Oracle® Health Sciences Omics Data Bank

Release Notes

Release 1.0.1

E27535-02

April 2012

This document provides a complete, up-to-date description of Oracle Health Sciences Omics Data Bank (ODB). You should read and understand all tasks described here before you begin your installation.

The current version of the Release Notes for Oracle Health Sciences Omics Data Bank Release 1.0 is available on Oracle Technology Network. Ensure that you have the latest version before you begin.

This document contains the following sections:

Audience

This document is intended for the following audience:

- Scientists
- Bioinformaticians
- Biostatisticians
- Physicians
- Clinicians
- Researchers

Additional Product Information

Refer to the most recent version of your ODB user and implementation documentation for related information. For a list of available ODB documentation, refer to Implementation and User Documentation.

Features

Omics Data Bank (ODB) consists of several components:

- A data model for storing and querying genomic data including genes, proteins, pathways, variants, expressions.
- Scripts for uploading reference data from public sources into the model.
- Scripts supporting the upload of user result data in four popular formats including Variant Call Format (VCF), Mutation Annotation Format (MAF), Complete Genomics (CGI) masterVar format, gene expression tab-separated values.

The objective of the Omics Data Bank is to expand the reach of Oracle Health Sciences Translational Research Center into the omics domain towards the goal of enabling biomarker discovery and validation. ODB v1.0 and Cohort Data Model (CDM), part of Cohort Explorer v1.0 release, can be integrated to enable querying for patients based on attributes from either the clinical or omics domain.

The features of ODB include:

- A schema optimized for fast, real-time queries
- Querying for patients, specimens, genes, proteins, pathways
- Tight integration with CDM and the flexibility of integration with any clinical data mart
- Focus on storing human references and results, yet designed to concurrently store data from any species.
- 27 reference tables to store information on publicly established references:
  - species, DNA sources, chromosomes, genes
  - proteins
  - gene and protein components: exons, introns, beta sheets, alpha helices, and so on
  - pathways
  - literature, information, terminology
- 13 result data tables to model user results, including:
  - gene expression
  - variants—substitutions, insertions, deletions
  - copy number variations (CNV)
- Command line Java and PLSQL scripts for loading public reference data: Ensembl, SwissProt, HUGO, Pathwaycommons
- Command line configurable scripts in PLSQL enable the end user to upload results in popular result data formats including VCF, MAF, CGI, gene expression .tsv files
- Integration with Oracle Secure Files for secure access, traceability, and compression

By combining Omics Data Bank with the clinical data model underlying Oracle Health Sciences Cohort Explorer v1.0, the gap between the clinical and genomic data can be bridged.

For example, you can rapidly assess the number of patients qualifying for a given study based on their medical history; including diagnosis, medications or procedures already undergone as well as based on results from genome tests indicating presence of a particular mutation.

You can query for genes that belong to a particular pathway and review how the expression of these genes vary in tumors, healthy samples and samples collected after a treatment.

Multiple genomic features for patients can be queried together in order to narrow down the causes for the success or failure of treatment in particular patient sub-populations. This is the first step towards value-based, personalized health care.

Enhancement Introduced and Issues Resolved in Omics Data Bank 1.0.1

This section lists the enhancementsintroduced and issues resolved in Omics Data Bank 1.0.1 release:

- Relationships among multiple Result Tables in ODB schema have been altered to improve performance of loading and querying.
- To enable novel partitioning strategies, several new tables have been added, for example, W_EHA_STUDY, W_EHA_SPECIMEN, W_EHA_CHROMOSOME. New partitioning strategies significantly improved query performance.
- Several tables have been removed, for example, W_EHA_RESULT, W_EHA_ RSLT_LINK saving disk space without affecting functionality.
- Tables holding different result types, for example, sequencing, no-call have been split into multiple tables, resulting in more efficient storage use without compromising on requirements.
- Result loaders including MAF, VCF, CGI masterVar and gene expression have been rewritten to utilize external tables resulting in ten-fold or greater performance improvements when loading data.
- GVF loader has been rewritten to utilize external tables resulting in more than five-fold improvement in loading time without compromising functionality.

Data Model Overview

Oracle Health Sciences Omics Data Bank can be thought of functionally as two groups of tables. One set of tables are the *reference* tables which provide the metadata required to link results to specific portions of the genome. The second set of tables in the model are *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 Omics Data Bank with the Cohort Data Model (part of Oracle Health Sciences Cohort Explorer v1.0).

Reference Tables

Reference data consists of 27 reference tables designed to store data originating from public reference sources. The schema is designed in a star-like structure to allow real time querying through direct SQL and in the future, through pre-designed User Interfaces as well.

