Query Tab
  - Cohort Query Tab: Genomic Data (tab in accordion)
  - Query Tab: Relative Time Events: gene variant (radio button)
  - Cohort Viewer (top tab)
  - Cohort Viewer: Cohort List
  - Cohort List (PI attributes)
  - Cohort Viewer: Cohort Timeline
  - Cohort Reports
  - Cohort Viewer: Genomic Data Export
  - Single Patient Viewer: View Record
  - Single Patient Viewer: View Record (PI atrributes)
  - Single Subject Viewer: View Record (PI atrributes)
  - Single Subject Viewer: View Record (show patient data together)
  - Single Patient Viewer: View Record: Genomic Data Collected
  - Circular Genomic Viewer (Visquick)
  - Genomic Query
  - Genomic Query: Columns after searching for gene/variant coming from CDM (Patient Count, Specimen Anatomical Site columns)
  - Genomic Query - Gene/Variant Search-Result summary Export VIA GENE/VARIANT SEARCH
  - My Workspace: My Recent Queries
  - Short Cuts
  - My Cohort Lists
  - Queries or Lists shared with me
  - Queries or Lists shared with All
  - My Queries
  - Gene Sets
  - Manage Queries
  - Manage Gene Sets
  - Manage Cohort List

  - Jobs

8. trc-standard-report-group
   - Has access to Dashboard tab in Cohort Explorer and more specifically Standard Report subtab and all reports it contains
   - Can access the following screens:
     - Dashboard: Standard Reports
     - Dashboard (top level tab)
9. trc-custom-report-group
   - Has access to Dashboard tab and Custom Reports subtab in Cohort Explorer and can view all custom reports as selected by customer.
   - Can access the following screens:
     - Dashboard: Custom Reports
     - Dashboard (top level tab)
10. trc-admin-group
    - Has access to Navigator option Manage User Groups in Cohort Explorer
    - Can create new user groups etc for sharing cohort queries or lists
    - Can add or remove users from groups
    - Can manage user roles
    - Can access the following screens:
      - Manage User Group

If users only belong to the following groups, then limited functionality is available based on the assigned roles:

- trc-comics-limited-user-group
- trc-limited-user-group
- trc-standard-report-group
- trc-custom-report-group
- trc-basic-user-group

Users with the specific OBIEE license bundled with TRC have the roles listed in the previously configured in OBIEE. Additionally, if full OBIEE license is purchased, the following additional roles can be granted which are specific to OBIEE:

1. trc-basic-user-group
   - Able to view and execute OBIEE-based reports in the Dashboard page but cannot modify individual reports.
2. trc-developer
   - Can modify the OBIEE Web Catalog.
   - To publish the report, trc-Administrator should place it in a shared directory and assign appropriate access permissions.

If only a limited OBIEE licensed is purchased, neither the trc-developer nor the trc-administrator role is available. Without these roles, users in groups 1-5 are unable to build new reports or modify the content of existing reports. With a full OBIEE

license, users with the trc-developer role can build reports but the trc-administrator role is required to publish them.

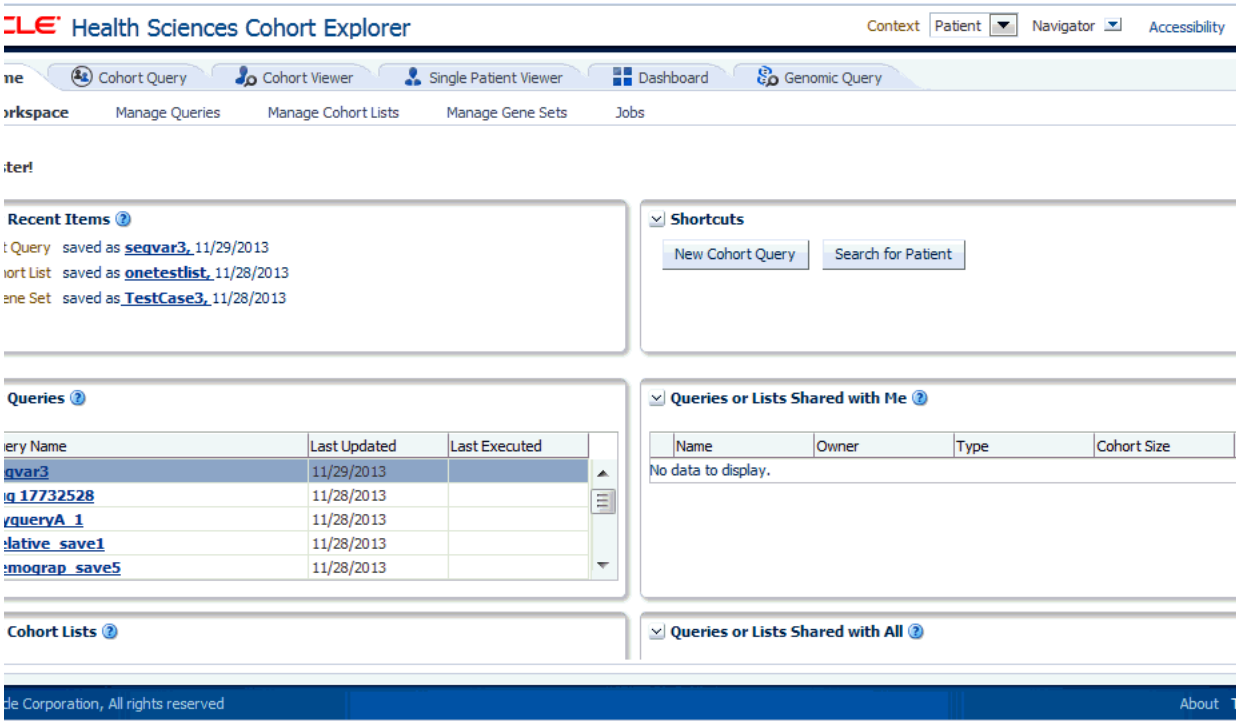
2.3 Patient vs Subject Context

> **Note:** For optimal display of TRC, ensure that your screen resolution is at least 1024 pixels.

Cohort Explorer application can be switched to run in either Patient or Study Subject context. In patient context, all queries are directed at patient tables in Cohort Data Model schema, while in the study subject context, the UI Query Engine queries subject tables in CDM.

By default, the Cohort Explorer context is set to Patient. The context can be set in the upper right hand corner of the screen as shown in the following figure.

Figure 2–1 Setting Context



When user switches context, the next login session will remember the last context set and keep it until it is changed again by the end user. Switching context will clear out loaded data in most of the viewer pages, however, it will not clear out criteria in Cohort Query page. Any criteria that are applicable in either context will be kept intact and criteria that are not applicable will be disabled. For example, Encounter criteria is not supported in Subject context thus when switching into Subject context, the criteria is disabled. If user switches back to Patient context, the criteria will be enabled again. If user elects to save the query, any disabled criteria will not be saved in the query. Also, each saved query is saved along with the context it was generated against. Queries in My Workspace page are displayed based on context they were generated

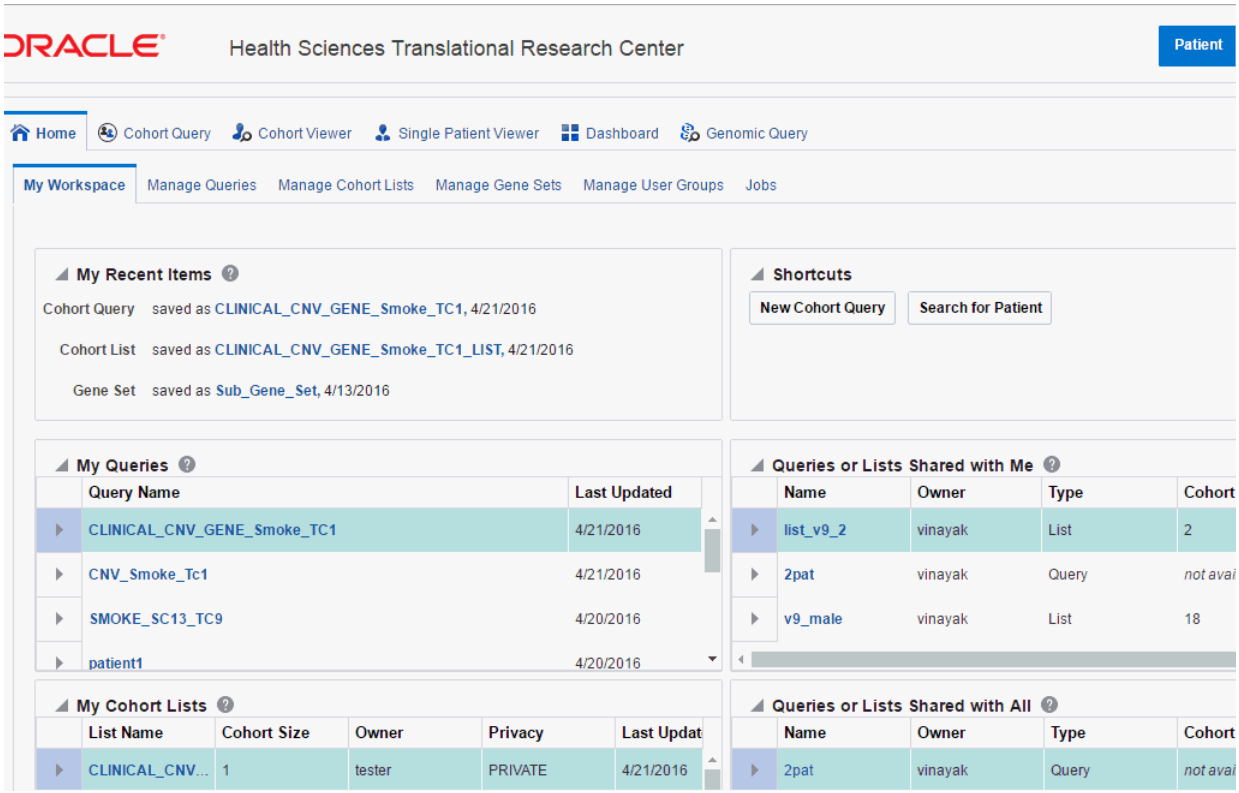
against so in Patient context, only Patient context saved queries should be shown, and similarly, in Subject context, My Workspace would only show Subject context queries.

For a Subject to be considered deleted and not visible for selection, a given Subject's delete flag in CDM should be set to *Y*. Otherwise if a Study is deleted, it will not cause any of the Subjects in this study to be deleted. Those subjects will still be visible, while the study will no longer be available for selection. To maintain data integrity, the Subject's delete flag should be set to *yes* alongside the deleted Study.

2.4 Landing Page

The landing page consists of all the options required to complete your clinical and omics tasks, and consists of 6 main tabs with multiple subtabs under each.

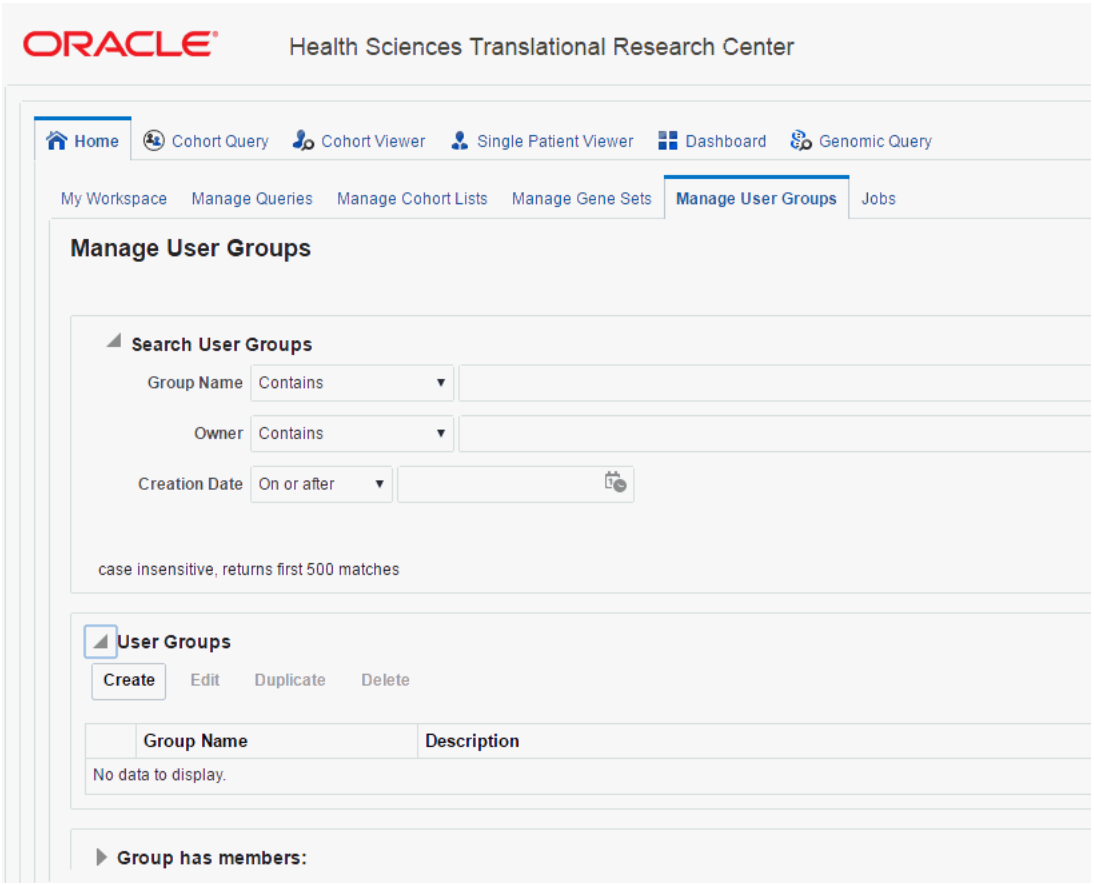
Figure 2–2 Landing Page



2.4.1 Manage User Groups

Manage User Groups, available under the Home tab, is a set of utility user interfaces, custom built for users of Cohort Explorer in order to enable collaboration. Any user can create user groups and assign users to those groups in order to simplify sharing. Instead of sharing queries or lists with each user individually, setting up a list of user makes it easier for end user to share multiple items with a group of individuals. Note that only the user that is the owner of a given list can modify members of a given list. However, any user can elect to utilize a given list for sharing and see members of a list, even if he or she is not the owner.

Figure 2–3 Manage User Groups



3

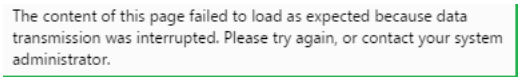
Cohort Explorer

This chapter contains the following topics:

Note: If the Oracle Health Sciences Cohort Explorer user runs into issues resulting in Internet Explorer 11 (IE 11) incompatibility, refer to the *Oracle® Health Sciences Translational Research Center Installation Guide* for upgrade patch instructions to resolve them.

Note: All fields marked with a magnifier glass icon support both autocomplete or manual search. The autocomplete feature is used when you enter a value in the text box, and the manual search is used when you click the magnifier glass icon.

If you enter a value in the text box, the system will have to return the result from the autocomplete. If you click the magnifier glass icon while the system is processing the autocomplete search, the system will throw the following error. To prevent the error, wait for the autocomplete results to return before clicking on the search magnifier glass icon.

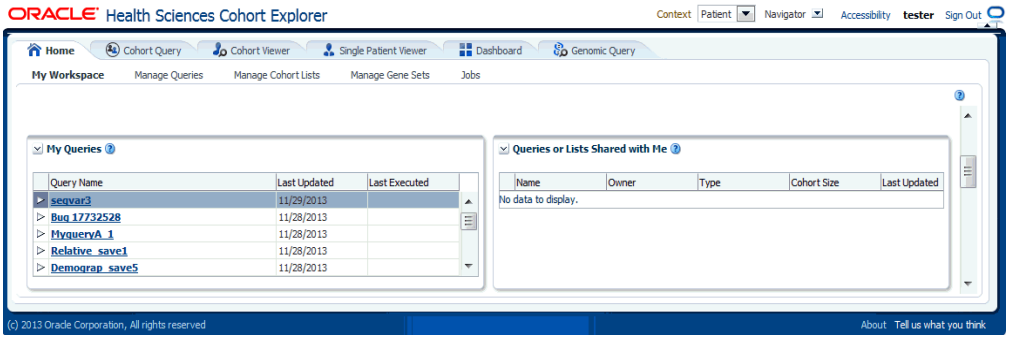


3.1 My Workspace

The My Workspace page is the default page for Oracle Health Sciences Cohort Explorer (CE) and Oracle Health Sciences Omics Data Bank (ODB) user. Depending on the license purchased, you may have access to most or all of the reports on this web page. With a standalone CE license, you can only view the Cohort Queries and Cohort

Lists details. If an ODB license is also available, the My Gene Sets details are also displayed.

Figure 3–1 My Workspace Screen



3.1.1 My Recent Items

This lists the most recent saved cohort query, cohort list, and gene set, which provides the quick access to the recent work.

3.1.2 My Queries

This lists the most recent saved patient cohort queries, sorted by the last updated date. Clicking the selected query will load it in the Cohort Query window.

3.1.3 My Cohort List

This lists the most recently saved cohort lists sorted by the last updated date. Clicking the selected list will load it in the Cohort Viewer/Cohort List window.

3.1.4 My Gene Sets

The My Gene Sets report lists the most recently updated gene sets. Clicking the selected gene sets will load it in the Manage Gene Sets/Manage window.

3.1.5 Shortcuts

This offers shortcuts to certain functionalities. Clicking **New Cohort Query** displays the Cohort Query Window with no criteria loaded. Clicking **Search for Patient/Subject** opens the View Record window.

3.1.6 Queries or Lists Shared with Me

This lists all queries and lists that are shared with you. Clicking the selected query will load it in the Cohort Query window. Clicking the selected list will load it in the Cohort Viewer/Cohort List window.

3.1.7 Queries or Lists Shared with All

This lists all public queries and lists. Clicking the selected query will load it in the Cohort Query window. Clicking the selected list will load it in the Cohort Viewer/Cohort List window.

3.2 Manage Queries

This section helps you search through queries available for a specific user using various options like Query Name, privacy, creation date and last updated. You can view a list of queries, which has either created by or has been shared with you. Once you see a list of queries after searching, you can perform Load, Edit, Delete operations. You can also make accessibility changes to the queries like Private, Shared with All and Shared.

3.2.1 My Queries

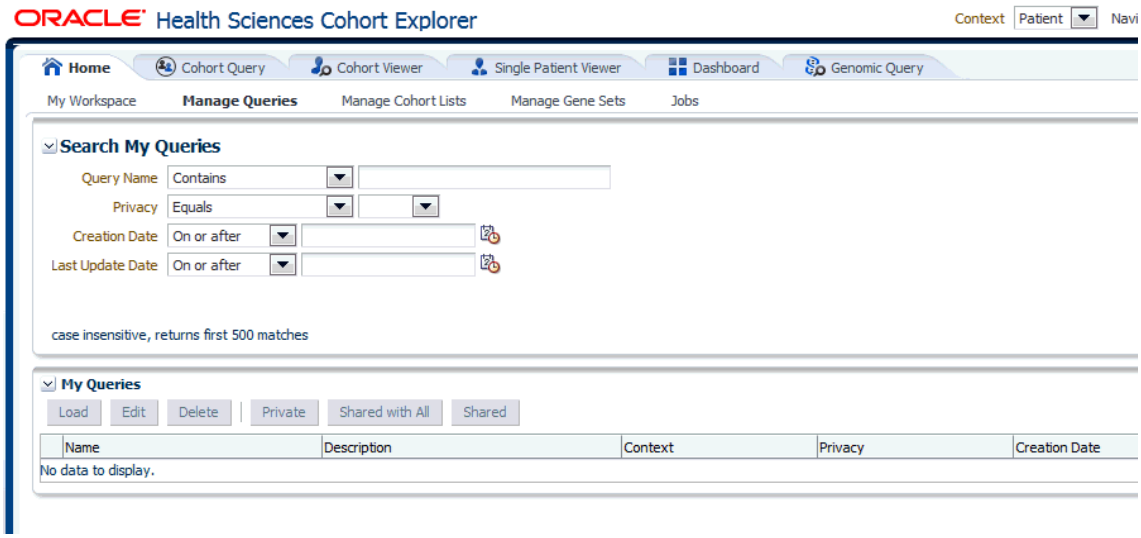
This section lists all the saved queries that you have access. You can filter the queries based on the search criteria given in the Search My Queries section. The query list displayed is irrespective of patient or subject context. You can load a query into the Cohort Query screen, edit a query, and delete a query.

3.2.2 Sharing Queries

If a shared query is selected, a list of users and groups is displayed in the table with whom it is shared.

While loading a query, if any criteria are already present in the Cohort Query section, it prompts you to append to the existing criteria or to clear and load the query. In case of different context, it prompts you to switch to the other context and load.

Figure 3–2 Manage Queries



3.3 Manage Cohort Lists

This section helps you search through a saved cohort list available for a specific user using various options like List Name, privacy, creation date and last updated. You can view a list of cohorts, which has either been created by or has been shared with you. Once you see a list of cohorts after searching, you can perform Load, Edit, Delete operations. You can also make accessibility changes to the queries like Private, Shared with All and Shared.

3.3.1 My Cohort Lists

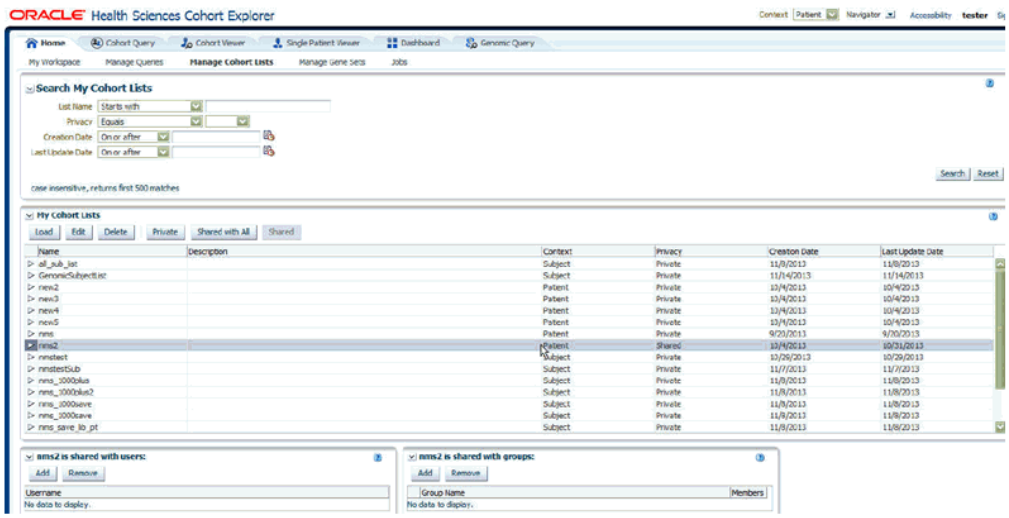
This section displays all the patient or subject lists that you created. You can filter the lists based on the search criteria given in the Search My Cohort Lists section. The list is irrespective of patient or subject context, as the context is shown as one of the columns. You can load a list into the Cohort List screen, edit a list, and delete a list.

3.3.2 Sharing Cohort Lists

If a shared list is selected, a list of users and groups will be shown below the table with whom it is shared.

While loading a list, if any list is already present in the Cohort List section, it prompts you to clear and load the list. In case of different context, it prompts you to switch to the other context and load.

Figure 3–3 Manage Cohort Lists



3.4 Manage Gene Sets

Gene Set is intended to be used as a way to collect genes into groupings or lists. Frequently, there will be a list of genes you work with regularly. A set of such genes might be as small as a couple of genes or large, consisting of hundreds of genes. Often you may keep several such sets, each characterizing a group of genes with particular attributes, for example, transcription factors, genes involved in some regulation mechanism, genes that have been implicated to contribute to a particular characteristic and so on. The concept of Gene Set lets you group genes into convenient *collections* for reuse.

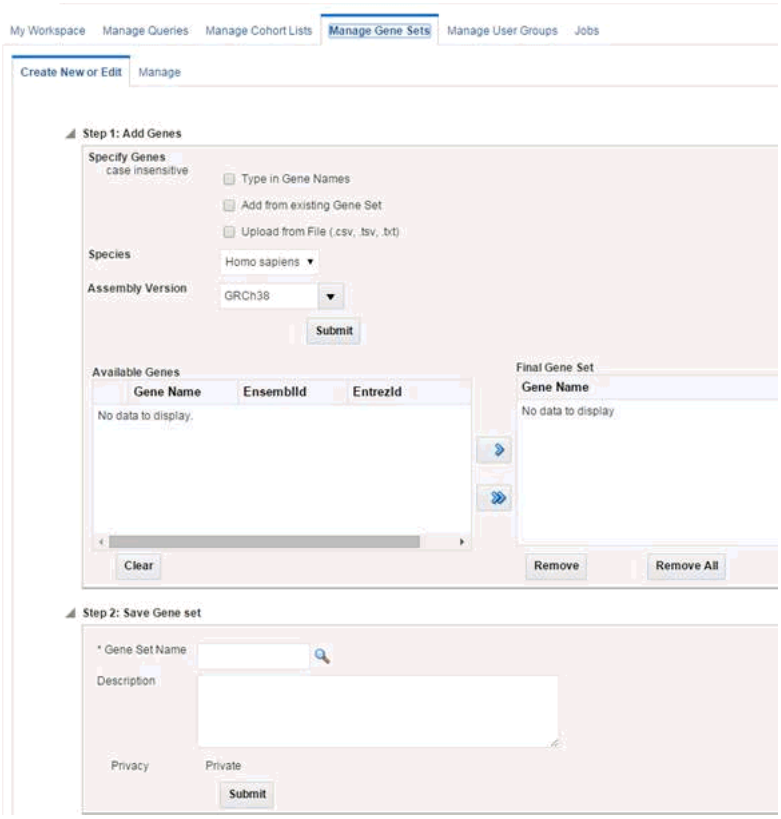
3.4.1 Create New or Edit Gene Set

The Create New or Edit gene sets is a web interface that helps you group genes as a set. You can elect to group a few genes for quick search retrieval or for use in a cohort query. For example, if you have a set of 10 genes that you plan to work with or always search for results based on genes from within this particular array of 10, you can create a new Gene Set to collect these 10 genes into one group. A gene can be part of many different Gene Sets. Furthermore, you can create many different gene sets, each gene set with a different combinations of genes.

Note:

1. There are no restrictions regarding which genes can be included in any particular Gene Set. You may choose to mix genes from multiple species, or assembly versions as the system does not enforce any such constraints.
2. Gene Sets currently are Private only and cannot be shared among users.
3. Gene Set names are not case- sensitive.
4. The limit for cut-and-paste option is 512 characters when genes are compared using Contains or Starts With option.
5. The file size limit for 'upload file' option is less than 5MB when genes are matched using Equals. When using Contains or Starts With option, the list of genes cannot be longer than 512 characters.

Figure 3–4 Create New or Edit Gene Set



The Create New or Edit Gene Set screen is used for the following:

1. Creating a new Gene Set or
2. Editing an existing Gene Set

From this area, you can choose to:

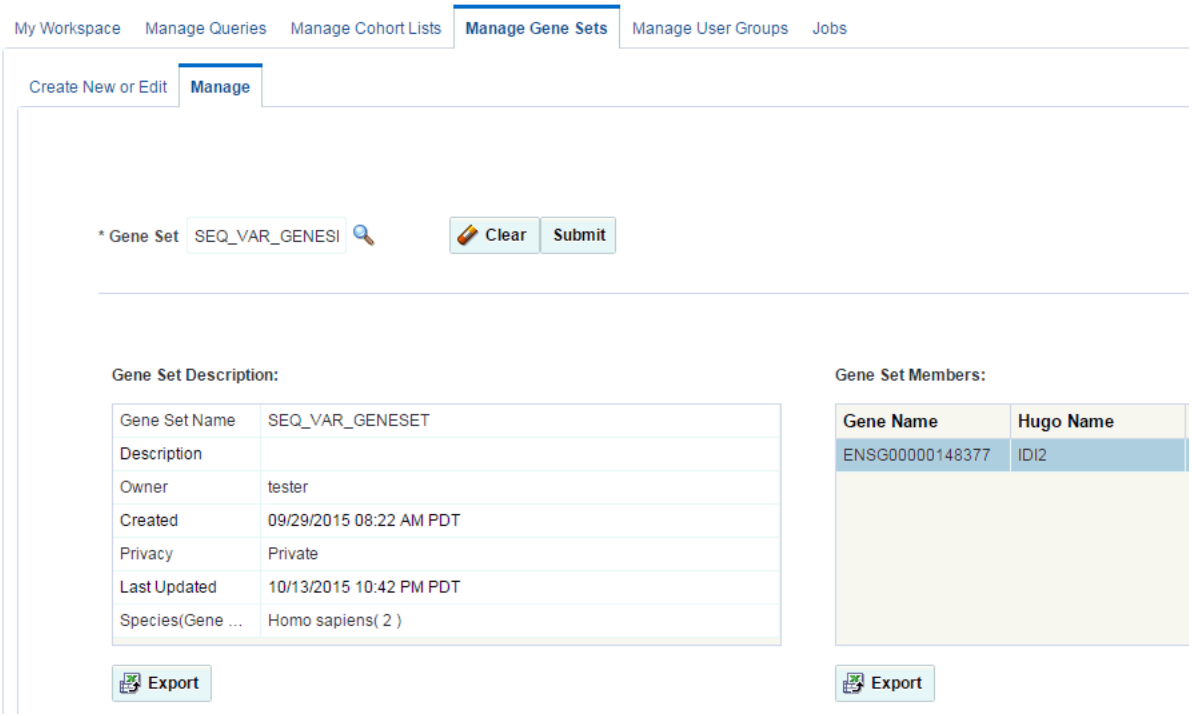
- Search for genes to be included in the gene set in ODB by using their Ensembl or HUGO names. You can enter multiple gene names separated by space, comma or semicolon. Additionally, you must specify the species and Assembly version that a given gene is linked to. Once you click **Submit**, the matching genes appear on the left hand side pane. To add a gene to the Gene Set, use the arrow button to move the gene from the left panel to the right hand side.

- Search for genes based on an existing Gene Set. First, select the **Gene Set** to add genes from, then click **Submit**. The genes from the existing Gene Set appear in the left hand side panel.
- Upload genes from a file. Select a text file from your desktop where the genes are delimited by comma, space or tab. Then click **Submit** and the genes that match genes in ODB will be loaded into left hand side panel.
- You can select to move all genes from the left hand side panel to the right hand side panel. The genes in right hand side panel are those that will comprise a gene set.

Once you are satisfied with the selection of genes, you can either preserve the selection as a new Genet Set or save over an existing one. You can give the Gene Set a name and Description, and click **Submit**. The system saves the new Gene Set and a confirmation box appears.

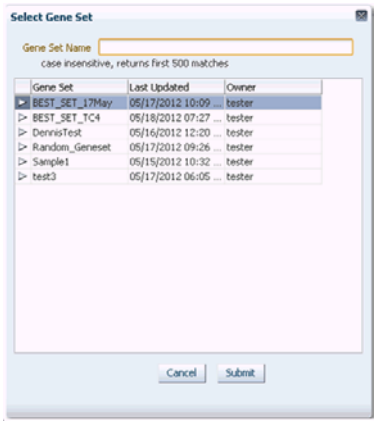
3.4.2 Manage Gene Set

Figure 3–5 Manage Gene Set



The Manage Gene Sets screen enables you to view the individual genes (members) included in a given Gene Set as well as the set's metadata. You can select a specific Gene Set to view by using its name and click **Submit** to view the Gene Set's detailed information.

Figure 3–6 Select Gene Set



Each Gene Set's metadata includes the name, description given by the user, owner of the gene set, privacy setting (currently all gene sets are Private only), the creation and update dates and summary of how many genes are in a gene set.

You can also export or print any of the Gene Set's data.

Finally, you can delete a Gene Set using the **Delete** button. A confirmation box appears to ensure you are performing the proper action.

> **Note:** Any Gene Set deletion is a soft-delete. A soft delete means that the actual Gene Set record is still in the database, but it can only be restored with the help of a System Administrator.

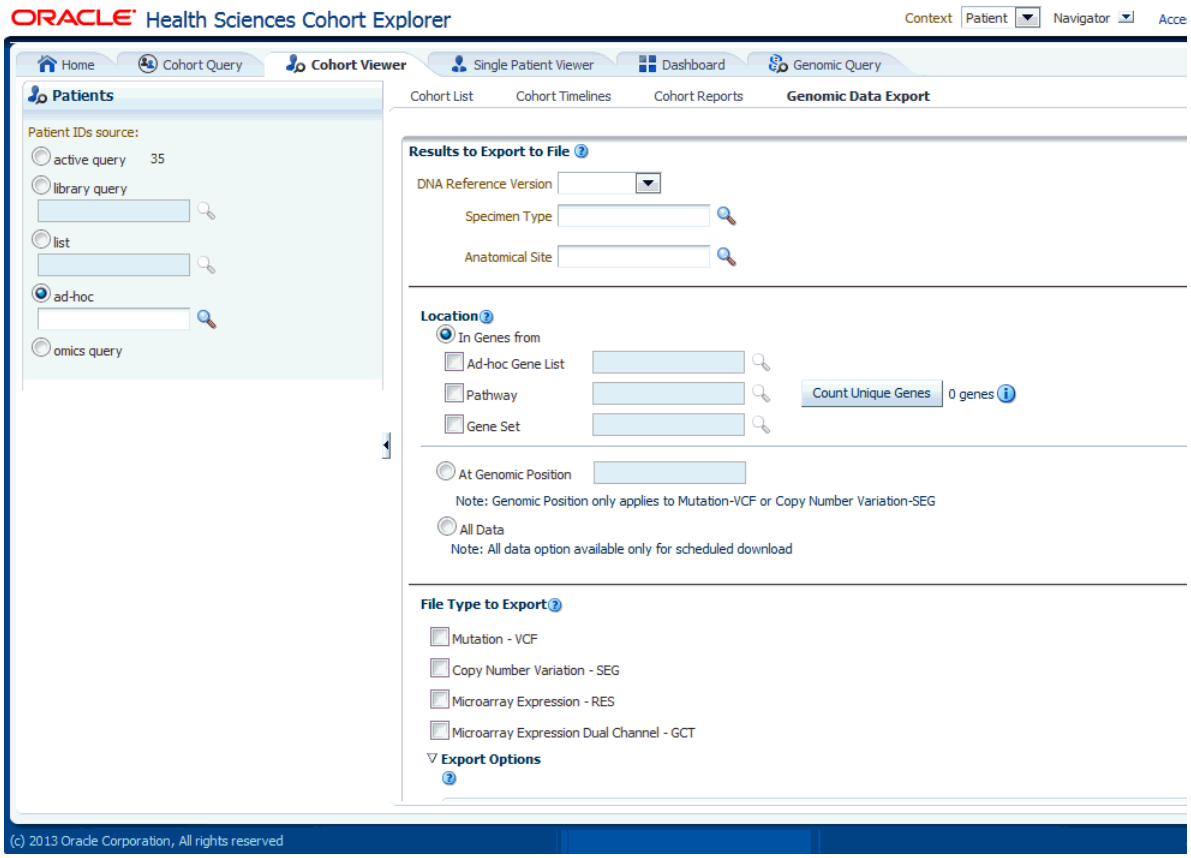
3.5 Jobs

Currently, jobs are scheduled only for Genomic Data Export. To schedule a job, navigate to **Genomic Data Export** screen under the Cohort Viewer tab.

> **Note:** Any job executed prior to applying TRC 3.1 will have its status changed back to 'Scheduled'.

The following screen demonstrates how to create and schedule the job.

Figure 3–7 Create and Schedule Job



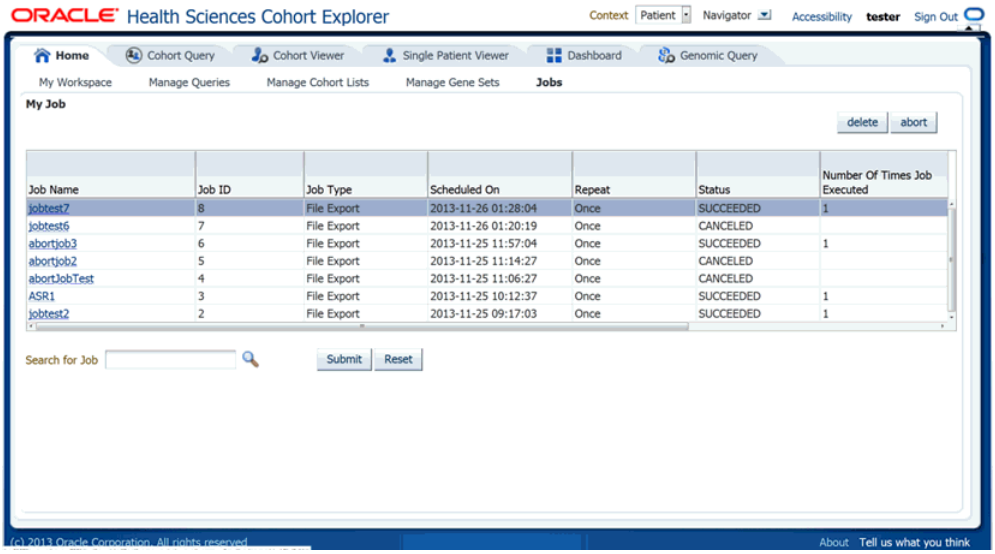
To create and schedule a job, perform the following steps:

1. Select the patient IDs source from one of the following options:
   - active query
   - library query
   - list
   - ad-hoc
   - Omics query
2. Select the DNA reference version.
3. Select the location for the gene from the following options:
   - Ad-hoc Gene List
   - Pathway
   - Gene Set
4. Select one of the following file formats to export:
   - Mutation-VCF
   - Copy Number Variation-SEG
   - Microarray Expression-RES
   - Microarray Expression Dual Channel-GCT

5. Select the **Schedule** export option. This prompts for the job name and description.
6. Click **Submit**. The job ID is created.

To view the information of this job and its progress, navigate to My Workspace, and click the **Jobs** tab.

Figure 3–8 Job Lists



Only the authorized user can see the job lists. The following table describes the columns and buttons in the My Job page.

Table 3–1 Columns and Buttons in My Jobs Page

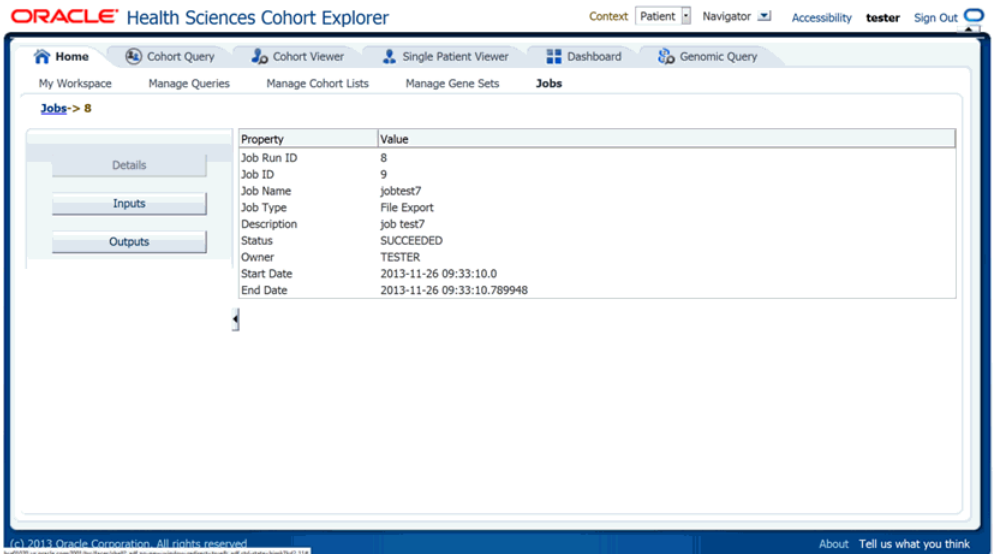
Column Name	Description
Job name	Job name specified when scheduling the job.
Job ID	Job ID generated while scheduling the job.
Job type	Only File Export job type is supported.
Scheduled on	Date and time on which the job is scheduled. The format is yyyy-MM-dd HH:mm:ss.
Repeat	You can schedule the job to run only once or recurrent.
Status	You can set the status to succeeded, failed, or canceled.
No. of times job executed	Count of how many times the job is run.
First execution started	Date and time when the job is scheduled to run. The format is yyyy-MM-dd HH:mm:ss.
Last execution successfully completed	Date and time when the job is last executed. The format is yyyy-MM-dd HH:mm:ss.
Search	You can search job by name or job ID.
Reset	You can reset your search criteria.
Delete	You can delete a job. A confirmation box appears to make sure you are performing the proper action. This is a soft delete, which means the job is still in the database but only the delete flag set to **Y**.
Abort	You can abort a job before it runs. You can also abort a queued or scheduled job but cannot abort a succeeded or canceled job.

3.5.1 Job Details

This section provides complete information of a job.

To view the job details, click the **Jobs** tab and click **Details**.

Figure 3–9 Job Details



The following table describes job details.

Table 3–2 Job Details

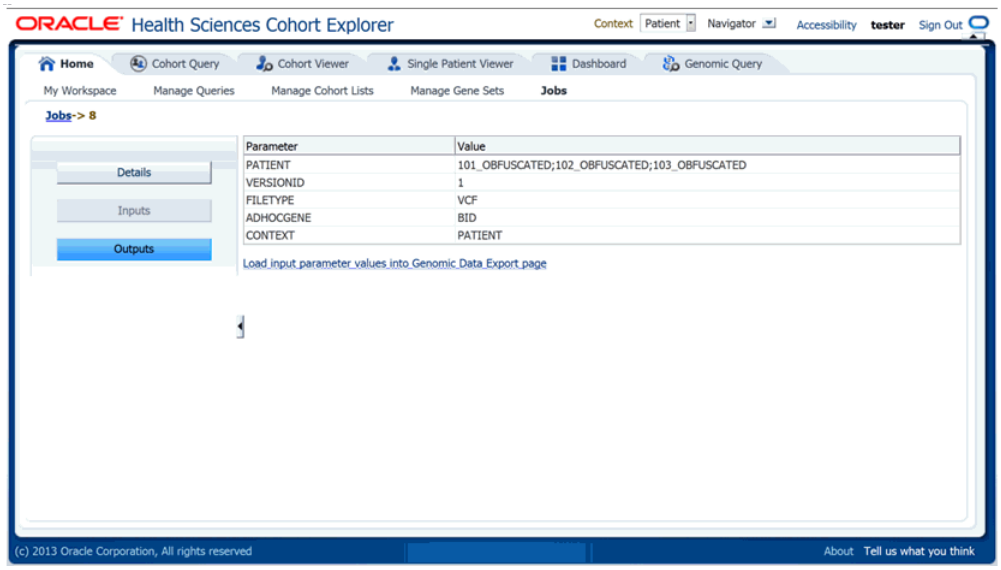
Column Name	Description
Job run ID	Job instance ID generated when job is scheduled.
Job ID	Job ID generated while scheduling the job.
Job name	Name of the job.
Job type	Only File Export job type is supported.
Description	Description of the job.
Status	Status of the job (succeeded, failed, or canceled).
Owner	User who created the job.
Start date	Date and time the job is scheduled to run. The format is yyyy-MM-dd HH:mm:ss.
End date	Date and time when the job is last executed. The format is yyyy-MM-dd HH:mm:ss.

