ORACLE
HEALTH SCIENCES

An Oracle White Paper
February 2014

Personalized Medicine: Informatics Challenges on the Road to the Clinic

ORACLE

Disclaimer

The following is intended to outline our general product direction. It is intended for informational 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.

Table of Contents

Executive Overview

Personalized medicine has the potential to revolutionize patient care. For benefit realization, it requires the integration of the complete biomarker lifecycle from discovery to the targeted treatment of patients. However, many organizations face the following informatics challenges with integration:

(1) Limitations in storage, computing power for analysis, and information access:

- Scalability challenges in storage and computing power have been mitigated with improvements in storage devices, data compression techniques, distributed computing architectures, cloud computing, and compression algorithms

- The challenge remains in the downstream information access

(2) A lack of biologist-friendly software applications to replace the user-unfriendly custom scripts is crippling collaboration

- Biologists need flexible and powerful software, but most of the current software solutions have limited graphical user interfaces

- Although retraining biologists is an option, their time is better spent on biological questions—software should adapt to the needs of the biologists

(3) An urgent need for standardizing data across omics and clinical data realms for cross-study comparisons:

- The proliferation of data formats in molecular and clinical areas, despite improvements in standardization, has created a mundane data preprocessing task for biologists and bioinformaticians

- Standardization is critical for focused translational research

(4) Systematic management of public reference and annotations data:

- Translational research needs to go beyond biomarker discovery to reach its goals
- The challenge is to distill the deluge of information into actionable information

(5) A false dichotomy between enterprise and open source software applications:

- Because open source software is innovative and dynamic and enterprise software is stable and robust, they complement each other well

Introduction

The objective of personalized medicine is to tailor treatment according to the molecular profile of each patient. It puts the concentration of preventive and therapeutic treatments on those who will most likely benefit, sparing expense and side effects for those who will not. One of the major challenges in modern, molecular-based disease research is patient genetic variability, which affects both the efficacy and safety of the resulting therapeutic treatments[1-5]. This challenge is overcome with individual patient omics profiles generated by a growing number of high-throughput molecular platforms. Today, personalized medicine is closer to reality than ever before through targeted treatment; [6-7]however, the substantial increase in data correspondingly requires scalable systems to continue to effectively manage the data and to remain current with advancing technology. For instance, data production is expected to surpass current informatics capacity as a result of a continuous expansion in next-generation sequencing (NGS), new achievements in single-molecule sequencing[8], and the need for a whole-genome approach recently validated by the ENCODE project[9].

It is pragmatic to assume that whole-genome sequencing of hundreds of thousands of individuals over the next few years will keep the scalability challenges of storage and processing power at the forefront of the informatics challenge. However, while evidence proves these hardware issues are surmountable, the more fundamental and harder-to-solve problems still need progress. These problems include:

- Multiplicity of data formats

---

[1] Hill AVS, GS Cooke, "Genetics of Susceptibility," *Nat Rev Genet,* 2(12), 967–977 (2001).
[2] Relling MV, W.E. Evans, "Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics," *Science,* 286(5439), 487–491 (1999).
[3] The Cancer Genome Atlas Research Network: Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways," *Nature,* 455(7216), 1061–1068 (2008).
[4] The Cancer Genome Atlas Research Network: Integrated Genomic Analyses of Ovarian Carcinoma," *Nature,* 474(7353), 609–615 (2011).
[6] S.L. Close, Mega JL, Wiviott SD *et al*, "Cytochrome P-450 Polymorphisms and Response to Clopidogrel," *New England Journal of Medicine,* 360(4), 354–362 (2009).
[7] G. Bollag, P. Hirth, J.Tsai *et al*, "Clinical Efficacy of a RAF Inhibitor Needs Broad Target Blockade in BRAF-Mutant Melanoma," *Nature,* 467(7315), 596–599 (2010).
[8] E.E. Schadt, S. Turner, A. Kasarskis, "A Window into Third-Generation Sequencing," *Human Molecular Genetics,* 20(4), 853–853 (2011).
[9] "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature,* 489(7414), 57–74 (2012).

- Complex biological inter-relationships inadequately captured by human-readable flat file formats
- Low-quality and non-normalized clinical data
- Paucity of methods and models to translate molecular research data into clinically actionable insights

These crucial informatics challenges must be addressed to channel personalized medicine from an academic research pursuit within university medical centers to broader adoption by the mainstream healthcare system.

Challenges

This section discusses the major informatics challenges hindering the potential role and power of translational medicine from bench to bedside.

Scalability: Storage, Computing Power, and Information Access

The two main scalability issues in omics research are the storage of high-volume data and the processing power needed to analyze it. From the beginning of the genomics era (marked by the first human genome sequence project) to the introduction of NGS technology, the speed of data creation has continued to surpass Moore's law[10]. Today, a single research project can generate hundreds of terabytes of data, creating challenges in storage, backup, and disaster recovery. On the processing side, most of the analytic algorithms involve complex mathematic operations on vast amounts of data. This can take days of processing time per genome, even on reasonably sized computer clusters, before accounting for the time necessary for quality control and reanalysis.

Scalability challenges in storage and computing power have been mitigated to some extent in the past decade with improvements in storage devices, data compression techniques, distributed computing architectures, and cloud computing[11]. For example, the Large Hadron Collider project at CERN manages 15 additional petabytes of data annually[12]; the Wellcome Trust Sanger Institute manages 16,500 cores and stores 16 petabytes of short-read DNA data[13]; EMBL-EBI reported managing 14

[10] S.D. Kahn, "On the Future of Genomic Data," *Science,* 331(6018), 728–729 (2011).

