Oracle Health Sciences
Translational Research Center:
A Translational Medicine Platform
to Address the Big Data Challenge
Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.
Abstract

As organizations move to advance translational research to achieve personalized medicine, researchers and clinicians face a barrier in the informatics arena. There is a shortage of fully-integrated informatics solution that let researchers and clinicians integrate, store, and analyze clinical and omics data from diverse sources, generated in-house as well as by public consortiums. Many researchers and clinicians also have to rely on bioinformaticians to perform mundane data management tasks in order to validate a simple hypothesis. Oracle Health Sciences Translational Research Center provides a complete and scalable informatics solution for translational research, with centralized data storage and analysis across genetic information areas (genomics, transcriptomics, and proteomics), vendor platforms, biological data types, and clinical data sources. It also offers flexible deployment options such as, an on-premise, a HIPAA-certified Software as a Service (SaaS), and a hybrid of these two. Through an intuitive user interface, Oracle Health Sciences Translational Research Center lets researchers stratify patients, and clinicians evaluate treatment response for similar patients in a self-sufficient manner, ultimately shortening the biomarker development cycle and accelerating the adoption of personalized medicine.
Key Areas that Oracle Health Sciences Translational Research Center will Support Your Institution

- A productized solution with quick-to-value time and long-term support
- High quality standardized clinical data
- Integration and analysis of cross-vendor and cross-modality omics data irrespective of scientific approach and technology
- Seamless integrated solution, which frees bioinfomaticians from mundane data management tasks to scientific innovation
- Analysis of internal datasets in the context of ever increasing large public domain data sets
- Traceability features fulfill data governance and auditing requirements
- Optimized for fast real-time queries with combination of hardware and software features, and scalable petabytes date records
- Flexible deployment options supporting on-premise, HIPAA-certified SaaS, and hybrid options

Introduction

Rapid growth of genomics data, increased patient expectations of quality of care, and financial pressure on efficient operation have created fundamental challenges and opportunities for healthcare and life sciences organizations to develop and implement therapies based on biomarkers. One of the big promises of personalized medicine is an increase in drug efficacy when compared to the current state. In 2008 and 2009, success rates for new drugs in Phase II clinical trials from 16 pharmaceutical companies decreased from 28% to 18%. Of these failures, 51% were due to insufficient efficacy and 19% due to preclinical safety. Although many organizations, including healthcare providers, insurers, pharmaceutical companies, and diagnostic laboratories, realize the potential of personalized medicine, they still face a critical barrier in inadequate implementation of informatics solution.

The main obstacles to adoption of the new paradigm are:

- Lack of seamlessly integrated, productized, and scalable infrastructure to support data management, analysis, and reporting.
- Vendor and modality-specific omics data, which are difficult to interrelate, and isolate data in clusters (that is, silo omics) that are – by themselves – insufficient to provide biological insights.
- Difficulty in jointly analyzing public domain data and data generated in-house.
- Disparate clinical data sources requiring heavy post processing.
- Lack of flexible and HIPPA-compliant deployment models.

As the cost of DNA sequencing continues to decrease and technological advances make sequencing both easier and more robust, the bottleneck in translational research has shifted from data gathering to data analysis. Due to this, many researchers are finding insurmountable analytical hurdles in testing simple hypotheses without the ad hoc assistance of a bioinformatician. For bioinformaticians, due to poor infrastructure design, it has been hard to build tools to enable researchers to be self-sufficient in their standard analyses. This has created an environment where bioinformaticians are required to perform mundane tasks, diverting their time and focus away from challenges and innovations. This ad hoc analysis paradigm often leads to scattered data and analysis files across multiple storage devices. As a result, it is difficult to reproduce results and transfer knowledge to external collaborators in accordance with regulatory requirements. In the past, some institutions have contracted consultants to build customize solutions, which has taken a long time to build. The customize solution also becomes a maintenance challenge due to high cost long-term contrast for external expertise and disconnect in knowledge between contractors and in-house resources.