3.5.2 Job Inputs

This section provides the details of parameters provided when scheduling a job.

To view the details of the job created, click the **Jobs** tab and click **Inputs**.

Figure 3–10 Job Inputs



The **Load input parameter values into Genomic Data Export page** link directs you to the genomic data screen with these values loaded. This occurs only when the job context matches the applications context. If not, the following note is displayed in place of link:

**Note: Selected Job cannot be loaded as its context is different from application context.

3.5.3 Job Outputs

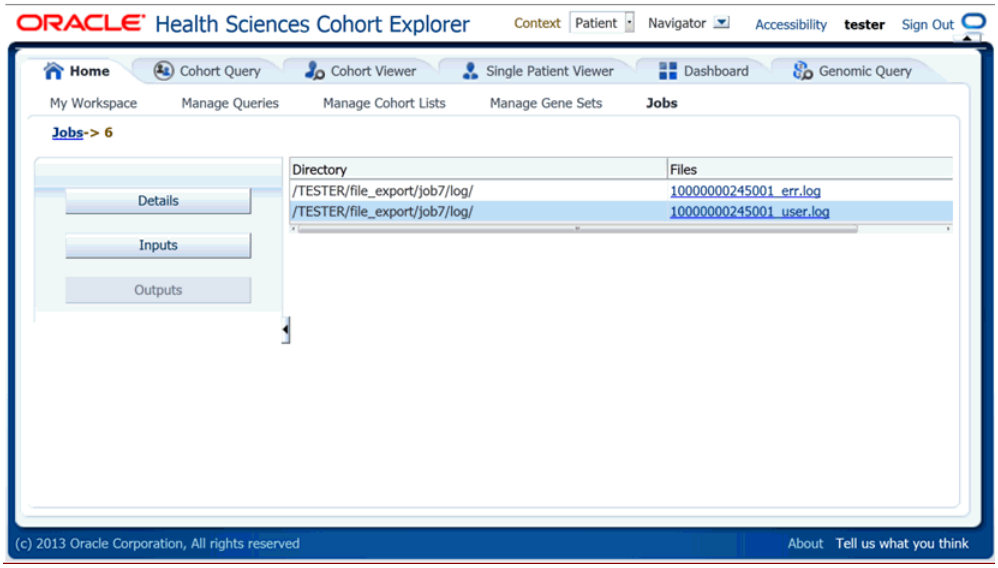
This section shows the result files of a job. The following files are generated:

- Error log file - contains the log file of the application if any exception or error occurred while running the job.
- User log file - contains the log file, which shows the error occurred while generating the export files.
- Admin log file - is for DB administrator.

You can download error and user log files but cannot download the admin log file.

Note: When you try to export a file that does not have any Patients or Subjects linked to the specimen, an empty file is generated in Schedule mode.

Figure 3–11 Job Outputs



3.6 Manage User Groups

Manage User Groups is a set of custom-built utility user interfaces that enable collaboration. You can simplify sharing, by creating user groups and assigning users to those groups. You can also conveniently share multiple items with a group of individuals by setting up a list of users. Note that

---

Note: Only the user who is the owner of a given list can modify its members. However, any user can utilize a given list for sharing and view the members of a list.

4

Cohort Query

This chapter contains the following topics:

4.1 Overview

Cohort Explorer in v3.1 contains numerous user interface enhancement to improve search for patients or subjects based on the context selection. The tools are organized within the **Cohort Query** tab, which is the second tab from left on the top row. In this tab, you create, run, and save queries that select subsets of patients or subjects from your patient or subject database.

You have a wide array of topics, each with unique data elements that you can select or specify to focus on the particular subset of patients or subjects you are looking for. Each time you specify criteria for a particular topic, CE recognizes your selection as a discreet statement and preserves the definition as part of the query. You can create as many criteria statements as you want for a single query to identify the patient or subject count or list according to your requirement.

Along with defining data criteria (for example, Gender = Male), you also configure the logic of a particular criteria statement to be either inclusion or exclusion. For each query you create (or view), the entire definition displays in the main window where you can make changes and re-run the query in an iterative way, to view the impact of your query design in real-time.

4.2 Cohort Criteria Selection

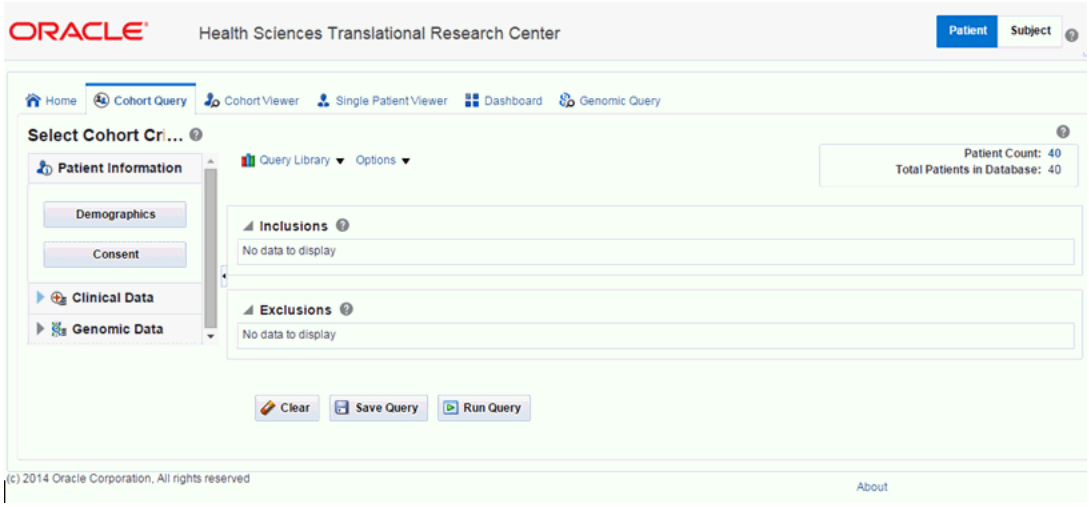
The query criteria are grouped into three broad data categories—Patient information, Clinical data, and Genomic data. Each category is further defined with multiple topics where you can drill into the specific parameters for that particular topic. The categories and topics are summarized in the following sections:

Table 4–1 Cohort Selection Criteria

Category	Topic
PATIENT INFORMATION	-
-	Demographics
-	Consent
CLINICAL DATA	-
-	Diagnosis
-	Clinical Encounter
-	Procedure
-	Medications
-	Patient History
-	Test or Observation
-	Specimen
-	Study
-	Relative Time Events
GENOMIC DATA	-
-	Sequence Variants
-	Copy Number Variation
-	Microarray Expression
	RNA-seq Expression

To display the topics for a particular category, select the arrow to the left of that category name. The left hand column slides and reveals the available topic areas for that category.

Figure 4–1 Inclusion and Exclusions

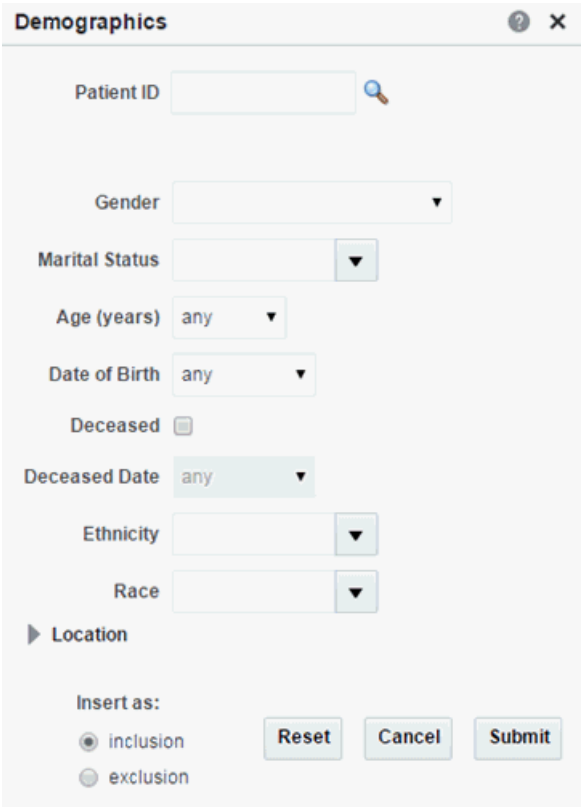


To add criteria to a query, select the topic that has the particular type of data you want to use for your definition. Each topic provides unique drop-down lists, searchable code lists or other parameters that are appropriate for that particular topic.

For example, within the **Demographics** tab you can specify the count of patients with a particular gender, age range or who live in a particular location. In contrast, the **Diagnosis** tab enables you to specify that you want patients with a specific diagnosis (by name or by code), which may have a particular onset date or date range.To specify Female patients between the ages of 40 and 45

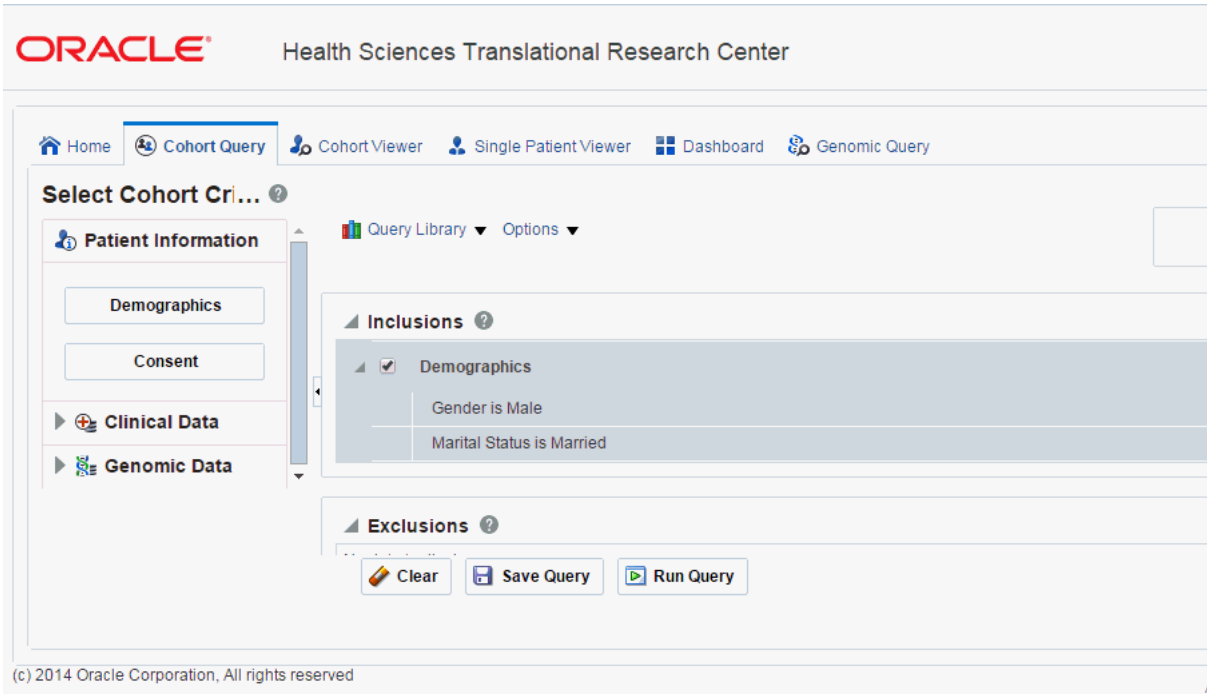
1. Click the arrow next to **Patient Information** to display the **Demographics** tab.
2. Click **Demographics.**
3. Select **Female** from the **Gender** drop-down list.
4. Enter the range in years in **Age (years).**
5. In the **Insert as** group, select **inclusion** or **exclusion** to confirm whether the definition is to be included or excluded.
6. Click **Submit.**

Figure 4–2 Demographics



At this point, the criteria you specified appears in its query statement in the appropriate section of the Query Patients page.

Figure 4–3 Query Statement

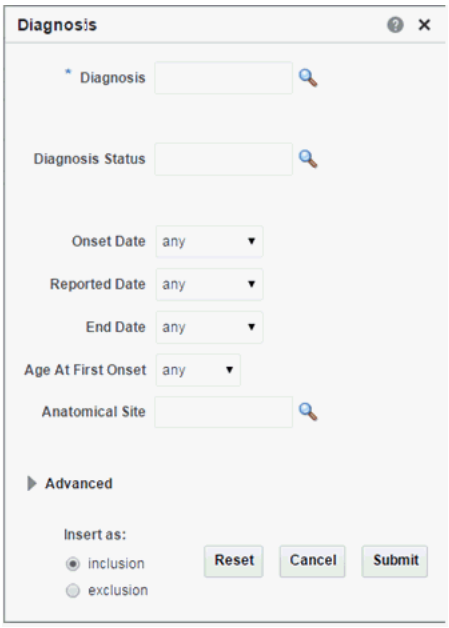


Each time you select your criteria, CE adds the definition statement to the display on the right. You can add as many statements of criteria to your Query as you want and you can see the query definition expand, each time you add a new statement. When you run the query, CE considers all the criteria rows in combination.

4.2.1 Required Fields for Criteria

Most of the criteria prompts include required fields. These are identified with an asterisk. For example, you are required to indicate one or more Diagnoses for a patient along with a date or onset or other parameters.

Figure 4–4 Diagnosis



If you submit a criteria statement without addressing a required field, the system prompts you to complete it.

The criteria selection topics are summarized in the following:

4.2.2 Patient Information or Subject Information

4.2.2.1 Demographics

Figure 4–5 Patient Information Demographics options

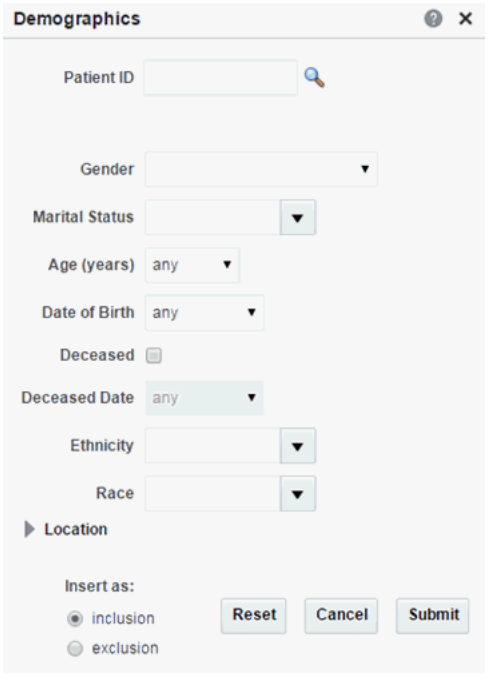


Table 4–2 Demographics Details

Field Name	Definition	Sample Value or Values
Patient ID or Subject ID	Double-blinded unique identifier for the Patient. Also known as the Oracle ID	-
Gender	Patient's gender	Male, Female
Marital Status	Patient's marital status	Married, Single, Separated, Divorced
Age (in Years)	Patient's chronological age	-
Date of Birth	Patient's date of birth	-
Deceased Date	Patient's decease date	-
Ethnicity	Code to reflect Patient's ethnicity	-
Race	Code to reflect Patient's race	-
City	Code for Patient's City of residence	-
State	Code for Patient's State of residence	-
Zip Code	Code for Patient's Zip code	-
County	Code for Patient's County	-
Country	Code for Patient's Country	-

4.2.2.2 Consent

Figure 4–6 Consent

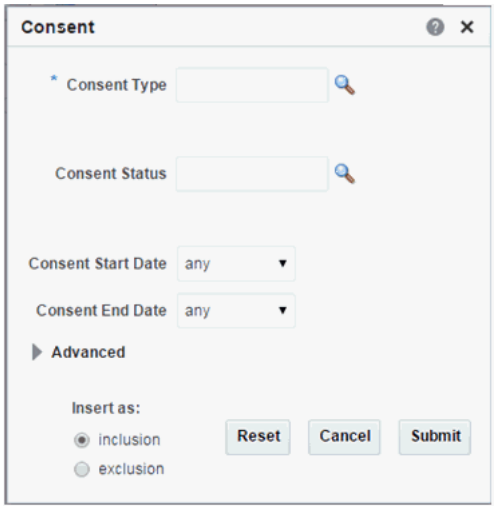


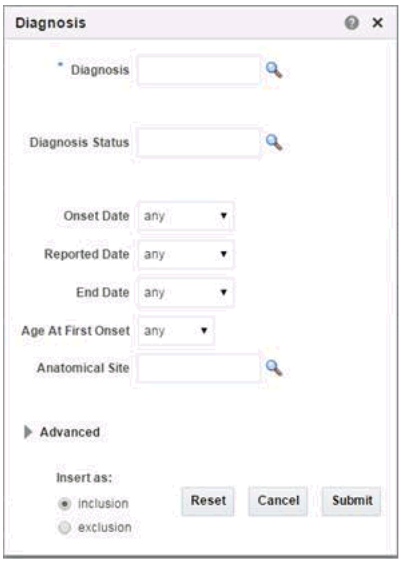
Table 4–3 Consent Screen Fields

Prompt Heading	Definition	Sample Value
Consent Type Code	Authorization for specified medical care.	Procedure Consent, Specimen Consent
Consent Status	The status of the consent form.	Active, Pending, Refused
Consent Start Date	The period start date of the patient's consent.	-
Consent End Date	The period end date of the patient's consent.	-

4.2.3 Clinical Data

4.2.3.1 Diagnosis

Figure 4–7 Diagnosis



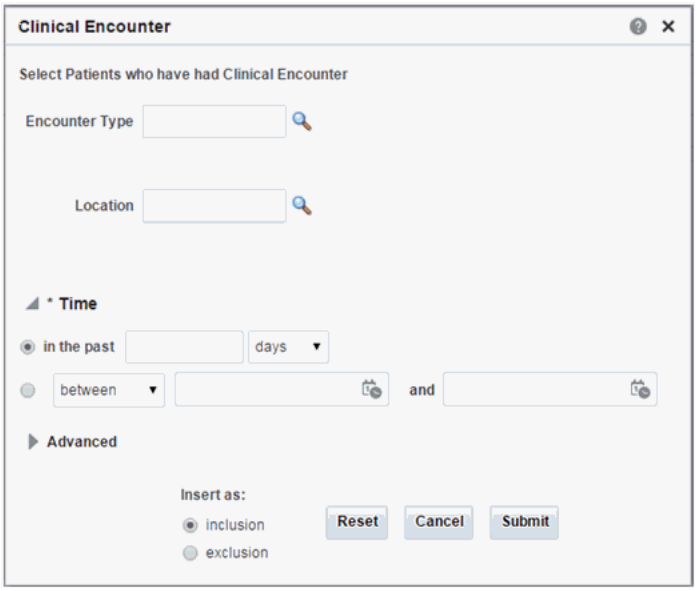
Note: Click the magnifying glass icon for **Diagnosis,** to search either by Diagnosis Code or Diagnosis Name.

Table 4–4 Diagnosis Screen Fields

Prompt Heading	Definition	Sample Value
Diagnosis	Code that classifies the patient's clinical condition.	Most commonly identified with ICD codes.
Diagnosis Status	Code that reflects the status of the Diagnosis	Active, New, Recurring
Onset Date	Date of the onset	-
Reported Date	Date when the diagnosis was recorded by a service provider	-
End Date	Date of resolution.	-
Age at First Onset	Age of the patient at first onset (years)	-
Anatomical Site	Anatomical site or sites related to the diagnosis	-

4.2.3.2 Clinical Encounter

Figure 4–8 Clinical Encounter



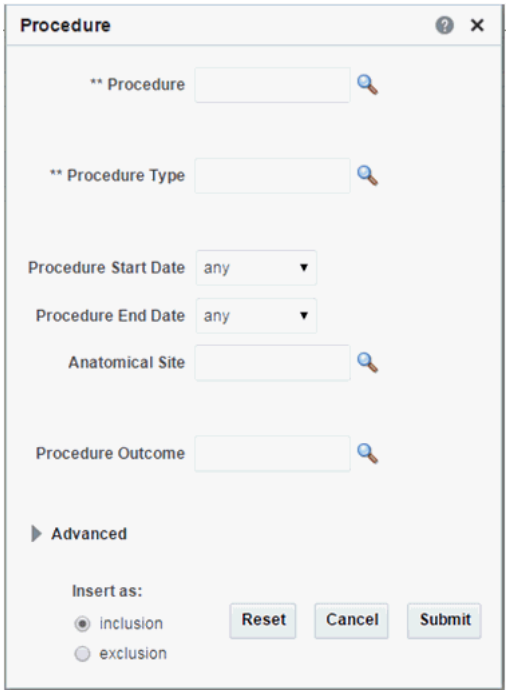
Note: Click the magnifying glass icon for Encounter Type, to search either by Encounter Code or Encounter Name.

Table 4–5 Clinical Encounter Screen Fields

Prompt Heading	Definition	Sample Value
Encounter Type	Type of clinical encounter that individual has undergone, valid values may be inpatient, outpatient	Inpatient, Outpatient
Location	Facility where it took place, generally name of the hospital, clinic, doctors office and so on	Sequoia Hospital, Medical Associates, Mass General Hospital Ear and Nose Dept
Time	Date when encounter took place	-
Datasource	Name of the data system actual clinical data is coming from	EMR1, EMR2

4.2.3.3 Procedure

Figure 4–9 Procedure



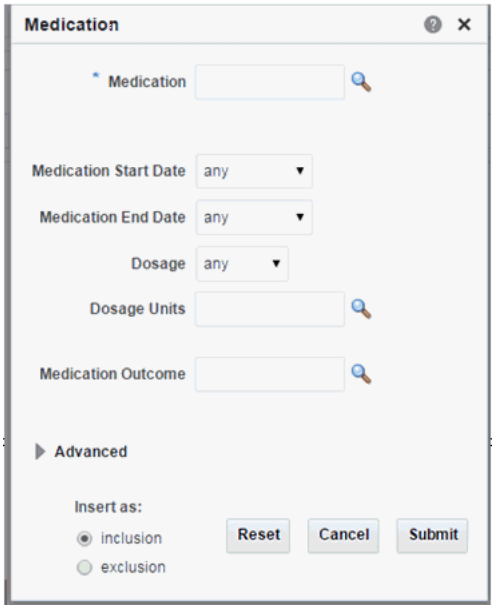
Note: Click the magnifying glass icon for **Procedure,** to search either by Procedure Code or Procedure Name.

Table 4–6 Procedure Screen Fields

Prompt Heading	Definition	Sample Value
Procedure	A discreet intervention performed by a clinician. Procedures are most commonly identified with CPT codes	-
Procedure Type	A sub categorization of procedures	Surgical, Radiology, Diagnostic
Procedure Start Date	The date when the procedure began	-
Procedure End Date	The date when the procedure concluded	-

4.2.3.4 Medication

Figure 4–10 Medication



Note: Click the magnifying glass icon for **Medication** to search either by Medication Code or Medication Name.

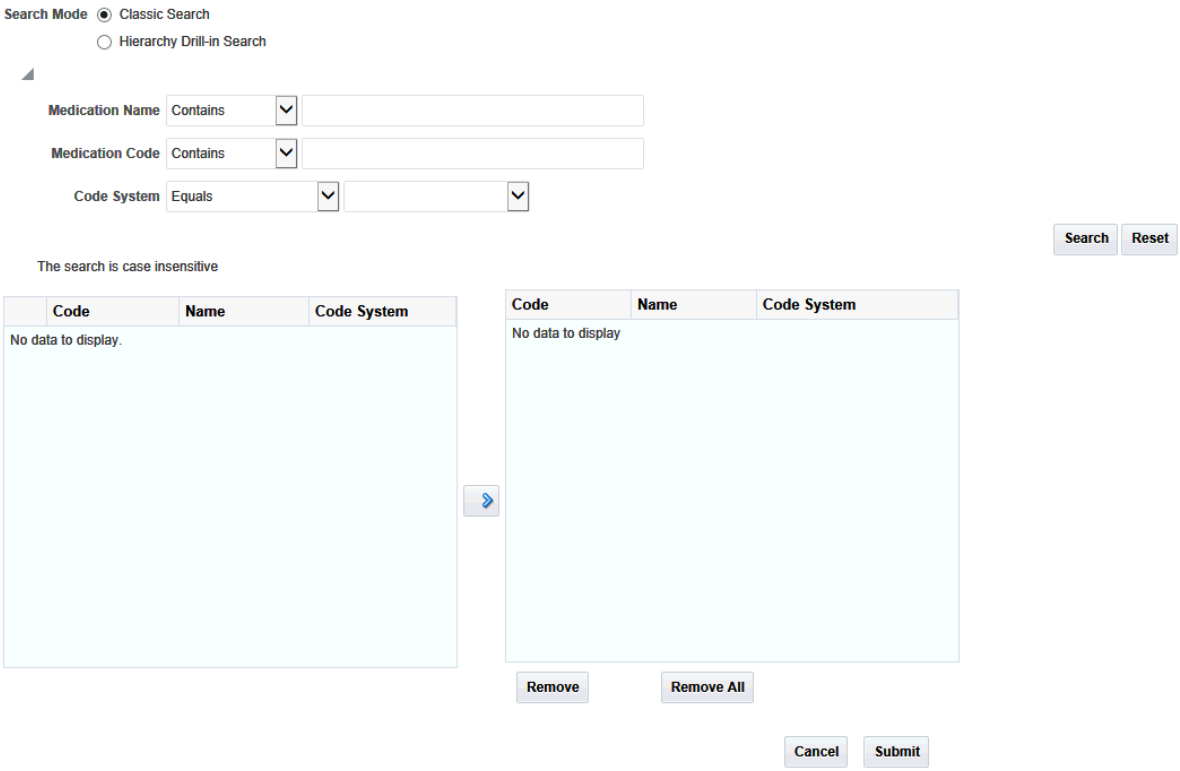
The Medication search menu supports the following two search modes:

Classic Search

Classic search lets users search for Medication codes based on medication name, medication code and code system. These parameters can be used with the options—Starts with, Ends with, Equals, Does not equal, Contains (default value), Does not contain, Is blank, Is not blank.

1. Navigate to **Cohort Query** > **Clinical Data** > **Medication**.
2. Click the magnifying glass icon next to the **Medication** field.
3. Select **Classic Search**.
4. Specify the search criteria either by using the above parameters individually or in combination, along with the available operators.

Figure 4–11 Classic Search



5. Click **Search**.
6. The Code, Code Name and Code System of the matching codes are displayed on the left. Expand the list items to view details about code hierarchy and the Code Description.

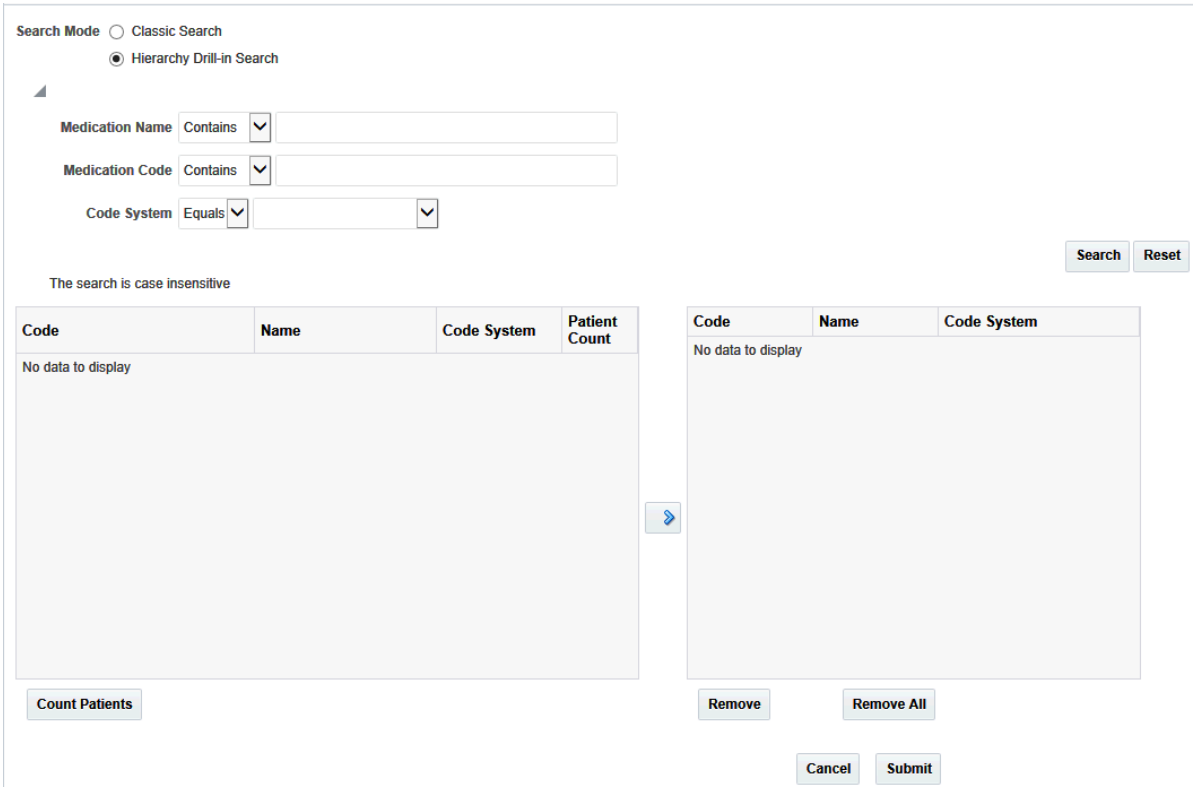
   If you have not specified any search criteria in the previous step, then all the codes at all the levels of hierarchy, from all the hierarchies are listed.
7. Select any number of rows on the left and copy them to the right hand side selection list using the **>** button.
8. Click **Submit** at the bottom of the screen. All the selected codes are used for defining query criteria in the Cohort Query.

Hierarchy Drill-in Search

Hierarchy Drill-in search lets you search for Medication codes based on medication name, medication code and code system. These parameters can be used with the options— Contains (default value), Starts with, Equals. The drop down list for Code System lets you select multiple Code Systems.

1. Navigate to **Cohort Query** > **Clinical Data** > **Medication**.
2. Click the magnifying glass icon next to the **Medication** field.
3. Select **Hierarchy Drill-in Search**.

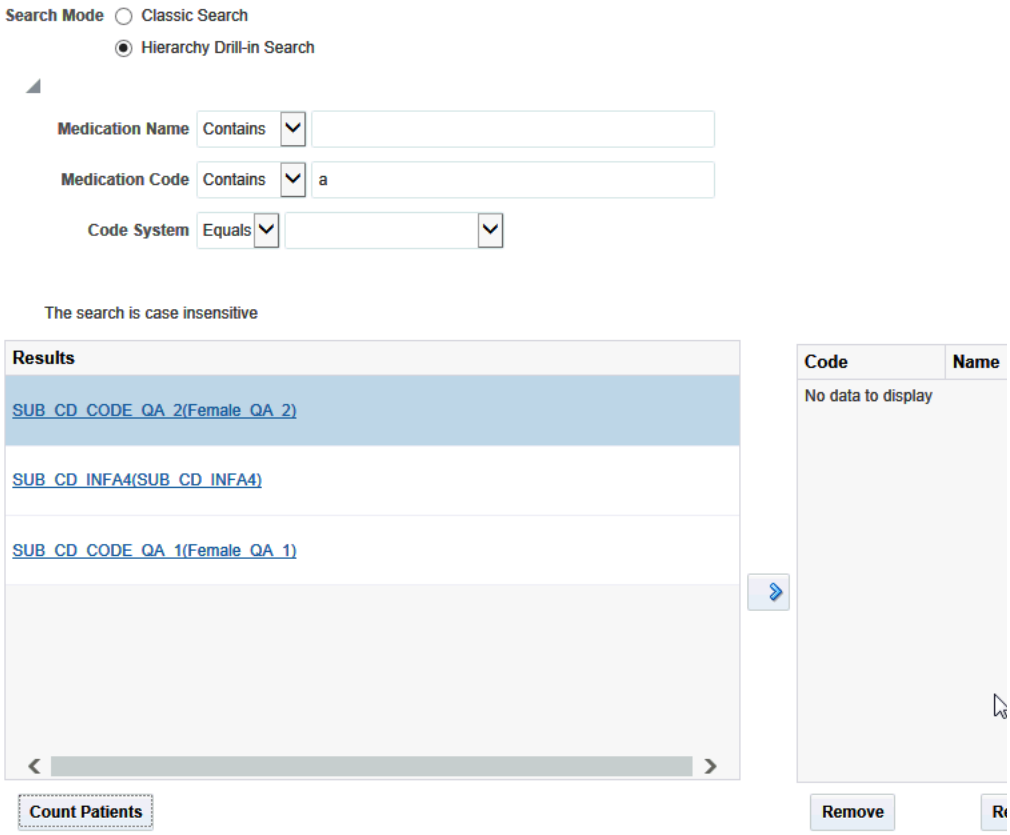
Figure 4–12 Hierarchy Drill-in Search



4. Specify the search criteria either by using the above parameters individually or in combination, along with the available operators.
5. Click **Search**.
6. The matching codes are displayed on the left. The entire hierarchy edge in which the matching code is present, is listed as a hyperlink. Click the hyperlink to view an indented list where each child level code is displayed at one indent level more than its parent code.

   If you have not specified any search criteria in the previous step, then all the codes at all the levels of hierarchy, from all the hierarchies are listed.

Figure 4–13 Matching Codes from Hierarchy Drill-in Search



7. To view the count of patients who were given the medication represented by the Code, select **Code(s)** and click **Count Patients**. The count is cumulative of the selected Code and all its child Codes. For example, if Second Level Child Code is selected, then the count corresponds to the Patients who were given the Medication represented by Second Level Child Code or Medications represented by any of the children of the Second Level Child Code.
8. Select any number of rows on the left and copy them to the right hand side selection list using the > button.
9. Click **Submit**. All the selected codes are used for defining query criteria in the Cohort Query.

Table 4–7 Medication Screen Fields

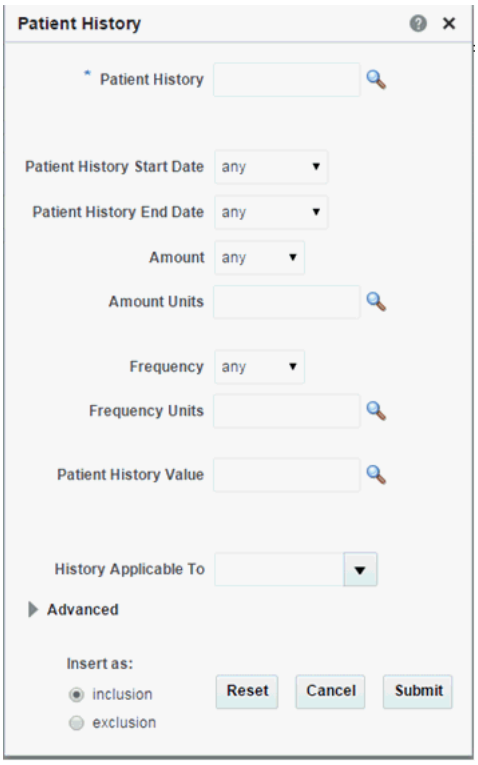
Prompt Heading	Definition	Sample Value
Medication	A pharmaceutical substance intended to provide therapeutic benefit.	NDC codes, RxNorm codes
Medication Start Date	The start date for the administration of the medication.	-
Medication End Date	The end date for the administration of the medication.	-
Dosage	The medication's dosage.	-
Dosage Units	Code for the medication dosage's units.	-

Table 4–7 (Cont.) Medication Screen Fields

Prompt Heading	Definition	Sample Value
Medication Outcome	Clinical outcome of taking a given medication as assessed by clinician or medical professional	Partial Recovery, Full Recovery, Adverse Reaction

4.2.3.5 Patient History

Figure 4–14 Patient History



Note: Click the magnifying glass icon for **Patient History** to search either by Medication Code or Medication Name.

Table 4–8 Patient History Screen Fields

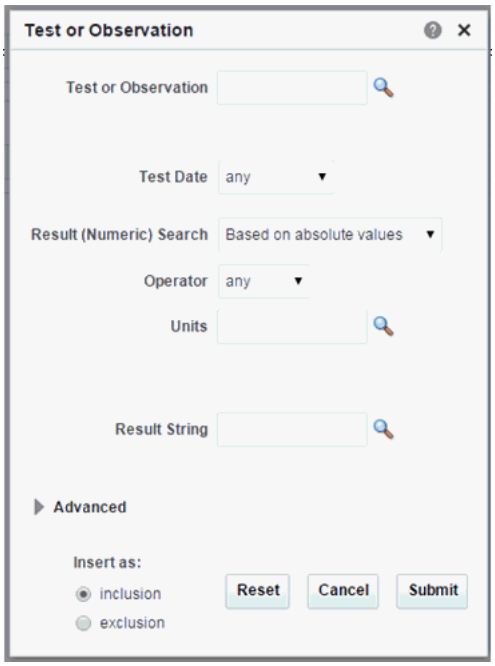
Prompt Heading	Definition	Sample Value
Patient History	A coded representation of the Patient History	Smoking, Obesity
Patient History Start Date	Date when this history was known to be active	-
Patient History End Date	Date when this history is known to no longer be active	-
Amount	Numerical value	For example, 1
Amount Units	Unit of Measure	For example, mmHg

Table 4–8 (Cont.) Patient History Screen Fields

Prompt Heading	Definition	Sample Value
Frequency	Numerical Value	For example, 12
Frequency Units	Unit of Measure	For example, TPDAY (times per day)
Text Value or Code	Value of history which is represented by text or coded.	Yes, No, Frequently, Rarely
History Applicable To	For Familial History, this will contain type of blood relative, which may have a history value	Father, Mother, Paternal grandmother, Paternal Grandfather

4.2.3.6 Test or Observation

Figure 4–15 Test or Observation



Note: Click the magnifying glass for **Test or Observation** to search either by Test or Observation Code, Test or Observation Name. The units will be filtered based on the selected Test or Observation.

Table 4–9 Test or Observation Screen Fields

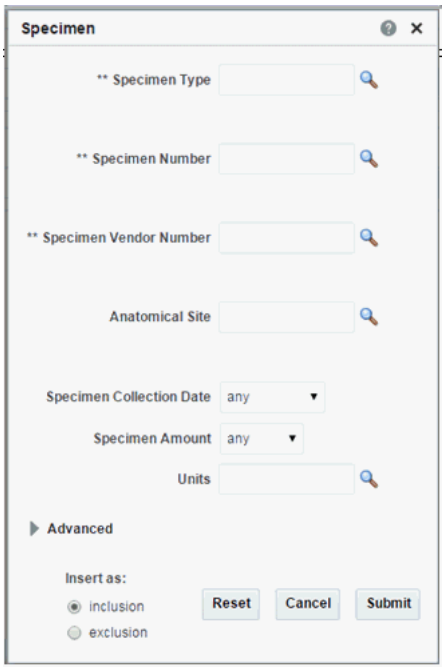
Prompt Heading	Definition	Sample Value
Test or Observation	Any kind of medical intervention performed to aid in the diagnosis or detection of disease	Blood Pressure
Test Date	The date the test was performed	May-12-2012

Table 4–9 (Cont.) Test or Observation Screen Fields

Prompt Heading	Definition	Sample Value
Result (numeric) Search	The Test result	140, 90
Operator	This search mode enables searching numeric results with UoMs based on reference range or absolute values.	
Units	The Test units of measure	mmHg
Result String (text)	Textual Test result, also Notes or remarks for the Test	Positive/Negative

4.2.3.7 Specimen

Figure 4–16 Specimen



Note: The three fields of **Specimen Type, Specimen Number,** and **Specimen Vendor** are used in combination to search for particular data. Click the magnifying glass icon next to each field, to specify the appropriate criteria.

Table 4–10 Specimen Screen Fields

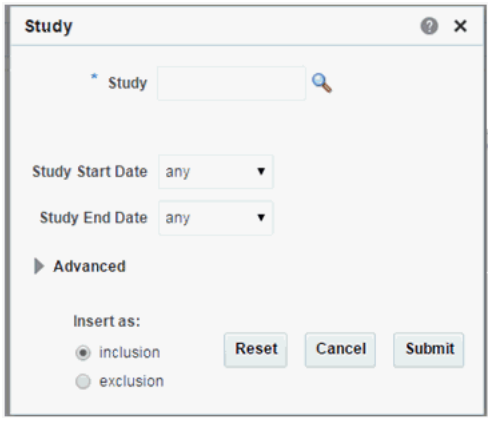
Prompt Heading	Definition	Sample Value
Specimen Type Name	A coded description for the type of Specimen.	Blood, Urine, Sputum
Specimen Number	A unique identifier for a specimen.	-

Table 4–10 (Cont.) Specimen Screen Fields

Prompt Heading	Definition	Sample Value
Specimen Vendor Number	A unique identifier for a source of the specimen, namely lab the specimen was analyzed.	Harvard-CancerInstitute01, MIT-Whitehead-Lodish5
Anatomical Site	The target site of the intervention.	-
Specimen Collection Date	Date the specimen was collected.	-
Specimen Amount	The amount of the specimen collected.	-
Units	The units of measure for the specimen collected.	Mg/dL , millimoles/liter

4.2.3.8 Study

Figure 4–17 Study Screen



Note: Click the magnifying glass icon for **Study,** to search either by Study Code or Study Name.