[11] B. Langmead, M. Schatz, J. Lin, M. Pop, S. Salzberg, "Searching for SNPs with Cloud Computing," *Genome Biology,* 10(11), R134 (2009).

[12] Worldwide LHC Computing Grid, 2012; available at: public.web.cern.ch/public/en/LHC/computing-en.html

[13] K. Ambrose, "DataBase File System (DBFS) at the Wellcome Trust Sanger Institute," (2012).

petabytes, doubling every nine months; and the Sequence Read Archive (SRA) surpassed 100 terabases of open-access NGS reads[14].

Advances in reference-based compression algorithms have achieved 5- to 54-fold compression rates compared to standard methods[15]. Some complex algorithms may require thousands of cores running for weeks. However, with continuous algorithm optimization, a fresh read alignment and *de novo* assembly of a human genome can be achieved in one to two hours and one to two days in a standard server machine[16-17]. Cloud computing has also brought the possibility of small and midsize institutions accessing the sizeable computing power necessary to perform intense computations on-demand, without having to invest in their own expensive hardware and support infrastructure[18].

However, there is a neglected scalability challenge centered on downstream information access, which must be flexible enough to adapt to novel research questions and approaches. It must also be able to cover hundreds of thousands of patients and control subjects with rich clinical data and omics data across multiple modalities (for example, genomic, transcriptomic, and proteomic). A researcher may want to find samples with mutations on a specific locus, excluding samples with copy number gains for a paralogous gene, to assess the correlation among molecular traits and treatment sensitivity[19]. In this scenario, the current approach of combining annotation pipelines and a postscripting search of flat files is difficult to scale and requires constant redesign of the whole pipeline of accompanying scripts as the information access criteria evolve through scientific iterations.

Biological data are inter-related, as nucleotide sequences span introns and exons, some of which get transcribed into mRNAs, of which some then get translated into proteins. A gene family may be differentially expressed in different tissues through alternative splicing or in the same tissue across different patients. In addition, only certain genomic variants, such as nonsynonymous substitutions, can change the amino acid sequence and affect protein structure. Thus, the different omics domains within each patient are not independent sources of data, but inter-related and interconnected. The data must therefore be stored and accessed in a way that reflects these relationships so comparisons among patients can be made from the protein to the transcriptome to the genomic level in a single operation.

---

[14] Y. Kodama, M. Shumway, R. Leinonen, "The Sequence Read Archive: Explosive Growth of Sequencing Data," *Nucleic Acids Research*, 40(D1), D54–D56 (2012).

[15] M. His-Yang Fritz, R. Leinonen, G. Cochrane, E. Birney, "Efficient Storage of High-Throughput DNA Sequencing Data Using Reference-Based Compression," Genome Research, 21(5) 734–740 (2011).

[16-17] CLC Bio, 2012; available at clcbio.com; Novoalign, 2012; available at novocraft.com/main/index.php

[18] DNAnexus: Cloud-Based NGS, 2012; available at dnanexus.com/

[19] J. Bean, C. Brennan, J-Y Shih *et al*, "MET Amplification Occurs with or Without T790M Mutations in EGFR Mutant Lung Tumors with Acquired Resistance to Gefitinib or Erlotinib," *Proceedings of the National Academy of Sciences*, 104(52), 20,932–20,937, (2007).

Relational databases, which are collections of databases perceived by their users as a collection of tables[20], are suited to this and can ensure scalability and security while maintaining flexibility. Oracle Health Sciences Omics Data Bank models interconnections among entities of molecular biology[21] (see "Reference and Annotation Data Management to Capture Ever-Improving Knowledge" below for more details about Oracle Health Sciences Omics Data Bank). Despite billions of records in the database and advanced high-performance hardware and an optimized database schema design, it is possible to achieve retrieval speeds of a few seconds even when querying millions of patients and hundreds of thousands of whole genome sequences.

The Need for Biologist-Friendly Software Applications

To gain insights from genomic data, biologists require powerful and flexible data analysis software. Under the current paradigm, these software applications are typically developed by bioinformaticians who have in-depth knowledge of one or more areas of genomics, genetics, biostatistics, and computer science, but limited experience in software engineering and interface design. Bioinformaticians spend most of their time focusing on improving the detailed functionality of the software they produce, often neglecting the user interface that is crucial for software adoption by the scientific community. As a result, most bioinformatics software is built with command-line or rudimentary user interfaces developed only as an afterthought, and which still require scripting skills for batch job processing and transformation from output to input formats among the subsequent programs in an analysis protocol. This severely confines the software utilization to a small group of skilled bioinformaticians and discourages its use by the intended audience—biologists at large.

Biologists need the rich functionalities necessary for testing novel hypotheses, but without the associated programming requirements that the most powerful informatics software still requires. The two common approaches to bridging this gap are to have biologists cross-train as bioinformaticians or to mandate bioinformaticians to support each step of the analysis workflow. For the former, new courses, training programs, and initiatives have sprung up in the last few years[22]. But because of the time and energy involved in training, the latter approach is now more typical, employing postdoctoral researchers trained in bioinformatics to create custom scripts as they progress. This leads to a collection of scattered data files and single-purpose programs that ultimately hamper the collaboration and knowledge transfer critical to help advance the field.