Most informatics infrastructure today focuses on a single omics technology vendor (for example, Illumina, Life Technologies, and so on) or in a single omics data modality (for example, genome versus transcriptome). While useful, this approach provides a fragmented picture of the underlying biological processes. Overwhelming evidence has shown that an integrated approach is the key to identify root cause of a disease based on gene structure, expression, and regulation across multiple omics modalities. For instance, an integrated analysis of cross-modality glioblastoms (GBM) data, including DNA copy number, gene expression, and DNA methylation aberrations, helped dissect genome-wide regulatory mechanisms for further investigation into the identification of candidate biomarkers for GBM tumors and potential therapeutic targets. A holistic view of cross-vendor and cross-modality data is thus critical for the development and delivery of targeted medicine.

The research community has devoted significant resources to make large-scale genomics data available to the public. For example, The Cancer Genome Atlas project (TCGA) and the International Cancer Genome Consortium (ICGC) offer comprehensive sets of complementary resources for researchers to understand the complete repertoire of mechanisms contributing to tumor initiation and progression. Similarly, the 1000 Genomes project provides a detailed catalogue

---


of human genetic variation. These efforts have provided a great opportunity for researchers to combine public data with the data generated in their own laboratories, to stimulate the testing of new types of hypotheses, increase statistical power of tests, and sub-sequentially generate new biological insights. Given these public data, researchers and clinicians can theoretically map their own patients' molecular profiles against the public data to predict disease progression and treatment responses. However, this approach requires daunting manual steps or complex programming skills in the current informatics infrastructure. For example, for verifying whether a particular mutation occurs in TCGA GBM patients, a researcher needs to pass through many files on the TCGA website to identify the relevant ones, to download them to the local storage, to write script (using bioinformatician's help), and then to perform the analysis and prediction. As a result, most public data are grossly under-utilized. Furthermore, this approach also leads to multiple copies of the same public data to reside in an institution as researchers in different labs independently manage their own copy, multiplying data storage, and labor costs.

On the clinical front, many translational research projects disproportionately invest in molecular profiling technologies compared to the clinical information collection systems, which is commonly done through a project-specific case report form. Without heavy and consistent post-processing, clinical data reuse and comparisons across projects are impossible. As translational research moves toward the bed-side, the sources of clinical data will continue to expand beyond case-report forms, and will include electronic health records, state or national registries, and records from various departments such as accounting, pharmacy, and pathology. From an informatics perspective, translational research tools need to support the integration and standardization of clinical data from disparate sources, with the help of consistent terminology and units of measure. Once these data is standardized, it will not only benefit translational research but also the analysis of institutional-wide metrics such as operating resource and physician performance metrics.

Institutions in different healthcare and life sciences areas have divergent informatics needs: from on-premise infrastructure to integrated and externally-managed solutions such as software as a service (SaaS). Given the sensitive nature of clinical records and increased regulation of omics data, the hosting system must be compliant with the Health Insurance Portability and Accountability Act (HIPAA). For example, the system must be able to encrypt sensitive data and audit each user's actions on Protected Health Information (PHI) with time stamp. Yet, there are limited options available in HIPAA-certified system. As both data and analysis standardize, it is likely that healthcare and life sciences institutions will continue to shift from on-premise to SaaS, or to a hybrid model, where only a relatively small set of crucial data and analyses are managed in-house. The remaining data and analysis are managed in the cloud, with the continuous synchronization between local copies and the SaaS solution.
Given the above barriers in informatics, we therefore believe a complete informatics solution for translational research must:

• be a productized solution with quick-to-value time and long-term support
• have a scalable system for data management, analysis, and reporting
• contain a user interface that can be managed without programming experience
• have a seamless integration between public domain data and data generated in-house, as well as an integration between public algorithms and algorithms developed in-house to define the best practices
• support standardization at many levels, from clinical data, to vendor technologies, and omics modalities
• have a flexible yet regulatory-complaint informatics infrastructure deployment model.

Oracle Health Sciences Translational Research Center addresses all of these requirements. It standardizes data sets and brings them together on a common storage, leveraging the economics of scale and keeping track of lineage information. Oracle Health Sciences Translational Research Center also captures biological relationships among data in the repository, allowing it to adequately represent data across vendor platforms and omics modalities. Since Oracle Health Sciences Translational Research Center is agnostic to data generating vendor platforms and can capture relationships among cross-modality data, both public data and data generated in-house can be analyzed on a common platform to achieve biological insights. Hosting a large amount of data from diverse sources, Oracle Health Sciences Translational Research Center has also been optimized to handle petabyte size data by combining software and hardware features to support extreme performance.