Table 4–11 Study Screen Fields

Prompt Heading	Definition	Sample Value
Study (Name or Identifier)	A reference to a particular study.	NOT-A-STUDY (any genomic data that is not explicitly tied to study would be categorized there), Glioblastoma Study
Study Start Date	The start date for the study.	-
Study End Date	The end date for the study.	-

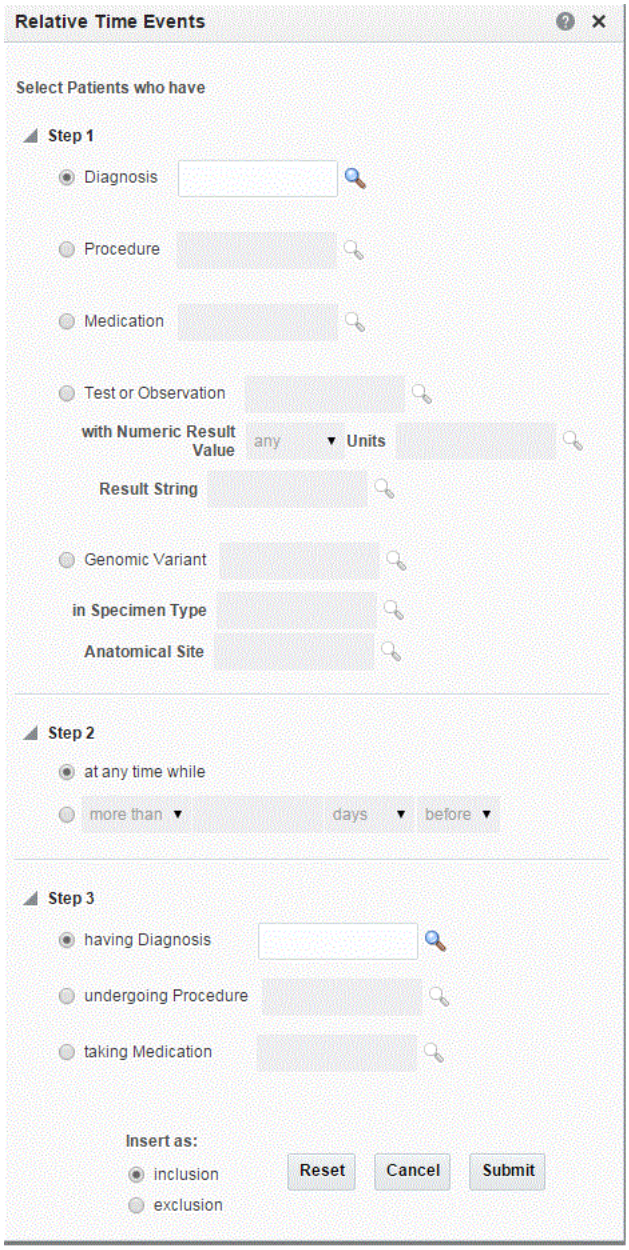
4.2.3.9 Relative Time Events

So far, we have covered clinical and other patient related criteria along with their metadata. For example, a user can specify a diagnosis and a range of dates for a diagnosis given to the patient to search for cohort within the particular time period.

However, frequently, the events that drive cohort or patient set searches are relevant to finding patients that may have certain event dependencies such as being diagnosed with a disease after taking a medication.

CE enables you to find such patients based on relative time event dependencies. With this criteria set you can search for patients who have a history of one clinical (or genomic change) event relative to another. In other words, you can specify a search for patients that took a certain medication or a set of medications within 30 days before being diagnosed with a specific disease. Or who had a procedure performed on them two months after starting a medication. The intent of the interface is to give the user a natural language interface to specify the criteria. On the backend, a complex query is generated by the Query Engine; however the user stays within the user friendly front end interface. A figure of the Relative Time Events criterion is shown below.

Figure 4–18 Relative Time Events



This is a three-step process:

1. Specify the event data to start your event.
2. Specify the relative time condition. This is how the events in Step 1 and Step 3 relate to each other.
3. Specify the related event to associate with the event in Step 1.

The result is a search criteria that is interpreted as: Search for patients that experienced Event X (at any time while, certain amount of time before or after) Event Y. The idea is that X is the original event, Y is the associated event, and the time frame describes the time relationship between the two events, for the given patient.

The events listed for selection in Step 1 or Step 3 are present in Clinical or Genomics Data categories.

However, it is important to explain what the Relative time conditions are in Step 2 and how their meaning translates into SQL:

- at any time while: the meaning of this relative time condition is to find any events that overlap with each other in terms of their occurrence. For example, if we have two distinct events with start and end dates, as long as at some point in time both events were true, this condition is satisfied.

  Example 1:

  Event in Step 1 - Started on May 1, 2010 - ended on June 29, 2011

  Event in Step 3 - Started on June 10, 2011 - till present

  The two events have overlap between June 10-29, 2011 thus the condition is satisfied.

  Example 2:

  Event in Step 1 - Started on Dec 1, 2010 - ended on Dec 12, 2011

  Event in Step 3 - was performed on August 3, 2011

  The two events have overlap on August 3, 2011 thus the condition is satisfied.

  Example 3:

  Event in Step 1 - Started on May 1, 2010 - ended on June 29, 2011

  Event in Step 3 - Started on June 10, 2009 - till present

  The two events have overlap between May 1, 2010 and June 29, 2011 thus the condition is satisfied.

  The dates used for time comparing each event type are as follows:

  – Diagnosis: Reported Date - End Date (if null, use present date)
  – Procedure: Start Date - End Date (if null, use present date)
  – Medication: Start Date - End Date (if null, use present date)
  – Test or Observation: Test Date (single day event only)
  – Gene Variant: Specimen Collection Date for end date use present date

- (more than / less than / exactly) [ ] (days / weeks / months / years) (before / after): the meaning of this relative time condition is to find any events that have start or occurrence dates that relate to each other with the time period specified by the condition.

  Examples:

  Start date of Event in Step 1 is more than 5 weeks before Start date of Event in Step 3

  Occurrence, for example, Test Date of Test or Observation in Step 1 is less than 1 month before Start Date of Medication in Step 3

  Gene Variant Specimen Collection Date in Step 1 is exactly 1 day after Procedure Start Date in Step 3

  The dates used for time comparing each event type are as follows:

  – Diagnosis: Reported Date

- Procedure: Start Date
- Medication: Start Date
- Test or Observation: Test Date
- Gene Variant: Specimen Collection Date

4.2.4 Genomic Data

4.2.4.1 Microarray Expression

Microarray expression is one of the growing set of genomic criteria that can be used to stratify patients. The genomic data driving this filter comes from the ODB model, specifically the W_EHA_RSLT_GENE_EXPR table and related tables. Gene expression data is associated with patients through the specimen used for genomic study analyses collected during study participation or other genomic testing.

The upper section of the criteria gives you the choice to specify Specimen Type and Anatomical site to further filter the results.

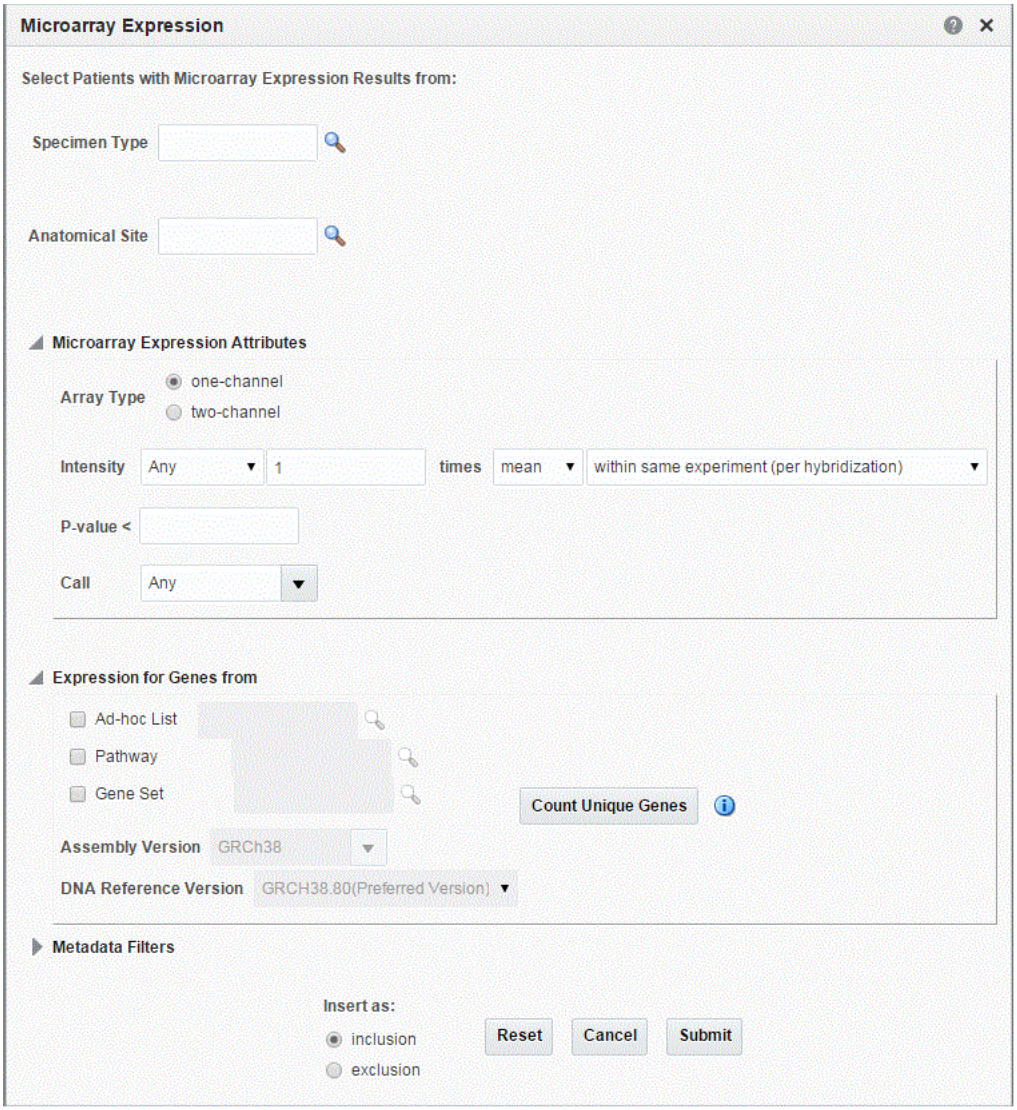
In the popup for Gene Expression, you are given two options (Array data types to choose from). For each option, you must specify the criteria when selecting patients based on their gene expression data.

- **One-channel** – Single channel data can be filtered based on the following data fields:
  - On Intensity; as a cutoff based on aggregates or on values, and across an experiment or multiple ones.
  - P-value; as a cutoff on values.
  - Call; on any of the three call types.
- **Two-channel** – Here data can be filtered on the following data field:
  - Log2Ratio – as a cutoff on values

In **Expression for Genes From**, select at least one unique gene from any of the three specified sources:

1. Ad-hoc list
2. Pathway
3. Gene Set

Figure 4–19 Microarray Expression



Note: Click the magnifying glass icon for **Specimen Type**, to search either by Specimen Name or Specimen Code. This also applies to the other search criteria. Criteria identified with * are required.

Table 4–12 Microarray Expression Screen Fields

Prompt Heading	Definition	Sample Value
Specimen Type	The specimen type description that identifies the specimen used for genomic tests.	Normal sample, tumor sample.
Anatomical Site	Anatomical site name or code corresponding to the specimen collected.	Left lung, kidney, small intestine

Table 4–12 (Cont.) Microarray Expression Screen Fields

Prompt Heading	Definition	Sample Value
Intensity	This and the following two fields are specific to Single Channel data. This field is a Gene expression value range specification. Data driving this selection is in the W_EHA_RSLT_ GENE_EXP table in ODB and is generally expected to be normalized now to take full advantage of this criteria selection. Additionally, the data values should all be positive. The criteria conditions allow you to narrow down patients based on upregulated gene expression (using *greater than 1.0 times mean* condition either across columns (*same hybridization*) or rows (*single result file*) of the data in the particular experiment.	Down regulated gene in hybridization example - Intensity is less than 2.0 times mean in the gene expression within a particular hybridization.
P-value	Significance value associated with the specific experiment, optional.	P-value < 0.00001
Call	Call made on the particular value, if present. All are taken unless specified. Optional.	P - Present, A - Absent, M - Marginal
Log2Ratio	Specific to dual channel differential gene expression relating the difference between, for example, control sample and the tested sample intensity. Data for this selection in found in the **W_EHA_ RSLT_2CHANNEL_GXP** table.	Float type integer representing a log base 2integer value of a fraction.
Expression for genes from*	Selection of genes that are to be used for patient stratification based on the expression. At least one of the below criteria must be specified.	N/A
Ad-hoc List: Gene	List of one or more genes.	-
Pathway	Reference to a pathway stored on the reference side of the ODB model. This in turn corresponds to a list of genes that are to be used to compare their Intensity values.	-
Gene Set	User-defined collection of genes that you can reference across any UI instead of having to build ad-hoc lists of genes each time.	MyGeneSet1, GlioblastomaSmithLa bGenes
Assembly Version	Represents GRCh assembly version. Default selection is last loaded version.	GRCh38 or GRCh37
DNA Reference Version	Represents the Ensembl reference version for getting gene annotations. This is related to 'Assembly Version' and by default it shows the preferred DNA Reference Version which is set in ODB.	Ensembl release 70

For Metadata Filter details for Microarray Expression, see Metadata Filters on page 4-35.

4.2.4.2 Sequence Variants

The Variants Criteria Selection interface falls under the category of Genomic criteria and enable the user to query for patients based on results present in the **W_EHA_RSLT_SEQUENCING** and related tables in the ODB model.

In the Sequence Variant screen, you have the option to select Specimen Type and Anatomical site for associating the selected variant criteria in this screen. Currently, the following 5 main options are available for searching:

- having Variants in selected Genes: this enables you to search variants in specific genes.
- having selected Genomic Variants: this enables you to search for specific known variants by their Cosmic or dbSNP identifiers.
- having Variants within specified Genomic Region: this enables you to search variants in specified genomic location like chr1:19094593-29302393 or whole of chr1.
- having Zygosity: this enables you to search variants with specifiec zygosity.
- having Genotypes: this enables you to search for specific genotypes like AT, or AA, or wildtype (same as reference), and so on.

Additionally there are other parameter, attributes and metrics associated with variants that you can use for further filtering on specific variants. These options are:

- Specifying variants by their attributes.
- Specifying variants by their non-synonymous substitution scores
- Specifying quality metrics filters depending on the sequencing file type like VCF, MAF and CGI masterVar.

The filtering options have been categorized into 2 steps. Step 1 as shown in Figure 4–20 has all mandatory filter options for creating a query. Step 2 shown in Figure 4–21 has the optional filters.

Figure 4–20 Mandatory Filters

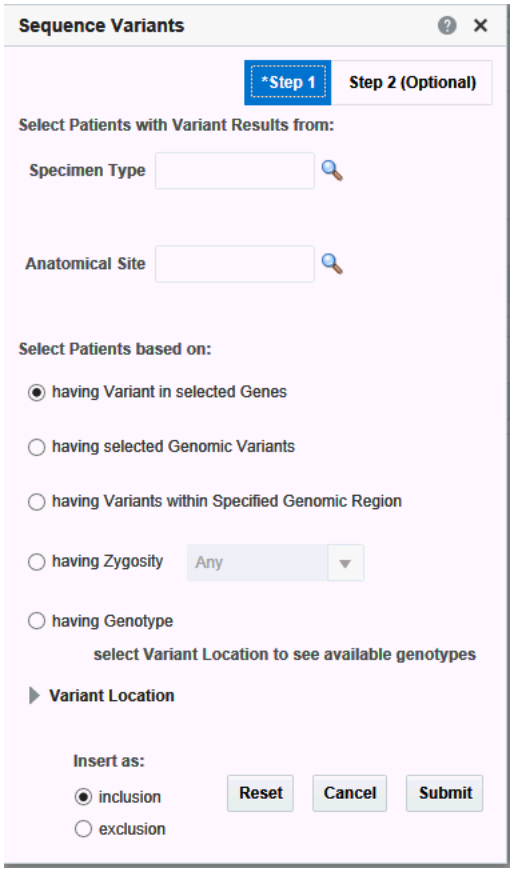


Figure 4–21 Optional Filters

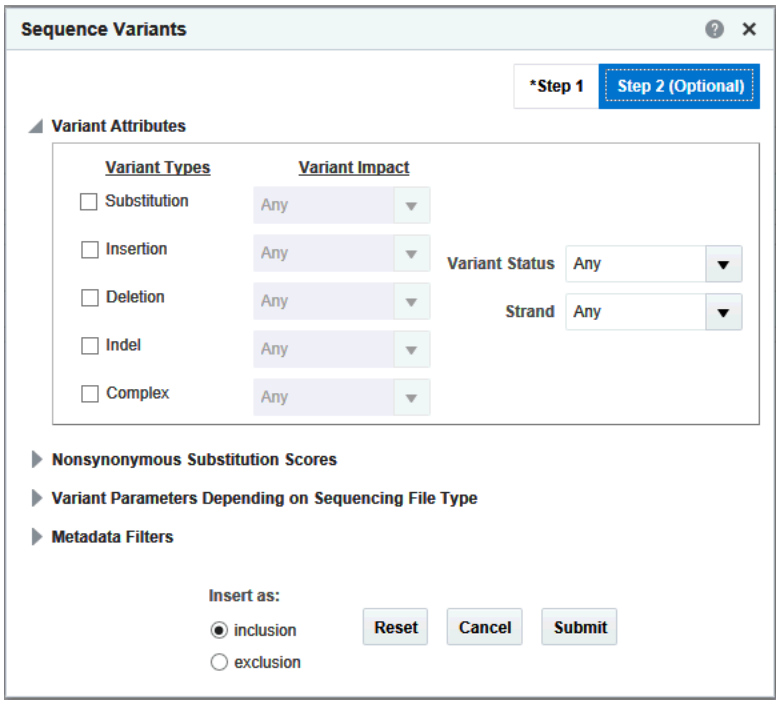


Table 4–13 Variants Screen Fields

Field Name	Definition	Sample Value
Specimen Type, Anatomical Site	For more information on Microarray Expression, refer to description under Microarray Expression section on page 4-22.	As in Microarray Expression section.
Having Variants in selected Genes	Variant mode selection - being able to specify variants in specified genes	N/A
Having selected Genomic Variants	Variant mode selection - being able to specify known variants by their identifier. At least one mode of variant specification must be given.	N/A
having Variants within Specified Genomic Region	Variant mode selection - being able to specify variants in a specific genomic location like chr1:1234324-3434333, chr7:1000, chr3 or chr2:1000+200	N/A
Genomic Variant ID	Specify known gene variant reference identifiers such as dbSNP or Cosmic. Allows for selecting multiple values. Assembly Version helps specify variants in a selected assembly version. Enables multiple selection, except for 'having Genotype' option where only single selection is allowed.	rs56289060, 905944
DNA Reference Assembly Version	Helps specify variants in a selected reference assembly version. Allows multiple selection, except for *having Genotype* option where only single selection is allowed.	GRCh37

Table 4–13 (Cont.) Variants Screen Fields

Field Name	Definition	Sample Value
DNA Reference Version	Represents the Ensembl reference version for getting gene annotations. This is related to 'Assembly Version' and by default displays the preferred DNA Reference Version, which is set in ODB.	Ensembl release 70
At Genomic Position	Helps specify a genomic region in which variants has to be searched.	chr7, chr7:1000, chr3:1000-2000, chr2:1000+200
Genotype	List available genotype values for the selected genomic position or genomic variant based on the position. Also displays wildtype base (same as reference base). User can select a combination of two genotypes to search or just one genotype and search for patients with these selected genotypes.	N/A
Additional Variant Information (Optional) are used in the context of either of the radio buttons selection above	Additional criteria used to filter down variants based on variant type and variant impact features.	N/A
Variant Type and corresponding Variant Impact	You can select any of the variant types supported along with additional variant metadata such as each variant type's impact on the resulting protein.	Variant Type: Substitution, Insertion, Deletion, Indel, Complex
	Note: Not all Variant Impact annotations apply to each Variant Type.	Variant Impact: Synonymous, Missense, Nonsense, Unknown,
Variant Status	Specifies which variant types to consider - whether the variant should be known or novel. Default considers all variants.	Known, Novel
Strand	Gene transcription direction attribute. By default, all directions of transcription are included.	+ means forward, - means reverse
Ad-hoc List: Gene; Pathway; Gene Set	For more information on Microarray Expression, refer to description under Microarray Expression section on page 4-22.	As in Microarray Expression section.
At Genomic Position	Specify genomic location for the variants to occur in, the format should be chr#:from-to or chr# if entire chromosome to be used.	chrX:13000-120000, chr7

Table 4–13 (Cont.) Variants Screen Fields

Field Name	Definition	Sample Value
Non-synonymous Substitution Scores	Data for this section is loaded into the reference side of the ODB model from Ensembl based on either Polyphen algorithm or SIFT algorithm. The prediction value can be specified numerically or alternatively can be specified using Polyphen or SIFT specific annotation. **Note**: SIFT or Polyphen predictions are only available for Known variants	With Polyphen, prediction between 0 and 1 or labeled as benign, Probably Damaging, possibly damaging and unknown With SIFT, prediction between 0 and 1 or labeled as Deleterious and tolerated
Variant Parameters Depending on Sequencing File Type	User can select to specify more detailed filtering criteria based on data coming from 3 different sequencing file formats such as VCF, MAF, CGI masterVar. As each input file formats uses different metadata to describe stored entities, depending on the sequencing input format selection, the user can elect to specify: VCF: Variant Call Format, Format.GQ range - user can specify upper or lower numeric values for this parameter MAF: Mutation Annotation Format, Score - user can specify upper or lower numeric values for this parameter Somatic Status Somatic Score Allele Read Count Reference Read Count Total Read Count RMS Base Quality RMS Mapping Quality AD/DP Ratio CGI masterVar: Complete Genomics masterVariation format. The available fields to search by are ■ Allele Zygosity ■ Score VAF ■ Score EAF ■ Allele read count ■ Reference read count	See appropriate file formats documentation for appropriate value ranges (VCF 4.2 format, MAF 2.0-2.2 format, Complete Genomics masterVar format). For example, Allele Zygosity for CG masterVar includes het-alt, hom, half, het-ref options.

For Metadata Filter details for Microarray Expression, see Metadata Filters on page 4-35.

4.2.4.2.1 At Genomic Position

You can opt to specify genomic data selection using genomic co-ordinates. You specify the chromosome region in a standard format for the Variation and CNV data. You can

specify a complete chromosome or part of a chromosome as criteria. Currently, only one chromosome region at a time is implemented for search.

The following chromosome region formats are supported.

- CHR15:10000-200000: Considers region between 10000 to 200000 in chromosome 15.
- CHR15:1,200,000+5000 - Considers 5000 bases upstream from 1,200,000 position in chromosome 15.
- CHR15 - Considers whole of the chromosome 15.
- CHR15:1000 - Considers 1000th nucleotide position of chromosome 15.

4.2.4.3 Copy Number Variation

The Copy Number Variation criteria selection interface falls under the category of Genomic criteria where you can query for patients based on results present in the **W_EHA_RSLT_COPY_NBR_VAR** table and related tables in the ODB model. Currently, this table contains data from two formats—from Complete Genomics and SEG and VCF files with CNV data. Both of these have log2 ratio stored in ODB.

For CNV, you can optionally select specimen type and anatomical site.

Next select CNV Result Type, which represents data from numeria (array based) and categorized sequencing based platform.

You can then filter results based on the list of Copy Number Variation Attributes. For example, for numeric based CNV Result Type, SNP log2 Ratio and for categorized based, gain, loss, equal to indicate Amplification, Deletion or no change in the copy number of a given gene or gene region.

Finally, you should specify the location of Copy Number Variation which is the gene or genomic position of interest.

Figure 4–22 Copy Number Variation - Numeric Based

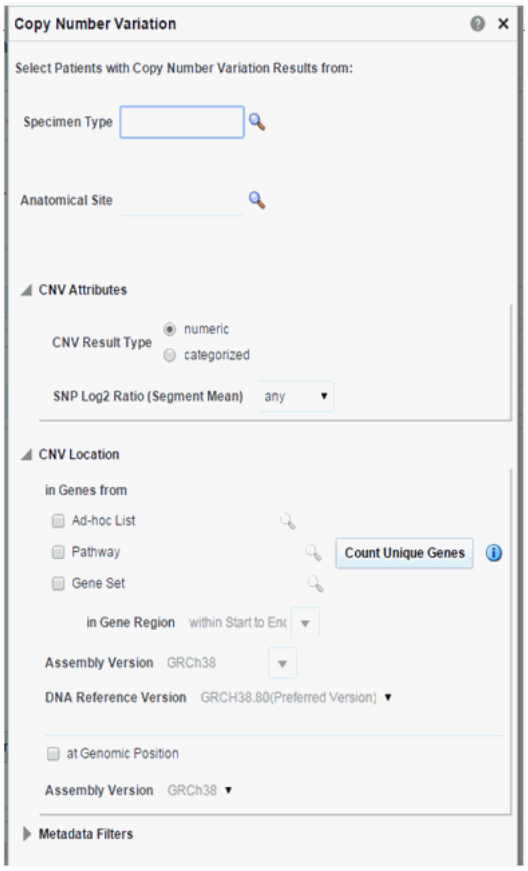


Figure 4–23 Copy Number Variation - Categorized Based

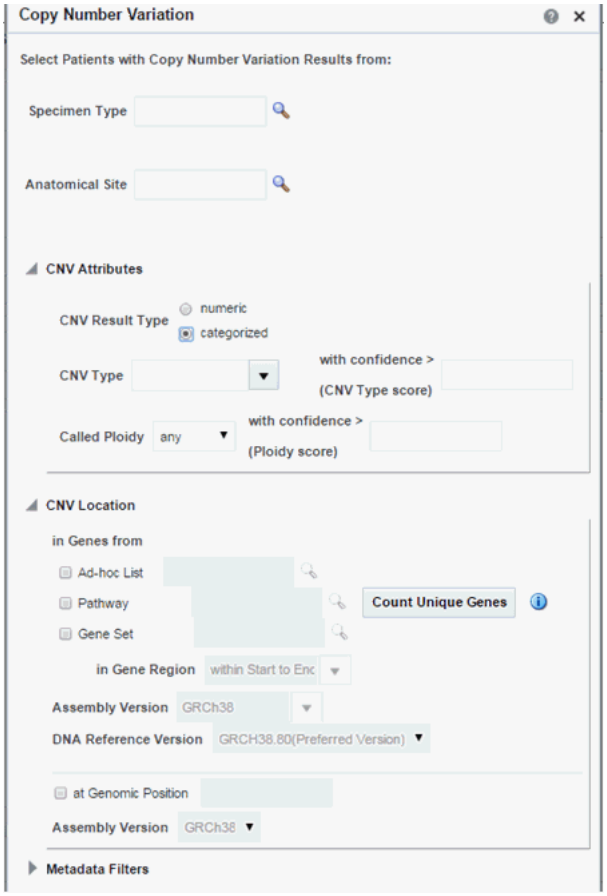


Table 4–14 Copy Number Variation

Prompt Heading	Definition	Sample Value or Values
Study*, Specimen Type, Anatomical Site	For more information on Microarray Expression, refer to description under Microarray Expression section on page 4-22.	As in Microarray Expression section.
CNV Result Type	Search CNV results either belonging to array based platform like Genome_Wide_SNP_6 array or sequencing based CNV data from complete Genomics.	Array based or sequencing based
SNP Log2 Ratio (Segment Mean)	Values for segment mean in the form of a range. You can also specify a single value in Log2 Ratio and search for results with segment mean greater than the specified value.	Numeric value. It can accept negative values.
CNV Type	Copy Number Variation attribute indicating whether it is an amplification - gain, deletion - loss, or no change - equal.	Gain, Loss, Equal
With confidence > (CNV Type score)	Copy number variation confidence score associated with CNV Type. Score is populated from the source file, and depending on the scoring method, the range can vary.	Numeric value, range can vary depending on source

Table 4–14 (Cont.) Copy Number Variation

Prompt Heading	Definition	Sample Value or Values
Called Ploidy	Values for ploidy can be given more specifically as a range, either upper and/or lower bound can be specified. For example, for duplication, called ploidy can be specified as 2.	Range of Ploidy to be selecting based on, e.g. for duplication, it can be specified as between 1.5 and 2.5
With confidence > (Ploidy score)	Confidence score associated with Called Ploidy. Score is populated from the source file, and depending on the scoring method, the range can vary. The higher the confidence, the more confidence is that the ploidy score is correct, lower range can be specified	Numeric value, range can vary depending on source.
CNV Location: in Genes from*	Selection of genes that are to be used for patient stratification based on Copy Number Variation. At least one of the below criteria must be specified.	N/A
Ad-hoc List: Gene; Pathway; Gene Set	For more information on Microarray Expression, refer to description under Microarray Expression section on page 4-22.	As in Microarray Expression section.
At Genomic Position	Specify genomic location for the variants to occur in: the format should be chr#:from-to or chr# if the entire chromosome is to be used.	chrX:13000-120000, chr7
Assembly Version	Represents GRCh version. Default selection is the last loaded reference version.	GRCh38
DNA Reference Version	Represents the Ensembl reference version for getting gene annotations. This is related to *Assembly Version* and by default displays the preferred DNA Reference Version, which is set in ODB.	Ensembl version 70

For Metadata Filter details for Microarray Expression, see Metadata Filters on page 4-35.

4.2.4.4 RNA-seq Expression

The RNA-seq Expression criteria selection interface falls under the category of Genomic criteria where you can query for patients based on results present in the W_EHA_RSLT_RNA_SEQ table and related tables in the ODB model.

As for Gene Expression, you can optionally select specimen type and anatomical site. Next, you can filter results based on the RPKM values, Raw Counts, Median length and strand.

Finally, you should specify the location for searching RNA-seq expression results which is the gene or genomic position of interest.

Figure 4–24 RNA-seq Expression

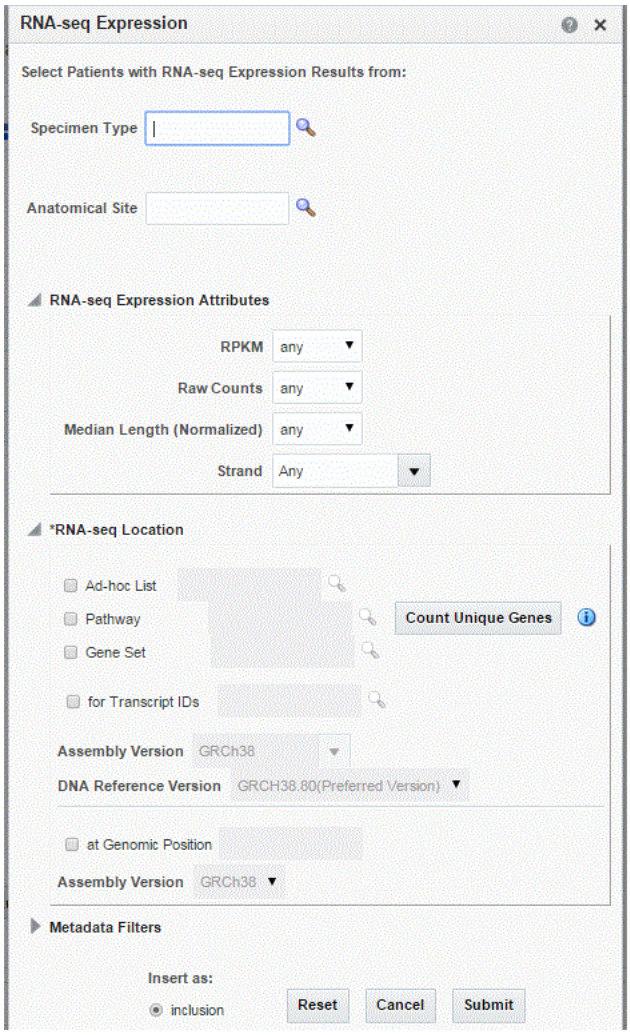


Table 4–15 RNA-seq Expression

Prompt Heading	Definition	Sample Value or Values
Specimen Type, Anatomical Site	For more information on RNA-seq Expression, refer to description under Microarray Expression section on page 4-22.	As in Microarray Expression section.
RPKM	Represents 'Reads Per Kilobaseq exon Model per million mapped reads', calculated expression intensity values in positive float or zero.	N/A
Raw Counts	Represents raw read counts in positive floating point values or a zero if unavailable.	N/A
Median Length (Normalized)	A normalized region length calculation in positive float or zero.	N/A
Strand	Strand of stored gene.	N/A

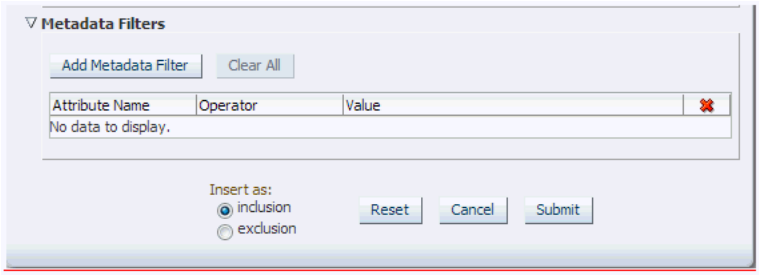
Table 4–15 (Cont.) RNA-seq Expression

Prompt Heading	Definition	Sample Value or Values
RNA-seq Location: in Genes from*	Selection of genes that are to be used for patient stratification based on RNA-seq expression. At least one of the below criteria must be specified.	N/A
Ad-hoc List: Gene; Pathway; Gene Set	For more information on RNA-seq Expression, refer to description under Gene Expression section.	As in Microarray Expression section.
For Transcript IDs	Search for Ensembl Transcipt ID	N/A
At Genomic Position	Specify genomic location for the variants to occur in, the format should be chr#:from-to or chr# if entire chromosome is to be used.	chrX:13000-120000, chr7
Assembly Version	Represents GRCh version. Default selection is the last loaded reference version.	GRCh38
DNA Reference Version	Represents the Ensembl reference version for getting gene annotations. This is related to *Assembly Version* and by default displays the preferred DNA Reference Version, which is set in ODB.	Ensembl version 70

4.2.4.5 Metadata Filters

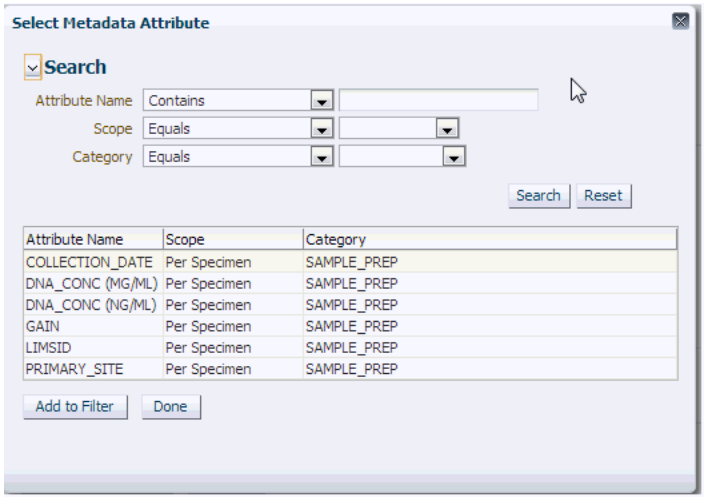
All genomic criteria screens like Sequence Variant, Copy Number Variation, Microarray Expression and RNA-seq Expression have additional filter criteria based on the metadata associated with the specimens or patients. This option is present at the bottom of each of the genomic criteria screens.

Figure 4–25 Metadata Filters



Once you expand the **Metadata Filters** option, click **Add Metadata Attribute** to open the Select Metadata Attribute dialog.

Figure 4–26 Select Metadata Attributes



Then search based on **Attribute Name**, **Scope** or **Category** to get a list of attributes associated with Metadata. Select attribute to add as Metadata Filter and assign a value for the added filer.

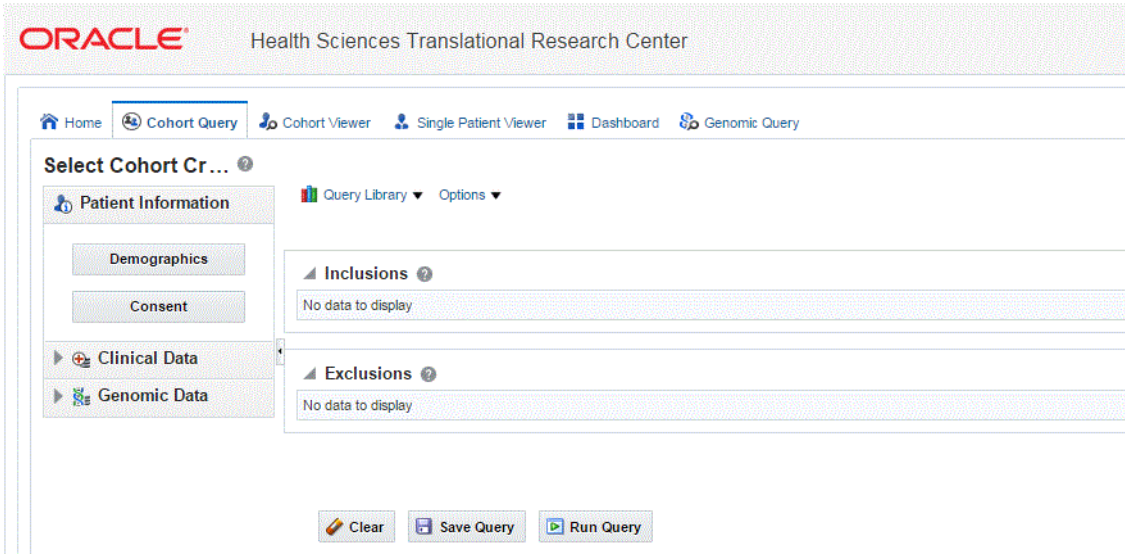
Table 4–16 Metadata Filters

Prompt Heading	Definition	Sample Value or Values
Attribute Name	Represents the metadata qualifier tag from W_EHA_QUALIFIER table.	N/A
Scope	Scope is retrieved from the table W_EHA_QLFR_TABLE. Based on the internal mapping scope is shown as 'Per Result' for value in table as 'W_EHA_RSLT_FILE_SPEC_QLFR' and 'Per Specimen' as scope value for value in table as 'W_EHA_RSLT_SPEC_QLFR'.	N/A
Category	Represents the metadata qualifier category tag from W_EHA_QLFR_CATEGORY table.	N/A
Value	Based on the selected metadata attribute this value datatype would change. If a numeric attribute is selected then a numeric value is given as input. Similarly, for date attribute, date would be as input and for character attribute, character would be as input.	N/A

4.3 Patient or Subject Count

Based on the Cohort Criteria, the application displays Patient or Subject Count. Total number of patients or subjects in the database is also displayed below that. This count is updated or refreshed each time you log into the system and select the **Query Patients** tab.

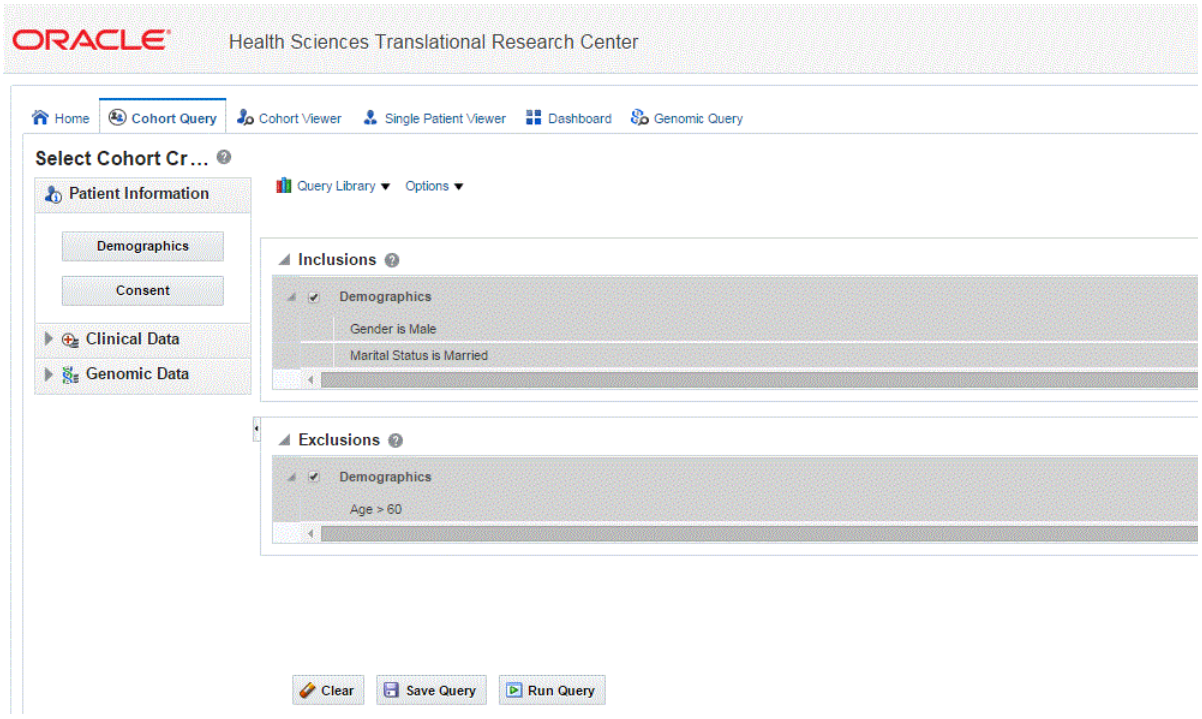
Figure 4–27 Patient Count



4.4 Inclusion and Exclusion Criteria

Once you have added several criteria statements to your query, you can select the arrow to the left of the criteria statement to expand or collapse its detail. You can also expand or collapse the display for the Inclusions or Exclusions criteria. This enables you to focus your attention on one aspect of your query as required.

Figure 4–28 Criteria



If you have specified multiple rows of criteria, CE considers the logic in combination when you run the query. The Query logic is slightly different depending on whether the statements are designated as either Inclusion or Exclusion.

- Each Inclusion statement is combined with an implied AND. For example,

  Demographics where Gender = Male

  AND

  Diagnosis where Diabetes onset date on or after March 1st 2001.

- Each Exclusion statement is combined with an implied OR. For example,

  Demographics where Age is less than 30

  OR

  Diagnosis where Asthma onset date prior to October 1st 1990.

Figure 4–29 Icons

You can continue to adjust the details of an existing statement without having to open the criteria selections on the left. Instead, you can select the icons on the right end of each statement.

- The pencil icon displays and edits your criteria selection
- The gray cross removes any particular statement.
- The plus icon is to add another set of criteria from the same topic. For example, clicking the plus icon from within a **Demographics** tab prompts you to select additional Demographics criteria.