It is better for biologists to spend their time answering scientific questions than retooling themselves as bioinformaticians, especially since the field is rapidly evolving. Similarly, bioinformaticians cannot be

[20] C.J. Date, "An Introduction to Database Systems," *Pearson/Addison Wesley*, (2004).

[21] W. Ou, J. Sheldon, Oracle Health Sciences Translational Research Center, "A Translational Medicine Platform to Address the Big Data Challenge," Oracle (2012).

[22] Bioinformatics Training Network, 2013; available at bionet.org

the gatekeepers of all project analyses without becoming the bottleneck due to the iterative nature of scientific research. Fundamentally, more biologist-friendly software is needed, which requires the involvement of both software engineers and user interface designers to expose the right amount of complexity in the underlying bioinformatics algorithms. Of course, the biologist-friendly software design requires a trade-off to balance simplicity against flexibility and robustness of the analysis, although Oracle believes this would be largely offset by directly engaging scientists in their own analysis with the entire biological context that the development of biologist-friendly software would bring forth.

To maximize software utility for biologists, software engineers must consolidate the technical specifications from bioinformaticians and usability design requirements from biologists. The Galaxy Project is certainly a good step in the right direction. Among many utilities, the Galaxy Project allows for variant calling, data format conversion, extraction of sequence features, and multivariate analysis, all through a user-friendly web interface[23-25]. Oracle Health Sciences Cohort Explorer, with its user-friendly interface, lets clinical scientists and biologists stratify patients and explore their clinical and omics data.

[23] B .Giardine, C. Riemer, R.C. Hardison *et al*, "Galaxy: A Platform for Interactive, Large-Scale Genome Analysis," *Genome Research*, 15(10, 1,451–1,455 (2005).

[24] D. Blankenberg, G. Von Kuster, N. Coraor *et al*, "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists," *Current Protocols in Molecular Biology*, (89), 19.10.11–19.10.21 (2010).

[25] J. Goecks, A. Nekrutenko, J. Taylor, "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences," *Genome Biology*, 11(8) (2005).

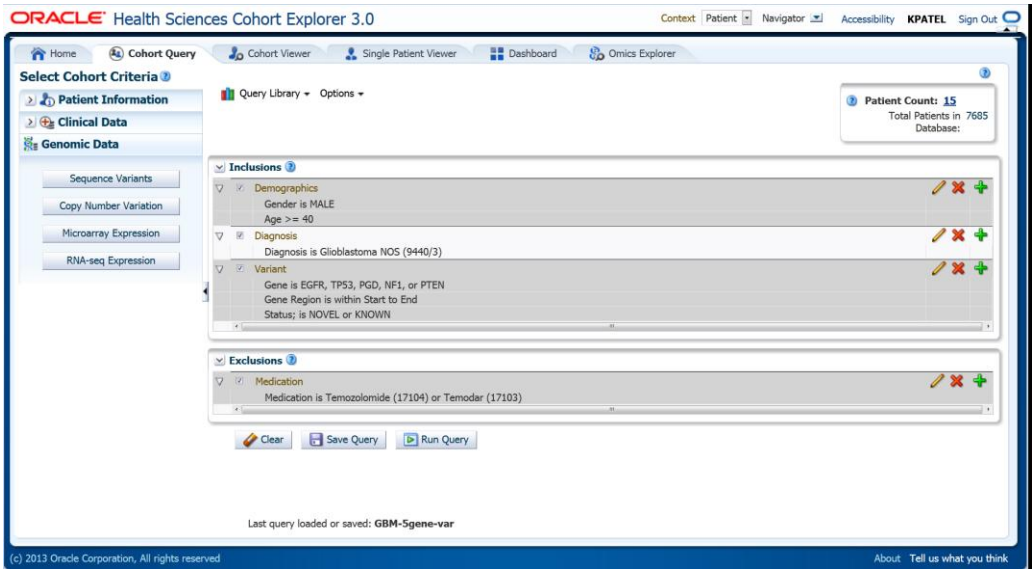

Figure 1. Oracle Health Sciences Cohort Explorer lets you stratify patients based on a list of inclusion and exclusion criteria across clinical and omics characteristics.