Oracle Health Sciences Translational Research Center was developed in collaboration with strategic development partners of repute such as, Erasmus Medical Center, Inova Health System, Merck & Co., Moffitt Cancer Center, Oregon Health and Sciences University, and Roche Ltd. Its productized approach ensures that product features support best practices in translational research and accelerate time-to-value. Oracle Health Sciences Translational Research Center is available as either on-premise or HIPAA-certified SaaS deployment, or hybrid options.

Oracle Health Sciences Translational Research Center Features

Intuitive Patient Stratification Interface with Both Clinical and Omics Criteria

Finding subjects with certain characteristics is a fundamental step in biomarker discovery and clinical decision support. Oracle Health Sciences Translational Research Center provides an interface that makes this step easy without the need for any programming or scripting skills. With a few clicks, users can specify both clinical and omics criteria of a desired patient group.
Figure 1 illustrates a use case where the user wants a group of male patients with non-small-cell-lung carcinoma (NSCLC) with a mutation in the EGFR gene, but no copy number gain of the MET gene. There are four criteria specified in the interface:

- Gender criterion in the Patient Information category
- Diagnostic criterion in the Clinical Data category
- Mutation criterion in the Genomics Data category
- Copy number variant criterion in the Genomics Data category.

Once all the criteria are specified, the user can obtain the number of patients who meet the criteria along with detailed information of selected patients in either a tabular or an event timeline presentation.

Cross-Vendor and Cross-Modality Omics Data Repository and its Biological Context – Simple yet Comprehensive

A critical factor in the success of an integrated approach to translational research is the ability to interrelate omics data. The relational tables within Oracle Health Sciences Translational Research Center is designed to follow the Central Dogma of Molecular Biology, following the series of biological events in the cell, from DNA sequence, to splicing, to transcription to RNA, to finally translation, and then to protein.
There are two sets of relational tables modeling omics data in Oracle Health Sciences Translational Research Center: the reference and the result tables. The reference tables store the reference genome builds and associated annotations, while the result tables store each specimen’s data generated by laboratories and data from public data repositories (for example, TCGA). Selected entities in the reference tables are shown in Figure 2.

- **DNA_STRAND table** stores each species’ genome build.
- **GENE and GENE_COMPONENT tables** store genes and their sub-components, as well as additional annotations.
- **PROTEIN and PROTEIN_COMPONENT tables** store information about proteins, their sub-components, and annotations. As shown in Figure 2, Oracle Health Sciences Translational Research Center captures the fact that only subset of gene components translate into protein, accounting for alternative splicing events.
- **PATHWAY table** models pathway memberships
- **VARIANT table** stores the differences compared to a reference genome, including single nucleotide polymorphisms (SNP), insertions, and deletions.

The generality of the reference tables can accommodate data from many reference sources such as Ensembl, NCBI, HUGO, and so on. To facilitate adoption, a set of data adapters are included as part of the product for reference source data and these details are provided in Section 3.2. Although many existing genomics data processing solutions are restricted to store known variants,
this is insufficient for users of the Next Generation Sequencing (NGS) technology, who commonly discover new polymorphic sites. Thus, Oracle's Oracle Health Sciences Translational Research Center solution models both known and novel variants by combining variant attributes in the VARIANT table and coordinate information in the DNA_STRAND table.

These reference tables not only provide biological context, but also serve as an anchor for results data acquired from different platforms. For example, each sample's sequence variant data are linked to the VARIANT table to access the biological context. It is also cross-referenced to dbSNP, the Catalogue of Somatic Mutation In Cancer (COSMIC), and other public annotation databases. Each sample's gene expression data are linked to the GENE table, and each sample's copy number variant data are linked to DNA_STRAND table. With this anchor design, it is straightforward to find patients with the EGFR mutation but not MET copy number gain. This biology-focused nature of the Oracle Health Sciences Translational Research Center also makes extensions to new types of omics modalities easy, adapting to continuous innovation in molecular technology.

**Variant Impact Assessment Accelerates Finding of Actionable Variants**

Exome and targeted sequencing projects routinely report hundreds of thousands of variants, while whole-genome projects report millions of variants. Since researchers often start with non-synonymous changes, that is, which alter amino acid sequence, the Oracle Health Sciences Translational Research Center offers two sets of impact assessment for each variant both capturing multiple transcripts per gene.