  Within an Inclusion statement, distinct criteria statements related to the same topic is considered with an OR criteria.

  For example,

  INCLUDE: Diagnosis = Diabetes AND (Demographics = Male OR" Demographics = Married)

  Within an Exclusion statement, distinct criteria statements related to the same topic will be considered with an AND.

  For example,

  EXCLUDE: Diagnosis = Diabetes OR (Demographics = Male AND Demographics = Married)

This function can be used to simplify your query structure, where you can address each data topic and category one at a time rather than having the same topic listed in both the Inclusion and Exclusion sections.

Note: Oracle recommends that you prepare the query definition in a structured format (line-by-line), prior to creating it in CE. Visualizing your definition may help you recognize a simple way to organize it and give you a tool to validate the accuracy for the data you input in your system.

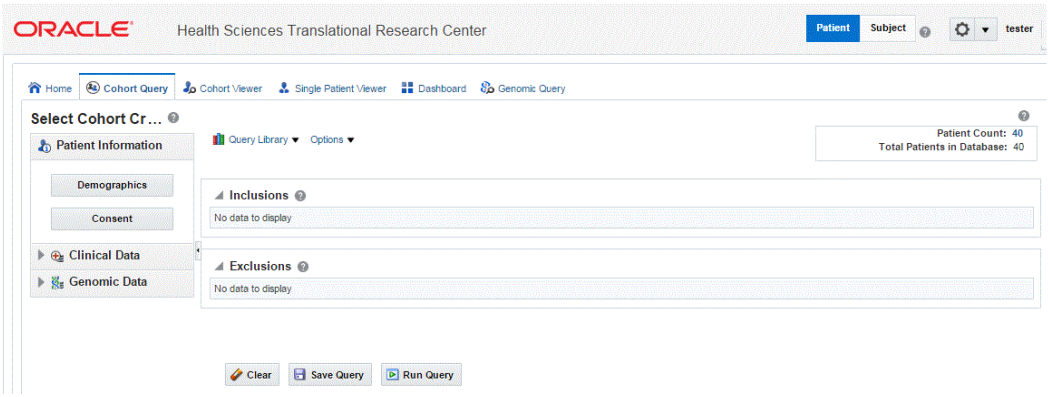
4.5 Final Patient or Specimen Count

Once you have defined at least one criteria statement, you can run the query. However, you can add as many additional criteria rows as you want. When you run a query, the system counts the number of patients that match the criteria as a subset of the Initial total count. CE also displays a count of the number of specimens for the patients identified by the query criteria. As you edit the criteria, the Patient Count and Specimen Count are updated each time you run the query.To run a query and view the counts, perform the following steps:

1. Select at least one row of criteria defined in the **Inclusions** or **Exclusions** options.
2. Click **Run Query**.
3. Click **Patient Count** to view the count of patients.

To the left of the Initial Patient Count display, there is a **Options** drop-down menu. If you select the down arrow, you can change the default display of the Specimen Count and hide the count.

Figure 4–30 Patient Count



4.6 Query Library

Once your query definition is complete, you may want to save it. When you save it, you create a unique name and have the option to enter a narrative description. CE preserves the criteria logic, but does not retain any information about the patients that were counted (or listed) when you last ran the query. As you create and save queries over time, you establish a valuable knowledge base that you want to reference in the future.

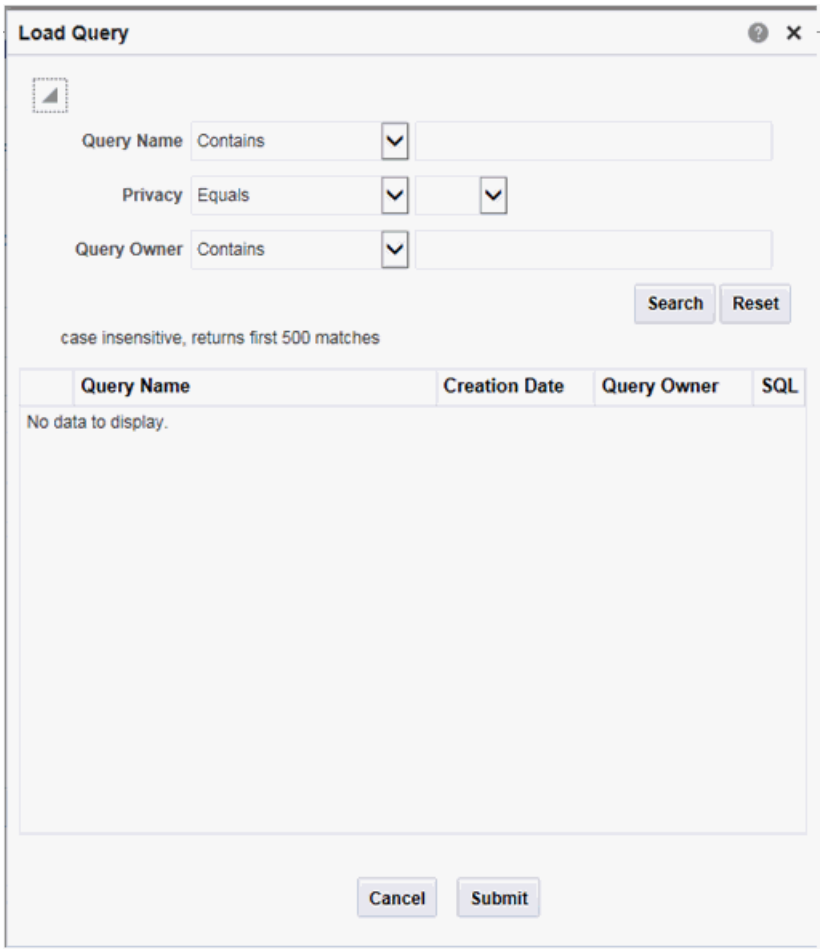
You can open a query at a later date to review your logic and possibly reuse all or part of it in a new query. In the upper left corner of the main window, there is a **Query Library** drop-down menu. This menu provides the ability to save the name of a query you have defined as well as to access a query that you have previously saved.

4.6.1 Load Query

To rerun a query and edit definition of the query from the **Query Library** menu, perform the following steps:

1. Select **Load Query**.
2. Enter all or part of a query name in the right hand text box.
3. Search for your query by entering your criteria from the **Query Name** drop -down menu.
4. Select the required query.
5. Click **Submit**. The query definition is displayed on the main page.

Figure 4–31 Load Query

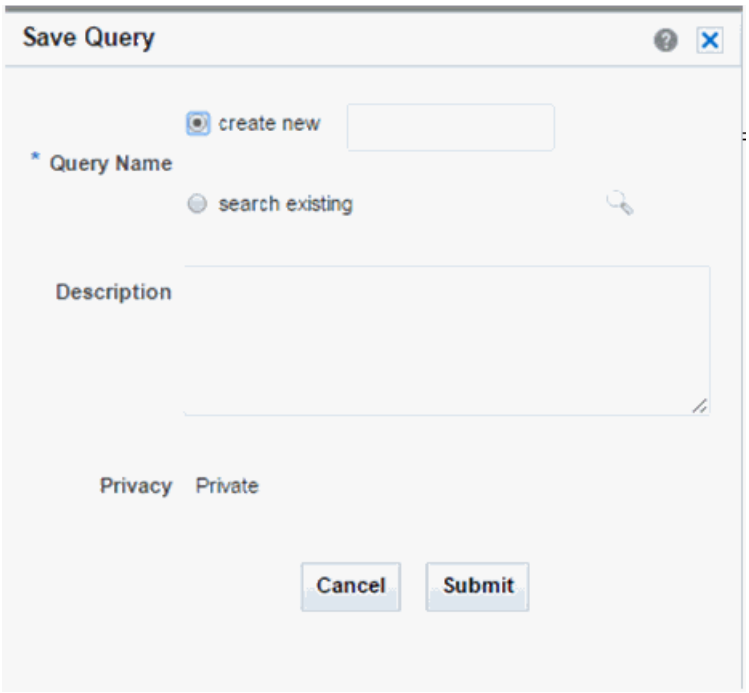


4.6.2 Save Query

Once you have at least one query criteria statement, you can select **Save Query** from the **Query Library** menu. You have the option to save as a new query (with a new name), or overwrite an existing query from the library.

If you enter a new query name and that name is already recognized in your library, the system prompts you to determine if you want to overwrite your existing query. If not, you have the option of entering a new query name.

Figure 4–32 Save Query



Note: When you save a query, the only information that is preserved is the criteria definition itself. The criteria statements are saved as logical elements. It is important to understand that neither the patient count nor any of the underlying patient data is saved with a query. Each time the query is opened, it must rerun to view current patient counts. Therefore, patient counts for the same query are expected to change over time.

5 Cohort Viewer

This chapter contains the following topics:

5.1 Cohort Viewer

In CE 3.1, a set of cohort viewers are supplied that enables the user to view patients or subjects in a variety of formats. You can view patient or subject details as a tabular list, or in a timeline view. You can drill into each single patient or subject details and see them all in one page. Furthermore, if Omics Data Bank model is licensed, you can look at patients or subjects genomic data in a circular genome viewer (using Visquick) or export patients or subjects data into formats acceptable by the Integrative Genome Viewer. The following sections describe the viewer options available in more detail.

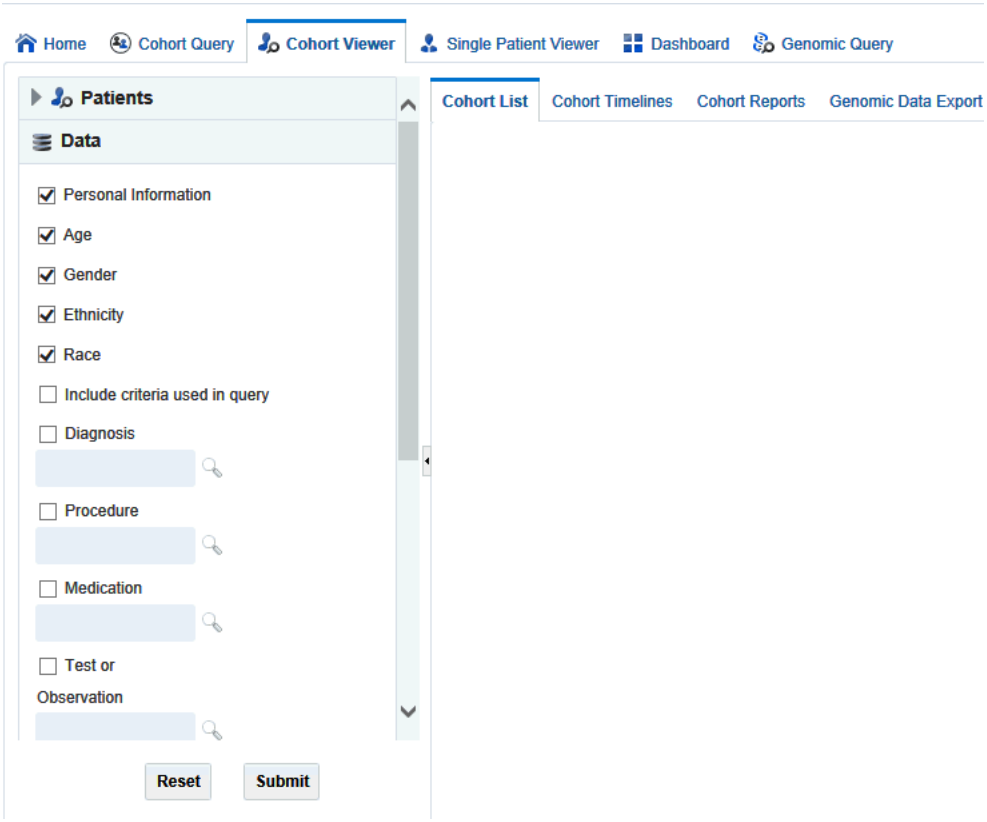
5.2 Cohort List Viewer

Once you run a query and can see the patient or subject count, you may want to review the data for those specific patients or subjects. To view a list, select the next tab to the right, the Cohort Viewer tab. This tab displays one row of data for each patient or subject represented in a query count. To the left of the Cohort List main window, there is a separate pane where you can select and filter the data you want to view for your **Cohort List**.

5.2.1 Patients

By default, this tab displays the option to list patients or subjects for the current active query, that is, the query currently loaded in the **Cohort Query** tab. To view this list, select **Submit**.

Figure 5–1 Show Patients or Subjects



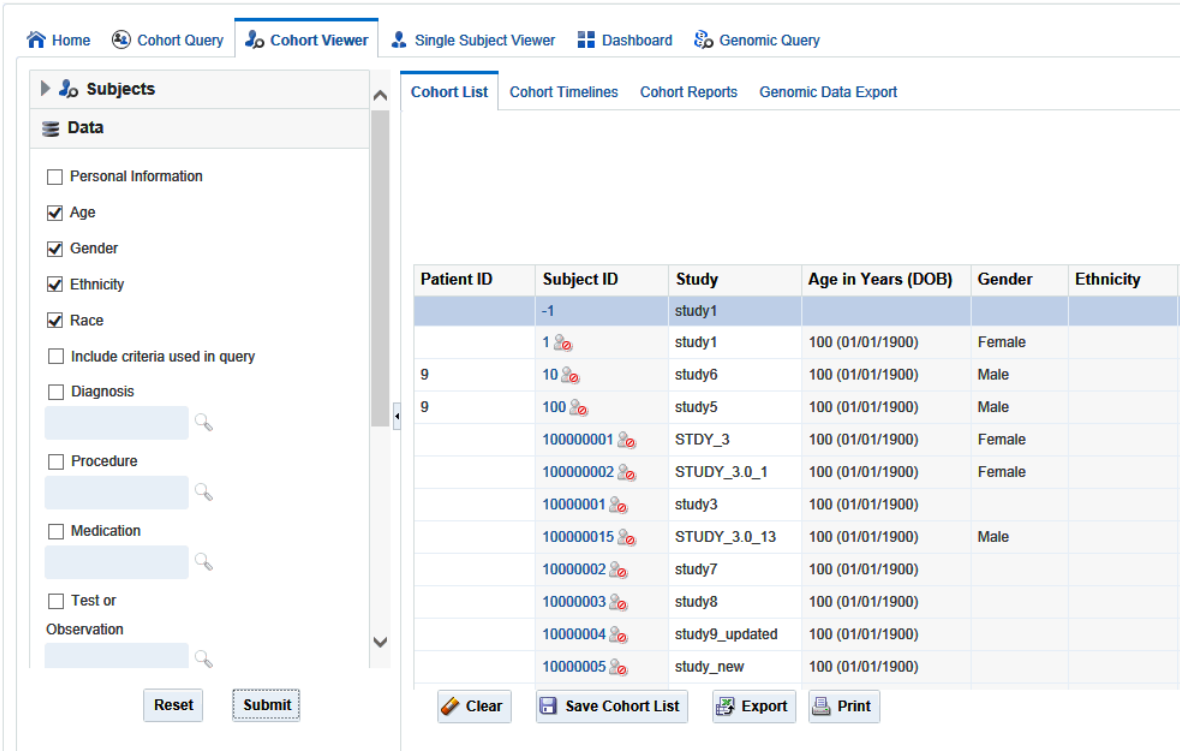
Alternatively, you can to view a patient or subject list for a query from the library, saved patient or subject list, or from the current omics query. You can specify one or more patient or subject (study) ID's on an ad-hoc basis if you have specific patients or subjects you want to examine.

5.2.2 Patient or Subject Data

Select the arrow to the left of **Data** to display the list of check boxes for data topics. These topics reflect the query selection criteria from the Clinical Information category. To view the data from the selected topics, perform the following steps:

1. Select one or more of the check boxes for the data you want to display in your list.
2. Select **Submit**.
3. The system reruns the query and adds relevant patient or subject data for each patient or subject listed.
4. To remove data, clear the appropriate boxes and select **Submit**.
5. To export the list to a Microsoft Excel sheet, click **Export**. To display the date in the format dd/mm/yyyy, use the formatting option in Microsoft Excel.

Figure 5–2 Cohort List



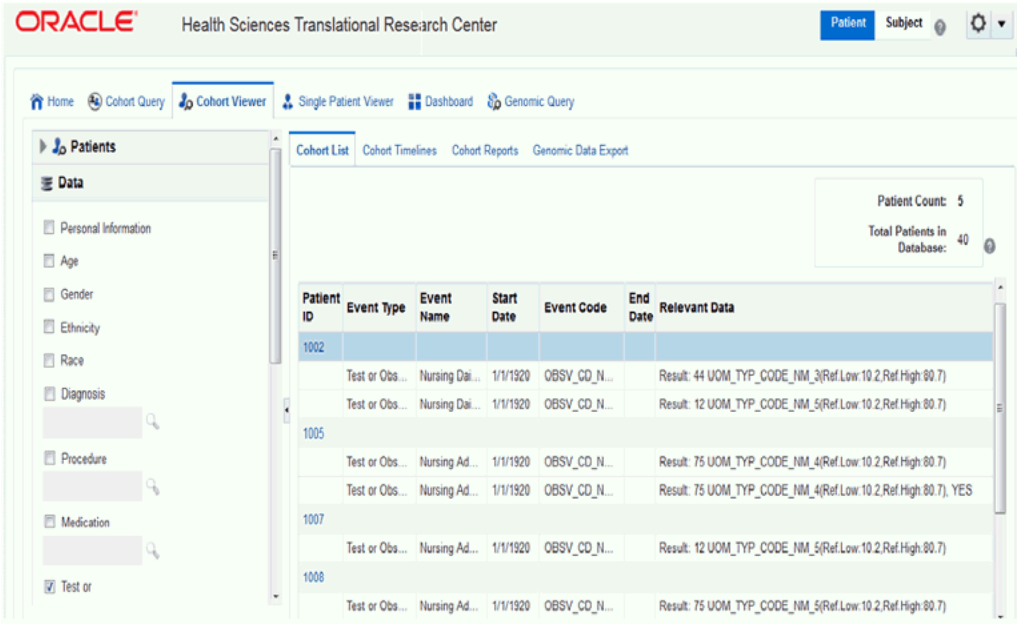
Note: A patient or subject may have more than one row of data for a particular data category (a patient or subject can have more than one procedure or medication). As a result, the patient or subject details display may show multiple rows or records for each patient or subject. Selecting several data categories to display may significantly limit your ability to review multiple patients or subject at the same time. This feature intends to help you visualize how the query criteria are manifested in the actual clinical information.

5.2.3 Displaying Reference Range Values

Cohort List supports displaying Reference Range values (if available) along with numeric results of Observation.

If the selected observation event data has the reference high and reference low range values along with numeric results, then the reference range values are displayed along with the numeric result values in the Relevant Data column of the cohort list.

Figure 5–3 Cohort List with Reference Range Values for Numeric Results of Observation



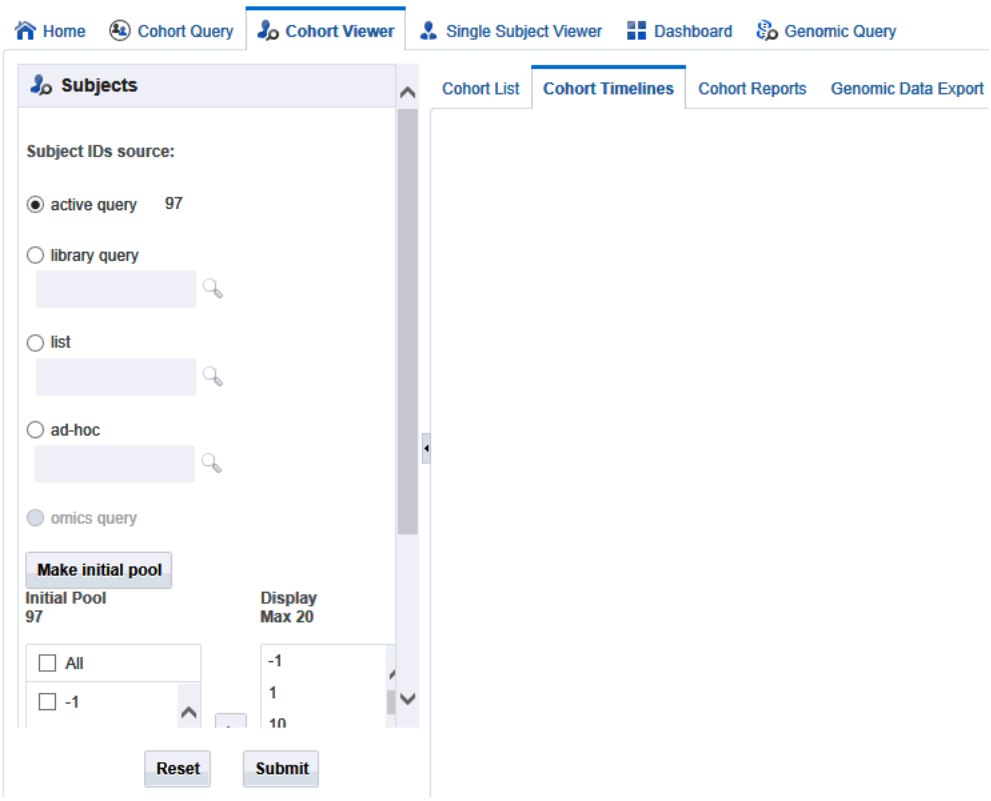
5.3 Cohort Timelines Viewer

To view patient or subject data in a list, CE has a Timelines Viewer which provides a method to view data for a small subset of patients in a more visual way. You can select data topics such as Diagnoses, Procedures, Medications to view the details and the system displays when specific activities occurred, in context with each other. Or, the view displays that a procedure was performed before or after a particular diagnosis was identified or a particular medication was taken by the patient or subject.

5.3.1 Selecting Patients or Subjects

The first step is to select the patients or subjects. This is similar to how you select patients or subjects for the **Cohort List**. The selection for the viewer is done by selecting patients or subjects either from the current active query which displays by default, from a query in the library, or by entering the **Patient or Subject ID** number for the patients or subjects you want to examine.

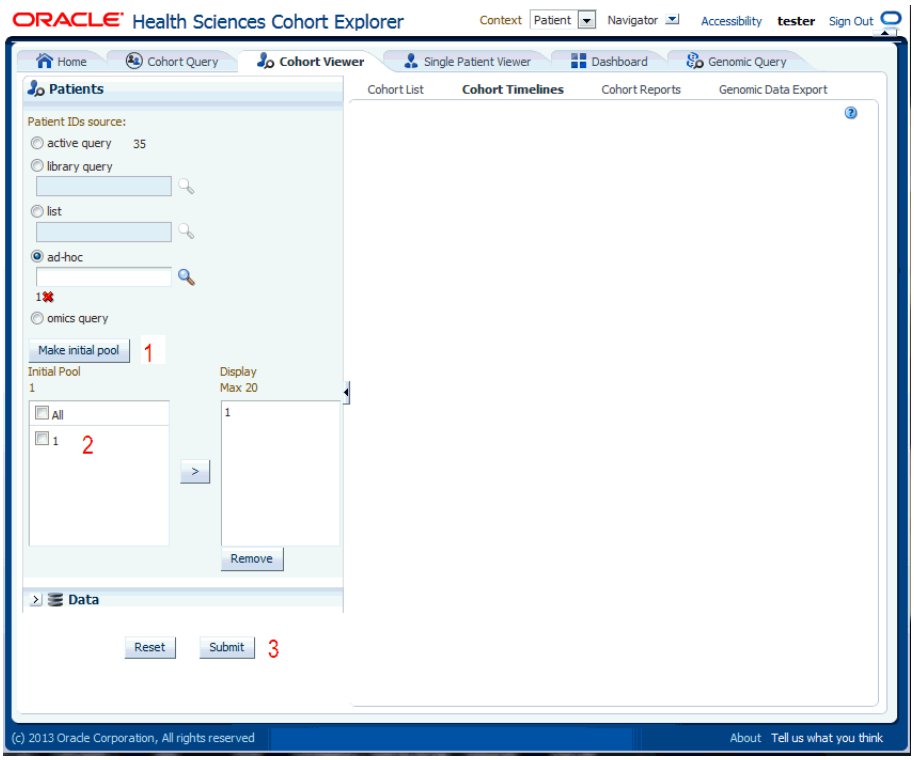
Figure 5–4 Selecting Patients or Subjects



To select patients or subjects to view perform the following steps:

1. The upper portion of the **Patients or Subjects** filter is where you specify patients or subjects from the active query, or library query, or particular **Patient or Subject ID**. Once you have chosen where the patients or subjects come from, you select the **Make Initial Pool** button (item 1 in Figure 5–5), and the patients or subjects to be added to the left hand list below the button. The last step is to select the patients or subjects to add to the **Initial Pool**.
2. The next step is to select up to 20 patients or subjects and move them from the **Initial Pool** to the **Display List**. This is done by clicking the right hand arrow (Item 2 in Figure 5–5). The **Display List** is what the system will reference for displaying the data in the timelines view.
3. Click **Submit** to view the corresponding data. However, you may select Clinical data for these Patients or Subjects, as outlined in the **Patient or Subject Pool** section.

Figure 5–5 Patient or Subject Pool

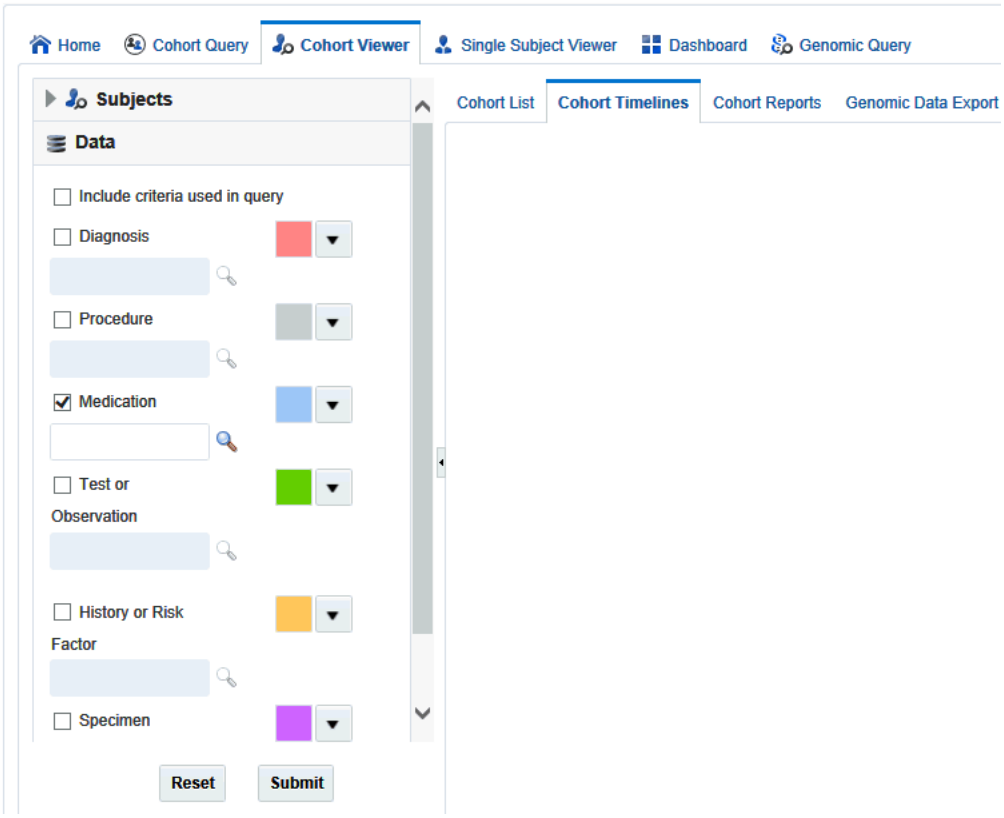


5.3.2 Selecting Data

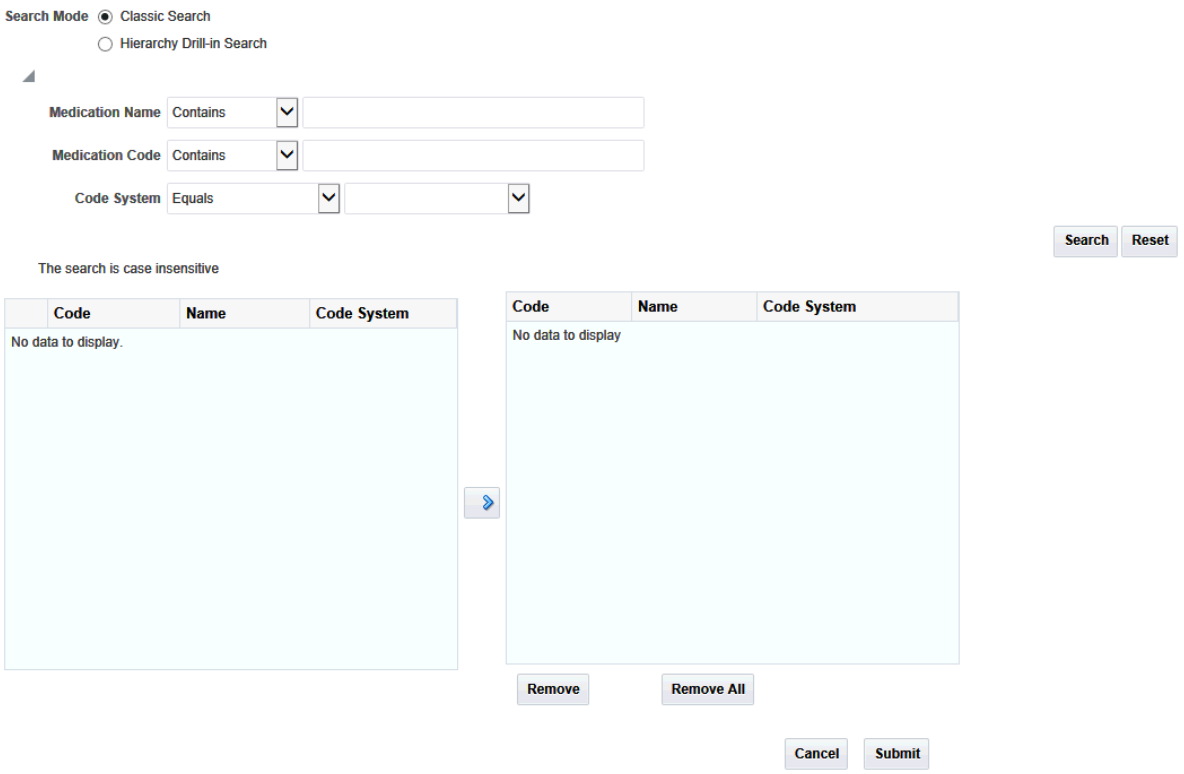
Once you have identified patients or subjects for the Display List, you then select the clinical information to be displayed in the viewer.

1. At the bottom of the left hand panel, click the arrow next to **Data**. The system displays a list of data topics.

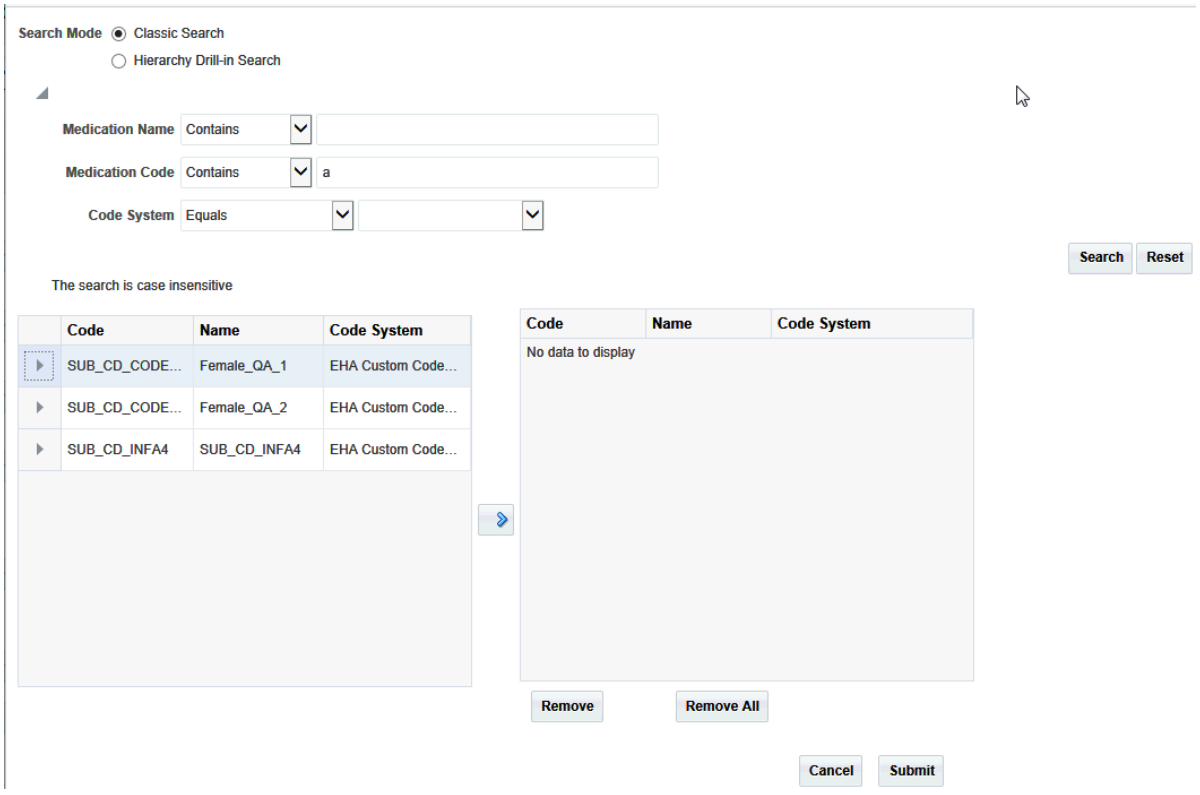
Figure 5–6 Selecting Data



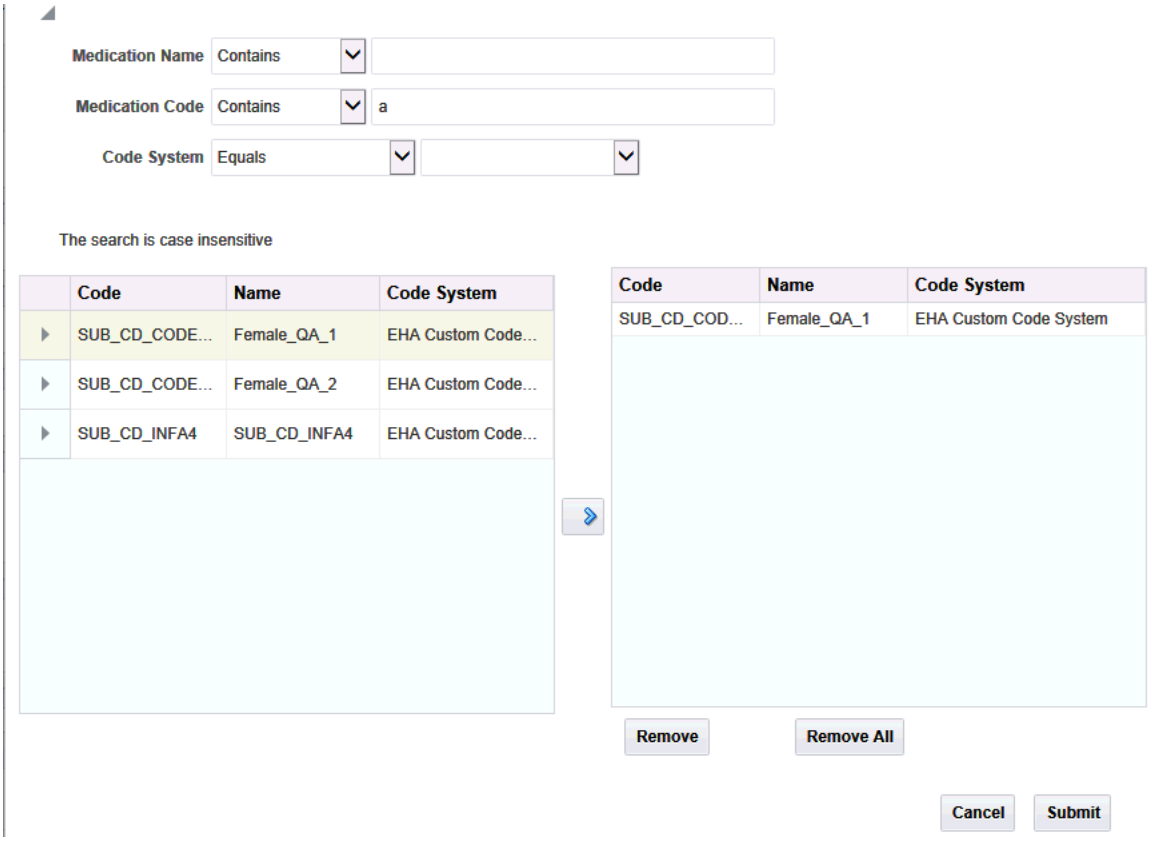
2. Select the box for each data topic you want to view. Then select the magnifying glass icon for each selected topic to search or specify the particular criteria. A popup search window is displayed.



3. For each topic, enter the Name, Code, or Code System.
4. Click **Search**.



5. Select one or more items and use the right hand arrow to move them to the right hand box.



6. Click **Submit**.

   Each data topic has a distinct color associated with it, because the data will be aligned in a timelines sequence, and the color serve as a visual separator between different data for the same patient or subject. CE has assigned default colors, but you can select any one of the drop down arrows to change that topic's color to suit your needs.

5.3.3 Displaying Patient or Subject Data

Once you have selected and specified the data you want to view, select **Submit** at the bottom of the list. The system displays a linear view of the selected data in the main window of the timelines viewer.

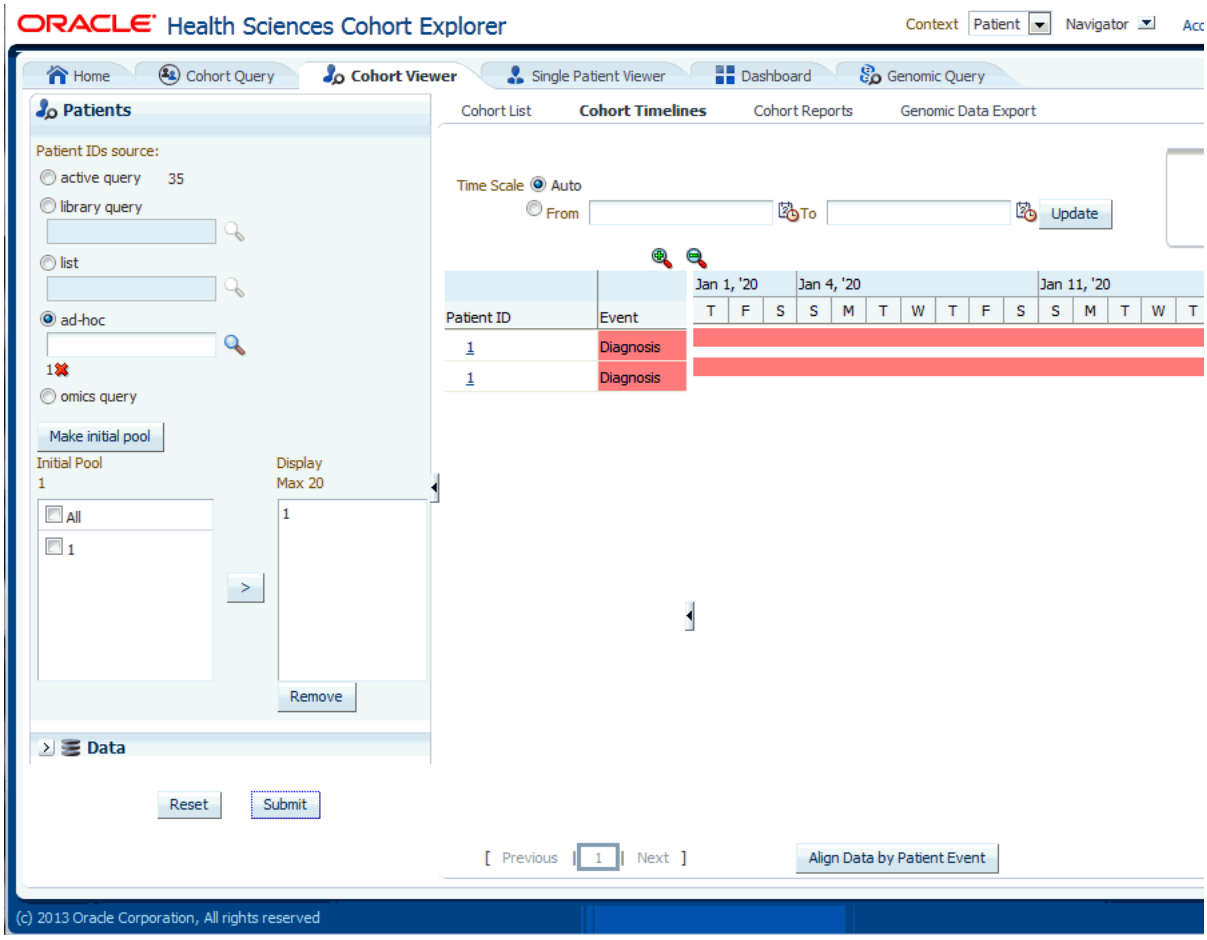
The display is subdivided into two sections, with a narrow vertical divider. The left side has attributes about the selected patients and the right side shows the colored layers of the selected clinical data, in a timeline view. You can move the divider left or right by hovering the cursor on the white vertical line. Select and hold down your mouse and then drag the divider to the desired position.

Also, you can hide the display of the patient or subject and data selection pane as well as the left hand side of the timelines view, by clicking the small left hand arrow on the right hand side of the respective area.

In Figure 5–7, the first column is the patient ID. Clicking it will navigate to the single patient viewer.

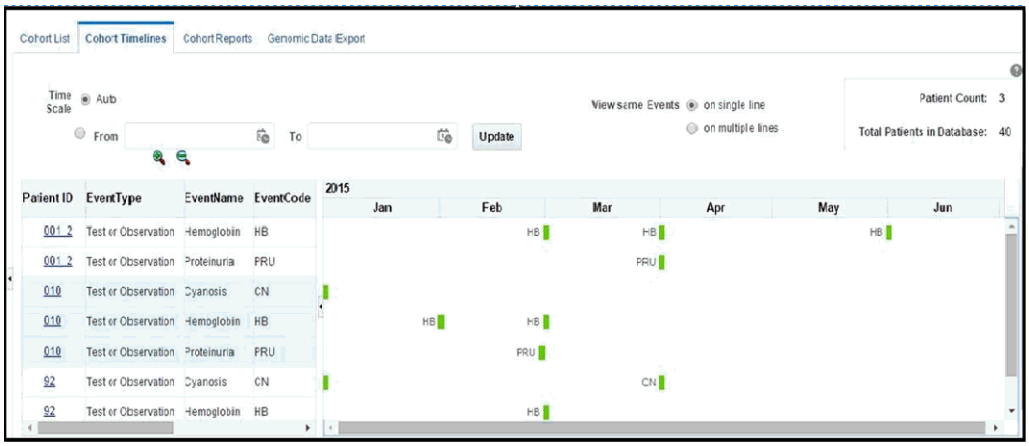
Note: If you are using Mozilla Firefox, click the patient ID. For Microsoft Internet Explorer and Google Chrome, double-click the patient ID.