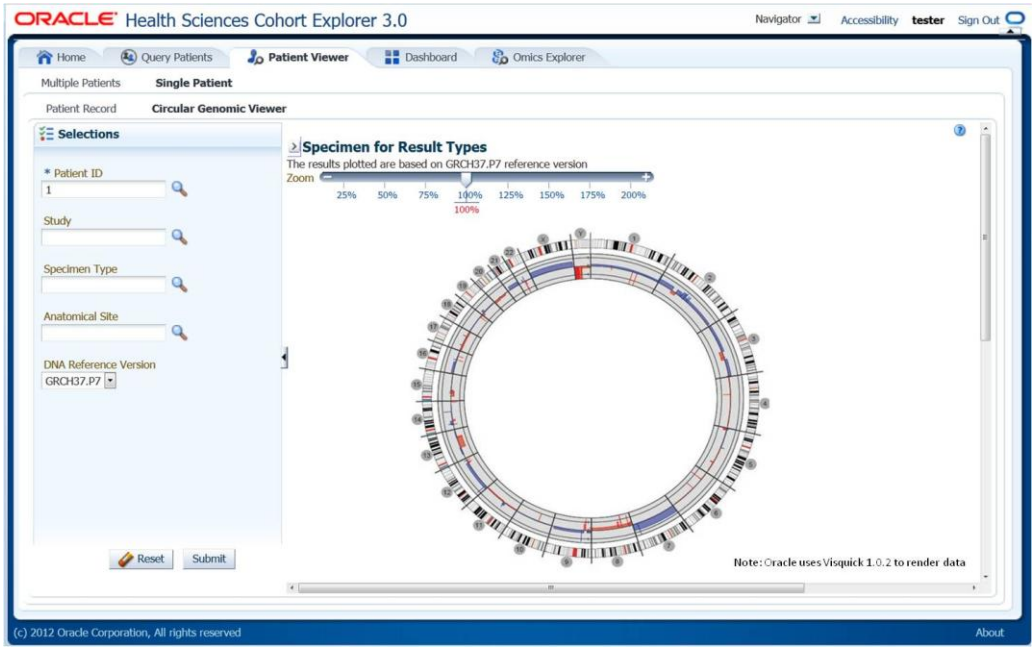

Figure 2. In Oracle Health Sciences Cohort Explorer, you can explore a patient's omics characteristics through the user interface with an embedded third-party Visquick library[26].

[26] Visquick library, Institute for Systems Biology. Available from: systemsbiology.org/visquick

Data Standardization for Both Omics and Clinical Data

Due to the nature of biological processes and the technology platforms applied, different types of omics data are stored in different formats. For example, one format may be designed for quantitative gene expression data while another is designed for nucleotide sequences. This has created a challenge while attempting to concurrently analyze data across multiple omics domains. It is compounded by the fact that some equipment vendors use proprietary formats that must be converted to the appropriate input format for the next step in the analysis protocol[27].

Within some NGS areas, the bioinformatics community has created data format standards, such as the sequence alignment map (SAM) for DNA-sequence multiple-alignment files[28] and the variant call format (VCF) for sequence variants[29]. However, multiple variations of these basic formats exist, all of which must flow from one step of the analysis to the next, often requiring manual intervention. For instance, in the Cancer Genome Atlas project, sometimes the same copy number information is presented in one column with notation x|y, but in others it appears separately in two columns[30]. Although these differences in formatting may seem trivial, bioinformaticians often deal with such mundane formatting issues rather than working on scientific problems.

As platform and analysis software continue to evolve, a timely update is necessary for any comprehensive bioinformatics solution to cope with the multitude of format versions. The need for this constant omics data update cycle is often underestimated during projects, and as new technologies generate new kinds of data sets, the need for this activity will only grow.

In translational research, it is easy to be enticed by the power of omics technology; however, it is of limited value without access to high-quality standardized clinical data. This is necessary to accurately characterize phenotypes and assess treatment outcomes. Evidently, data from both clinical trials and clinical practice must integrate, and pharmaceutical companies are increasingly prioritizing this integration—for example, pharmaceutical firms need to understand how their drugs perform within a real-world setting.

From an information technology perspective, this is an integration of data from two different systems—electronic data capture systems in clinical trials and electronic health records (EHRs) in clinical practice. Furthermore, clinical data is often inconsistent and incomplete, especially when gathered for different purposes and sourced from multiple systems. Any data integration approach

---

[27] L. Stein, "Creating a Bioinformatics Nation," *Nature*, 417(6885), 119–120 (2002).

[28] The Sam Format Specification Working Group, "The SAM Format Specification," (v1.4-r985).

[29] VCF (Variant Call Format) version 4.1, 2012; available at 1000genomes.org/wiki/analysis/variant-cell-format Oracle_HS_TRC_MDS_Whitepaper_093013_WBC_v3.docx

[30] The Cancer Genome Atlas Data Portal, 2012; available at tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp

must provide a method to separate nonvalidated, uncertain, or poor-quality data from consistent, accurate, complete, and standardized data. Particularly useful for doing this are source-system agnostic data structures that facilitate the easy loading and managing of semantics, relationships, and business rules from multiple source systems[31]. The data and associated metadata can then be standardized and loaded into data storage devices and made available for downstream applications. All these features are critical to support queries and analyze information from multiple, disparate, heterogeneous systems (such as demographic, diagnostic, laboratory, phenotypic, life style, claims, and so on), which will make comprehensive, focused clinical research possible.

Reference and Annotation Data Management to Capture Ever-Improving Knowledge

Omics data generated from a specimen cannot be interpreted in isolation and requires reference genomes and annotation information to determine the functional impact or gain insight from a holistic picture. For example, a variant found in a specimen needs to be interpreted with respect to the coding regions to determine its downstream impact or to search against previously reported mutations to determine its likelihood of being a somatic mutation.

As our knowledge of the genome continues to improve, many annotation data sources update every two to three months, and some even weekly. To capture the new knowledge, many series-design annotation pipelines have to rerun fully or partially when the annotation data updates to a newer version. The postannotation results are commonly kept in a flat file, which is named to reflect the versions of various annotation data sources shown in Figure 3(a). While this series workflow can handle a handful of annotation data sources, it is not scalable. Furthermore, different labs in an organization tend to have varying preferences for annotating data sources and the order in the series annotation workflow, resulting in many silos of data deposits and multiple workflows with small discrepancies. The file naming system is also meaningful to a small group of advanced users, leaving a large group of users unable to make use of the data to validate their scientific hypotheses.