1. **Variant impact:** For each sequence variant occurs at the coding region of a gene, Oracle Health Sciences Translational Research Center computes the type of effect it has on the protein sequence. For example, missense, nonsense, frameshift, and so on. This computation includes both known variants and novel variants reported in an experiment.

2. **Effect of amino acid substitutions on protein function:** Oracle Health Sciences Translational Research Center predicts amino acid changes, as a result of a sequence variant, that affect protein function using the methods of Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping (PolyPhen).7, 8

With the help of these two approaches, users can easily narrow millions of sequence variants down to a manageable set of candidate variants to conduct further analysis.

---


Embracing Community Standards and Efforts

The bioinformatics community has developed standards and formats for omics data from various modalities. Especially in the area of NGS, one to two data formats dominate at each level of data analysis. Oracle Health Sciences Translational Research Center includes multiple data adapters from common genomics data formats (details are in Section 3.2). Oracle Health Sciences Translational Research Center also supports the integration of the Visquick library for a circular genomics viewer\(^9\) and of the Integrated Genomics Viewer (IGV).\(^10\) As shown in Figure 3, users can specify a particular specimen and its available molecular data to be shown in the circular genomics viewer. If multiple omics modalities are selected, the viewer shows each one in a concentric circle. In IGV, users can specify specimen, genomics data modality, and genome region of interest at the control panel, and export a file in one of the IGV-supported formats.

![Figure 3. Oracle Health Sciences Translational Research Center’s integration with the Visquick library, providing a cross-modality circular genomics viewer for a selected specimen](image)

---

\(^9\) Visquick: Javascript project providing a library of configurable SVG and Canvas-based visual tools 2012, Institute of Systems Biology.

Data Lineage Across Multi-Level Result Files – Supporting Traceability

In our solution, lineage is maintained among the entities in the underlying data and their corresponding raw data. For example, Oracle Health Sciences Translational Research Center stores the file name and directory of aligned reads, and then links this entry to the sequence variant results that are derived from these aligned reads. If a user finds an interesting sequence variant, a variant with a particularly low quality score, or an SNP that prompts doubt about the accuracy of the raw data, user can trace it back to the corresponding sequence alignment and raw reads. Oracle Health Sciences Translational Research Center also lets user export the alignment files for further examination in a genome browser. This feature is particularly important for oncologists in the clinical decision support setting due to the extremely heterogeneous nature of tumor samples. Tracing back to the raw data helps clinicians to confidently differentiate mutations from noise in the data.

Consistent, Correct, Complete, and Standardized Clinical Data

Clinical data are often inconsistent and incomplete, especially when sourced from multiple and diverse systems and gathered for different purposes. Any data integration approach must thus provide separation of data from non-validated, uncertain, or poor quality sources from consistent, correct, complete, and standardized data. Data management by Oracle Health Sciences Translational Research Center is achieved through source-system agnostic data structures that facilitate the ease of loading and managing of semantics, relationships, and business rules from multiple source systems. The data are then standardized, and loaded along with associated metadata into a comprehensive warehouse available to downstream applications. Oracle Health Sciences Translational Research Center can also simultaneously refer to multiple terminologies fitting into diverse use cases. All of these features are critical to support queries and analyze information from multiple, disparate, heterogeneous systems – demographic, diagnostic, laboratory, phenotypic, life style, claims and more – to enable comprehensive, focused clinical research.

Scalability to Strive Big Data Problems in Translational Research

When targeting therapies, large amount of information must be analyzed and distilled down to drive decisions. Oracle Health Sciences Translational Research Center is capable of integrating millions of patient records and hundred thousand whole genome sequences. To manage petabytes data records, Oracle Health Sciences Translational Research Center is benefit from the development expertise in jointly optimizing software and hardware features. Especially, leveraging the Oracle's Exadata technologies including smart scan, storage index, and flash cache, and our proposed partition method, Oracle Health Sciences Translational Research Center ensures sub-second response times for many queries in scenarios with hundred thousand whole genome sequence data. It also provides scalability, which ensures that as the needs for genotype-phenotype analyses expand, the research platform can scale up and continue to deliver results.
Data loading speed from external system to Oracle Health Sciences Translational Research Center was benchmarked with the 69 publicly available data sets from the Complete Genomics Inc.’s website\textsuperscript{11} and the approximately one thousand samples in the 1000 Genomes project.\textsuperscript{12} It is able to process hundreds of whole-genome samples per day. This robust performance lets Oracle Health Sciences Translational Research Center handle a large amount of initial data load and accommodate the ever-increasing speed at which DNA sequencing data is obtained.