Figure 5–7 Timelines



5.3.3.1 Selecting the Timeline Mode

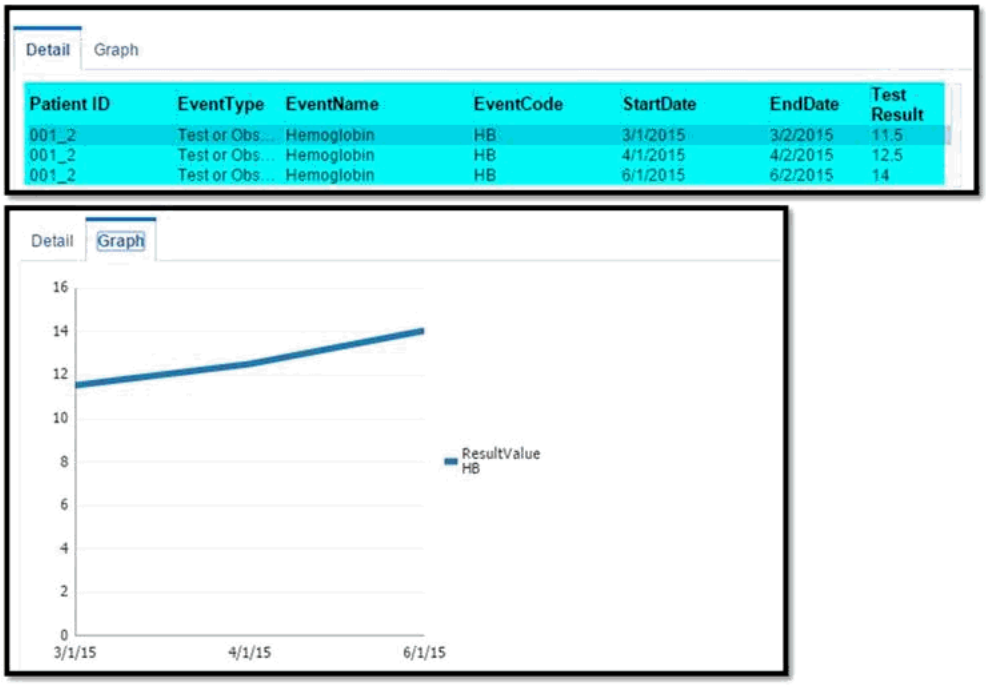
Figure 5–8 View Same Events



Using the option **View same Events**, you can display the events in one of the following modes:

Single Line

This is the default mode of displaying data on the Timelines tab.



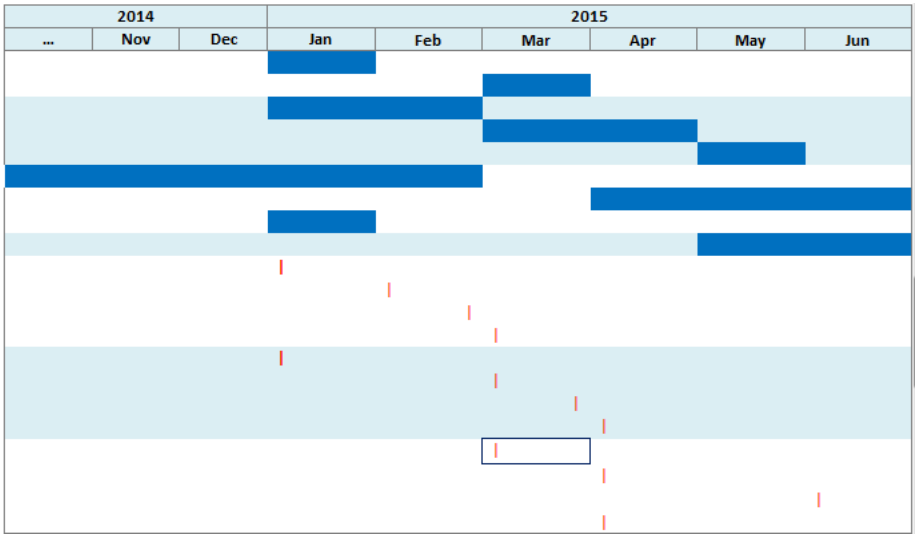
In this mode, all instances of a repeating event are displayed on the same line and separate events are displayed on separate lines. The events are considered to be repeating on the basis of Patient, Event Type and Event Code. For example, if a patient P1 has taken the medication M1 from 01-March to 31-March and then from 15-August to 30-September, then the Medication Event M1 is considered to be repeating for the patient and this mode displays the two occurrences of the medication event in the same line in the timeline.

If there is an overlap of events, for example, the start date of an event is the same as or earlier than the end date of the preceding occurrence of the event, then such overlapping occurrences are displayed in separate lines in the timeline. However, the left hand table displays a single row corresponding to all occurrences of the event.

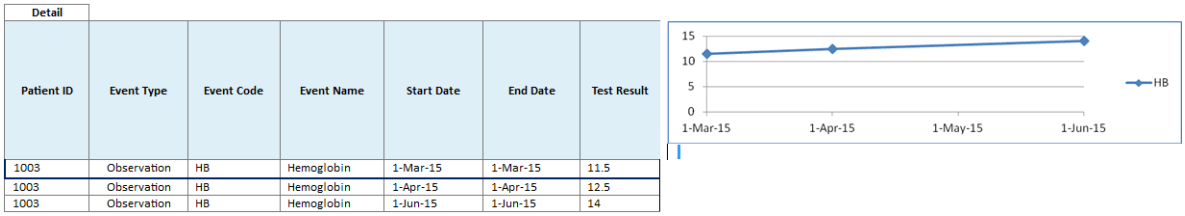
In the single line mode, the tabular display on the left side uses alternating background color to visually delineate data of different patients or subjects. All of the rows corresponding to a patient or subject, for a particular Event Type, have one background color and the next set of rows are in a different color.

Multiple Lines

Patient ID	Event Type	Event Code	Event Name	Age in Years	Gender Name
1001	Medication	M1	Medication 1	40	Male
1001	Medication	M2	Medication 2	40	Male
1002	Medication	M1	Medication 1	35	Female
1002	Medication	M2	Medication 2	35	Female
1002	Medication	M1	Medication 1	35	Female
1003	Medication	M3	Medication 3	50	Male
1003	Medication	M3	Medication 3	50	Male
1003	Medication	M2	Medication 2	50	Male
1004	Medication	M1	Medication 1	25	Female
1001	Observation	CN	Cyanosis	40	Male
1001	Observation	HB	Hemoglobin	40	Male
1001	Observation	PRU	Proteinuria	40	Male
1001	Observation	HB	Hemoglobin	40	Male
1002	Observation	CN	Cyanosis	35	Female
1002	Observation	HB	Hemoglobin	35	Female
1002	Observation	PRU	Proteinuria	35	Female
1002	Observation	CN	Cyanosis	35	Female
1003	Observation	HB	Hemoglobin	50	Male
1003	Observation	HB	Hemoglobin	50	Male
1003	Observation	HB	Hemoglobin	50	Male
1003	Observation	PRU	Proteinuria	50	Male



In this mode, all of the instances of events are displayed on different lines, even if the events are repeating. For example, the two occurrences of the medication event mentioned above are displayed in separate lines in this mode.



Detail

Patient ID	Event Type	Event Code	Event Name	Start Date	End Date	Test Result
1003	Observation	HB	Hemoglobin	1-Mar-15	1-Mar-15	11.5
1003	Observation	HB	Hemoglobin	1-Apr-15	1-Apr-15	12.5
1003	Observation	HB	Hemoglobin	1-Jun-15	1-Jun-15	14

Following are some details about timelines:

- If an event has both Start Date and End Date values available, then the timeline uses the Event Code to mark the start of an event.
- A tooltip on the timeline provides additional information about the event.
- Upon selecting an event in the timeline, the **Detail** table below the timeline lists all the occurrences of the event and provides additional information such as the start and the end date of the occurrences.

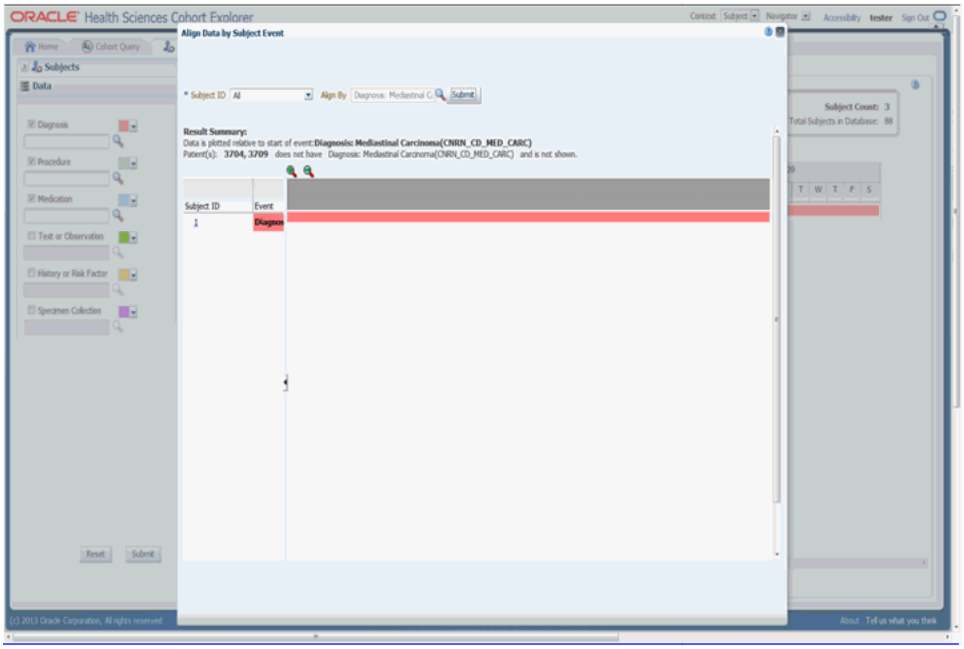
  Select only one Timeline Event at a time to render the corresponding Detail Table. If you select multiple Timeline Events, then the UI is likely to stop responding and may become unusable.
- For Test or Observation events with numeric result values, the **Detail** table displays the results of all the occurrences of the event. If two or more occurrences of Test or Observation event have numeric results available, then the **Graph** tab displays a line graph for those numeric result values.
- Since the Test or Observation event has only one date event, the user interface uses dummy End Date values that are 1 day later than the corresponding Start Date values. This is to support proper rendering of graph. The Detail table also displays these dummy End Date values.
- Events are not displayed on the left hand table or on the timeline in the following conditions:
  - When an event's Start Date is null.
  - When an event has only one Date attribute (for example, Observation has only Observation Date) and the attribute's value is null.
- If an event's End Date is null, then the System Date is used for displaying the event.

5.3.4 Align Data by Patient or Subject Event

Above the data display, there are two functional controls:

- The **Time Scale** option enables you to choose a specific date range for viewing the data. The default auto setting displays the data without a specific time reference.
- The **Align Data by Patient or Subject Event** tab enables you to designate a particular piece of clinical data to serve as the anchor, around which the remaining data is adjusted in the viewer. For example, if you choose a particular procedure as the anchor for alignment, medications or diagnostic tests be adjusted in the display to show how long before (or after) the anchor they were administered.
- When you select a particular diagnosis or medication code to align patients, all the event codes at different hierarchical nodes are considered for the selected event code.

Figure 5–9 Align Data by Subject Event



5.3.5 Including Criteria Used in Query Option

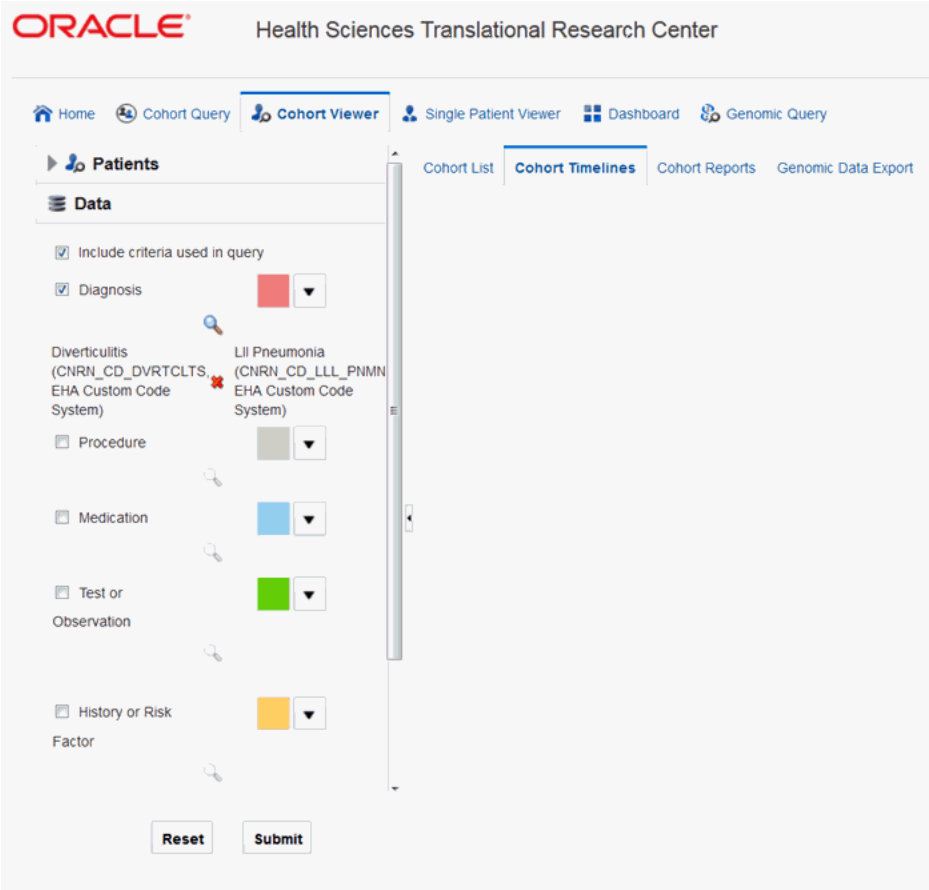
Using the **Include criteria used in query** check box, you can view the data used to define the inclusion criteria of the currently active query. The criteria defined in the cohort query will be displayed in the data section of the cohort timeline.

If the inclusion criteria of the active query is not based on any data topics supported on Cohort Timeline, a message is displayed to select at least one of the data elements supported by Cohort Timeline.

Note: The data topic *Clinical Encounter* can be selected only in conjunction with other data elements that are linked to Clinical Encounter, for example Diagnosis, Procedure and so on.

For example, in the Figure 5–10, **Diagnosis** has been defined as inclusion criteria in the cohort query screen. Once you select it, the corresponding diagnosis criteria is displayed. Then click **Submit** to view the corresponding timeline events.

Figure 5–10 Include Criteria Used in Query



5.4 Cohort Reports

This section lets you view various cohort reports.

> **Note:** In the Active Query mode, if no filters have been selected in Cohort Query, then the reports display data of only the first 10,000 patients or subjects. This limit is configurable and can be changed using the DEFAULT_ACTIVE_QUERY_LIMIT property in the TRC.properties file.
>
> However, irrespective of the patient or subject selection option, if the count exceeds the value specified by the GENOMIC.REPORTS_MAX_PATIENT_COUNT option of TRC.properties file, a warning message is displayed. You can either continue to plot with a large number of patient or subjects (might impact performance) or change the selected cohort to a smaller number of patients or subject.

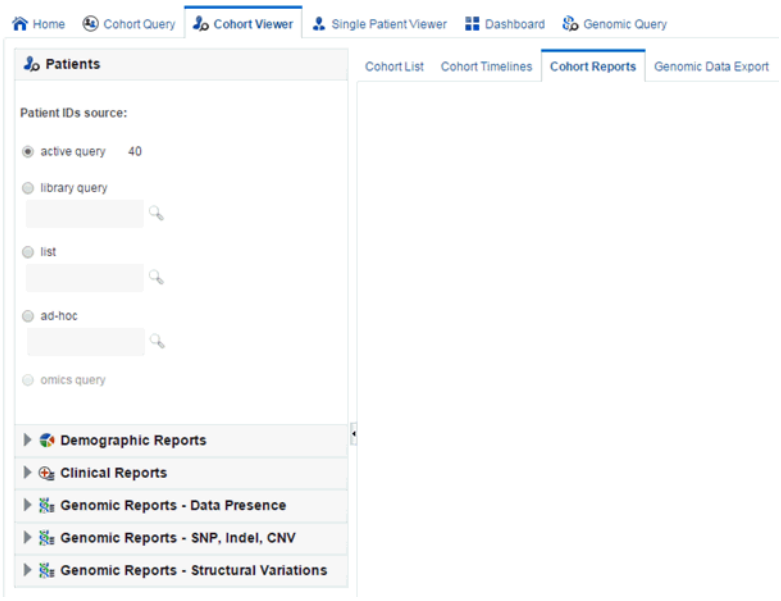
5.4.1 Demographic Reports

You can view demographic reports for a patient list. The queries for these reports can be based on patient list in the cohort query, the library, patient list, ad-hoc query or omics. Perform the following steps to view demographic reports:

1. Navigate to **Cohort Viewer** > **Cohort Reports** tab.

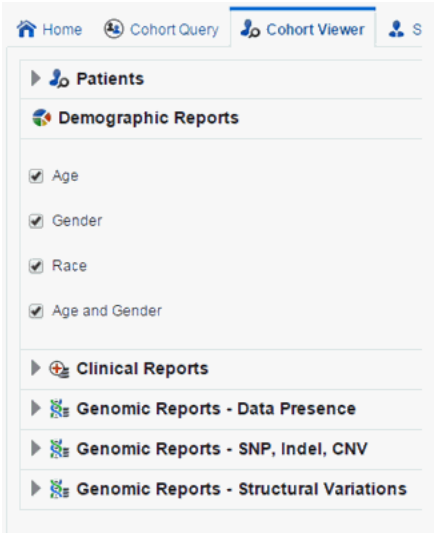
2. Expand the **Patients** node on the left.

Figure 5–11 Patient Node



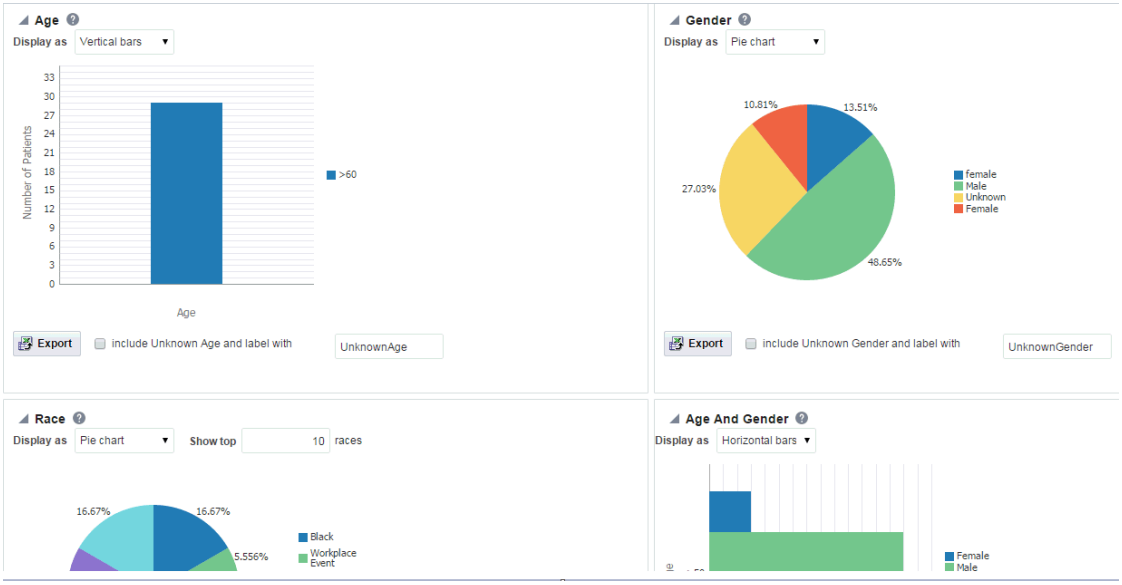
3. Select the **Patient IDs Source**.
4. Expand the **Demographic Reports** tab.

Figure 5–12 Select Demographic



5. Select the check boxes for the data you want displayed in your list.
6. Select **Submit**.
7. The system reruns the query and adds relevant patient or subject data for each patient or subject listed. Depending on you selection, reports age, gender, race and/or age and gender will be displayed. The data will be displayed as pie charts or bar graphs depending on your settings.

Figure 5–13 Demographics Report



5.4.1.1 Handling Unknown Data

If any missing or unknown data is present, a check box is provided at the bottom of each graph to include missing or unknown data as shown in Figure 5–13.

For the Age, Gender and Race graphs, if the check box is selected, the graph is refreshed with the unknown data. Here unknown data refers to age, gender or race details that are not defined for a particular patient.

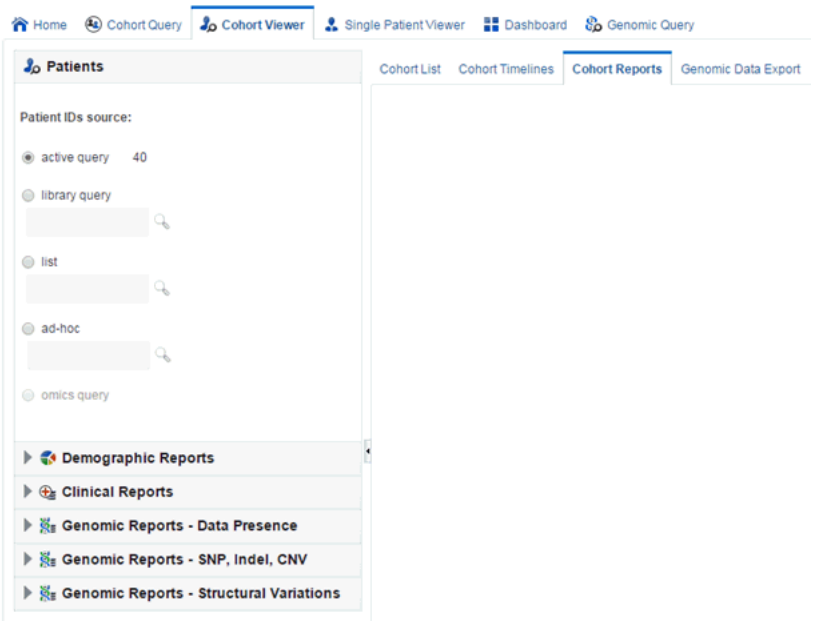
For the Age and Gender graph, the unknown age for different genders is combined into one group and the known age is in other group. Both groups are shown in the graph.

5.4.2 Clinical Reports

You can view clinical reports for a patient list. The queries for these reports can be based on patient list in the cohort query, the library, patient list, ad-hoc query or omics. Perform the following steps to view a clinical reports:

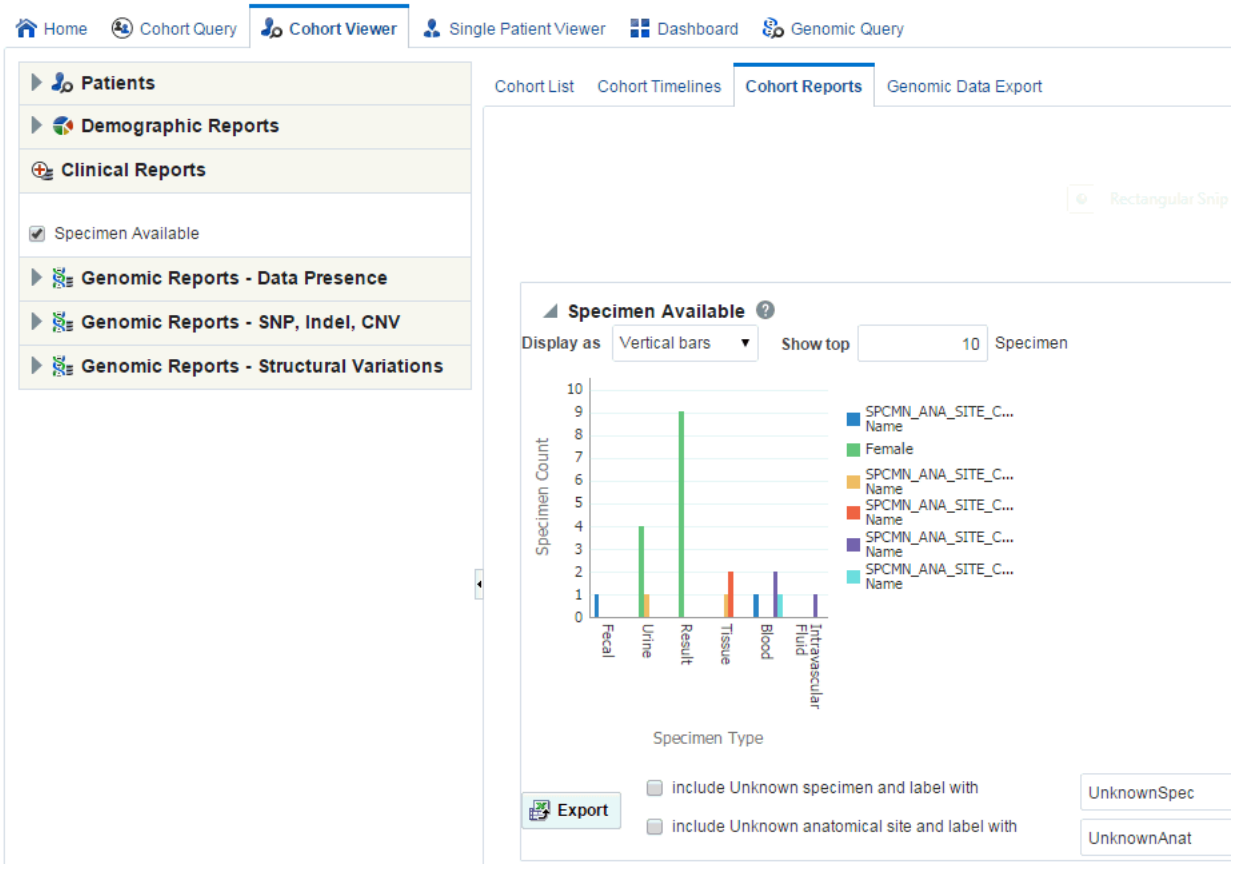
1. Navigate to **Cohort Viewer** > **Cohort Reports** tab.
2. Expand the **Patients** node on the left.

Figure 5–14 Patient ID Source



3. Select the **Patient IDs Source**.
4. Expand the **Clinical Reports** tab.
5. Select the **Specimen Available** check box.
6. Select **Submit**.
7. The system reruns the query and adds relevant patient or subject data for each patient or subject listed. Depending on you selection, reports age, gender, race and (or) age and gender will be displayed. The data will be displayed as pie charts or bar graphs depending on your settings.

Figure 5–15 Clinical Report



5.4.2.1 Handling Unknown Data

If any missing or unknown data is present, a check box is provided at the bottom of each graph to include missing or unknown data as shown in Figure 5–15.

If **Include Unknown Specimen and Label With** or **Include Unknown Anatomical site and Label With** are selected, the graph is refreshed with the unknown data. Here unknown data refers to the specimen and anatomical site data that is not defined for a particular patient.

5.4.3 Genomic Reports

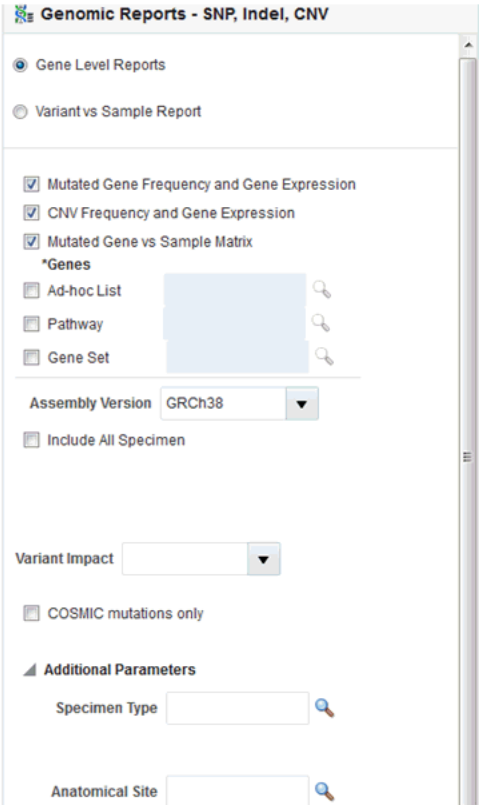
5.4.3.1 Changes in Genomic Reports

The following changes have been added to Genomic Reports in TRC 3.1:

- An additional option to **Include All Specimen** is provided in all reports. This enables you to toggle while taking into account specimens, without genomic data, when generating reports.
- Genomic Reports data is now filtered by Assembly; with a filter drop down provided for Assembly version.
- A new node has been added to display a report on genomic data presence.

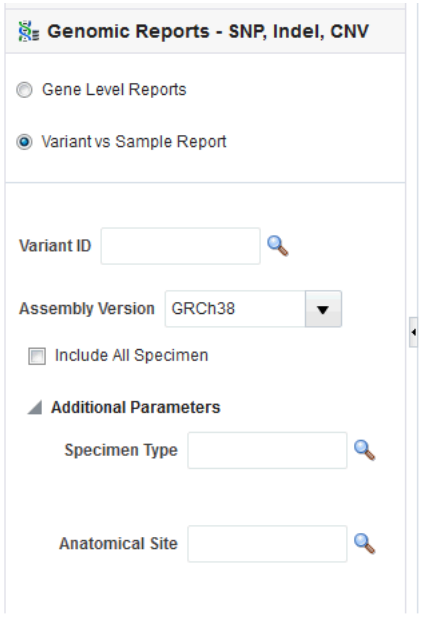
- The Genomic Report—SNP, Indel, CNV—now has two reporting options, for **Gene Level Reports** and **Variant vs. Sample Reports**.
- Mutated Gene vs. Sample Matrix has additional variant filters; a multi-select drop-down list Variant impact filters, and a check box to limit to Comic mutations only.

Figure 5–16 SNA, Indel, CNV Gene-Level Filter Panel



- The **Variant vs. Sample Report** provides the option of reporting the presence of variants in specimen samples by a user input of Variant IDs.

Figure 5–17 SNA, Indel, CNV Variant vs Sample Filter Panel



5.4.3.2 Data Presence

The Genomic Reports - Data presence lets you view a breakup of the patient cohort as follows:

- Patients in the cohort with specimens collected that have accessible genomic data
- Patients in the cohort with specimens collected but no genomic data
- Patients in the cohort with no specimens collected

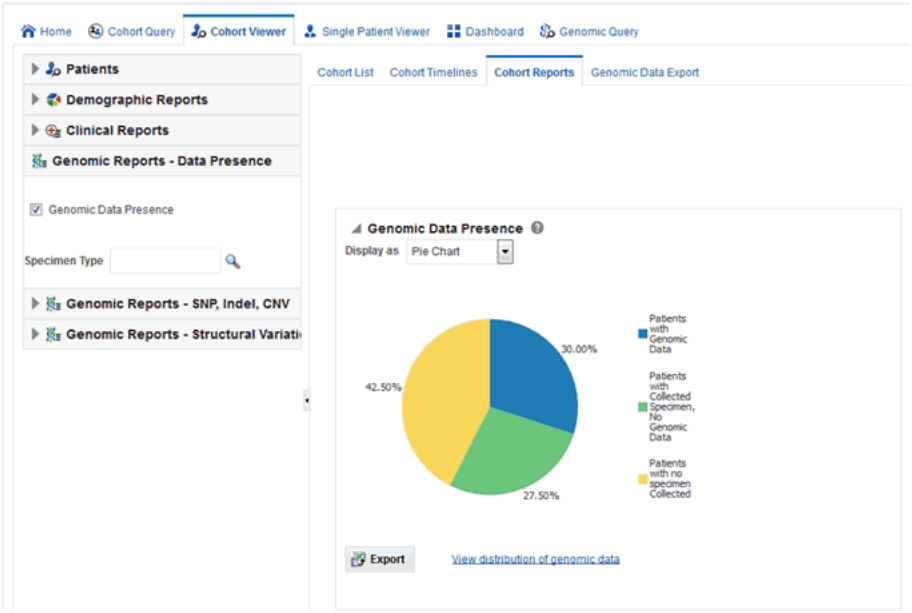
You can view the Data Presence plot as Pie Chart, Horizontal Bars, Vertical Bars, and Table. The report also provides an option to view the distribution of present genomic data along the following available genomic data types:

- Sequencing
- RNA-seq
- Gene Expression
- Copy Number Variation

Perform the following steps to view a Genomic data presence report:

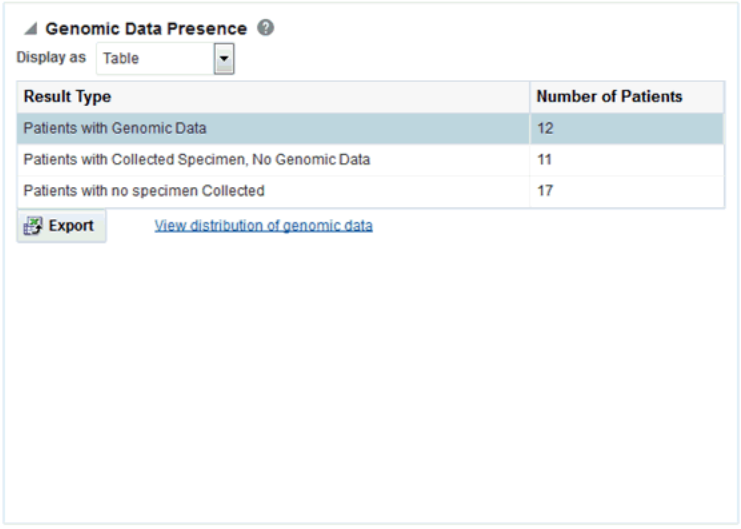
1. Expand the **Genomic Reports - Data Presence** node in the left panel.
2. Select the **Genomic Data Presence** check box. An optional **Specimen Type** input field appears to allow for filtering the initial cohort on specimen type.
3. Click **Submit** to generate the report.

Figure 5–18 Genomic Data Presence



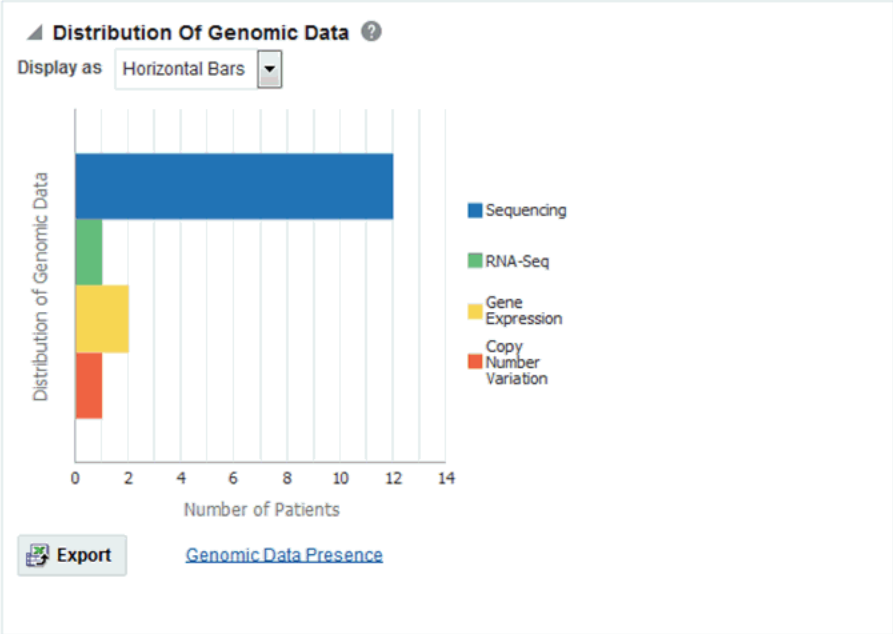
4. To change the display type after generating the report, select display option from the **Display as** drop-down list.

Figure 5–19 Data Presence - Table Display



5. To view a breakup of present genomic data by data type, click **View distribution of genomic data**.

Figure 5–20 Distribution of Genomic Data



6. To export the graphical plot type displays as a PNG image, and the table type displays as a Excel spreadsheet, click **Export**.

5.4.3.3 SNP, Indel and CNV

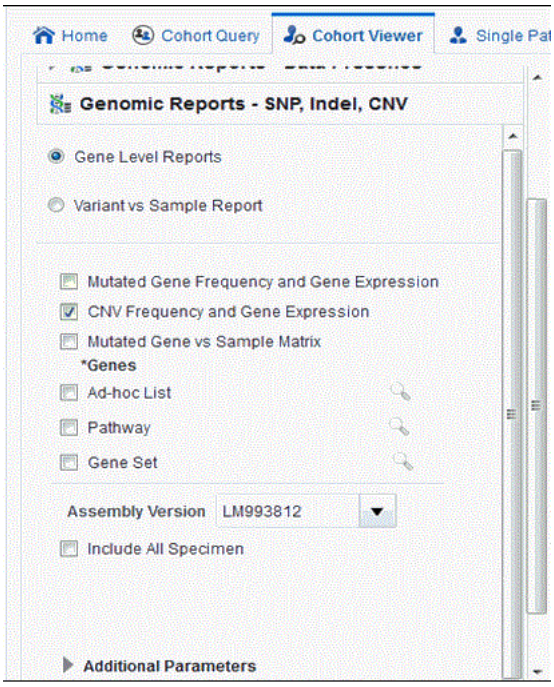
5.4.3.3.1 Gene Level Reports - Mutated Gene Frequency and Gene Expression

Genomic Reports - SNP, Indel and CNV display the available SNP Indel genomic reports based on the selected cohort of patients or subjects, if in subject context.

First, you must select the source of patient or subject identifiers as shown in Figure 5–21. The source of patients or subjects is same for all cohort viewers and consists of one of the following option:

- active query from Cohort Query interface
- saved query from a query library
- saved list of identifiers
- ad-hoc list of identifiers
- list of patient or subject IDs based on a query performed through the Genomic Query tab

Figure 5–21 Select Source of Patient or Subject Identifiers

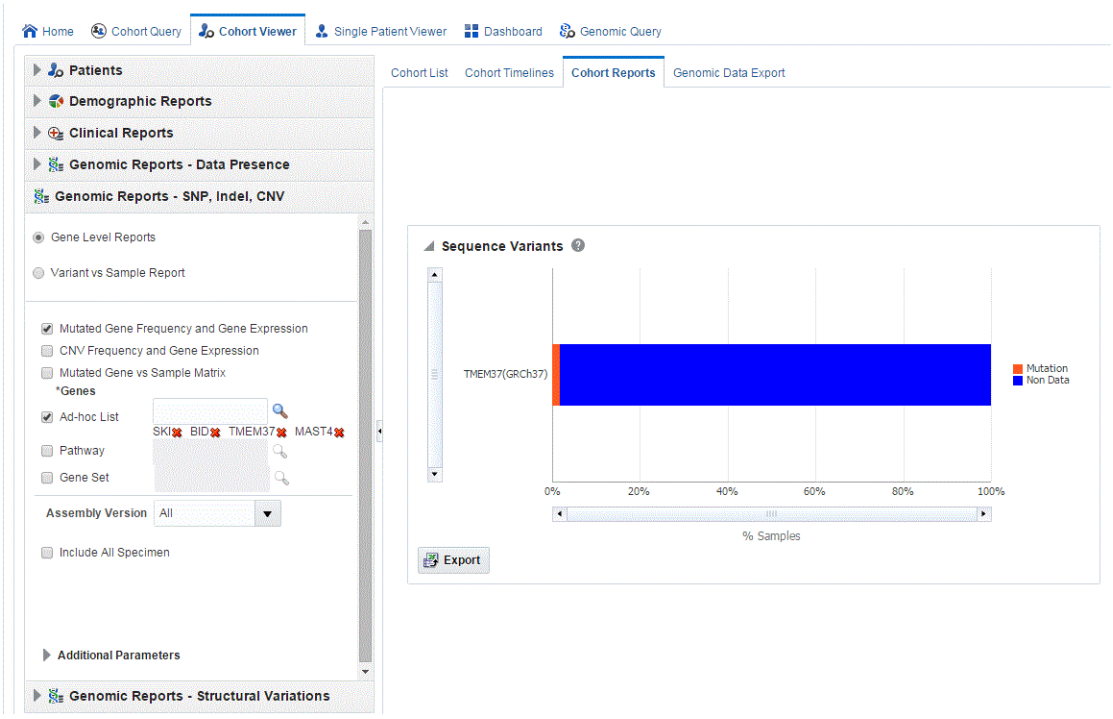


Next, you must show the Mutated Gene Frequence and Gene Expression report under Genomic Reports - SNP, Indel CNV category as shown in Figure 5–22. You can also opt to add additional parameters such Specimen Type, Anatomical Site, Assembly Version which will only consider results linked to the selected categories. Also, the check box **Include All Specimen** which when unchecked (this is the default value) will return a result which has only genomic data. If this option is selected, the result will include specimen without genomic data. Once you click **Submit**, a histogram report will show the percentage of samples for the relevant cohort which have sequence variants information within the selected genes.

You can display the results as horizontal bars, vertical bars, or as a table. You can also export the results into pdf if bars are exported or into Excel if table is exported.