To meet these challenges, Oracle proposes a scalable design for managing a large number of reference and annotation data sources while maintaining flexibility and intuitions for end users to access the data. A schema is designed that can manage multiple reference data sources such as Ensemble genome, dbSNP, HapMap, COSMIC, and so on. A graphical illustration of the design is shown in Figure 3(b). The schema also establishes business keys between the corresponding entities. For instance, the schema links any identical variants reported in dbSNP and COSMIC to the same genome coordinate and the same sequence variant description that follows the nomenclature published by the Human Genome Variation Society (HGVS).

[31] M. Samwald, A. Coulet, I. Huerga *et al*, "Semantically Enabling Pharmacogenomic Data for the Realization of Personalized Medicine," *Pharmacogenomics*, 13(2), 201–212 (2012).

For result data, the sequence variants detected in a specimen are indexed with respect to the aligned reference genome and the genomic coordinate in the database. The linkage and the index are designed to achieve high performance in an environment with trillions of variants from hundreds of thousands of specimens. Furthermore, multiple versions of an annotation database—such as dbSNP version 132, 137, and so on—can be stored concurrently in the schema. Hence, each user can dynamically choose a preferred set of annotation data and the preferred version(s) to interpret the result data at the time of query.

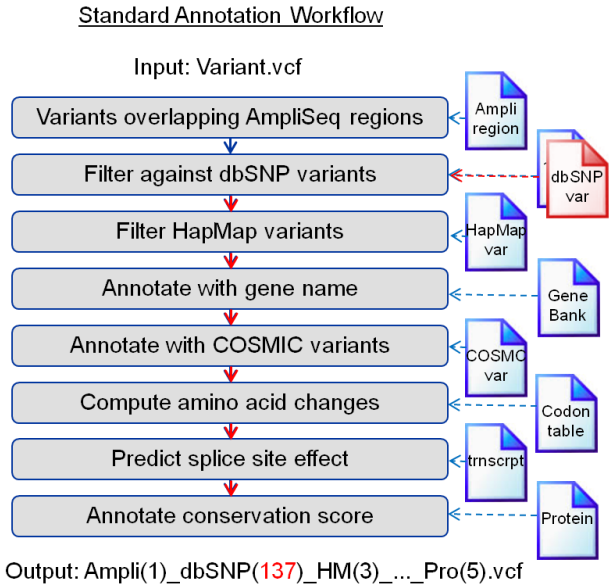

Figure 3(a)

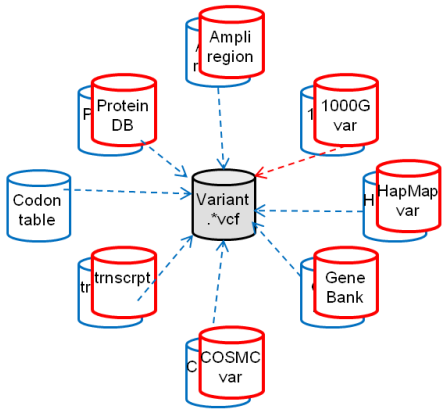

Figure 3(b)

Figure 3(a): Standard annotation workflow; Figure 3(b): Oracle Health Sciences Omics Data Bank design.

If there are multiple sets of sequence variants aligned to different versions of the reference genomes—that is, GRCh36 and GRCh37—an end user can aggregate search feedback across the different versions of reference genomes. This feature is important for clinical scientists who want to understand the scientific impact in a broader sense while not drilling into the bioinformatic details. As illustrated in Figure 4, if a user tries to identify specimens with rs6671243 variants from genome builds GRCh36 or GRCh37, the proposed system will first identify the associated dbSNP versions to each genome build, from which it obtains the associated genome coordinates and the variant descriptions—that is, chr1:916214 and chr1:926351, respectively. The system then searches for the result sequence variants using one of the two genome coordinates and sequence variant descriptions according to the result alignment's reference genome build version, followed by result aggregation. If a user is only interested in a particular genome build, they can further specify that in the query time, and one execution branch will be executed as shown in Figure 4.

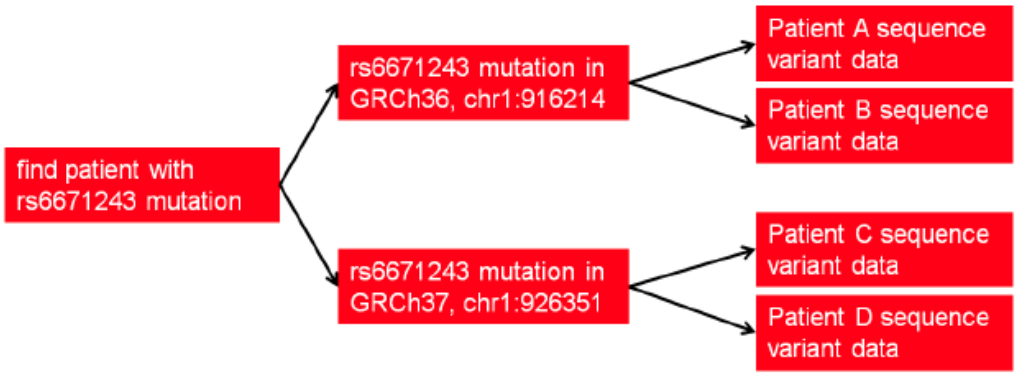

Figure 4. The underlying workflow to aggregate data from genome builds GRCh36 and GRCh37.