The reference schema is designed to be populated from 4 distinct sources:

- The gene information is loaded from EMBL files which are stored online in the Ensembl database (http://www.ensembl.org). Ensembl is a joint project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute. This online database maintains references to other online database projects (dbSNP, NCBI, Cosmic, etc.) and provides references to each of these databases. The model loads this cross reference information to allow queries to use specific database references if needed.
- The protein information source is the online SwissProt database (http://www.ebi.ac.uk/uniprot/). This database project is a consortium of various groups including the European Bioinformatics Institute.
- The HUGO Gene Nomenclature Committee (http://www.genenames.org/) provides reference seed information needed for identifying gene locations. The HUGO gene names are needed to find various cross references as well as the correct chromosome number for each gene. The HUGO Gene Nomenclature Committee is the authoritative group for all gene names.
- The last reference source is the Pathway Commons (http://www.pathwaycommons.org,) which collects information on pathways and proteins/genes participating in each pathway. The coverage of pathways as reference is limited to information available in the gsea format files.

Note: Oracle provides loader scripts to auto-populate reference tables but does not ship the schema pre-populated with reference data.

The table below contains a list of Reference Tables. However, this is not an exhaustive list.

Table	Description	Benefits
W_EHA_SPECIES	The species table stores information about each genome stored in the database. The current model allows for any number of species genomes to be loaded.	Allows handling of multiple species in the same schema. Customizations can be done on per species basis; for example, promoter offset.
W_EHA_DNA_SOURCE	The DNA_SOURCE table stores multiple records for different reference DNA strands for each species. Each cell in the species has a copy of this reference DNA. There are buffers of DNA considered to be the reference for each organism. These reference strands are then used to map detected variations for each organism tested.	Allows for quick determination of chromosomes for each variant. A source of reference buffer for each species.

Table	Description	Benefits
W_EHA_GENE	Each chromosome of the DNA strand has many different genes. Each gene has a starting position and an ending position. The entire size of the gene does not create the protein directly, but there are recognized sections of the DNA that scientists agree should be considered as part of the gene. The GENE table has fields for how the Ensembl database refers to the gene, as well as the recognized gene name.	Links gene Ensembl name with HUGO name. Common source of reference for multiple, dependent tables including GENE_ STRUCTURE, XREF, and GENE_COMPONENT.
W_EHA_HUGO_INFO	This table is very important for storing reference seed information needed for identifying gene locations. The EMBL files reports each gene with a LOCUS_TAG which uses the name registered with the HUGO Gene Nomenclature Committee (**http://www.genenames.org/**).	Stores identifiers for genes coming from different source databases. Keeps track of all gene names and synonyms. Matches each gene to a chromosome.
W_EHA_PROTEIN	Most of the known genes produce types of protein molecules. The PROTEIN table contains the amino acids that comprise each protein molecule. The amino acid sequence comes from Ensembl while the SwissProt files give much more information about the protein molecule. There are more descriptive names for each protein which are not stored in the EMBL file (such as, insulin). The SwissProt file provides links to cross references as well as literature references.	Keeps track of a protein's amino acid sequence and summary for a quick lookup.
W_EHA_VARIANT	The VARIANT table is used to record known variants that differ from the reference. Most of these variants are well documented and compiled from other research. When results are uploaded, sometimes novel variants are detected and there are no known references for this variant. These results generate new VARIANT records which may be of interest to researchers when they look for novel discoveries.	Stores both known and new variants and provides for a quick matching and lookup of variants and their location on the DNA sequence.

Table	Description	Benefits
W_EHA_PATHWAY	Pathway is used to describe a series of interactions in a cell. Many different types of biological pathways exist; including genetic, metabolic, signaling, and so on. This table is used to store publicly available pathways. A given gene or protein may belong to many of such pathways.	Intended to be used to store pathways. However, it can be used to name any collection of genes and/or proteins.

Reference Loaders

The reference tables listed in the previous section are not pre-populated with reference data at the time of shipping. However, all necessary scripts and utilities are provided that allow you to easily populate the Reference tables. Necessary steps for populating Reference tables are described in the Oracle Health Sciences Omics Data Bank v1.0 Programmer's Guide. There are four loader scripts:

- W_EHA_HUGO_INFO table loader, which populates only gene reference tables from HUGO
- W_EHA_VARIANT and related tables loader, which populates variant information from Ensembl gvf files
- W_EHA_PATHWAY and related tables loader, which populates pathway and pathway member information from http://www.pathwaycommons.org
- W_EHA_DNA_SOURCE, W_EHA_GENE, W_EHA_PROTEIN and related tables loader, which populates the DNA sequence, gene, and protein data from Ensembl and SwissProt databases

Result Tables

Result data consists of 13 tables designed to store result data coming from user files. Supported results include sequencing data such as simple variants, copy number variation, no-call results, and gene expression data. The schema is designed to enable efficient querying for patients or specimens with any combinations of genomic attributes in ODB and clinical attributes in Cohort Data Mart. The schema is designed to allow linking to files in the SecureFile system.

The table below contains a list of Result Tables. However, this is not an exhaustive list.

Table	Description	Benefits
W_EHA_RESULT	Any result record stored in RSLT child tables (for example, SEQUENCING, GENE_EXP) has a mirror record in the RESULT table. This table is intended to store derived higher level results. In addition, this table links to the metadata tables that contain information about the type of results, specimen database source, type of result file.	Allows linking different types of parent results to child results (Level 3-4 data). Links to the specimen data source and identifier from the clinical datamart; for example, Cohort Explorer datamart.