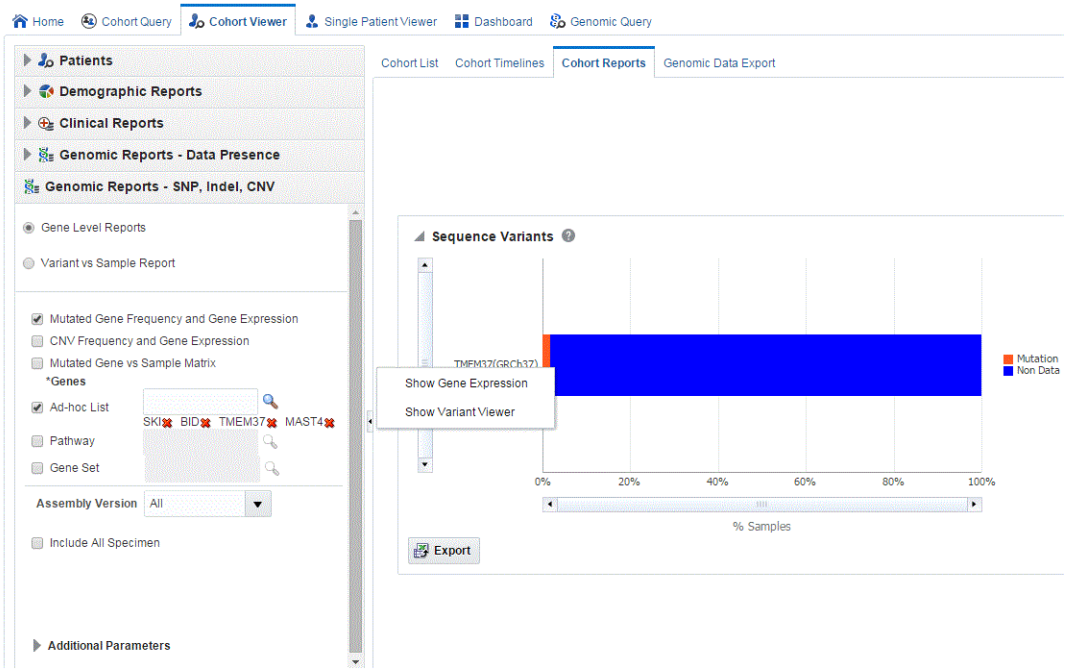
Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to `http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34`.

Figure 5–22 Show the Sequence Variants



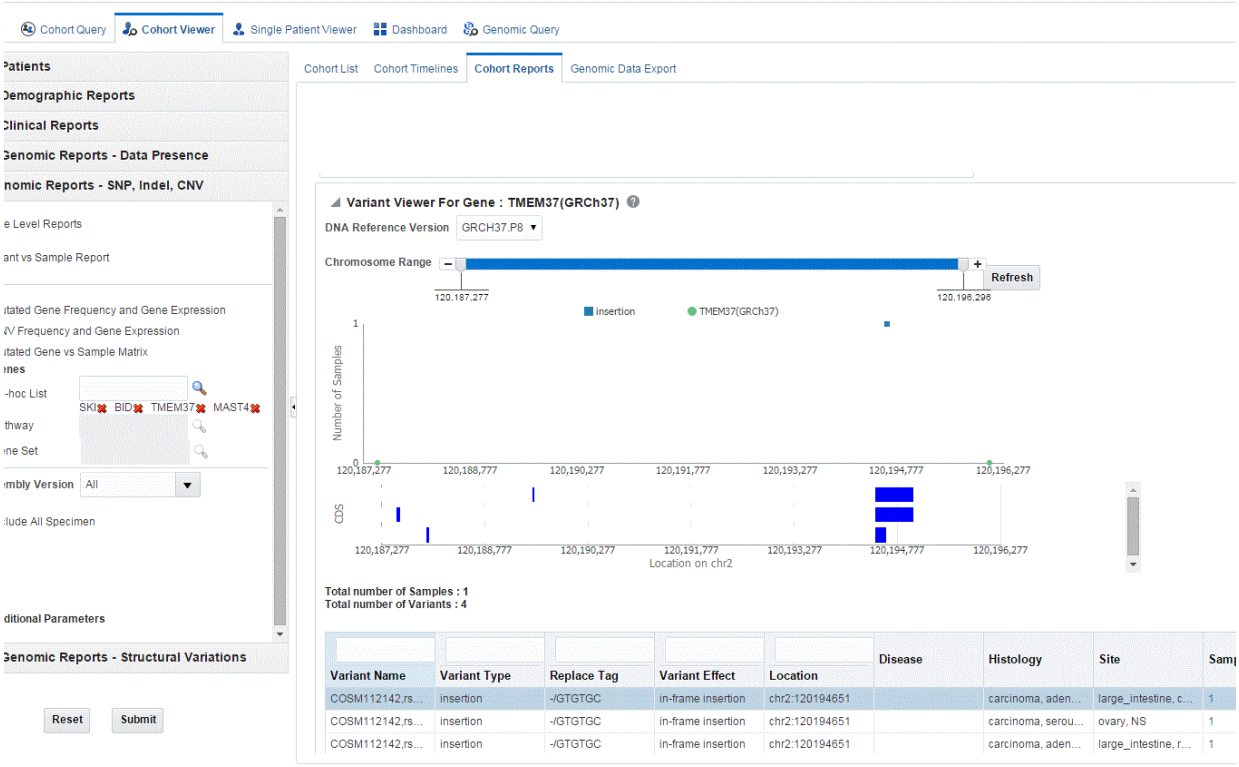
On clicking the histogram that is displayed for the different selected genes, a popup is displayed to either get the Gene Expression plot or the Variant Viewer to display details of the variant.

Figure 5–23 Popup



Selecting **Show Gene Expression** link in the popup, displays the Gene Expression details for Single Channel or Dual Channel. In the Single Channel Expression, you are required to provide hybridization details to get the information. If the selected cohort and genes do not have any sequencing data associated with them, the plot is not rendered and gene expression information cannot be plotted for that selection. The gene expression plot is displayed only when the specimens of the selected cohort have both sequencing and gene expression data.

When you click **Show Variant Viewer** plot, a plot similar to the following is rendered:

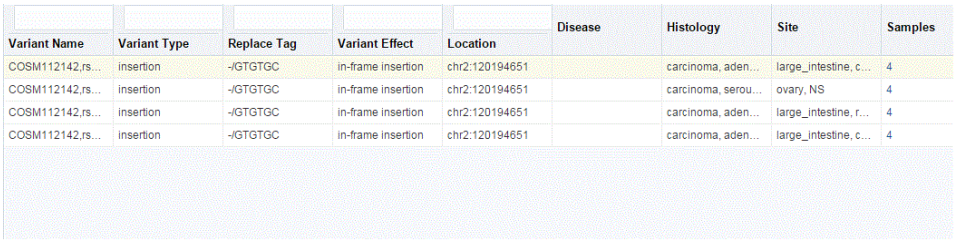


The variant viewer displays variants based on the annotation (gene coordinates) of the selected reference version. If the selected gene coordinates of the DNA Reference version change, then there can be a difference in the specimen and variant counts in the report.

This displays necessary information about all variants present in that gene in the samples belonging to the selected cohort of subjects or patients. The Mutation plot displays the total number of specimen (on *y*-axis) having a particular mutation (on *y*-axis) falling in the selected gene for belonging to a specific reference. The *x*-axis contain the range of gene region including the flanking region.

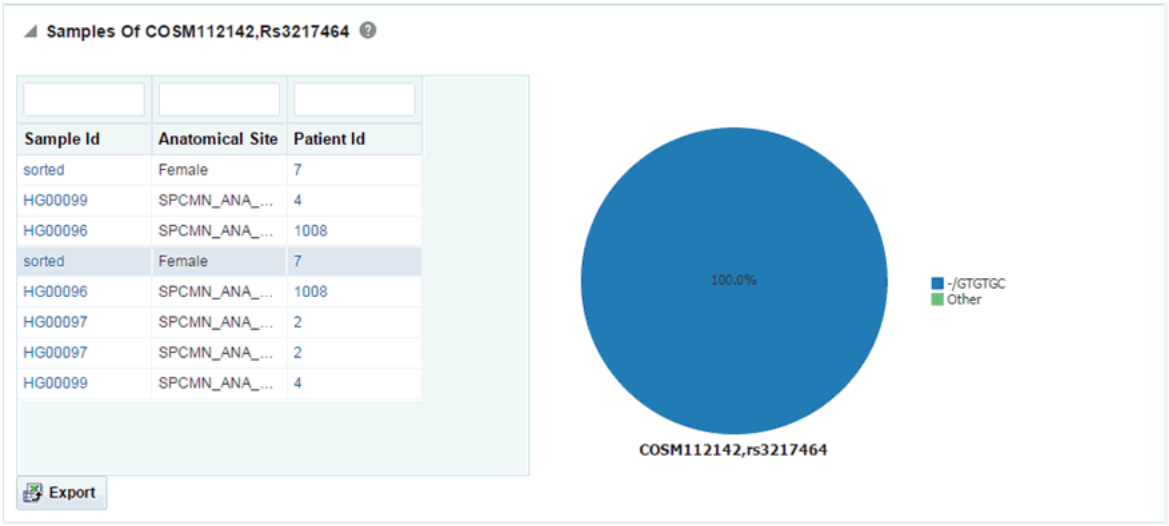
The CDS plot displays bars of CDS regions for each of the Ensembl Transcripts belonging to the selected gene of a selected reference version. Each row in the CDS plot represents one Transcript ID of the selected gene and reference version.

The following table represents the details of each variant like Variant name, Variant Type, Replace Tag, Variant Effect if present, Location, Disease if loaded, Histology, Site Total number of samples, Patients. The results are displayed in descending order of specimen count, such that each row contains unique variants.



Variant Name	Variant Type	Replace Tag	Variant Effect	Location	Disease	Histology	Site	Samples
COSM112142,rs...	insertion	-/GTGTGC	in-frame insertion	chr2:120194651		carcinoma, aden...	large_intestine, c...	4
COSM112142,rs...	insertion	-/GTGTGC	in-frame insertion	chr2:120194651		carcinoma, serou...	ovary, NS	4
COSM112142,rs...	insertion	-/GTGTGC	in-frame insertion	chr2:120194651		carcinoma, aden...	large_intestine, r...	4
COSM112142,rs...	insertion	-/GTGTGC	in-frame insertion	chr2:120194651		carcinoma, aden...	large_intestine, c...	4

You can drill down to each mutation by clicking **Number of samples** present in the samples column of the table. Once clicked, a table and a pie chart are displayed below the table. The table represents the details like Sample ID, Study (only in subject context), Anatomical Site and Patient or Subject ID, which contains the selected mutation, number of samples which do not contain this mutation (labeled as Other) from the group of samples having mutations for the selected gene and selected reference version.



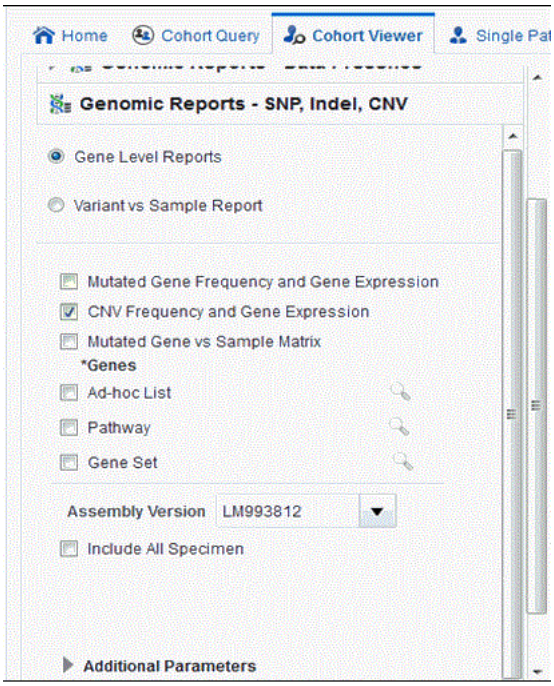
5.4.3.3.2 Copy Number Variation Frequency and Gene Expression

CNV Frequency and Gene Expression is displayed on the Genomic Reports -SNP Indel genomic reports available to the end user based on the selected cohort of patients or subjects, if in subject context.

First, you must select the source of patient or subject identifiers as shown in Figure 5–24. The source of patients or subjects is same for all cohort viewers and consists of one of the following option:

- active query from Cohort Query interface
- saved query from a query library
- saved list of identifiers
- ad-hoc list of identifiers
- list of patient or subject IDs based on a query performed through the Genomic Query tab

Figure 5–24 Select Source of Patient or Subject Identifiers

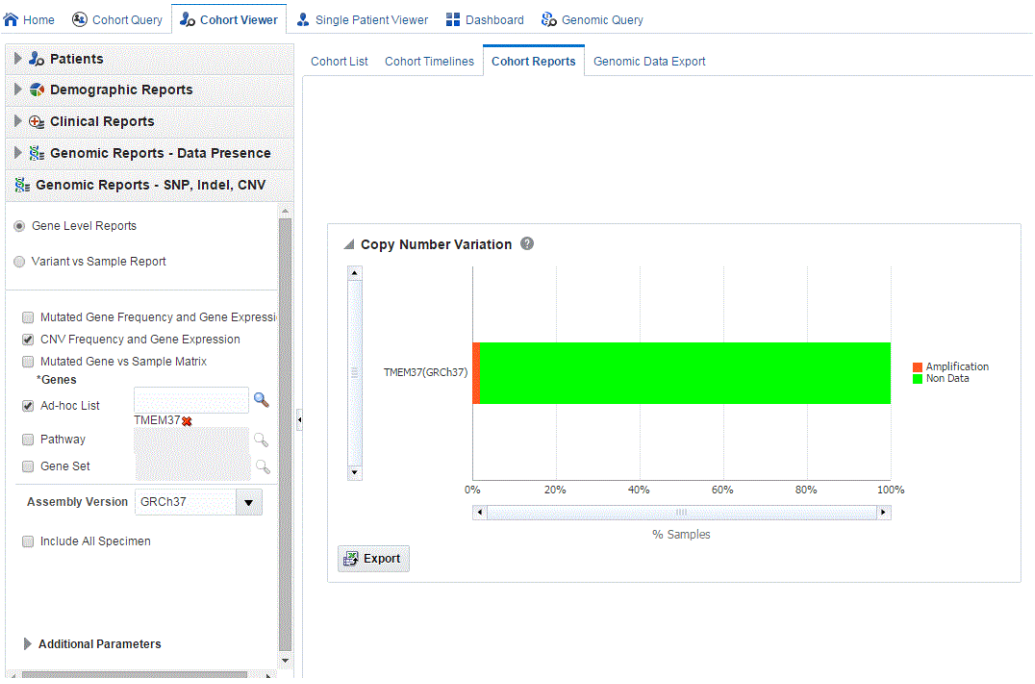


Next, select to show the Copy Number Variation report under Genomic Reports - SNP, Indel CNV category as shown in Figure 5–25. You can also opt to add additional parameters such Specimen Type, Anatomical Site, Assembly Version, which will only consider results linked to the selected categories. Also, the check box **Include All Specimen** which when unchecked (this is the default value) will return a result which has only genomic data. If this option is selected, the result will include specimen without genomic data. After you click **Submit**, a histogram report will show the percentage of samples for the relevant cohort which have copy number variants information within the selected genes.

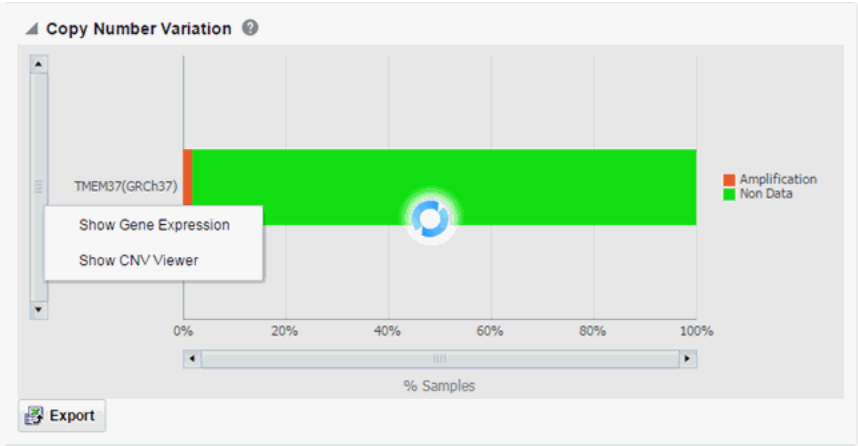
You can display the results as horizontal bars, vertical bars, or as a table. You can also export the results into pdf if bars are exported or into Excel if table is exported.

Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to `http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34`.

Figure 5–25 Show the Copy Number Variations

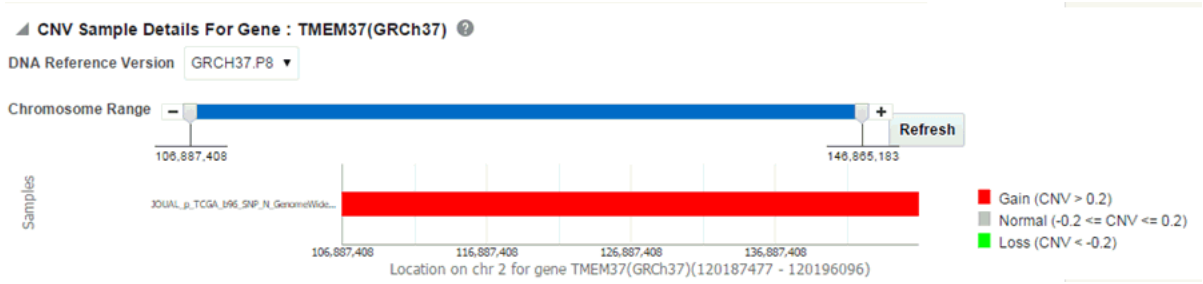


On clicking the histogram that is displayed for the different selected genes, a popup is displayed to either get the Gene Expression plot or the CNV Viewer to display details of the variant.



Selecting **Show Gene Expression** in the popup, displays the Gene Expression details for Single Channel or Dual Channel. In the Single Channel Expression you must provide hybridization details to get the information. If the selected cohort and genes do not have any sequencing data associated with them, the plot is not rendered and gene expression information cannot be plotted for that selection. The gene expression plot is displayed only when the specimens of the selected cohort have both CNV and gene expression data.

When you click **Show CNV Viewer** plot, then the plot similar to the following is rendered:



This displays necessary information about all CNV variants that are present in that gene in the samples belonging to the selected cohort of subjects/patients. The CNV plot displays each specimen (on *y*-axis) and its values falling in the selected gene for belonging to a specific reference.The *x*-axis plots the range of gene region. The plot is color coded with CNV with Gain (CNV > 0.2) is Red, Normal (-0.2<CNV<0.2) in grey and Loss (CNV < -2) in Green.

The CNV table gives the details about the CNV data like Samples, File Type, Chromosome, Anatomical Site, Start Position, End Position, CNV Value and Patient or Subject ID.

Sample Id	File type	Anatomical Site	Chromosome	Startposition	Endposition	CNV Value	Patient Id
JOUAL_p_TCGA...	CNV .seg file	SPCMN_ANA_SI...	2	106887408	146865183	0.205	6

5.4.3.3.3 Mutated Gene vs Sample Matrix

The Mutated Gene vs Sample Matrix plot displays a high level pictorial view of the presence of specific variants in genes of interest across various specimens of patients or subjects in a cohort. You can see if a particular gene of a specimen has selected mutations or no-mutations or no-data loaded. The same plot lets you view CNV data on the genes for the specimens.

The specimens are grouped together for each patient or subject and are ordered based on their collection date.

You can also view *patient_id:specimen_number(specimen_vendor_number):collection_dt:no of vairants* in the tooltip for each data point.

Variant selection is based on variant impact, implying that the depiction of mutations on genes for each specimen is based on the variants having selected variant impacts. For example, if you select *frameshift* mutation for EGFR, then the query searches for variants causing frameshift impact for EGFR gene and reports as *mutation* if it finds any such variants. If there is only non-variant information, then it reports as non-mutation in the plot. If there is no data, then it shows no-data in the plot.

Following are the definitions of various types of data in the plot.

- Mutation: If any of the variants are present in the selected gene of a specimen, then it is reported as mutation.
- No Mutation: If none of the variants are present in the selected gene of a specimen and there is non-variant information available for that gene, then it is reported as No Mutation.
- No data: If there are no variants and also no non-variant information and there is no-call data for the selected gene of a specimen or there is no-data loaded for this specimen, then it will be reported as no data.

- Amplification: If there is CNV data, which has seg_mean value more than zero, then it is reported as amplification.
- Deletion: If there is CNV data which has seg_mean value less than zero, then it is reported as deletion.

To view this plot:

1. Select the source of patient or subject identifiers as shown in Figure 5–24. The source of patients or subjects is same for all cohort viewers and consists of one of the following:
   - Active query from Cohort Query interface
   - Saved query from a query library
   - Saved list of identifiers
   - Sd-hoc list of identifiers
   - List of patient or subject IDs based on a query performed through the Genomic Query tab
2. Select to show the Mutated Gene vs Sample Matrix report under Genomic Reports - SNP, Indel CNV category. Select genes as shown in Figure 5–25. Data is filtered using the Variant Impact or the cosmic mutation selected.You can also add additional parameters such Specimen Type, Anatomical Site, Assembly Version, which will only consider results linked to the selected categories. Also, the check box **Include All Specimen** which when unchecked (this is the default value) will return a result which has only genomic data. If this option is selected, the result will include specimen without genomic data.
3. After you click **Submit**, a report displays the percentage and details of samples for the relevant cohort which have sequence variants and copy number variants information within the selected genes.

Figure 5–26 Mutated Gene vs Sample Matrix Report

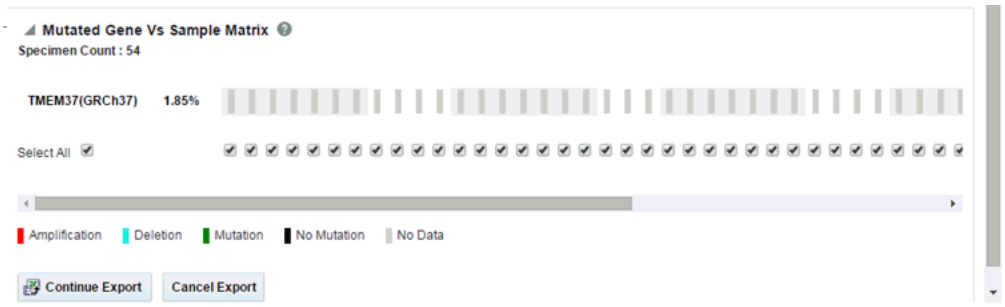
Each bar represents a specimen, which is grouped patient wise and in the order of the specimen collection date. The Sequence Variants and CNV information for each specimen is displayed. The percentage of specimens with mutation on the gene is also mentioned. Export functionality is provided at each specimen level.

For improved performance, data for only 10 genes is displayed at a time on the plot. Use the **Next** and **Previous** links to retrieve information for other set of genes. The export functionality also includes only the data of the 10 genes that is displayed in the above plot.

If the selected genes do not belong to the assembly provided for plotting, then a message stating that no CBIO data is available for selected criteria is displayed.

Note:

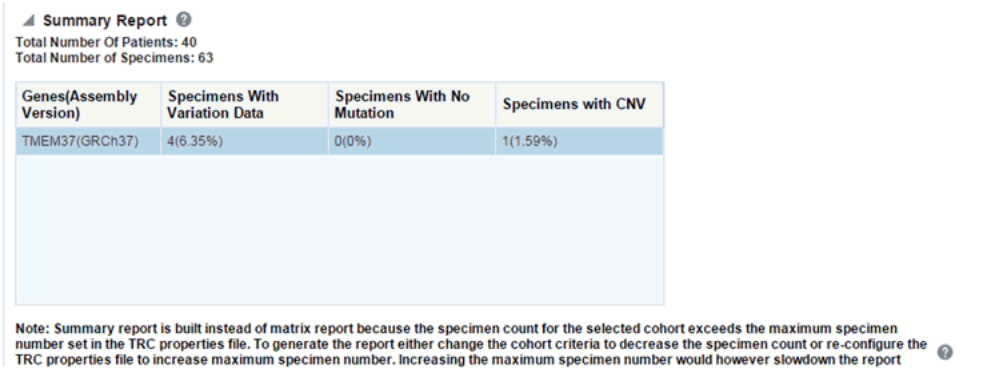
- In rare cases, less than 10 genes might be displayed because some of the selected genes may have multiple identifiers as they are placed in different chromosomes.The data for these is clubbed together as a single gene record on the plot.
- The best performance is observed for 10 genes and specimen count less than 300. As the data increases, the performance decreases linearly. Also, performance degrades with the presence of non-variant data.



Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34.

There is a limit to the number of specimens that can be seen in the gene matrix plot. This is an application level parameter called 'MAX_SPEC_REPORT' that is defined in the TRC.properties file. The default value of this parameter is 1000. If the number of specimens in the cohort used for this analysis is greater than the specified value, then the following warning message is displayed:

If you continue from this warning message, the report will display the summary statistics instead of the matrix plot. You may have to decrease the cohort size based on some criteria like specimen type or anatomical site or otherwise use cohort query to view the matrix plot. Alternatively, you can also change the default value of MAX_SPEC_REPORT parameter to a desired value and rebuild the plot. However, rebuilding maybe affect the performance of the plot generation.



The summary report only displays the percentage of specimens with different categories of data as shown in the image above.

5.4.3.4 Variant Level Reports

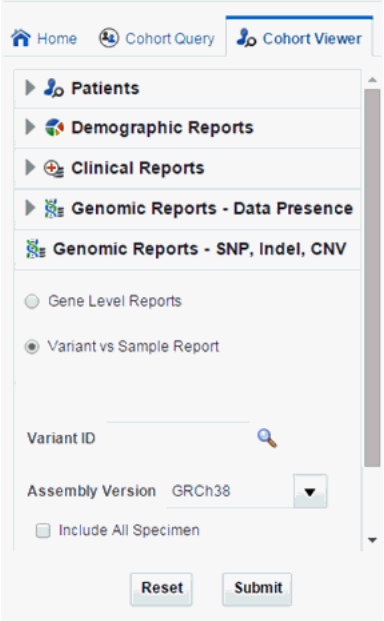
The **Variant Level Reports** option is displayed on the Genomic Reports - SNP Indel genomic reports available, based on the selected cohort of patients or subjects, if in subject context.

It provides a high level pictorial view of the presence of specific variants of interest across various specimens of patients or subjects in a cohort.

The specimens are grouped together for each patient or subject and are ordered based on their collection date. You can also view *patient_id:specimen_number(specimen_vendor_number):collection_dt* in the tooltip for each data point.

To generate the plot:

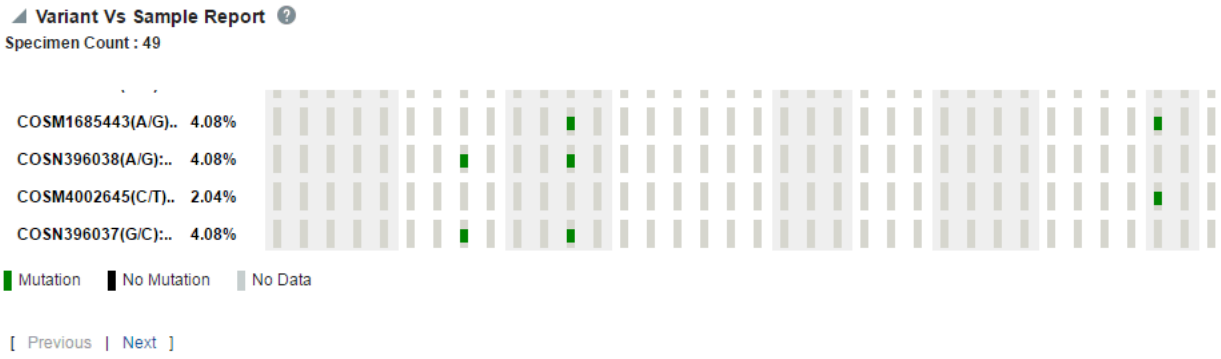
1. First, select the source of patient or subject identifiers as shown in Figure 5–24. The source of patients or subjects is same for all cohort viewers and consists of one of the following:
   - Active query from Cohort Query interface
   - Saved query from a query library
   - Saved list of identifiers
   - Sd-hoc list of identifiers
   - List of patient or subject IDs based on a query performed through the Genomic Query tab
2. Select to show the Variant Level Reports under Genomic Reports - SNP, Indel CNV category. Select variants as follows.



You can also opt to add additional parameters such Specimen Type, Anatomical Site, Assembly Version, which will only consider results linked to the selected categories. Also, the check box **Include All Specimen** which when unchecked (this is the default value) will return a result which has only genomic data. If this option is selected, the result will include specimen without genomic data.

3. After you click **Submit**, a report displays the percentage and details of samples for the relevant cohort, which have mutation within the selected genes.

Figure 5–27 Variant vs Sample Report



Each bar represents a specimen, which is grouped together patient wise and in the order of the specimen collection date. The mutation information for each specimen is displayed. Each row represents one variant and the label represents reference_ id, assembly version and the variant replace_tag value to maintain the uniqueness of the variant. The percentage of specimens for each mutation is also mentioned.

For improved performance, data for only 10 genes is displayed at a time on the plot. Use the **Next** and **Previous** links to retrieve information for other set of genes. The export functionality also includes only the data of the 10 genes that is displayed in the above plot.

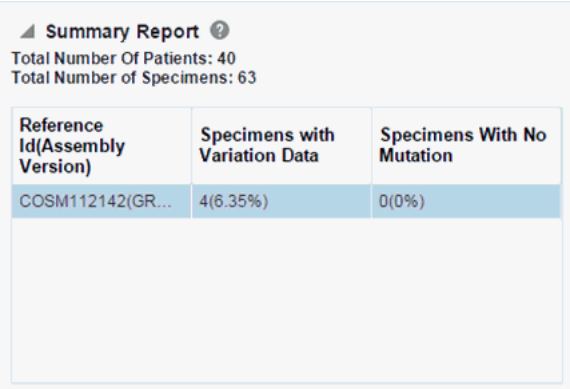
If the selected variants do not belong to the assembly provided for plotting, then a message stating that no CBIO data is available for selected criteria is displayed.

Note: The performance of the variant matrix report degrades if variants are encountered in the intergenic region, which will result in the query not being able to use partitioning correctly.

Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34.

There is a limit on the number of specimens that can be seen in matrix plot. This is an application level parameter called 'MAX_SPEC_REPORT' that is defined in the TRC.properties file. The default value of this parameter is 1000. If the number of specimens in the cohort used for this analysis is greater that the specified value then the following warning message is shown below:

If you continue from this warning message, the report will display the summary statistics instead of the matrix plot. You may have to decrease cohort size based on some criteria like specimen type or anatomical site or otherwise using cohort query to view the matrix plot. Alternatively, you can also change the default value of MAX_SPEC_REPORT parameter to a desired value and rebuild the plot but the performance of the plot generation may be effected. The summary report will only display the percentage of specimens with different categories of data as shown below.



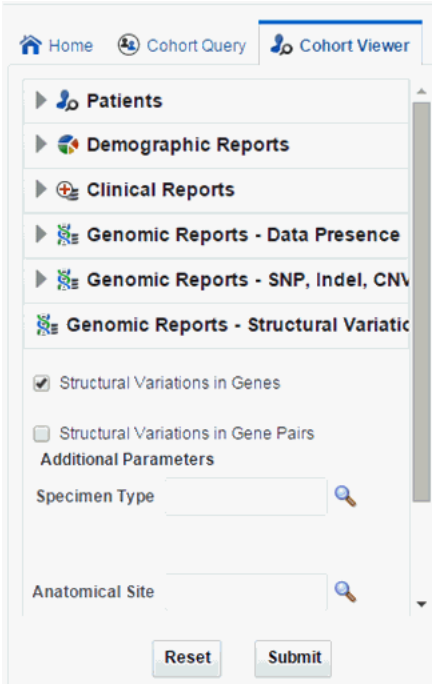
5.4.3.5 Structural Variations in Genes

Structural Variations in Genes is one of the genomic reports available to you based on the selected cohort of patients or subjects, if it is in the subject context.

First, you must select the source of patient or subject identifiers as shown in Figure 5–28. The source of patients or subjects is same for all cohort viewers and consists of one of the following option:

- active query from Cohort Query interface
- saved query from a query library
- saved list of identifiers
- ad-hoc list of identifiers
- list of patient or subject IDs based on a query performed through the Genomic Query tab

Figure 5–28 Select Source of Patient or Subject Identifiers

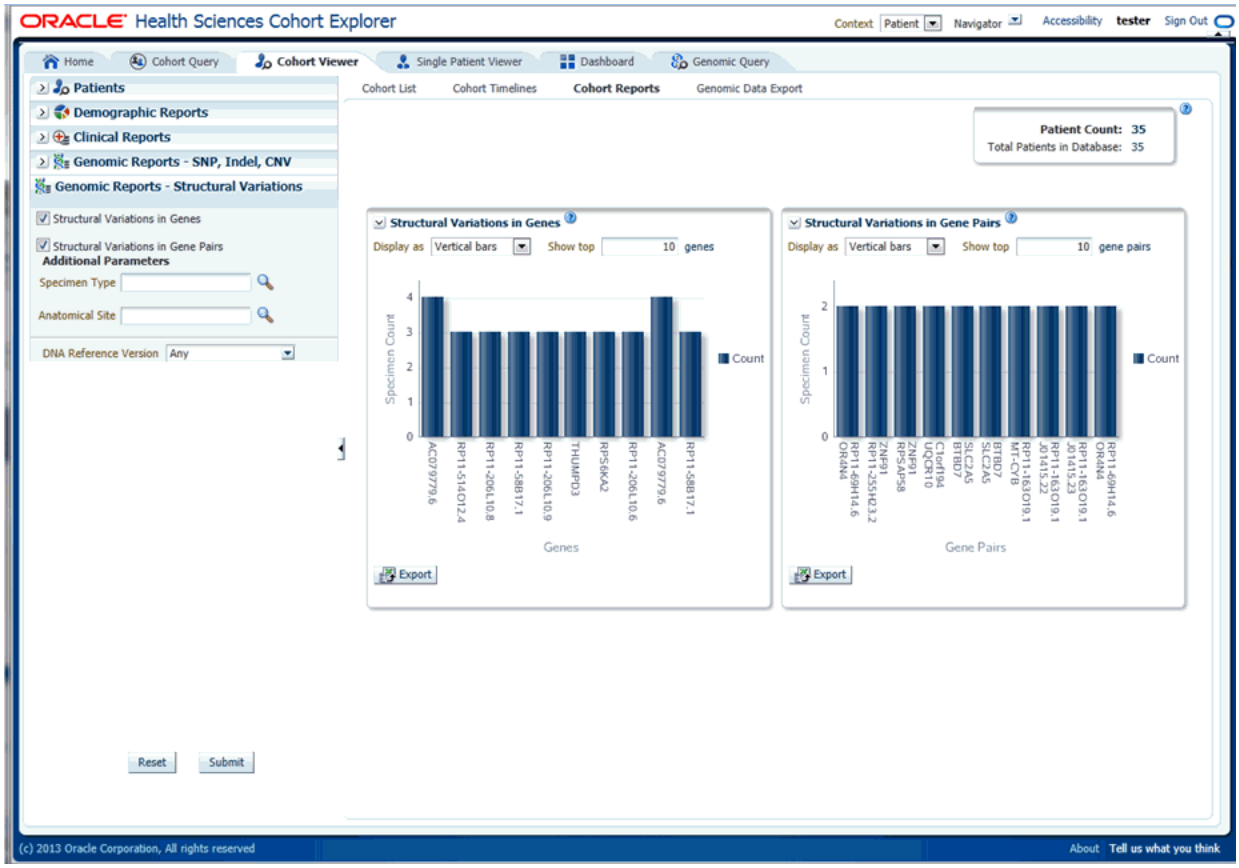


Next, you must show the Structural Variations (SV) in Genes report under Genomic Reports - Structural Variations category as shown in Figure 5–29. You can also add additional parameters such as specimen type, anatomical site, DNA version, which will only consider results linked to the selected categories. Once you click **Submit**, a histogram report shows the occurrence of structural variations involving genes sorted from the gene involved in most SVs for a given cohort, and subsequently display genes with decreasing frequency of Structural Variations. By default, top 10 genes are shown and the you can increase or decrease the number of genes displayed.

You can display the results as horizontal bars, vertical bars, or as a table. You can also export the results into pdf if bars are exported or into Excel if table is exported.

Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34.

Figure 5–29 Show the Structural Variations



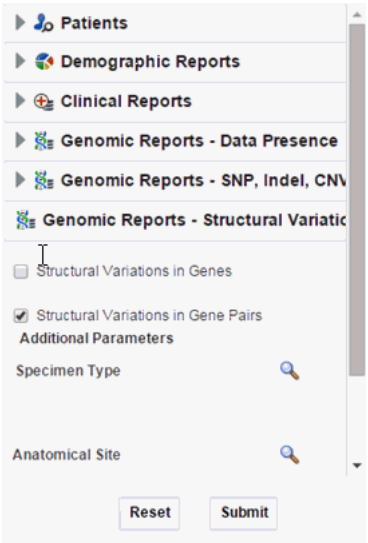
5.4.3.6 Structural Variations in Gene Pairs

Structural Variations in Gene Pairs is one of the SV genomic reports available to you based on the selected cohort of patients or subjects, if it is in the subject context.

First, you must select the source of patient or subject identifiers as shown in Figure 5–30. The source of patients or subjects is same for all cohort viewers and consists of one of the following option:

- active query from Cohort Query interface
- saved query from a query library
- saved list of identifiers
- ad-hoc list of identifiers
- list of patient or subject IDs based on a query performed through the Genomic Query tab

Figure 5–30 Select Source of Patient or Subject Identifiers

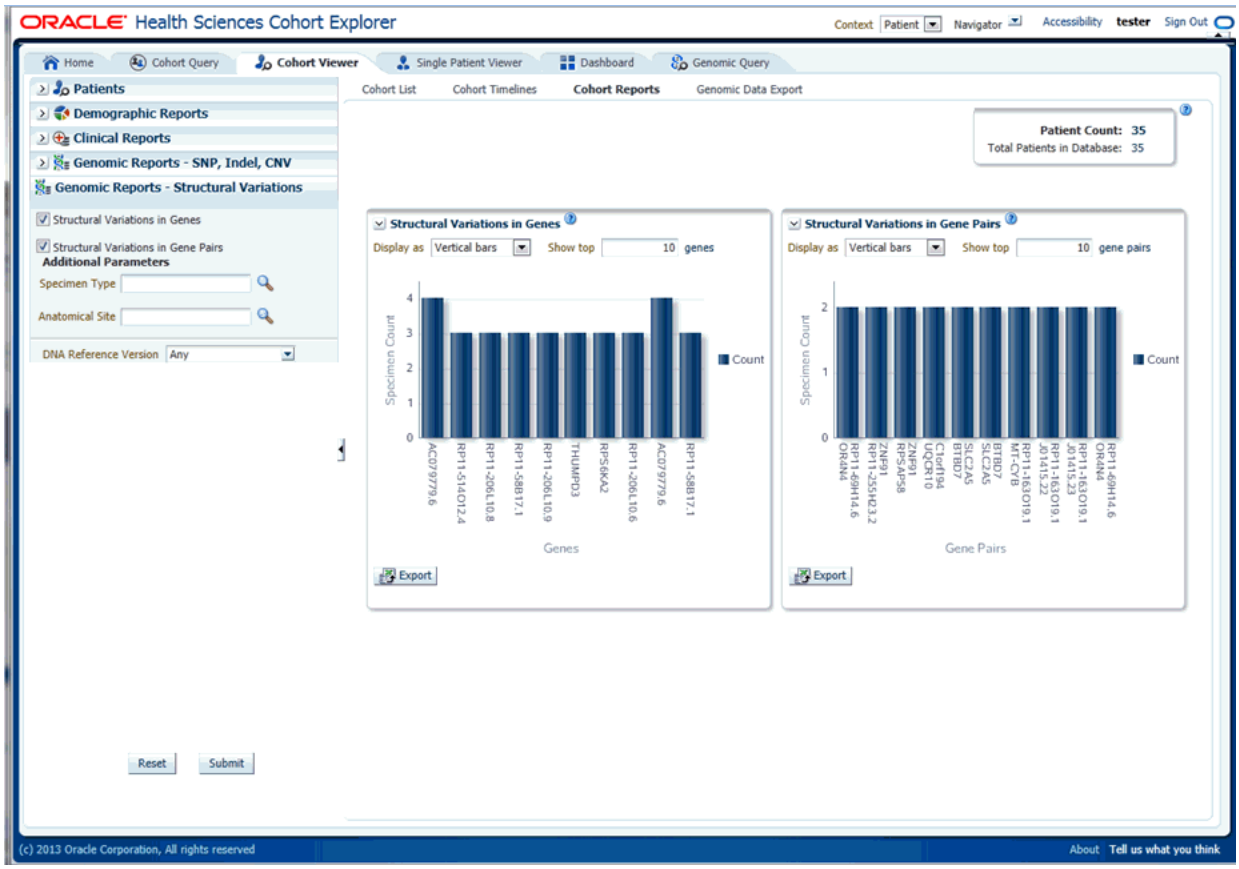


Next, you must show the Structural Variations (SV) in Gene Pairs report under Genomic Reports - Structural Variations category as shown in Figure 5–31. You can also add additional parameters such as specimen type, anatomical site, DNA version, which will only consider results linked to the selected categories. Once you click **Submit**, a histogram report will show the frequency of occurrence of structural variations in a cohort among gene pairs. The histogram is automatically sorted from the gene pair with most SVs as per cohort and the incidence decreases or at best stays the same for each subsequent gene pair. By default, top 10 gene pairs are shown and the user can elect to change the default number of bars shown.

You can display the results as horizontal bars, vertical bars, or as a table. You can also export the results into pdf if bars are exported or into Excel if table is exported.

Note: When opening the Excel file, you may receive a warning from Excel stating that the file is in a different format than specified by the file extension. This warning can be safely ignored. For more information, refer to http://docs.oracle.com/cd/E23943_01/web.1111/b31973/af_table.htm#autoId34.