Fundamental benefits of the Oracle Health Sciences Omics Data Bank design include:

- Dynamic reference and annotation data management in contrast to steady design in the series annotation pipeline
- Straightforward data aggregation across genome builds

At the time of this writing, this is the only schema in the market that can handle multiple versions of reference genome builds and multiple versions of annotation data sources in a dynamic design.

False Dichotomy Between Enterprise and Open Source Software

Thanks largely to the growing and innovative academic bioinformatics community, a multitude of easily accessible open source analysis algorithms exist and are critical components for most translational research studies. However, the data still requires vast amounts of preprocessing and postprocessing as it moves from one step to the next in the analysis workflow. Moreover, many open source software solutions are developed and tested within a bioinformatician's own limited computing environment. As a result, when the software is shared with others working in a different environment, it often requires extensive troubleshooting, such as identifying and installing missing libraries and configuring file access permissions.

During a typical project, researchers try different combinations of analysis methods and parameters, which generates a large number of scripts and files spread across multiple storage devices with similar names (for example, a product of adding file name suffixes based on the programming parameters used) and create serious traceability issues after the project is completed. This is an increasingly important issue, not just for ensuring high-quality and replicable research with a number of high-profile

reports in the press[32], but also for complying with the Clinical Laboratory Improvement Amendments (CLIA).

Typical open source software developed by the academic research community is innovative, dynamic, and rapidly evolving as appropriate for novel analyses. However, due to its disposable nature, this type of open source software lacks many features commonly found in enterprise-class software designed for broad use within a large organization and industry, making open source and enterprise-class software a good complement to each other. These features include data access security, scalable data integration, track-and-trace audibility, and the overall satisfaction that comes from knowing the software has been developed and tested under a robust software development lifecycle with a cohesive roadmap appropriate for handling sensitive patient data. The latter is increasingly important with the growing need for regulatory compliance and the increasing role of the Food and Drug Administration (FDA) in biomarker applications[33].

Taking advantage of the innovative nature of open source software, Oracle Health Sciences Cohort Explorer integrates with the cutting-edge genomics visualization library, Visquick, to provide end users with the tools to help better understand a patient's genomic characteristics, as seen above in Figure 2.

With a solid technical foundation in data access security, enterprise software lets organizations specify every user's access permissions for each level of data, and can thus control the access to protected health information to only direct caregivers. It can also keep detailed data access logs, as required by the Health Insurance Portability and Accountability Act (HIPAA). Several enterprise software solutions also provide robust analysis workflow management[34]. This enables researchers to incorporate open source codes with the enterprise software, keeping track of the data and algorithm versions and the movement between output and input files. For instance, if a researcher finds a DNA variant call suspicious, the enterprise software can help him or her trace the DNA read distribution at that locus as well as the corresponding quality score.

Enterprise software also follows strict development cycles and rigorous testing across various operating systems, without the common difficulties seen in open source development, such as missing depending libraries and formal documentation that is not up to date or is missing. In enterprise software,

---

[32] C.M. Micheel, S.J. Nass, G. Omenn (editorials), "Evolution of Translational Omics: Lessons Learned and the Path Forward. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials," Institute of Medicine of the National Academies, *The National Academies Press*, Washington DC (2012).

[33] "Guidance for Industry: E16 Biomarkers Related to Drug or Biotechnology Product Development: Context, Structure, and Format of Qualification Submissions," US Department of Health and Human Services Food and Drug Administration, Maryland (2011).

[34] Oracle Life Sciences Data Hub

functional requirements and technical designs are well documented and critical for any third-party auditing for regulatory compliance certification. Moreover, since many code contributors have specific interests, open source consortiums often find it hard to manage software's future functional requirements and prioritization. In contrast, the ownership of the future development of enterprise software is clear, and numerous enterprise software solutions have dedicated channels to address users' support requests for existing or upcoming releases.

Enterprise software does not limit the flexibility required for innovative analysis using open source code. Instead, it provides the framework to ensure that the result of an innovation is securely captured in a way that can be reproduced and audited, thereby providing consistent tools to researchers across laboratories and organizations.

Future Perspectives

As discussed previously, the vast majority of platforms, tools, and applications labeled as solutions for translational research or translational medicine are unable to bridge the gap between basic biomarker discovery and the ultimate target of translational activities—molecular therapies with direct application to human beings in a regulated clinical setting.

Oracle Health Sciences Translational Research Center and its constituent component Oracle Health Sciences Omics Data Bank provide a scalable infrastructure for conducting biomarker discovery and validation and also form a foundational platform that can serve as the core of a molecular decision support system useful in a clinical environment.

In comparison to a data management and analysis platform that's for research use only, a system suitable for molecular decision support (MDS) in a clinical setting must allow the ability to:

1) Track, store, and retrieve raw and analyzed data (that is, data denoting genomic variation in individual clinical samples) that is compliant with CLIA and CAP regulations

2) Categorize, filter, and interpret genomic variants to produce a data set that is clinically meaningful in a repeatable manner

3) Combine molecular test data with phenotypic data (such as data from electronic health records) in real time to best inform clinical variant interpretation and treatment recommendations

4) Report clinically significant test outcomes and treatment recommendations clearly, concisely, and meaningfully to patient-facing clinicians who may have varying levels of expertise in genetics and genomic medicine applications

The following section discusses the suitability of using Oracle Health Sciences Translational Research Center in a clinical environment and details the additional pieces of functionality that, when added to the centralized data management capabilities of Oracle Health Sciences Translational Research Center bridge the "last mile" to direct clinical application.