Table	Description	Benefits
W_EHA_RSLT_ SEQUENCING	This table contains sequencing results, more specifically, variant information such as insertions, deletions or substitutions. Each variant record in the SEQUENCING table is linked to its reference in the W_EHA_VARIANT table. Loading scripts are provided to parse and load data into this table from three types of files: VCF, MAF or Complete Genomics (CGI) masterVar.	Stores variant information from different sources. Linked with reference, allows for fast identification of novel variants.
W_EHA_RSLT_NOCALL	This table is designed to store no-call information coming from the Complete Genomics masterVar result file and the provided loading script is automatically loads no-call results.	Allows the storing of no-call results. This functionality is typically not supported in known data models.
W_EHA_RSLT_COPY_NBR_ VAR	This table is intended to store copy number variation (cnv) results. Currently this table needs to be populated manually by the end user; however, there are plans to write auto-loading scripts from common formats.	Assuming cnv results are populated, it allows the end user to query for combinations of copy number variation and other types of results.
W_EHA_PROBE	PROBE table stores probe identifiers, and other data such as the sequence identifying each probe. PROBE table links each probe to a corresponding gene. With an additional XREF table, probes can be annotated with any relevant cross reference information. Loader scripts are provided allowing the user to populate these tables from a standardized format.	Probes from different platforms and vendors can be stored and easily linked to genes. Designed to store probe metadata.
W_EHA_RSLT_GENE_EXP	This table stores gene expression results and quality metrics. Loader scripts exist, allowing the user to upload their results from properly formatted files.	Easily upload results from gene expressions coming from standard platforms, and represented via probes.
W_EHA_DATASOURCE	This table is used to store information about clinical data sources for specimens collected that have genomic data in ODB. This table is pre-seeded with CDM (Cohort Data Model) information, allowing ODB to seamlessly integrate clinical and genomic patient data.	Allows linking to clinical information about a specimen, including patient information. Can be used to integrate with any clinical datamart, not only CDM.

Table	Description	Benefits
W_EHA_RSLT_FILE	This table is used to store the path-to-source file and other information when linked with the RSLT_FILE_TYPE table. Built-in support allows for linking to both external file systems as well as SecureFiles.	ODB data model supports either regular ('external') file system storage as well as SecureFile storage of input result files. Metadata about each file is stored in the linked RSLT_ FILE_TYPE table including type, version, vendor etc

Result Loaders

To support easy loading of user results, the Omics Data Bank includes four loaders designed to load data from popular result formats into the result tables. Sequencing data loaders are written in PLSQL and support the loading of variant data from VCF, MAF, CGI masterVar files. Additionally, for CGI masterVar, the loader supports the loading of no-call results. For gene expression, the included PLSQL loader is designed to load either gene probe information or gene expression results corresponding to probes.

Secure File Support

The ODB schema provides placeholders to be able to link to files stored in the Secure File system. The W_EHA_RSLT_FILE table allows for differentiating between regular ('external') and SecureFile storage option.

> **Note:** Oracle tests that all entities needed for the Secure File system link are addressed in the schema. If you wish to insert result records in SecureFile, you must manually insert the linkage attributes (for example, FILE_CONTENT_ID).

Integration with Oracle Health Sciences Cohort Explorer Data Mart

Oracle Health Sciences Omics Data Bank v1.0 is designed to seamlessly integrate with Oracle Health Sciences Cohort Explorer (OHSCE) v1.0.

CDM is part of Oracle Health Sciences Cohort Explorer v1.0 and consists of patient-relevant dimensions and associated measures. CDM is designed to optimize patient-centric queries. When integrated with ODB, it enables querying for patients or specimens based on combinations of clinical and genomics attributes. For further information on Cohort Data Mart, refer to the Oracle Health Sciences Cohort Explorer v1.0 documentation online.

Instructions for Installing Oracle Health Sciences Omics Data Bank

Refer to the *Oracle Health Sciences Omics Data Bank Secure Installation and Configuration Guide* version 1.0, for detailed instructions on how to install and set up the ODB application.

Implementation and User Documentation

The following ODB version 1.0 user documentation is available on Oracle Technology Network at http://www.oracle.com/technetwork/index.html:

- *Oracle Health Sciences Omics Data Bank Release Content Document* (Part E27537)
- *Oracle Health Sciences Omics Data Bank Secure Installation and Configuration Guide* (Part E27536)
- *Oracle Health Sciences Omics Data Bank Secure File Store Guide* (Part E27538)
- *Oracle Health Sciences Omics Data Bank Programmer's Guide* (Part E27509)
- *Oracle Health Sciences Omics Data Bank Electronic Technical Reference Manual* (available on My Oracle Support at https://support.oracle.com/)

Disclaimer Regarding Third Party 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.

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 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.

Oracle® Health Sciences Omics Data Bank, Release 1.0.1
E27535-02