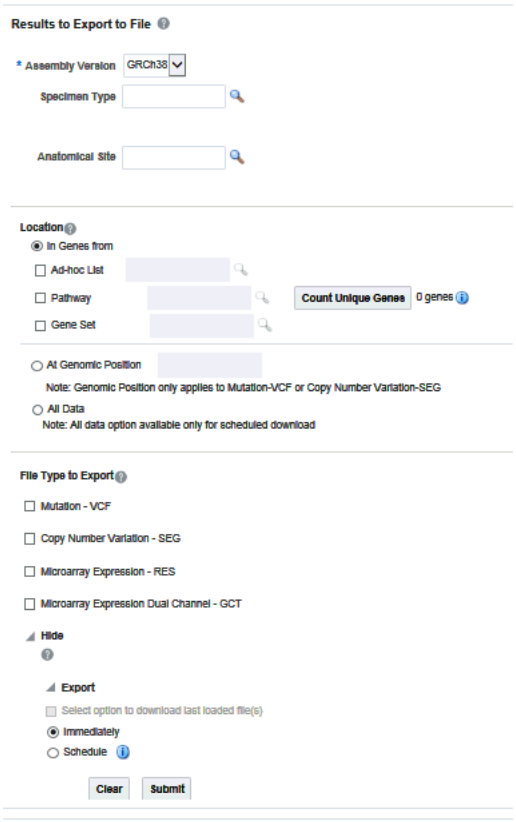
Figure 5–31 Show the Structural Variations



5.5 Genomic Data Export

The Genomic Data Export page is used to export the genomic data for patients or subjects filtered based on Study, Specimen type and Anatomical Site in a specific file format. Currently, exporting variation data from sequencing platform, single and double channel gene expression and Copy Number Variation data in VCF, SEG, RES and GCT file formats are supported. These formats are supported by the IGV browser.

Figure 5–32 Genomic Data Export

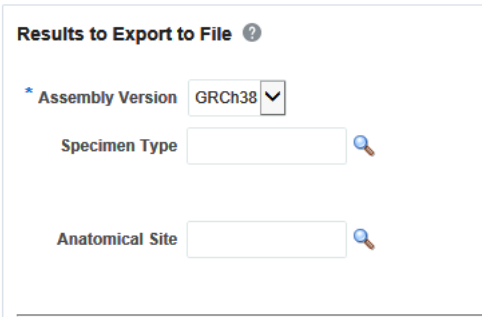


5.5.1 Selecting Patients or Subjects

You can download data for patients or subjects already selected in an active query or from a query library or from ad-hoc list of patients ID. There is no upper limit on the number of patients or subjects to be selected, however the performance slows down as more and more patients are selected.

5.5.2 Selecting Results to Export

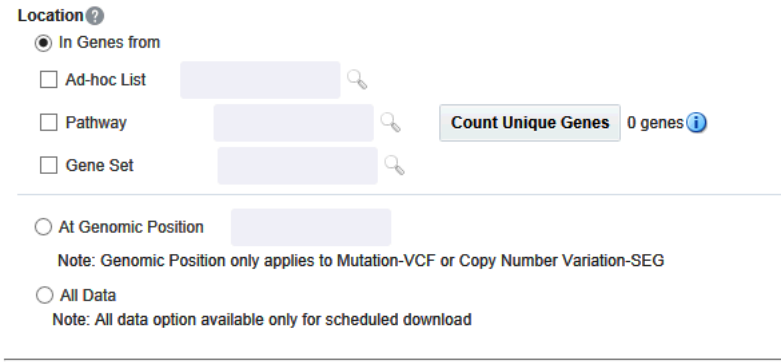
Figure 5–33 Selecting Results to Export



After selecting Patients, select the **Assembly Version**, **Specimen Type**, and **Anatomical Site**. These selection criteria will help you filter out patients based on the requirements. **Specimen Type** and **Anatomical Site** also have multiselect options. Currently only one version of data can be exported at a time.

5.5.3 Selecting Location

Figure 5–34 Selecting Location



You can download genomic data for either a list of genes or pathway or a gene set for a defined region in chromosome. You can also export the genomic data for a specific chromosome region and also the complete genomic data for the patient using the **All Data** option.

On Exadata, the code takes advantage of chromosome based partitioned data for VCF and SEG export. This enables more accurate results to be exported, including intergenic result. On non-Exadata systems, only the results that lie within any gene boundary are exported.

5.5.3.1 In Genes From

You can select genes from one or more of the provided three options. Using Ad-hoc List, you can select one or more genes. Using Pathway, you can select one or more pathways which in turn will get the list of genes associated with the selected pathway internally for querying. With Gene Set, you can use the user-defined collection of genes.

The genomic data to be downloaded is based on the above selected genes.

5.5.3.2 At Genomic Position

You can alternatively download genomic data using the genomic co-ordinates. You specify the chromosome region in a standard format for the Variation and CNV data to be exported. The Gene Expression - RES and Gene Expression - GCT download option would be disabled for genomic region criteria. You can specify a complete chromosome or a part of chromosome as criteria. Currently, only one chromosome region at a time is implemented for search.

The following chromosome region formats are supported.

- CHR15:10000-200000: Considers region between 10000 to 200000 in chromosome 15.
- CHR15:1,200,000+5000 - Considers 5000 bases upstream from 1,200,000 position in chromosome 15.
- CHR15 - Considers whole of the chromosome 15.
- CHR15:1000 - Considers 1000th nucleotide position of chromosome 15.

5.5.3.3 All Data

The genomic location selection is **All Data** option, which is only available for Schedule download and not for immediate download. With this option you can download all the data available for the specimens belonging to the selected patients or subjects falling under the selected criteria.

5.5.4 Selecting File Type

This panel lists out four file type options to export. You can select all four options at a time. For Genomic Region criteria, the Gene Expression – RES, and Gene Expression Dual Channel options are disabled. Once you select the option and click **Submit**, the data is generated and a link is provided in the bottom panel separately for each result type.

Figure 5–35 File Type to Export options

5.5.4.1 Mutation - VCF

This option exports the sequencing variation data for the selected patients or subjects for either the selected genes, pathway, geneset or for a given chromosome region as selected in the previous option. VCF supports multiple specimens' data in a single file.

The metadata header gives the following information that differs based on the search criteria:

1. ##fileformat=VCFv4.1
2. ##fileDate: Date and time of the VCF file generated.
3. ##source=Omics Data Bank (ODB)
4. ##Total Number of patients included in this VCF file
5. ##Total Number of samples included in this VCF file
6. 7.##INFO=<ID=NS, Number=1, Type=Integer, Description=Number of Samples With Data>
7. ##FORMAT=<ID=GT,Number=1,Type=String,Description=Genotype>
8. ##FORMAT=<ID=GQ, Number=1, Type=Integer, Description=Genotype Quality>
9. ##FORMAT=<ID=GQVAF, Number=2, Type=Integer, Description=Genotype_ quality_X>
10. ##FORMAT=<ID=DP, Number=1, Type=Integer, Description=Read Depth>
11. ##FORMAT=<ID=AD,Number=.,Type=Integer,Description=Allelic depths for the ref and alt alleles in the order listed >
12. ##FORMAT=<ID=HQ, Number=2, Type=Integer, Description=Haplotype Quality>
13. ##FORMAT=<ID=BQ,Number=.,Type=Integer,Description=Average base quality >

14. ##FORMAT=<ID=MQ,Number=.,Type=Integer,Description=Average mapping quality >
15. ##FORMAT=<ID=SS,Number=1,Type=Integer,Description=Variant status relative to non-adjacent Normal,0=wildtype,1=germline,2=somatic,3=LOH,4=post-transcriptional modification,5=unknown>
16. ##FORMAT=<ID=SSC,Number=1,Type=Integer,Description=Somatic Score>

The following data types are imported to VCF file:

1. CHROM: chromosome
2. POS: position of the variation
3. ID: dbSNP ID or COSMIC ID associated with a variant
4. REF: reference allele
5. ALT: variant alleles
6. QUAL: not populated. Will have '.' specified in this column.
7. FILTER: is populated as PASS.
8. INFO: Not populated. Will have '.' specified in this column.
9. FORMAT:GT: genotypic data for each specimen.
10. FORMAT:GQ: genotype quality. If not value available in DB, then '.' is specified in the file.
11. FORMAT:GQX: mapped to GENOTYPE_QUALITY_X column.
12. FORMAT:DP: this stores the TotalReadCount for a specific variant.
13. FORMAT:AD: this stores the reference read count and Allele read count for a specific variant.
14. FORMAT:HQ: not populated as of now. Will have '.' specified in this column.
15. FORMAT:FT: this stores GENOTYPE_FILTER column value.
16. FORMAT:BQ: stores the RMS base quality.
17. FORMAT:MQ: stores the RMS mapping quality.
18. FORMAT:SS: stores the somatic status
19. FORMAT:SSC: stores the somatic status score value.
20. Flex field format: If any custom formats are available, they are also included in the export.

1000 Genomes VCF 4.1 conventions are followed while exporting variation data, however certain datatypes, which are non-standard, like BQ and MQ, may differ in convention for some customers since there is no standard way to represent them.

5.5.4.1.1 Handling Non-variant and No-call Data

If NON_VARIANT and (or) NOCALL records exist for any given position, the zygosity is checked to determine if the format information from these tables is used.

Note: For *het-ref* or *half* zygosity values, these other format fields are compared with the existing SEQUENCING information. This information is then used with zygosity to create the format string.

The NON_VARIANT data allows for GQ, GQX, MQ, BQ and the first reference read count of AD. The NOCALL data allows for all format fields to be compared. Both NON_VARIANT and NOCALL do not support exporting flex fields. The GT value of the format string reflects the stored zygosity as follows:

Zygosity	FORMAT string GT:GQ:GQX:BQ:MQ:AD:DP
het-ref	1/0:99:98:38:45:20:10,10
Half	1/.:99:98:34,34:45,45:20:10,5
Het-alt	1/2:99:98:43,44:56,67:20:0,10,10
Hom	1/1:99:98:34,34:45,45:20:0,19

If there are no result records for any specimen, the export displays "." with no other information for the format.

5.5.4.1.2 Handling Ambiguous Sequencing Data in Export

There could be cases where users reload genetic information multiple times for the same specimen. This may create ambiguous values for the different fields that exist in the VCF export file. The export code deals with such ambiguous numerical values that represent the quality (that is, GQ, GQX, AD, BQ, MQ). This code now computes minimum values and ensure that the value of least confidence is reported. There could be more complex cases, for instance, if there are 2 different alleles for the same position belonging to the same specimen, or variants with same position for same specimen with different zygosity. The export code uses MIN functions on all values including all the text fields. This allows for VCF export to create a valid file that can be loaded into genome browsers.

Alternatively, you can choose not to consider data from a specific specimen or a specific file using following methods:

- Using DELETE_FLG - A user may load results for a specimen more than once that can completely contradict previous results. Users can set the DELETE_FLG as 'Y' on W_EHA_RSLT_SPECIMEN and (or) W_EHA_SPEC_PATIENT or W_EHA_SPEC_SUBJECT to have previous loads excluded, and then reload the correct result files. When the user now exports the data, only the latest loaded specimen data is considered for export.
- Using FILE_URI - Oracle recommends using this method since you need not reload the data again as opposed to the above method. When there are multiple files loaded with contradicting data for the same specimen, user can set some files as obsolete by changing the W_EHA_FILE_LOAD.FILE_WID column. For example, if you have loaded the same specimen data 3 times and would like to consider the latest file loaded for export, then you must first identify the latest FILE_WID from W_EHA_FILE_LOAD table. Then change the FILE_WID of two old files in W_EHA_FILE_LOAD table to the latest FILE_WID. Now, all the three records belonging to the three file loads contain same FILE_WID, which represents the latest file load and only the latest file export data is exported.

Representing AD Values

Allele depth values represented under the AD datatype are in the order of the alleles represented in the GT. Refer to the following table with examples:

ALT	FORMAT	SAMPLE1	
G,C,T	GT:AD	1/2:0,4,6	0 represents reference_read_count 4 represents allele_read_count of 'G' 6 represents allele_read_count of 'C'
G,T	GT:AD	2/2:0,4	0 represents reference_read_count 4 represents allele_read_count of 'T'
G,T	GT:AD	1/0:10,5	10 represents referen ce_read_count 5 represents allele_read_count of 'G'

5.5.4.2 Copy Number Variation - SEG

The copy number variation data is exported in SEG format. Currently, CNV data from any array based system like Affymetrix Genome Wide SNP 6 array whose data is in SEG format while loading in ODB is supported. The main requirement for exporting CNV data is to have the SEG_MEAN value in the CNV table of ODB.

For exporting data that is not loaded from SEG files, for example, data from CGI CNV files or any other source of CNV data, users have to create their own loader. The loader is expected to calculate the SEG_MEAN value since this value is most important for export.

1. ID: specimen ID of the reported CNV segment
2. chrom: chromosome name
3. loc.start: start position of the CNV segment
4. loc.end: end position of the CNV segment
5. num.mark: for array based CNV data, this stores the number of probes details
6. seg.mean: this stores the segment mean value from SEG_MEAN column in CNV table.

5.5.4.3 Gene Expression - RES

RES is one of the gene expression formats supported by IGV browser. Currently, only microarray gene expression data is exported to this format. Following data types are imported to RES format:

1. Description: hugo name of a specific probe
2. Accession: probe ID
3. Intensity: intensity value of the associated probe
4. Call: call of the associated probe

5.5.4.4 Gene Expression Dual Channel - GCT

GCT is one of the gene expression formats supported by IGV browser. Currently, only AgilentG4502A platform microarray gene expression data is exported to this format. Following data types are imported to GCT format:

1. Description: Gene symbol of a specific probe
2. Accession: probe ID
3. Intensity: intensity value of the associated probe
4. Call: call of the associated probe

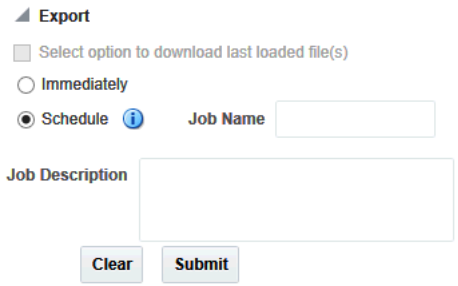
Note: The GCT file takes its gene symbol for the probe from the 2-channel composite element of ADF file. This is input into the ADF composite table in ODB. This value may not match with HUGO name in certain cases as TRC associates 2-channel records in the result table that has partial (which includes a flanking region set by the user) genomic coordinate. The coordinate overlaps between composite elements and gene segments in the reference. This may also result in some cases in more than one unique gene in the reference mapping to a gene composite element.

5.5.4.5 Export Options

Currently, you can export data in the following two ways:

- Select option to download last loaded file(s)
- Immediately, which is the default option
- Schedule

Figure 5–36 Export Options - Schedule Mode



The default option (immediately) gives you the file link on the same screen and you can click to download it immediately. The link provided has a specific naming convention: *<file type>*_ODB_*<date:MM-DD-YYYY>*_*<time:HH24*-MI-SS>*.*<file_type_extension>*. For example, RES_ODB_09-14-2014_04-26.res. A short description of the file stating data type and advise on the expected count of features is displayed below the created link.

The scheduler option runs the process as a job. You can track the status of the job from the **Home** > **Jobs** tab. This option is best suited for exporting large data set like **All Data**, whole chromosome variants and so on. For schedule option, you must provide a job name and description. Then click **Submit** to start the process. For more details, see Jobs on page 3-7.

There is a possibility of replicate and duplicate data in the database. This could be due to loading multiple files belonging to the same specimen_number. This can happen if the same library is sequenced multiple times or the data is reanalyzed, for example, the reads were realigned using the new reference version and hence new VCF or gVCF files are created for same sample. In this scenario, you can use the option to export VCF data only from last loaded files. For example if variation data has been loaded for a specimen in Jan 2015, Mar 2015 and July 2015, then using this option you can export data from the file loaded in July 2015 and it would not consider variants from the file loaded in Jan and Mar 2015.

Note: The Schedule Jobs option uses an asynchronous approach to store the file in DBFS. As an alternative to downloading the file using the link in the Cohort Explorer Jobs page, there are other ways to access DBFS. From a Linux OS, you can mount DBFS using *dbfs_client* application and then browse the directories. Windows OS does not support the FUSE interface and cannot mount DBFS directly. However, there is a *dbfs_client* application for Windows that can execute commands to access DBFS. The Windows version of *dbfs_client* lets you use the command line to execute normal directory commands. You can list the DBFS directories as well as copy data from DBFS to the local drive. The *dbfs_client* application is part of the standard Oracle client software.

For more information about using *dbfs_client*, see http://docs.oracle.com/cd/E11882_01/appdev.112/e18294/adlob_client.htm#ADLOB0006.

6

Single Patient or Subject Viewer

This chapter describes the single patient or subject viewer of TRC. It contains the following topics:

6.1 View Records

The View Record screen is designed to help you focus on the discreet medical history or genomic history of a particular patient. Even though the Cohort Query functions provide tools for selecting or examining cohorts of patients, it can be difficult to define the appropriate criteria without a closer examination of the data itself. With the Single Patient Viewer, you can drill into the data details of a particular patient, and locate the attribute or data element that is most pivotal as selection criteria for your cohort.

The viewer is designed to display data for one particular patient or subject, based on the unidentified Patient or Subject Identifier within the CDM database. The Patient or Subject Identifiers are displayed when you list patients or subjects from a query (**Cohort List** tab).

At the top of the screen, there is a search function for locating the particular patient or subject to view. The search is based on the unidentified Patient Identifier, to protect patient identity in patient context and Subject Identifier along with associated study in the Subject Context. After selecting the patient or subject you want to examine, the system displays all the available information about that patient or subject, organized in distinct section as follows:

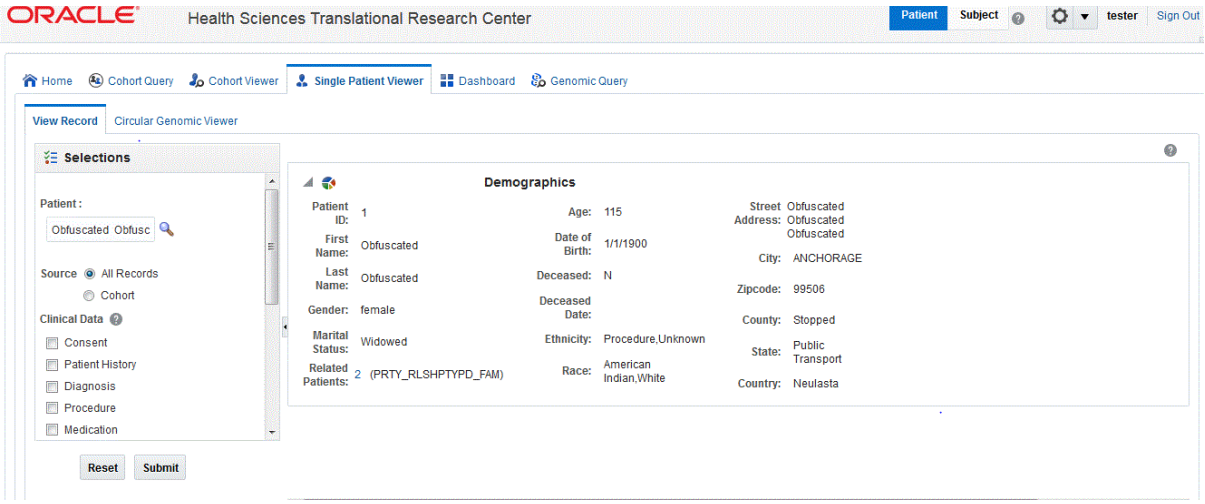
Clinical Data

- Demographics
- Consent Forms Signed
- Patient History
- Patient Diagnosis
- Procedures
- Medications Taken
- Test or Observation
- Specimen Samples Collected
- Clinical Encounter (only in patient context)

Genomic Data

- Specimen with Genomic Results
- Derived Files
- File Lineage
- Variants Found

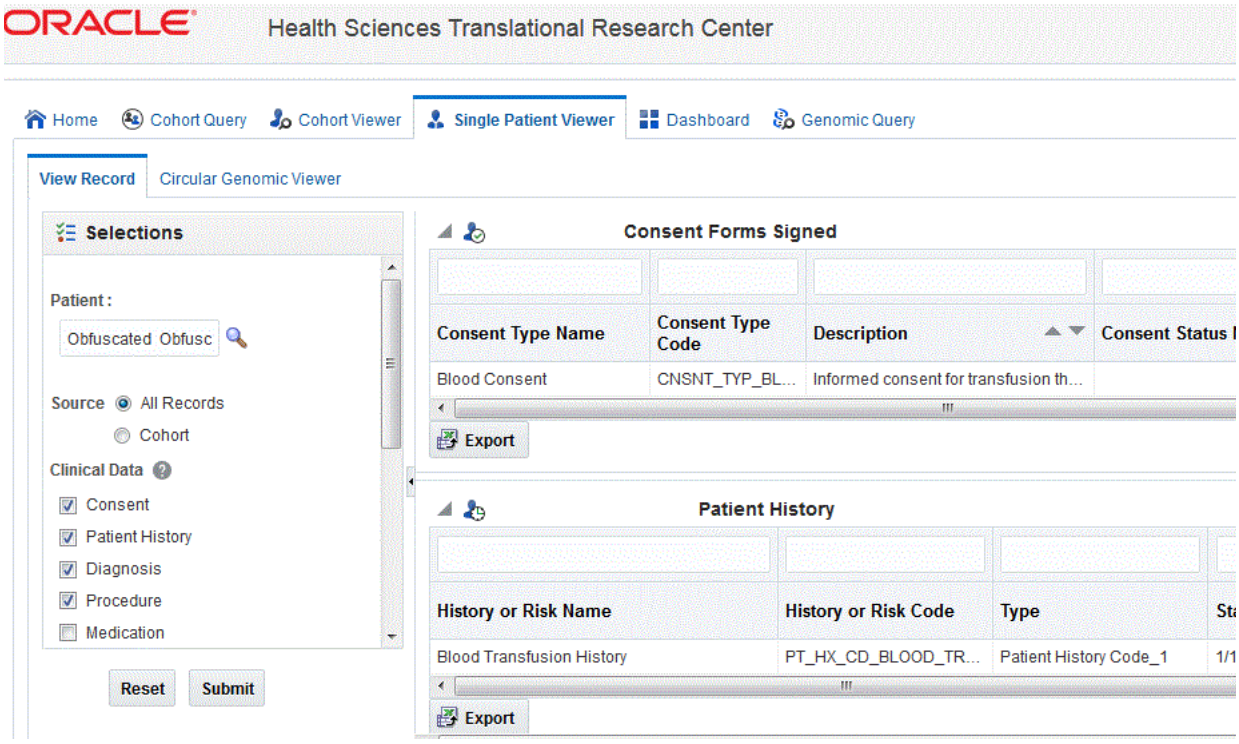
Figure 6–1 View Record



At the bottom of the page, there is a **Print** button, which prints the current screen.

You select the check box next to the name of any one section to show the display of that section.

Figure 6–2 Hide and Display Options



Below each section, there is an **Export** button that enables you to save the displayed data in a discreet file.

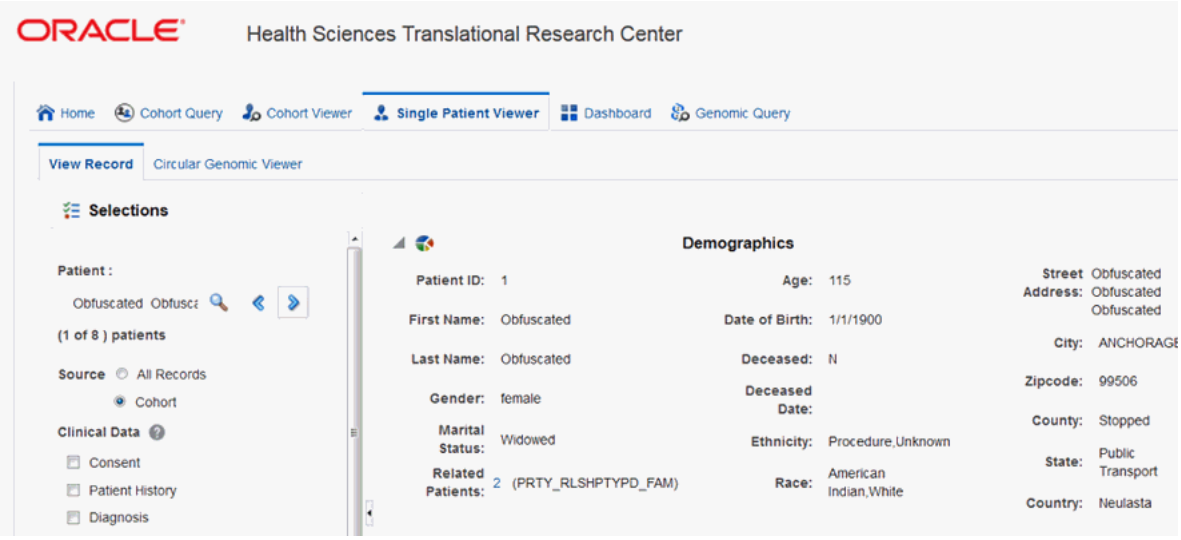
A patient or subject may have significant volumes of information, hence the sections are designed to display multiple rows. If you type a value in the blank section above any one particular attribute and press Enter, the system will filter the rows in that section, based on the value you enter. This lets you focus on one particular row of data. In addition, clicking the header of any column with sort the data within the column.

Note: The attributes that display for the various categories align with the data terms and definitions that are described in the **Cohort Query** tab, where you select criteria for selecting patients or subjects.

6.1.1 Navigating Through Selected Patients or Subjects

If you have a specified patient or subject count in the single patient or subject screen on the Cohort Query or Cohort List screen, you can navigate through all the selected patients or subjects clinical and genomic history by using the Previous [<] and Next [>] buttons. Under the **Patient** field, the current Patient's ordinal number (ordinal position of the Patient record within the selection) and the total number of available Patients or Subjects (in the Cohort) will be displayed. For example, in the following figure it is (*1 of 8*) patients.

Figure 6–3 All Records and Cohort Options



If you have navigated from the Cohort Timeline screen into the single patient or subject screen, then the total number of available patient or subject cohort will be the count of patients or subjects selected in the initial pool.

6.1.2 Source

When you log into the application and navigate to the single patient/subject screen, by default, **All records** source is selected. The Cohort option will be selected when you navigate to single patient or subject screen from any other screen with some patient or subject count. When the *Cohort* option is selected, you can scan through the patients or subjects selected in the previous screen. When the *all record* option is selected, you can search through all the patients available in the CDM database.

6.1.3 Clinical Data

The attributes displayed for each section are listed below:

- **Demographics**

  Gender, Marital Status, Age, Date of Birth, Deceased Date, Ethnicity, Race, City, Zip code, County, State, Country

  In case you have a special role (pi-user role) then additional columns will be displayed in the Demographics section Contact Info, Street Address, Related Patients or Subjects

- **Consent Forms Signed**

  Consent Type Name, Consent Type Code, Description, Consent Status Name, Status Code, Start Date, End Date, Data Source

- **Patient History**

  History or Risk Name, History or Risk Code, Type, Start Date, End Date, Frequency (Units), Amount (Units), Text Value of Code, Applicable To, Data Source

- **Patient Diagnosis**

Diagnosis Name, Code, Age at First Onset (in Years), Onset Date, Date Reported, End Date, Status, Anatomical Site Name, Anatomical Site Code, Data Source

- **Procedures**

  Procedure Name, Procedure Code, Procedure Type, Type Code, Start Date, End Date, Outcome, Anatomical Site Name, Anatomical Site Code, Data Source

- **Medications Taken**

  Medication Name, Medication Code, Description, Start Date, End Date, Dosage, Dosage Units, Outcome, Data Source

- **Test or Observation**

  Test or Observation Name, Test or Observation Code, Type Date, Numeric Result (Units), Result (text), Data Source

- **Specimen Samples Collected**

  Specimen Type Name, Specimen Type Code, Date Collected, Anatomical Site Name, Anatomical Site Code, Amount, Units, Data Source

- **Clinical Encounter**

  Encounter Id, Encounter Type, Start Data, End Date, Additional Details (clicking this provides more details about the encounter like Event Name, Event Type, Start Date, End Date, Location Name (Location Type)), Data Source

6.1.4 Genomic Data

Genomic data is displayed in four sections as follows:

Table 6–1 Specimens with Genomic Results

Column Heading	Definition	Sample Value or Values
Specimen Id	Specimen belonging to the selected patient or subject	HG00096
Specimen Vendor Id	Specimen Vendor for that specimen	Vcf
Version Label	Represents Assembly Version (DNA reference version against which this data was loaded)	GRCh37(V68)
Sequence Variants Results	Whether the sample has sequence variants results	Yes / No
Copy Number Variation Results	Whether the sample has copy number variants results	Yes / No
Single Channel Microarray Results	Whether the sample has single channel results	Yes / No
Dual Channel Microarray Results	Whether the sample has dual channel results	Yes / No
Rna-Seq Expression Results	Whether the sample has rna sequencing results	Yes / No

Derived Files (Level 3 or 4 results loaded into ODB schema tables for querying) and

File Lineage (Level 1 or 2 files linked to from ODB tables (frequently in binary format and not yet interpreted)

Note: If you have appropriate permissions, and if files are present in the middle tier accessible location, the path listed in the **File Name** fields for Genomic data are enabled to allow you to click and download the files directly from CE.

Table 6–2 Derived Files

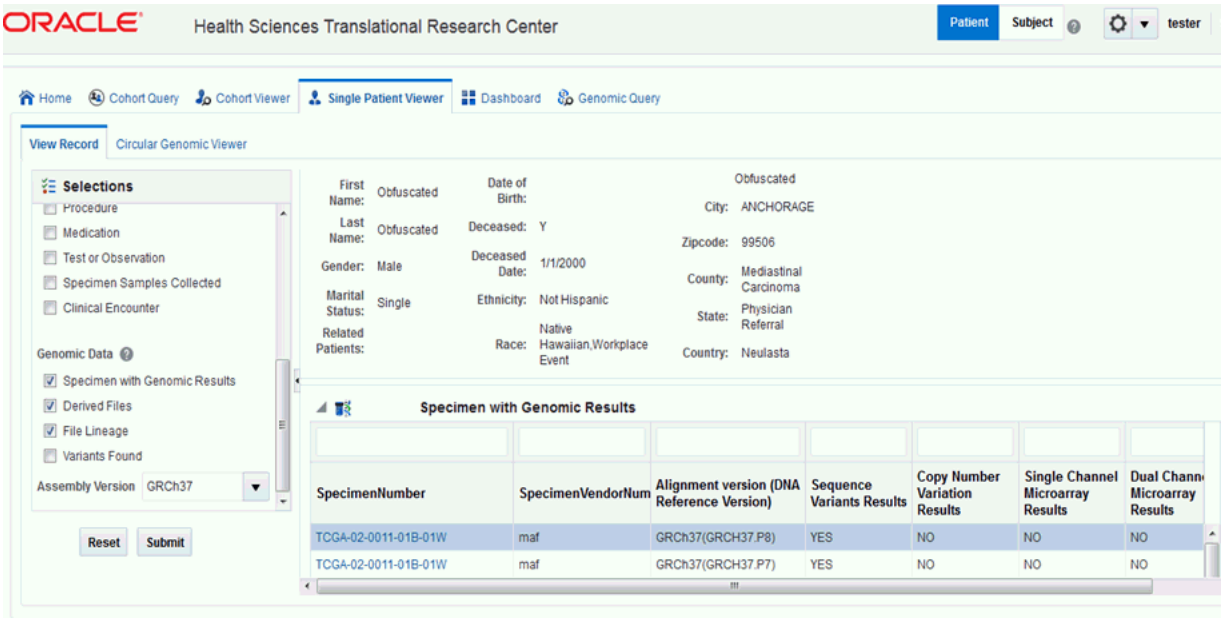
Column Heading	Definition	Sample Value or Values
Filename	Filename including path of the genomic file stored including path	C:/John_specimen01.vcf
File Size in MB	Size of the File in MB	Numeric, positive integer
File URI	URI of the file	File://trc/abc.maf
Alternate Filepath	The FTP path of the file	
File Type, Version	Type of file and Version	Variant Call Format, 4.1
Result Type	Type of result data in the file	Sequencing, Copy Number Variation, Gene Expression (2-channel or single channel)
Alignment Version (DNA Reference Version)	Represents Assembly Version (DNA reference version against which this data was loaded)	GRCh37(v68)
Total Number of Specimen in File	Total number of specimen that the file contains where not all specimen belong to the selected patient	Numeric, positive integer
Last Updated	When record was last updated	19-Mar-2012

Table 6–3 File Lineage

Column Heading	Definition	Sample Value or Values
Parent Filename	Parent Filename including path of the genomic file stored including path	C:/John_specimen01.BAM
File Size in MB	Size of the File in MB	Numeric
File URI	URI of the file	File://trc/abc.maf
Alternate Filepath	The FTP path of the file	
File Type, Version	Type of file and Version	Binary Alignment Map, 1.0
Alignment Version (DNA Reference Version)	Represents Assembly Version (DNA reference version against which this data was loaded)	GrCH37
Last Updated	When record was last updated	19-Mar-2012
Derived Child Files: File - Specimen Id, Vendor Id	Information about Derived Results files that have their lineage based on the particular Low Level file.	C:/John_specimen01.vcf - HG00096_1,HarvardLab1

Assembly Version

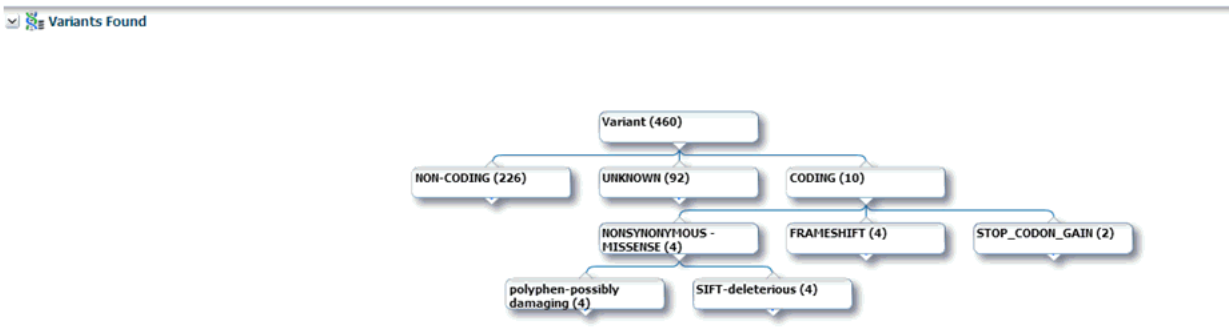
The Assembly Version drop-down is a multi-select component. By default, the last loaded assembly is displayed but this selection can be modified. The Genomic data is filtered out with this assembly version instead of the DNA reference version. One Assembly version can belong to multiple DNA reference versions. A new column **Alignment Version (DNA Ref Version)** has been added to the genomic data tables as shown in the following figure.



6.1.4.1 Variants Found

This displays the different variants available for a patient. These variants are grouped and displayed in a hierarchical structure with the count of the variants displayed for each type of variants.

Figure 6–4 Variants Found



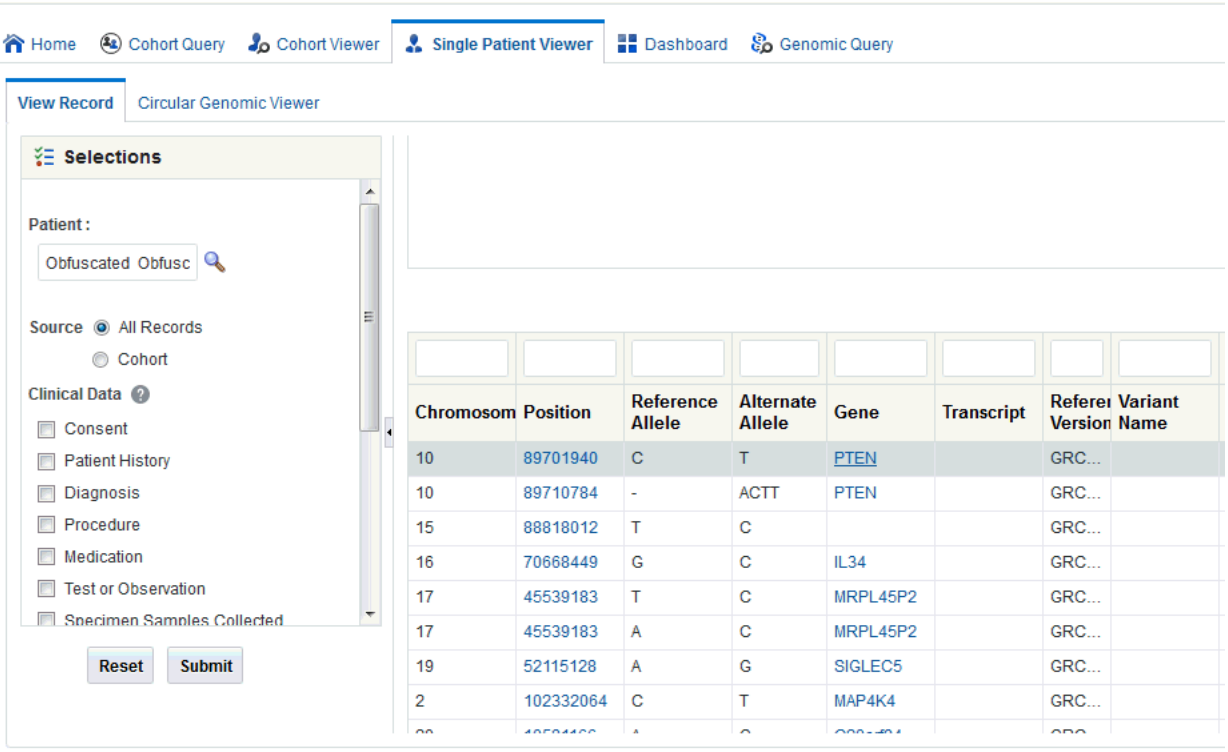
On selecting any of the nodes the details of the variants are displayed in a table as follows:

Table 6–4 Variant Details

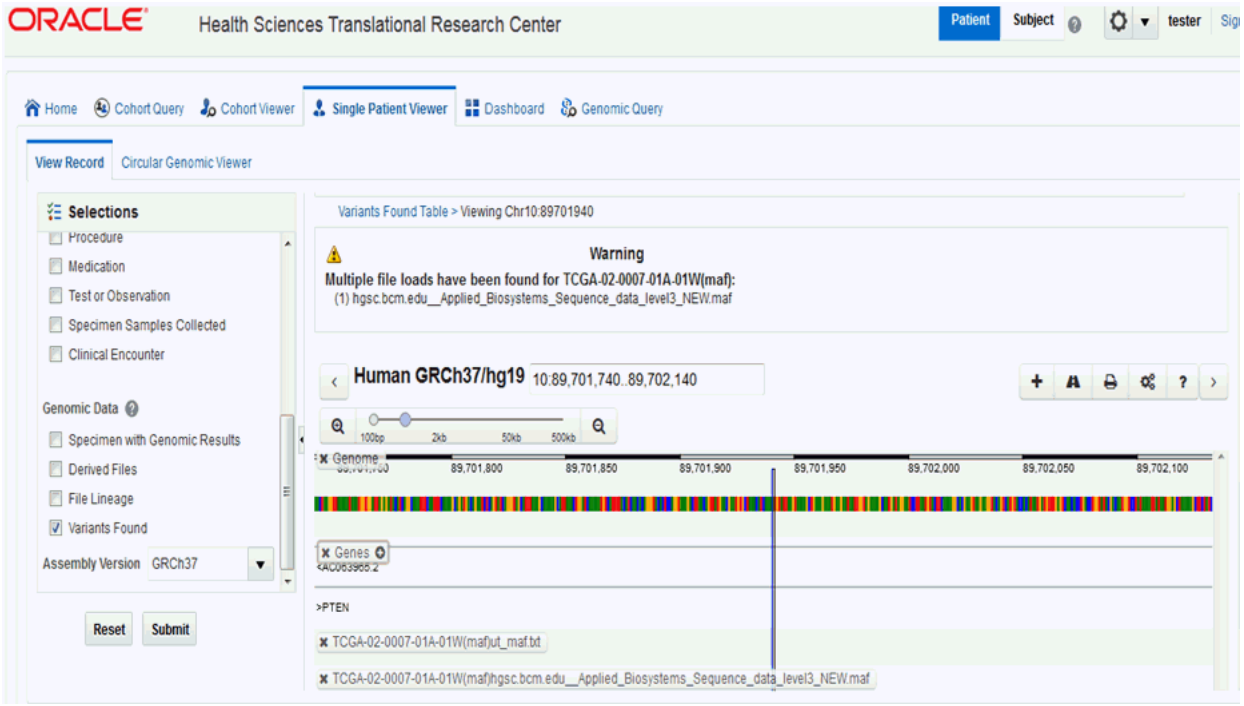
Column Heading	Definition	Sample Value or Values
Chromosome	Chromosome location of the variant	1
Position	The position of the variant within the chromosome	Numeric value
Reference Allele	The reference allele of the variant	C
Alternate Allele	The alternate allele of the variant	G
Gene	The gene containing the variant	BID
Transcript	The transcript name	ENSTXXX
Variant Name	The reference id of the variant	rs111
Variant Type	The type of variant	Substitution
Variant Status	Status of the variant	Known
Protein Name		
SIFT Impact	The SIFT impact of the variant	Intolerant
Polyphen Impact	The polyphen Impact of the variant	Damaging
Drug	The related drug	Clofazimine
Associated Disease	The disease associated to the variant	Anaemia
Histology		
Site		
Specimen Id	The specimen containing the variant	HG00096
Alignment Version	Alignment version of a variant	GRCh37

6.1.4.2 Dalliance Browser

The Dalliance browser is a third party tool that displays a graphical representation of the variant, chromosome or gene range. The variant detail table comprises of hyperlinks for variant reference ID, gene name and chromosome position. Clicking any of these will navigate to the Dalliance browser where you can see the particular VCF tracks around a specific variant, a specific gene, or in an explicitly specified chromosomal range. For example, in the following screenshot, clicking any of the hyperlinks in the **Gene** column will navigate to the Dalliance browser.



If a file is loaded multiple times for a particular specimen, then the following warning message is displayed. It lists all the multiple files names.



Following are some configurations required to plot the gene track in the Dalliance browser. You can create your own DAS server and the corresponding entries should be added in the TRC.properties file. Following is an example of the required entries, to

look up the Authority, University of California Santa Cruz (UCSC) name for the alignment and the public URLs for the Sequence and Genes tracks (reference tracks):

```
DALLIANCE.AUTHORITY_37=GRCh
DALLIANCE.UCSC_NAME_37=hg19
DALLIANCE.SEQ_URL_HG18=http://www.derkholm.net:8080/das/hg18comp/
DALLIANCE.SEQ_URL_HG19=http://www.derkholm.net:8080/das/hg19comp/
DALLIANCE.GENES_URL_HG18=http://www. derkholm.net:8080/das/hsa_54_36p/
DALLIANCE.GENES_URL_HG19=http://www. derkholm.net:8080/das/hsa_59_37d/
```

While integrating Dalliance with TRC, the server `http://www.derkholm.net:8080/das/` has been used. However, this has been shut down.