The high-throughput laboratory research-use-only (RUO) instruments that interrogate human samples at the DNA, RNA, and protein levels are (with little alteration) increasingly being submitted for approval as devices suitable for human diagnostic and prognostic use. Even absent such approval, microarray and next-generation instruments are already used widely in CLIA- and CAP-accredited labs—the responsibility for appropriate clinical use of the instruments simply falls to the CLIA lab director rather than the instrument vendor. (Instrument vendors cannot legally market an RUO instrument for clinical applications.) Since Oracle Health Sciences Translational Research Center stores and tracks variant call format and related files (for example, SAM/BAM), it follows that the platform can be used identically in a clinical setting, provided that the product facilitates compliance with regulations imposed by CAP and CLIA standards. This is indeed the situation—the security, data tracking, auditing, and archiving capabilities that Oracle Health Sciences Translational Research Center features allow its full use as a platform for the storage of "clinical grade" sequencing results.

It is critical to emphasize that the central data model of Oracle Health Sciences Translational Research Center —Oracle Health Sciences Omics Data Bank—is a *relational* database model. Individual samples are linked in a hierarchy to the variations they exhibit. The variations are linked to the annotations and functional effect that pertain to them. At the same time, samples are linked to normalized clinical data describing the phenotypic properties the samples possess. Due to the nature of relational database functionality, none of these links, once set, can be reset, reassigned, or inadvertently modified. This is in contrast to, for example, a repository of flat files and folders where related entities are stored solely based on similar identifiers and best-effort organization. While accessing data organized in this way may be acceptable for RUO applications, it is unacceptable in a clinical setting where the relationships among samples, sample properties, genomic variation, and annotations must be inviolate.

Through an intuitive user interface, Oracle Health Sciences Translational Research Center allows the annotation and filtering of genomic data using numerous available data sources (for example, dbSNP, HapMap, and so on). Access to a plethora of such data sources is ideal for research purposes, where any association between a variant and suspected metabolic effect or phenotype may be of interest for strengthening a hypothesis or merely increasing the probability of ruling out a false-positive finding. By contrast, clinical interpretation of data requires a more restricted view of variation, where the focus is on variants with a causal link to disease, or variants with a clear, known association to therapeutic modalities such as drug metabolism, efficacy, or susceptibility. Such clinically scoped data sets exist, often from commercial vendors who have invested considerable effort in curating public and private genomic databases, winnowing those down to consist of repositories containing only clinically actionable content.

These commercial data services are accessible through API or direct download of the entire data set. Both of these access scenarios are easily supported in Oracle Health Sciences Translational Research Center. In some instances, genomic data alone is sufficient to draw phenotypic conclusions or to take therapeutic action. A canonical example is sickle-cell disease, where the presence or absence of certain mutations and data on the heterozygosity or homozygosity of those mutations in an individual leads to a clear prognosis. By contrast, an individual with an alternate form of the CYP2D6 gene in and of itself

may be viewed as nonactionable. Nevertheless, in combination with a prescription for a drug with an uptake affected by the cytochrome p450 enzyme complex, the genomic data for this individual takes on new meaning. Properly combining such data elements and providing timely, actionable data to a physician (for example, a warning not to prescribe a certain medication) is the essence of a decision support system.

The addition of a rules engine (such as a formalized representation in software of process logic) leads to an action through combinations of data elements. The introduction of a rules-engine framework, complete with rules-authoring tools suitable for capturing and executing genetic analysis procedures and linking known genotype/phenotype truths, to Oracle Health Sciences Translational Research Center will provide the deterministic element essential to a molecular support system.

A final note on the pharmacogenomic example given above: Oracle Health Sciences Translational Research Center incorporates clinical phenotype data through a data model that is populated daily. As clinical actions (such as prescription writing, test ordering, vital-sign evaluation, etc.) take place in real time, any decision support system must likewise access such phenotype and related data in real time. Fortunately, well-established standards that originated in the medical environment exist for this purpose. Of primary interest is the HL7 messaging protocol, an incorporating framework designed to gather necessary real-time data, which is a fairly simple undertaking.

The final element necessary for a decision support system with true clinical utility is a reporting and notification system that communicates actionable recommendations to the system's end users in various formats appropriate to the differing goals and expertise of its consumers. As discussed, the final product of the genetic variant interpretation process is fully annotated, clinically meaningful variants combined with actionable recommendations (for example, additional tests to be ordered, suggested prescriptions, behavioral or dietary modifications, etc.). A report authoring environment that combines this information and has a user-friendly delivery to devices (tablets, desktop computers, etc.) with various form factors and modes of use and to independent systems (EMRs, CPOE systems) is essential.

Big Data Analytics Infrastructure Requirements

Earlier in this paper, the core technical challenges of molecular decision support (MDS) systems were outlined as follows:

1. Storage limitations and costs
2. Proliferation of data types and formats
3. Management and leverage of open source analytics components
4. Integration of molecular information with clinical support systems
5. Security and traceability

The following challenges are common with core drivers in the emerging big data analytics market:

1. Volume—TBs and PBs of molecular data
2. Variety—the need to integrate molecular data with many types of data
3. Velocity—ingesting data and delivering actionable analytics where and when needed

Big data technologies promise to address these requirements by dramatically reducing storage costs, expanding the types of data that can be integrated, and enabling the adoption of analytics evolving in the industry. While big data approaches introduce critical new capabilities, core enterprise disciplines must be leveraged including security, traceability, and backup.