**Multiple Secure Deployment Options to Satisfy Different Customer’s Needs**

Oracle Health Sciences Translational Research Center can be deployed in a variety of modes depending on organizational preference. In addition to an on-premise option, the solution can also be deployed as SaaS in a private cloud, or as a hybrid option. Data transparency is assured in the cloud environment. In particular, our cloud service customers are benefit from the standard formats we support, and at any point in time customers can export their data stored in Oracle Health Sciences Translational Research Center into the standard formats.

Translational research informatics solutions are often required to handle highly sensitive data, such as patient records, genomics data, and data potentially classified as PHI. Oracle Health Sciences Translational Research Center provides a SaaS environment certified by HIPAA regulations with independent auditing from third-party agencies. The data centers hosting the SaaS environment are also trusted by the world’s largest pharmaceutical companies to manage their clinical trials.

**Architecture Components**

Oracle Health Sciences Translational Research Center is composed of four fully integrated components, Oracle Healthcare Data Warehouse Foundation, Oracle Healthcare Analytics Data Integration, Oracle Health Sciences Omics Data Bank, and Oracle Health Sciences Cohort Explorer. Oracle Healthcare Data Warehouse Foundation and Oracle Health Sciences Omics Data Bank are the back-end data schemata storing clinical and omics data, respectively. Oracle Healthcare Analytics Data Integration is the back-end clinical data integrator between customers’ source systems and Oracle Healthcare Data Warehouse Foundation. Oracle Health Sciences Cohort Explorer is a Web-based application accessing both clinical and omics data stored in the back-end components.

\textsuperscript{11} Public Genome Data Repository, 2012, Complete Genomics: Mountain View, CA.

Oracle Healthcare Data Warehouse Foundation

Oracle Healthcare Data Warehouse Foundation is a comprehensive set of logical and physical data models for deploying pre-built business intelligence, analytic, data mining, and performance management applications from Oracle and its partners. While supporting both standard terminologies and proprietary data dictionaries, the solution includes an expanding list of more than 1,000 entities and 12,000 attributes spanning the clinical, financial, operational, and research domains. It stores data at their most granular level, thereby maximizing the potential for reuse by multiple analytics applications.

Oracle Healthcare Data Warehouse Foundation supports data standardization and cross-validation check, an important step when dealing with clinical data originating from diverse sources. For example, it can find and eliminate duplicate data while ensuring correct data attribute survivorship to create a “single source of truth”. With such a capability, the master data in the warehouse are always up-to-date, consistent and complete. Furthermore, all the transformations during the data standardization process are logged, providing a complete data lineage record and with permanent traceability.

Oracle Healthcare Data Warehouse Foundation is a main building block in Oracle Health Sciences Translational Research Center, supplying standardized and high quality clinical data critical for accurate disease classification and thereby enabling high confidence biomarker discovery and clinical decision support activities.
Oracle Health Sciences Omics Data Bank

Oracle Health Sciences Omics Data Bank is a critical component of the Oracle Health Sciences Translational Research Center solution, as a cross-platform and cross-omics modality data model that delivers the extreme performance required for querying vast omics data sets typical of today’s translational research studies. The solution design was based on the Central Dogma of Molecular Biology. In the latest release, the Oracle Health Sciences Omics Data Bank schema models multiple data modalities, including sequence variant, copy number variant, and gene expression through both microarray and sequencing technology (that is, RNA-seq). Oracle Health Sciences Omics Data Bank can also store data from multiple biological species.

Oracle Health Sciences Omics Data Bank includes out-of-the-box adapters for loading customer-generated molecular profiling data, as well as public domain data from sources such as Ensembl, Pathway Commons, TCGA, and the 1000 Genomes Project to further speed time to value. For data generated in-house, it provides adapters to Variant Call Format (VCF)\textsuperscript{13}, Mutation Annotation Format (MAF)\textsuperscript{14}, tab-delimited format for expression data, RNA-seq expression-exon data format\textsuperscript{15}, and MasterVar format from Complete Genomics Inc.\textsuperscript{16} Oracle Health Sciences Omics Data Bank is scalable to projects with hundred thousand whole genome sequences, exceeding the 5-year ambitions of most large organizations.