These can be manually customized in TRC.properties. To add any new alignments, add all the above code to TRC.properties. Make sure that you research the reference for where they can be found. It is possible to use downloaded files instead of public DAS server(s) for reference, but the client should host these files on a web server.

6.2 Circular Genomics Viewer

Circular genomic data viewer provides an interface for you to visualize the genomic data which includes variation, micro array expression, copy number variation, dual channel expression and rna sequencing. The system uses the VisQuick tool, which is a Javascript library built specifically for genomic data visualization.

You must select Patient ID or Subject ID (study), which are required fields. If the Patient or Subject ID has been selected in the View Record screen, it will be inherited here. Optionally, you can also select **Specimen Type and Anatomical Site** to add filter criteria. In addition, the DNA Reference Version selected is used to filter out the results and determine the cytoband to be used while rendering the circular genomic plot for any of the five data types. Based on the filter criteria, the matching specimen in any of the five result types are displayed. You can select specimens of different result types however only one specimen of each result type is allowed. By default, the cytoband of chromosomes is also plotted which is the outer most ring of the circular plot.

Figure 6–5 Selection Criteria

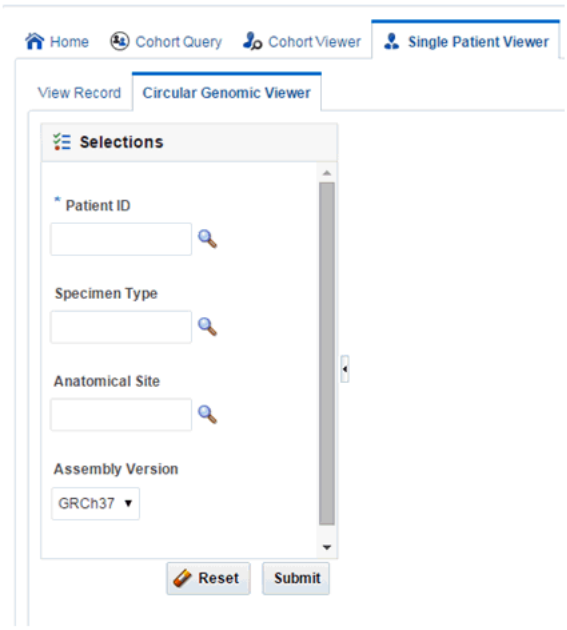
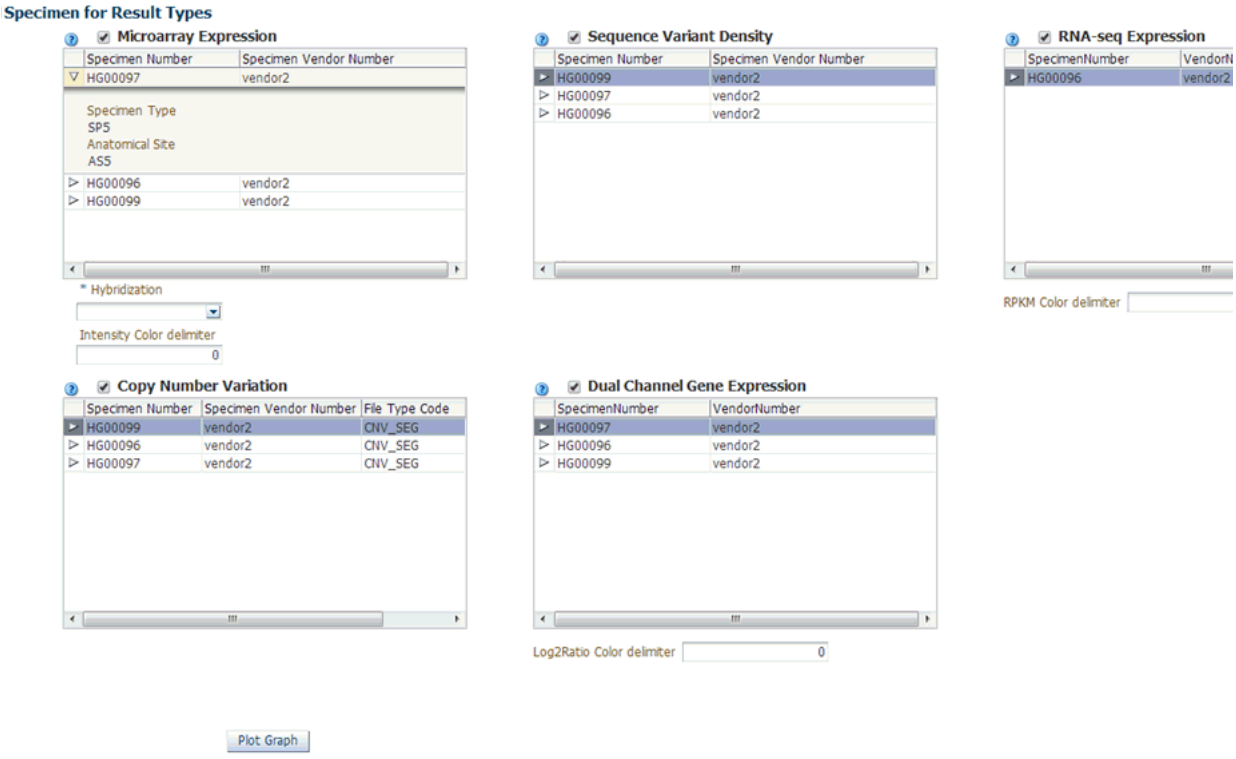


Figure 6–6 Specimen for Result Types



6.2.1 Selecting Data to Plot

6.2.1.1 Microarray Expression

The microarray expression panel displays all the specimens of the result type expression. You can select only one specimen at a time.

For the selected specimen from the panel, you can view the list of hybridizations available for that specimen. You can select maximum of two hybridizations from the multiple choice box, for which the plot is rendered.

The color delimiter value is used to render color to the data points in the plot based on the intensity value. If the intensity value is above the value defined in color delimiter then the data points have green color Otherwise, the points have red color.

6.2.1.2 Sequencing-Variant Density

This option enables you to plot the density of variants for every 100-kb region for a specimen. The sequence variant density panel displays all the specimens of the result type sequence variants. You can select only one specimen at a time.

6.2.1.3 RNA-Seq Expression

This panel displays all the specimens of result type RNA Sequencing.You can select only one specimen at a time.

The color delimiter value is used to render color to the data points in the plot based on the RPKM value. If the RPKM value is above the value defined in color delimiter then the data points have green color. Otherwise, the points have red color.

You can use the check boxes provided next to each of the result types to determine if the specimen selected will be plotted or not. You can select one or all of the result types for plotting of the graph. Once the specimens are selected, the circular genomic plot is displayed. The outermost circle in the plot is the Cytoband. The supported Cytoband versions are: hg-18 and hg-19.

Note: To display the cytoband in the Circular Genomic plot we need an entry mapping each DNA Reference Version to the Cytoband in the TRC_LOOKUP_CODE table in the APP_schema.

For each reference version we must insert a row with the following values:

- CODE_TYPE : TRC_REFVERSION_CYTOBAND
- CODE:cytobandHG19 (for HG19), cytobandHG18 (for HG18), or cytobandHG38 (for HG38)
- CODE_NAME :Name of the loaded DNA reference version (for example, V69)

You can hover over the plot to get details such as:

- Microarray Expression: Chromosome, Start Position, End Position, Value and Gene
- Copy Number Variation: Chromosome, Start Position, End Position and Value
- Sequence Variants: Chromosome, start position, end position, and value. The density value is calculated by: Total number of variants for 100kb region / 100000
- Dual Channel Microarray Expression: Chromosome, Start Position, End Position, Value and Gene
- Rna Seq Expression: Chromosome, Start Position, End Position, Value and Gene

6.2.1.4 Copy Number Variation

The panel displays all the specimens of the result type copy number variation. You can select only one specimen at a time.

The value for copy number variation depends on the CNV result type selected. If the selected specimen contains data from Genome_Wide_SNP_6 array, then the value for CNV is taken from segment mean stored in the database. If the selected specimen has data from complete genomics, then the value is calculated based on the calledPloidy value stored in the database. The value to be plotted is calculated using the following formula for CNV data from complete genomics:

log2(called_ploidy/expected_ploidy)

where,

expected_ploidy is 2 for chr1-22

expected_ploidy is 2 for chrX for females.

expected_ploidy is 1 for chrX outside the psuedo-autosomal region in males

expected_ploidy is 2 for chrX inside pseudo-autosomal region in males

The pseudoautosomal regions on chrX for 'NCBI build 37' as reported by Complete Genomic are 60000 - 2,699,519 and 154,931,043 - 155,260,559.

The pseudoautosomal regions on chrX for 'NCBI build 36' as reported by Complete Genomic are 0 -2,709,519 and 154,584,237 - 154,913,753.

For called_ploidy zero, there the log2 will be infinity, in such cases the final value is taken as -2.

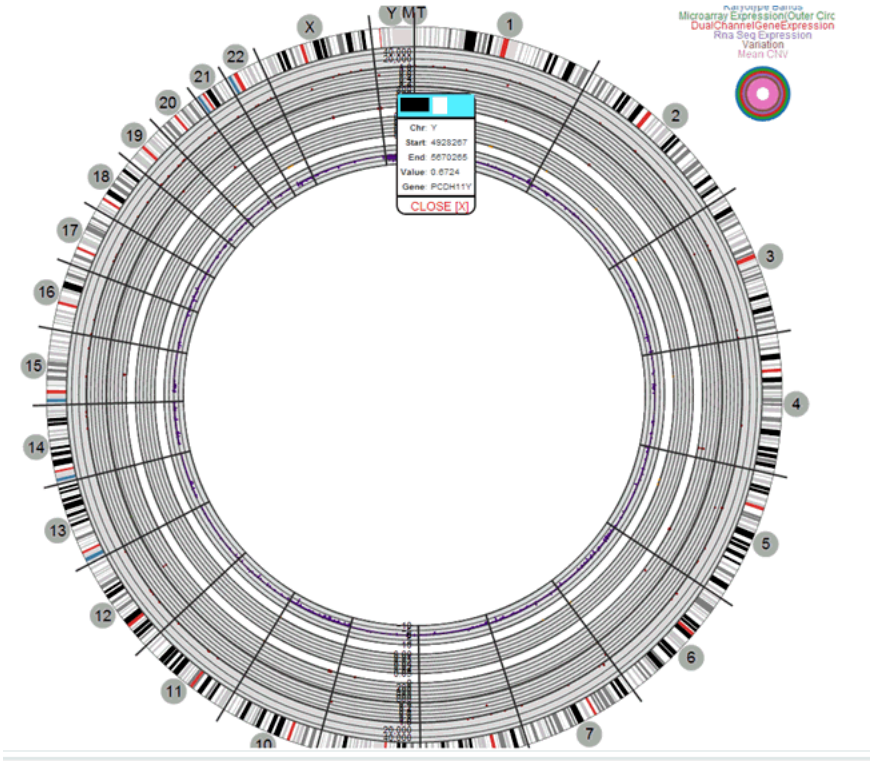
6.2.1.5 Dual Channel Microarray Expression

This panel displays all the specimens of result type Dual Channel Microarray expression. You can select only one specimen at a time.

The color delimiter value is used to render color to the data points in the plot based on the Log2Ratio value. If the Log2Ratio value is above the value defined in color delimiter then the data points have green color. Otherwise, the points have red color.

6.2.2 Circular Representation

Figure 6–7 Circular Genomic Plot



Note: In Internet Explorer 9, the compatibility mode should be turned off for Circular Viewer to run properly.

7

Dashboard

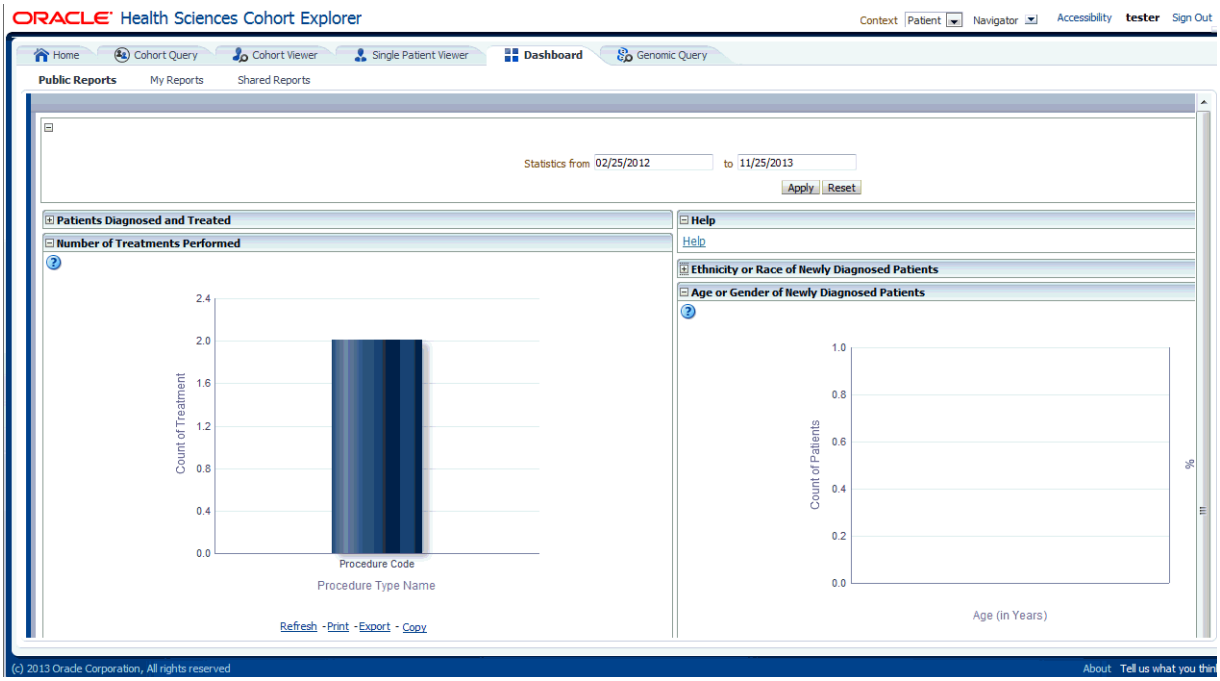
This chapter describes the TRC dashboard. It contains the following topics:

- Section 7.1, "Standard Reports"

7.1 Standard Reports

The Dashboard page displays seven reports. All of these reports have date range filtering criteria, that is **Statistics from** and **to** that you can adjust according to your requirements.

Figure 7–1 Dashboard Page



7.1.1 Selection of Date Range Using Statistics from and to Prompt

By default, all reports display the past one month's data from the current date. You can also view data for a date range of your choice.

For more information, see the prompt details in Statistics from-to (Prompt Set TRC-CEP-001) on page 4.

The following reports are shown and driven by data range in the prompt set:

- Patients Diagnosed and Treated (TRC-CER-001,TRC-SD-P004)
- Number of Treatments Performed (TRC-CER-005)
- Ethnicity or Race of Newly Diagnosed Patients (TRC-CER-002)
- Age or Gender of Newly Diagnosed Patients (TRC-CER-003)
- Biospecimen Samples Status (TRC-CER-006)

7.1.2 Patients Diagnosed and Treated (TRC-CER-001,TRC-SD-P004)

The Patients Diagnosed and Treated report displays the past patient data for the user-defined date range, which by default displays one month's data from the current date.

It is preconfigured during the initial setup or installation phase and can be altered only by a CE-Administrator or CE-Developer. The columns of the resulting patient count table are:

- Number of diagnosed patients: Number of patients diagnosed with a specific disease during the time period specified.
- Number of patients successfully treated: Number of patients that have been treated successfully during this time period.
- Number of consented patients: Number of patients that have consented to specific treatments (type of consent considered is customized based on your requirements) during the specified time period.
- Number of patients undergoing treatment: Number of patients that are or have been undergoing any type of treatment during the specified period of time.

For more information, see the report details in Patients Diagnosed and Treated (TRC-CER-001,TRC-SD-P004)

7.1.3 Number of Treatments Performed (TRC-CER-005)

The Number of Treatments Performed report counts the number of treatments performed during the specified from-to time period. The default being the past one month's data from the current date.

Treatments are categorized according to types and include Surgery, Chemotherapy, Radiation, and Gene Therapy and so on. Treatments are selected during the configuration phase by either the CE-Administrator or CE-Developer and cannot be modified by other users.

For more information, see the report details in Number of Treatments Performed (TRC-CER-005)

7.1.4 Ethnicity or Race of Newly Diagnosed Patients (TRC-CER-002)

The Ethnicity or Race of Newly Diagnosed Patients report is a pie chart view of how many patients of a particular race or ethnicity have been diagnosed with a disease during the specified time period. The default being the past one month's data from the current date. You can toggle between Ethnicity (by default) and Race to see the counts and percentages of patients.

For more information, see the report details in Ethnicity or Race of Newly Diagnosed Patients (TRC-CER-002)

7.1.5 Age or Gender of Newly Diagnosed Patients (TRC-CER-003)

The Age or Gender of Newly Diagnosed Patients report displays how many patients within a particular age category and gender have been diagnosed in the facility during the user specified time period. The default being the past one month's data from the current date. There following five age categories are configured:

- 0-18 years old
- 19-35
- 36-50
- 51-65
- 66+

Users with CE-Administrator or CE-Developer role can modify the age ranges and the number of age categories.

For more information, see the report details in Age or Gender of Newly Diagnosed Patients (TRC-CER-003)

7.1.6 Biospecimen Samples Status (TRC-CER-006)

The Biospecimen Samples Status report displays the counts of specimen samples that are collected during the specified time period. The default being the past one month's data from the current date. The counts are broken down by the Anatomical Site, listed along the rows.

The list of available sites is configured by the customer during the system installation. The Anatomical Site is a hierarchy where you can drill into the details of each Anatomical site to view its subcategories along with associated counts. The display columns are:

- Type of specimen: Multiple types of specimen collected for each anatomical site. For example, tissue, cell culture, liquid, and so on. All of the available types of specimen are listed for each site.
- Number of Specimen Collected in Facility: Displays the number of specimens collected for each listed Anatomical Site and Type during the specified time period. The default being past one month's data from the current date.
- Number of Patients Corresponding to Collected Specimen: Number of unique patients that correspond to the specimen collected in the previous column.
- Number of Patients Providing Specimen who Consented: Number of patients that have signed consent forms allowing usage of their samples for research. The exact type of consent considered vary by the facility and is configured during the initial system installation.

For more information, see the report details in Biospecimen Samples Status (TRC-CER-006)

CE includes the following prompt set:

- Selection of Date Range Using Statistics from and to Prompt

CE includes the following reports:

- Patients Diagnosed and Treated (TRC-CER-001,TRC-SD-P004)
- Number of Treatments Performed (TRC-CER-005)
- Ethnicity or Race of Newly Diagnosed Patients (TRC-CER-002)

- Age or Gender of Newly Diagnosed Patients (TRC-CER-003)
- Biospecimen Samples Status (TRC-CER-006)

7.1.7 Detailed Information on Each Report and Prompt Set

7.1.7.1 Statistics from-to (Prompt Set TRC-CEP-001)

You are prompted for a date range, where the default value is the past one month's data from the current date. This prompt is used to filter all reports on the Summary page.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Prompt Set Type

- Text Field

Location

- CE Dashboard, Summary page

Dimensions

- Diagnosis.Diagnosis Start Date
- Diagnosis End Date, Procedure. Procedure Start Date
- Procedure. Procedure End Date
- Specimen.Specimen Collection Date
- Medication. Medication Start Date
- Medication.Medication End Date
- Consent.Consent Start Date, Consent. Consent End Date
- Specimen.Anatomical Site Name

Supplementary Reports

None

Reports Referenced

- TRC-CER-001
- TRC-CER-003
- TRC-CER-004
- TRC-CER-005
- TRC-CER-006

Reports Referencing This Prompt

None

Prompt Descriptions

Table 7–1 describes the prompts in the Statistics from-to Prompt Set.

Table 7–1 Statistics from-to (Prompt Set TRC-CEP-001)

Prompt Page Number, Prompt Page Heading	Prompt Heading	Prompt Type or Default Value Layer	Measure or Dimension in PL Layer
1, NA	Statistics from to	Text Field or FROM 1 month back from current date TO current date	Diagnosis.Diagnosis Start Date, Diagnosis End Date, Procedure. Procedure Start Date, Procedure. Procedure End Date, Specimen.Specimen Collection Date, Medication. Medication Start Date, Medication.Medication End Date, Consent.Consent Start Date, Consent. Consent End Date

7.1.7.2 Patients Diagnosed and Treated (Report TRC-CER-001)

This report displays statistics on patients in a facility for the time period specified, the default being the past one month's data from the current date. These statistics include the number of Patients Diagnosed and Treated with disease during the time interval and how many have given consent to medical care.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Report Type

- Table

Location

- CE Dashboard, Summary page

Dimensions

- Diagnosis
- Treatment
- Consent

Supplementary Prompts

- Statistics from-to (Prompt Set TRC-CEP-001)

Reports Referenced

None

Reports Referencing this Report

None

Column Descriptions

Table 7–2 describes the columns in the Patients Diagnosed and Treated report.

Table 7–2 Patients Diagnosed and Treated (Report TRC-CER-001)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	Number of Diagnosed Patients	TRC-CEM-003
NA	Number of Patients Successfully Treated	TRC-CEM-006
NA	Number of Patients who Consented	TRC-CEM-005
NA	Number of Patients Undergoing Treatment	TRC-CEM-007

7.1.7.3 Number of Treatments Performed (Report TRC-CER-003)

This report displays statistics on the number and types of procedures or treatments that have been performed during specified time period. The default being the past one month's data from the current date.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Report Type

- Graph: Vertical Bars

Location

- CE Dashboard, Summary page

Dimensions

- Treatment

Supplementary Prompts

- Statistics from-to (Prompt Set TRC-CEP-001)

Reports Referenced

None

Reports Referencing This Report

None

Column Descriptions

Table 7–3 describes the columns in the Number of Treatments Performed report.

Table 7–3 Patients Diagnosed and Treated (Report TRC-CER-001)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	Surgery	TRC-CEM-010
NA	Chemotherapy	TRC-CEM-011
NA	Radiation	TRC-CEM-012
NA	Targeted (Gene Therapy)	TRC-CEM-013

7.1.7.4 Ethnicity or Race of Newly Diagnosed Patients (Report TRC-CER-004)

This report displays the ethnicity and race statistics for the patients in the facility who have been diagnosed during the time period specified in the dashboard prompt. The default being the past one month's data from the current date. The total count of patients in this report is based on the patients diagnosed during the period specified by the user.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Report Type

- Graph: Pie Type: Default Style: Default

Location

- CE Dashboard, Summary page

Dimensions

- Patient.Ethnicity
- Name Patient. Race Name

Supplementary Prompts

- Statistics from-to (Prompt Set TRC-CEP-001) Display: Ethnicity, Race

Reports Referenced

None

Reports Referencing This Report

None

Column Descriptions

Table 7–4 describes the columns in the Ethnicity or Race of Newly Diagnosed Patients report

Table 7–4 Ethnicity or Race of Newly Diagnosed Patients (Report TRC-CER-004)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	% Total Patients	TRC-CEM-015
NA	Number of Diagnosed Patients	TRC-CEM-014

7.1.7.5 Age or Gender of Newly Diagnosed Patients (Report TRC-CER-005)

This report displays the age and gender statistics for patients in the facility who have been diagnosed during time period specified in the dashboard prompt. The default being the past one month's data from the current date. The total count of patients in this report is based on the patients diagnosed during the period specified by you.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Report Type

- Graph: Pie Type: Default Style: Default

Location

- CE Dashboard, Summary page

Dimensions

- Patient.Ethnicity
- Name Patient. Race Name

Supplementary Prompts

- Name Patient. Race Name

Reports Referenced

None

Reports Referencing This Report

None

Column Descriptions

Table 7–5 describes the columns in the Age or Gender of Newly Diagnosed Patients report

Table 7–5 Age or Gender of Newly Diagnosed Patients (Report TRC-CER-005)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	% Total Patients	TRC-CEM-017
NA	Number of Diagnosed Patients	TRC-CEM-016

Table 7–5 (Cont.) Age or Gender of Newly Diagnosed Patients (Report TRC-CER-005)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	Male	Patient.Gender
NA	Female	Patient.Gender
0-18	NA	Patient.Age
19-35	NA	Patient.Age
36-50	NA	Patient.Age
51-65	NA	Patient.Age
66+	NA	Patient.Age

7.1.7.6 Biospecimen Samples Status (Report TRC-CER-006)

This report displays the statistics on biospecimen collected during the time period specified. The default being the past one month's data from the current date. These statistics include the number of samples collected, the types of samples and their anatomical site, the number of patients corresponding to these specimens and the number of patients that consented for these specimens to be utilized.

Audience

- Physician
- Researcher
- Clinician
- Limited Access User

Report Type

- Table

Location

- CE Dashboard, Summary page

Dimensions

- Specimen.Specimen Type Name
- Specimen.Anatomical Site Name

Supplementary Prompts

- Statistics from-to (Prompt Set TRC-CEP-001)

Reports Referenced

None

Reports Referencing this Report

None

Column Descriptions

Table 7–6 describes the columns in the Biospecimen Samples Status report

Table 7–6 Biospecimen Samples Status (Report TRC-CER-006)

Table Heading	Column Heading	Measure or Dimension in Presentation Layer
NA	Anatomical Site	Specimen.Anatomical Site Name
NA	Type of Specimen	Specimen.Specimen Type Name
NA	Number of Specimen Collected in Facility	TRC-CEM-018
NA	Number of Patients Corresponding to Collected Specimen	TRC-CEM-019
NA	Number of Patients providing Specimen who Consented	TRC-CEM-020

8

Genomic Query

This chapter contains the following topics:

8.1 Omics Data Bank

ODB is a product within the Translational Research Center product family. It consists of a data model to store data as well as a set of tools that is loader scripts to load data into the model. The ODB standalone license does not include any web-based user interfaces. However, if ODB is licensed in addition to CE, several web-based User Interfaces are enabled to query against the ODB model.

ODB can be functionally thought of in two groups of tables. One set includes the reference tables which provide the metadata required to link results to specific portions of the genome. The second set represents the result tables used to capture results and link each result to an object in the reference model, and link the results back to the patient. The patient link is accomplished by linking the ODB with the Cohort Datamart, which is a part of CE 3.1.

The following sections list user interfaces available for querying ODB when licensed together with CE.

8.2 Genomic Query

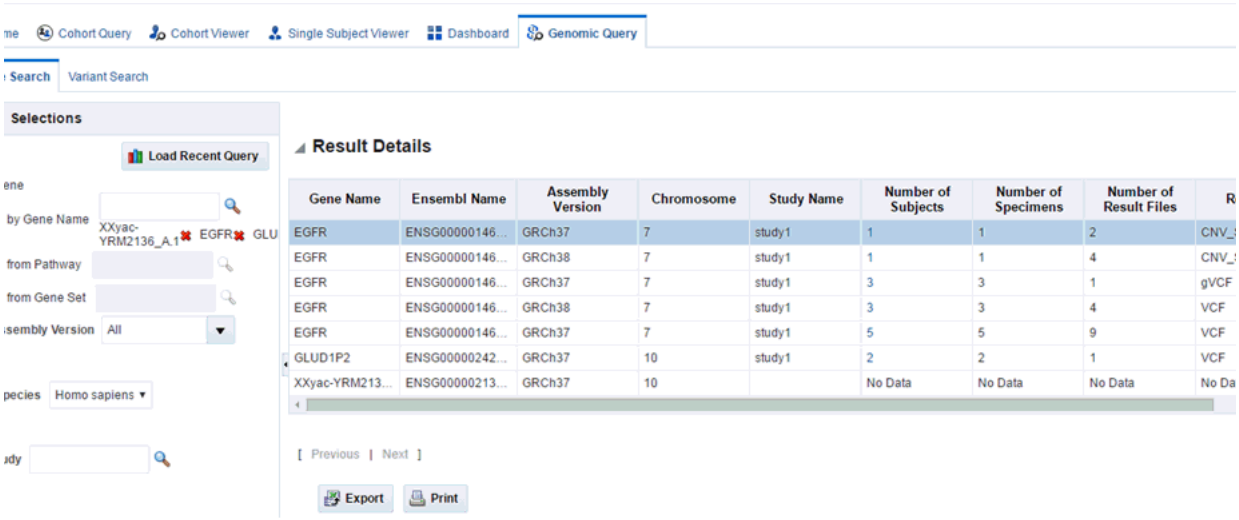
The **Genomic Query** tab can be utilized as a search interface into the patient genomic results in ODB. Two search modes are provided.

8.2.1 Gene Search

This tab lets you to specify one or more genes to be searched across results in ODB. You must specify Species, Assembly Version and Study that provides the output of the result data.Study is an optional criteria that you can provide in the subject context.

Assembly version restricts all the results associated against it. For **Gene Search**, select the Assembly version. Rather only the annotations of the selected genes are retrieved based on the Assembly version. The default Assembly version is the one that was last loaded in ODB, but you can select any one. If you select the Assembly version GRCh38, then the query retrieves gene co-ordinates for this Assembly version and a list of chromosomes present in this region.

Figure 8–1 Gene Search



There are currently three ways by which the genes may be specified:

- Specify one or more genes explicitly by name.

  Note: When you search for a gene (hugo_name), you may also see results for a different hugo_name. This is because of synonyms or aliases present in the gene cross-reference in the ODB data model.

  These synonyms or aliases originate from different data sources and are loaded from HGNC HUGO data source. For example, when data loaded from HGNC Hugo has results associated with it, if you select hugo_name BNIP3 in the genomic query, you may also see results of NIP3. If you search for A1CF gene, you may see an alias ACF64.

  These synonyms and alias do not appear when you search with pathway as input.

- Specify gene set. If you have a set of genes that you would like to search for, you can specify the gene set and all participating genes in the gene set is searched for results.
- Specify pathway to search for all genes belonging to a given pathway. Pathway information is on the reference side of the ODB model tables.

Once appropriate selections are made, click **Submit** to see the output information. This information includes:

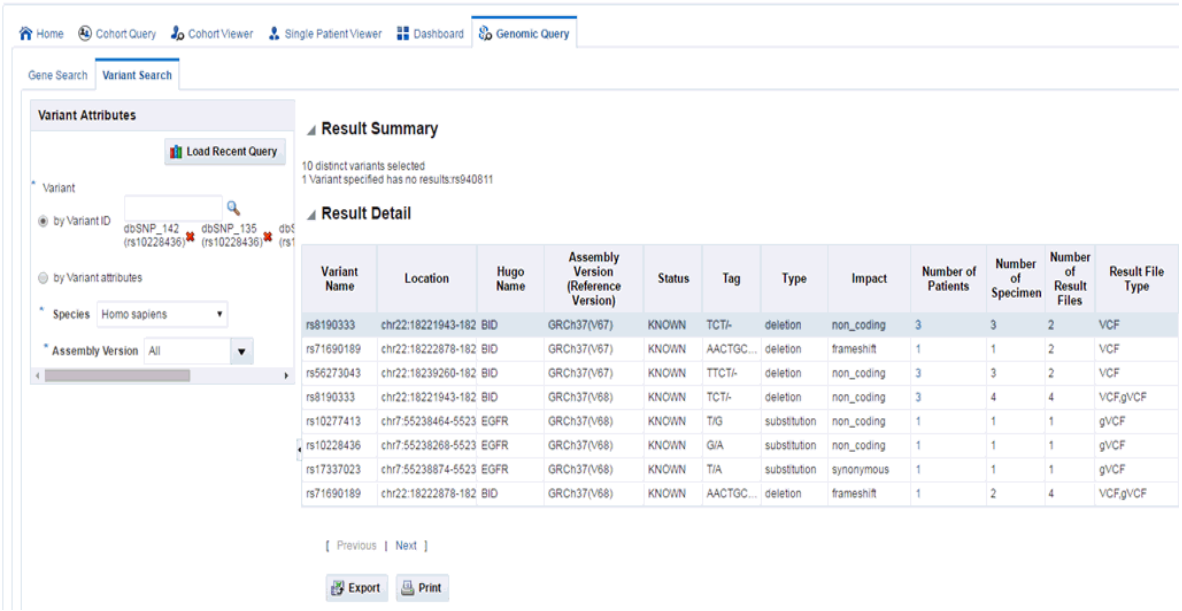
- Tabular list of metadata for each gene that is found and not found to have genomic results in ODB. Columns for which no data is found are labeled with *No Data*. The columns include:
  - HUGO gene name,
  - Ensembl gene name,
  - Assembly Version,
  - Chromosome

- the number of distinct patients (or subjects) found with any data for a given gene,
- the number of distinct specimen found with any data for a given gene,
- the number of matching result files present in ODB
- the type of result files present in ODB. For example, VCF, MAF, gene expression.

You can export the resulting table to a spreadsheet or print it. Click the Patient count to navigate to different screens like View Record, Circular Genomic Viewer, Cohort List, Cohort Timelines, Cohort Reports and Genomic Data export.

If the pathway, geneSet or genes have been selected from the ad hoc genes selection, the result details table will be paginated based on gene selection. The first page displays the result for 25 gene selection and the next page will be the result for next 25 genes and so on.

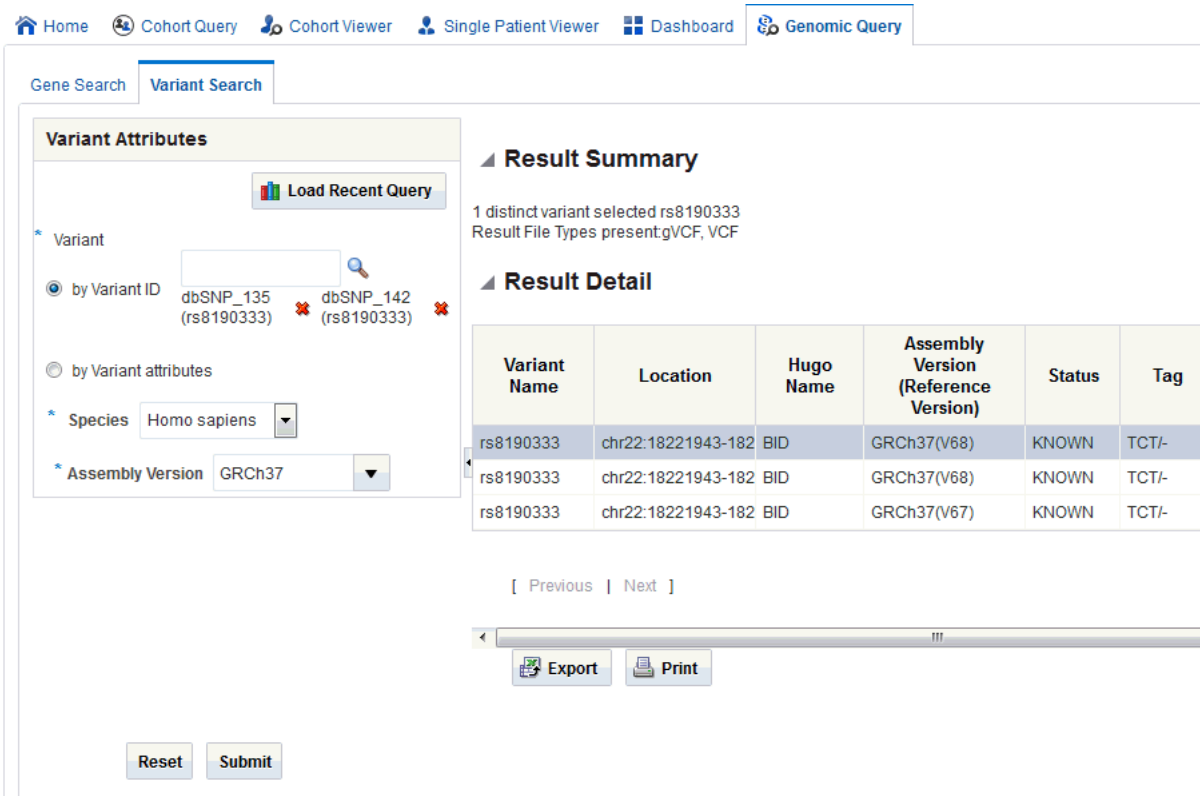
If there is no result data for the gene you are searching, then it appears in the table with No Data in the columns as specified in Figure 8–1.



8.2.2 Variant Search

This tab lets you to specify one or more variant identifiers to be searched across results in ODB. You must specify **Species** and **Assembly Version** that displays the result data. Study can be provided as an optional parameter in the subject context.

Figure 8–2 Variant Search



There are currently two ways that variants may be specified:

- Specify one or more variants explicitly by their variant identifier. Currently, the reference side of the schema once loaded will have Ensembl sourced variant identifier which include dbSNP and Cosmic identifiers. You can specify either one of those to try to find variants.
- Specify variants by attributes like location on a gene or DNA source and additional metadata. You can provide genes in 3 ways:
  - Specify one or more genes explicitly by name.
  - Specify gene set. If you have a set of genes that you would like to search for, you can specify the gene set and all participating genes in the gene set are searched for results.
  - Specify pathway to search for all genes belonging to a given pathway. Pathway information is on the reference side of the ODB model tables.
- You can specify Genomic Position by typing in chr#:from-to location. For example, to find variants in base pair region 1-200 on chromosome 7, you would need to type in chr7:1-200. Also, if you want to search for any variants on chromosome 7, you can just type in chr7. Additional options that can be specified include known or novel status on the variant, variant type such as insertion, deletion, substitutions, indel, complex. The strand is the directionality of the variant and it can be either plus (+) or minus (-).

Once appropriate selections are made, you can click **Submit** to see the output information. This information includes:

- Summary of the output: this displays how many variants were searched for, how many have results found and, if applicable, how many do not have any results present in ODB for the criteria selected, what types of result files are found for the variants if any.
- Tabular list of metadata for each variant that is found to have genomic results in ODB. The columns list includes the following option:
  - Variant Name, if the name exists,
  - Location of the variant,
  - Hugo Name
  - Assembly Version (Reference Version)
  - Status, whether the variant is known or novel,
  - the replace tag that tells the end user what exactly the variant is,
  - Type, whether this variant is an insertion, deletion and so on,
  - Impact, the impact of a given variant (only available for known variants),
  - Study, the study this variant has been found to have results for,
  - Number of Patients, the number of distinct patients found with any data for the particular variant,
  - Number of Specimen, the number of distinct specimen found with any data for the particular variant
  - Number of Result File, the count of type of result files present in ODB,
  - Result File Type, type of result files present in ODB, for example, VCF, MAF, gene expression with data for a given variant.

Note: On gene search selections: A variant may be present in alternate genes that overlap with the gene region of the selected gene in a given version. For any variant found, these alternate gene annotations are displayed to the user along with annotations from gene of interest.

You can export the resulting table to a spreadsheet or print it.

9

Benefits of Cohort Explorer and Omics Data Bank if Purchased Together

This chapter contains the following topics:

9.1 Benefits

Using Cohort Explorer together with Omics Data Bank you can perform the following functions:

- Easily and rapidly assess how many patients (or study subjects) fulfill any complex set of custom criteria coming from a combination of clinical, demographics, and omics domains.
- Stratify patients (or study subjects) based on comparative analysis of certain events occurring, for example, find patients who have a specific mutation within three months after undergoing gene therapy treatment.
- View details of a cohort of individuals, then drill down to the individual patient's medical history without revealing patient's identity for those users who do not have permission to see identifiable data. You can scan through each individual patient using the Next and Previous buttons.
- For Individuals with permission to see personally identifiable data for patients, they can see PI data of individuals they have access to, and obfuscated data for those patients they do not have permission to see.
- For a single patient or study subject, select categories of demographic, clinical, or genomic data to show all his/her available data in one page. A comprehensive view of all genomic results for patient's/subject's specimen is available with drill-in options to view those results in Circular Genomic Viewer. Furthermore, a hierarchical viewer shows variant profile of a given individual along with impact and location categorization of each mutation. For those patients/subjects who have lower level (for example BAM, SAM, and so on) files linked in with their interpreted results, user with appropriate permissions can see the file lineage and download file from the Cohort Explorer interface. You can also view the gene and variant level information in a Dalliance genome browser from the Single Patient Viewer.
- Quickly generate a variety of prebuilt demographic, clinical and genomic reports based on a custom cohort. You can use those reports for analysis and research into relevant features.

- Sequence Variants and Copy Number Variation cohort reports facilitate quick analyses of the number of specimen found with mutations, no-mutations or no results for a selected set of genes. You can select to drill further and view these results in a Genome-like Browser which shown chromosomal position of each mutation or CNV, respectively. You can drill further to see what specimen/patient exactly matched for a given feature.
- Using the Gene vs Specimen matrix report, you can view the high level matrix plot with mutation in selected genes for each specimen of the cohort. The matrix plot is color coded for each specimen based on mutation, non-mutation and no-data points. The plot also displays CNV information for the gene of the specimens, if available. Patients are stratified based on specimen number, which are ordered based on the collection date.

  Using the Variant vs Specimen matrix report, you can view the high level matrix plot with variants for each specimen of the cohort. The matrix plot is color coded for each specimen based on mutation, non-mutation and no-data points. Patients are stratified based on specimen number, which are ordered based on the collection date.
- Structural Variation histogram cohort report performs search for the most frequently structurally mutated genes or gene pairs.
- Evaluate the efficiency of your organization via the operational dashboards in the Dashboard page.
- Assess what biospecimen samples are currently available for a given set of patients or study subjects.
- Use Circular Genomics Viewer to get a glimpse on genomic data for a given patients/study subject. Data shown in circular viewer includes variants, single and dual channel gene expression, RNA-seq results and Copy Number Variation results.
- Export genomic data into other gene browser readable formats, based on patient sample analysis. Genomic data supported includes gene expression, copy number variation, mutation data. Select to either export the data immediately or schedule a job for asynchronous file generation. Once job is completed, you can download generated file(s) from under Home > Jobs page.
- Use Genomic Query interface to search for genes or gene variants across genomic result data in ODB.
- Identify how many patients/subjects and specimen have been found to have genomic results for genes or gene variants, and then drill into the patient/subject list to inspect those individuals for any other features.
- Keep track of your saved cohort queries, as well as cohort lists. Easily share those with other users or user groups using a simple web user interface.
- View all your queries, cohort lists, gene sets under My Workspace tab where you can also see any queries or lists that have been shared with you.
- Keep track of your favorite genes as a saved Gene Set.
- Turn on logging to track users and any activity within the application accessing data in either the Cohort Data Mart (CDM) or in the Omics Data Bank (ODB) model.

Index