With data management addressed, focus turns to enabling analytics and delivering actionable recommendations. Incumbent analytics platforms are challenged to meet the need. With the deluge of data, many departmental systems are not scaling. They were not designed to manage the converged needs of personalized medicine. More fundamentally, moving large volumes of data becomes impractical. Even if these two issues are addressed, isolated informatics systems are, by nature, not integrated with core delivery systems, thus delaying or preventing prescriptive analytics from reaching the point of care. Finally, the new sources of data require new analytics pipelines to be introduced.

To enable the promise of the Oracle Health Sciences Translational Research Center platform, an extensible, scalable, and secure infrastructure is required. By combining the capabilities of emerging big data technologies with the proven disciplines of enterprise data warehousing, Oracle's unified informatics solution addresses these needs.

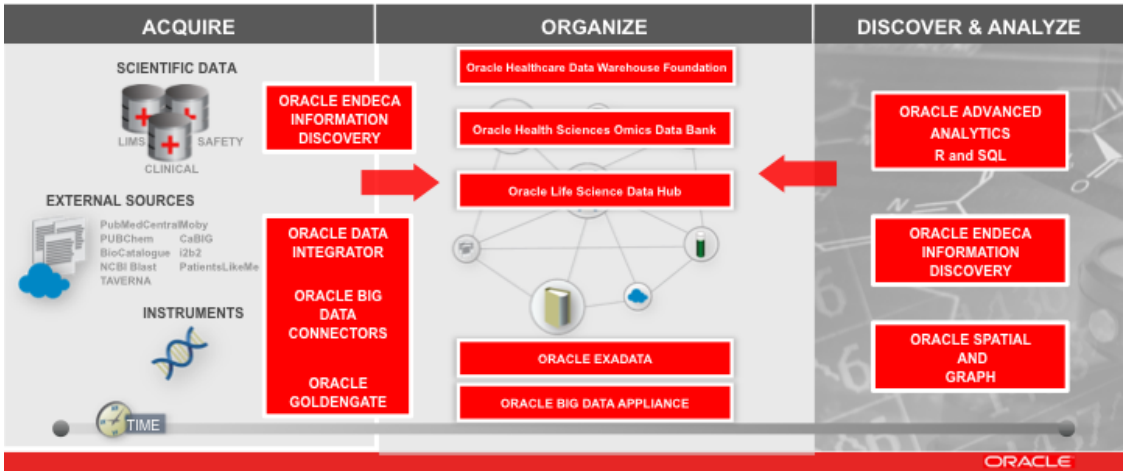

Figure 5. Oracle unified information architecture for Oracle Heath Sciences Translational Research Center

Oracle's unified information architecture allows molecular, structured, and unstructured data to be easily and cost-effectively integrated with the data warehousing and business analytics solutions to maximize the value for the needs of MDS.

The relational model at the heart of Oracle Health Sciences Translational Research Center is enabled by the Oracle Exadata component of the architecture. Oracle Exadata allows Oracle Health Sciences Translational Research Center to transparently leverage database and infrastructure innovations such as flash memory, high-speed networking, and compression. For the storage and processing of raw data (molecular, text, partner data, journals), Oracle Big Data Appliance provides a cost-effective platform for leveraging new storage economics that resides with a grid computing capability. Oracle Big Data Appliance supports the new molecular processing models written for the parallel Hadoop environment.

With interdisciplinary data converged and secure on the platform, analytics and decision support applications can be consolidated. By allowing analytics to run directly on converged data, the unified

information architecture improves analysis time. Substantial time savings come simply from eliminating the need to prepare and move data. Executing cohort-wide analytics close to the data improves the performance of the informatics.

The unified information architecture enables further analytics extensibility. By allowing open source (R) informatics to run on the platform, an MDS can leverage evolving informatics developed to exploit the new data. Most importantly, the architecture allows the informatics models to be securely reused by any platform that can access the database via a common query language (SQL).

Oracle Big Data Connectors and Oracle Data Integrator provide the tools to capture and organize a wide variety of data types from different sources including journal articles, doctor notes, and raw molecular data and clinical systems. Oracle Big Data Connectors allows you to extend Oracle Health Sciences Translational Research Center to a multitude of sources.

While ease of integration and access are critical enablers, security is mandatory. Oracle has enhanced its unified information architecture to extend enterprise-class security capabilities to all the data being managed in the pipeline.

Conclusion

As the speed of data production continues to increase beyond improvements in data storage and new technologies in raw processor speed, the next few years should continue to see dramatic improvements in the efficiency of analysis and data compression algorithms. The speed at which these new analysis tools are coded and released, combined with their increased complexity and sophistication, will only add to the IT training burden of biologists at the expense of energy that could otherwise be spent analyzing biological problems. The future direction of translational medicine will thus depend on how quickly the bioinformatics community can develop powerful end-to-end analysis tools that are biologist-friendly, of solid enterprise quality, and nimble and dynamic enough to accommodate the ever-increasing range of analysis tools produced by the open source community.

Personalized Medicine: Informatics Challenges on the Road to the Clinic
February 2014
Author: Jonathan Sheldon, PhD; Wanmei Ou, PhD; Andrew Boudreau
Contributing Authors:David Teszler

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together