Oracle Healthcare Analytics Data Integration

Oracle Healthcare Analytics Data Integration provides a rich platform for processing of data from diverse source systems and loading of the homogenized, consistent, complete, correct, and standardized analytics ready data into Oracle Healthcare Data Warehouse Foundation.

Oracle Healthcare Analytics Data Integration includes a set of interface tables that provide abstraction from Oracle Healthcare Data Warehouse Foundation, serve as a system of original record proxy, support versioning of data, and act as a persistent placeholder for incoming data from source systems. Foundation validation and integration rules are integral parts of Warehouse Integration, during which Oracle Healthcare Analytics Data Integration can identify exception conditions and provide comprehensive exception logs with traceability to erroneous data and configurable exception messages. Customers can also configure the integration rules to allow for suspension of data to accommodate late arriving data. Furthermore, Oracle Healthcare Analytics Data Integration provides a rich platform for processing of data from diverse source systems and loading of the homogenized, consistent, complete, correct, and standardized analytics ready data into Oracle Healthcare Data Warehouse Foundation.

\begin{itemize}
\item \textsuperscript{13} VCF (Variant Call Format) version 4.1. 2012; Available from: http://www.1000genomes.org/wiki/Analysis/Variant+Call+Format/vcf-variant-call-format-version-41.
\item \textsuperscript{16} Standard Sequencing Service Data File Formats, 2012, Complete Genomics Incorporated.
\end{itemize}
Integration supports data governance through data profiling, management of data quality, metadata management. With these features, customers can obtain data lineage and impact analysis, versioning of data, fulfilling auditing requirements.

**Oracle Health Sciences Cohort Explorer**

Oracle Health Sciences Cohort Explorer is a Web-based interface that provides scientists and clinicians with a dashboard to monitor patient population, based on commonly used metrics, with an agile query interface, and a detailed event viewer for each patient's longitudinal information. With this tool, clinicians and scientists can be self-sufficient and timely in their own analysis.

Oracle Health Sciences Cohort Explorer accesses integrated clinical and omics data from Cohort Data Mart (CDM) and Oracle Health Sciences Omics Data Bank. As a data mart off Oracle Healthcare Data Warehouse Foundation, CDM contains a subset of entities and attributes that are relevant to translational research. Due to the data standardization capability in Oracle Healthcare Data Warehouse Foundation, the clinical data in CDM are always ready for use by analytic tools such as Oracle Health Sciences Cohort Explorer without further post processing.

Oracle Health Sciences Cohort Explorer provides a highly interactive user-friendly interface for identifying and exploring patient sub populations based on clinical and omics information such as demography, diagnosis, treatment, specimen, mutation, and copy number variant, and the selected cohorts can be saved for future use. Figure 1. Oracle Health Sciences Translational Research Center patient stratification use case illustrates a patient stratification use case. Researchers and clinicians can further examine selected patients using the longitudinal viewer, and thus notice any inconsistencies in treatment schedules and treatment outcomes.

In addition to patient-focused cohort queries, Oracle Health Sciences Cohort Explorer also supports queries that are gene-centric and includes state of the art molecular data visualizers, which provide a holistic view of the multiple omics modalities stored in Oracle Health Sciences Omics Data Bank. Figure 3 illustrates an example. This integrated view is the perfect foundation for systems biology-based approaches to further the molecular understanding of disease in the discovery of new drug targets.

**Conclusions**

Oracle Health Sciences Translational Research Center is a scalable informatics solution for translational research. Its back-end data components, Oracle Healthcare Data Warehouse Foundation, Oracle Healthcare Analytics Data Integration, and Oracle Health Sciences Omics Data Bank, seamlessly integrate clinical and omics data from diverse clinical data sources as well as from vendor-specific and modality-specific omics data silos, providing standardized data readily use for the front-end application, Oracle Health Sciences Cohort Explorer. With a few clicks, researchers and clinicians can identify patients with specific clinical and omics characteristics of interest with full traceability support. This self-sufficient environment will significantly shorten the biomarker discovery cycle and transform clinical decision capability, opening the way for the practice of personalized medicine.