Applied & Translational Genomics 3 (2014) 105–108

Contents lists available at ScienceDirect

Applied & Translational Genomics

journal homepage: www.elsevier.com/locate/atg

A collaborative approach to develop a multi-omics data analytics platform for translational research

Axel Schumacher [a], Tamas Rujan [a], Jens Hoefkens [b,*]

[a] Genedata AG, Margarethenstrasse 38, 4053 Basel, Switzerland
[b] Genedata Inc., 750 Marrett Road, Suite 403; Lexington, MA 02421, USA

ARTICLE INFO

Keywords:
Omics
Integration
Data analytics
Data sharing
Translational research
Scalability

ABSTRACT

The integration and analysis of large datasets in translational research has become an increasingly challenging problem. We propose a collaborative approach to integrate established data management platforms with existing analytical systems to fill the hole in the value chain between data collection and data exploitation. Our proposal in particular ensures data security and provides support for widely distributed teams of researchers. As a successful example for such an approach, we describe the implementation of a unified single platform that combines capabilities of the knowledge management platform tranSMART and the data analysis system Genedata Analyst™. The combined end-to-end platform helps to quickly find, enter, integrate, analyze, extract, and share patient- and drug-related data in the context of translational R&D projects.

1. Introduction

A typical challenge that translational researchers in clinical settings, pharmaceutical R&D and academic centers face is the integration and analysis of clinical patient data with high-dimensional omics data, resulting from the examination of individual patient samples. Researchers need to analyze those integrated datasets in order to more efficiently develop new, safer, and more effective drugs. Pharmaceutical companies have therefore begun developing data management systems such as the open source platform tranSMART, a knowledge management system to store and share patient data (see also the article by TranSMART Foundation in next issue). tranSMART enables an integrated way for scientists to store and share curated phenotypic data such as clinical observations and severe adverse events (SAEs), biomarker data from gene expression, genotyping, unstructured text from sources such as journal articles and conference abstracts, reference data, and metadata (Athey et al., 2013). However, tranSMART is not an end-to-end solution for the analyses of such data, as it lacks, for example, the ability to perform cross-study analyses or other more sophisticated statistical analyses. Although tranSMART provides a limited set of analytical features for stratification of clinical data into molecular subtypes, Target Enrichment Analysis, and summary statistics, among others, it does not provide enough analytical capabilities to generate answers to typical questions in biomarker identification and validation and sample classification and prediction. Generally speaking, highly scalable enterprise solutions that fill the analytics gap in the translational value chain and that integrate platforms such as tranSMART were, until recently, still missing. One reason is that computational analysis and biological interpretation of high-dimensional omics data (i.e. genomics, transcriptomics, epigenomics, proteomics, and metabolomics) pose significant challenges. For example, 1) the statistical and visual interpretation of data derived from different technological platforms is exceptionally complex, in part due to the large number of parameters; and 2), to fully exploit and compare data from different studies and/or technologies, analytical platforms need intelligent systems that can interact with the tranSMART data warehouse bidirectionally, without the need of exporting data to different file formats during the process.

In a collaborative effort, with partners from the pharmaceutical tranSMART user community, we tackled those issues to fill the gaps in the translational value chain by integrating tranSMART with Genedata Analyst™, a multi-omics data analysis platform. It is capable not only to easily import integrated data from tranSMART for in-depth statistical analyses, but also to provide a bidirectional, interactive data-flow back to the tranSMART system, with the option to integrate additional data from other public repositories, such as COSMIC, dbGaP, GEO, ArrayExpress, TCGA, or dbSNP.

2. Results & discussion

Translational research relies on an integrated set of data, so that different questions can be postulated. tranSMART stores such integrated data sets from a variety of different established databases, targeting early stages in the R&D value chain such as clinical data collection. One example is OpenClinica (Franklin et al., 2011), an open source

* Corresponding author at: 750 Marrett Road, One Cranberry Hill, Suite 403; Lexington, MA 02421, USA.
E-mail address: jens.hoefkens@genedata.com (J. Hoefkens).

http://dx.doi.org/10.1016/j.atg.2014.09.010

clinical trial software used for Electronic Data Capture (EDC) and clinical data management. OpenClinica works primarily in the context of low-dimensional (clinical) sample information, and includes tools to aid in protocol design, creation and management of electronic Case Report Forms (eCRFs), monitoring of studies, and obtain real-time transparency across multiple studies and sites. Similarly, it is planned to connect tranSMART to REDCap (Research Electronic Data Capture), a secure web application for building and managing online research surveys and databases (Harris et al., 2009). Such open source tools promise to be a powerful combination: for example, data collected in OpenClinica can be further shared in the tranSMART environment.

Our focus is combining tranSMART's data warehouse capabilities with the advanced analytical tools found in Genedata Analyst to provide inter-disciplinary teams with an integrated research collaboration and data analysis platform that covers the translational research value chain, from the patient to drug development (Fig. 1).

To fully utilize the strengths of tranSMART, several functionalities are required that can be complemented by the strong algorithmic capabilities included in Genedata Analyst. In general, to be useful for translational research, any analytical platform has to meet the following requirements:

A. Scalability: Big data processing today is still inefficient in many areas, particularly due to 1) the diversity of data formats, use of unstructured data, and the speed with which it needs to be generated and processed. Also, 2) the sheer volume of data is a bottleneck as the amount of public and proprietary data is exponentially expanding; in fields such as genomics, the data collected by new instruments is outpacing Moore's Law by several orders (O'Driscoll et al., 2013). Due to their high-dimensional structure, multi-omics approaches inherently increase the already surging amount of big data in the life-sciences. As a consequence, analytical tools must be able to utilize unstructured data and must provide nearly unlimited scalability to enable interactive analyses of large data sets with billions of data points.
B. Open system/flexible plugin APIs: The goal of any analytical platform is to be system agnostic and able to integrate into an existing IT environment. As such, flexible plugin APIs are a core requirement, enabling the platform to be extensible and to provide users with a scalable data mining platform capable of easily integrating with in-house and third party data analysis tools and databases. Public APIs for data import, export, and statistical analysis give users the freedom to automate their own analyses and workflows.
C. Access to public data repositories: Integration of public data repositories with downstream analytical tools able to access to a wide variety of multi-omics data (preferentially with standardized use of ontology-derived nomenclatures) enables efficient data discovery and data sharing. Preferentially, the chosen platform should already contain an out-of-the-box integration with public omics data sources including GEO, ArrayExpress, and other resources.

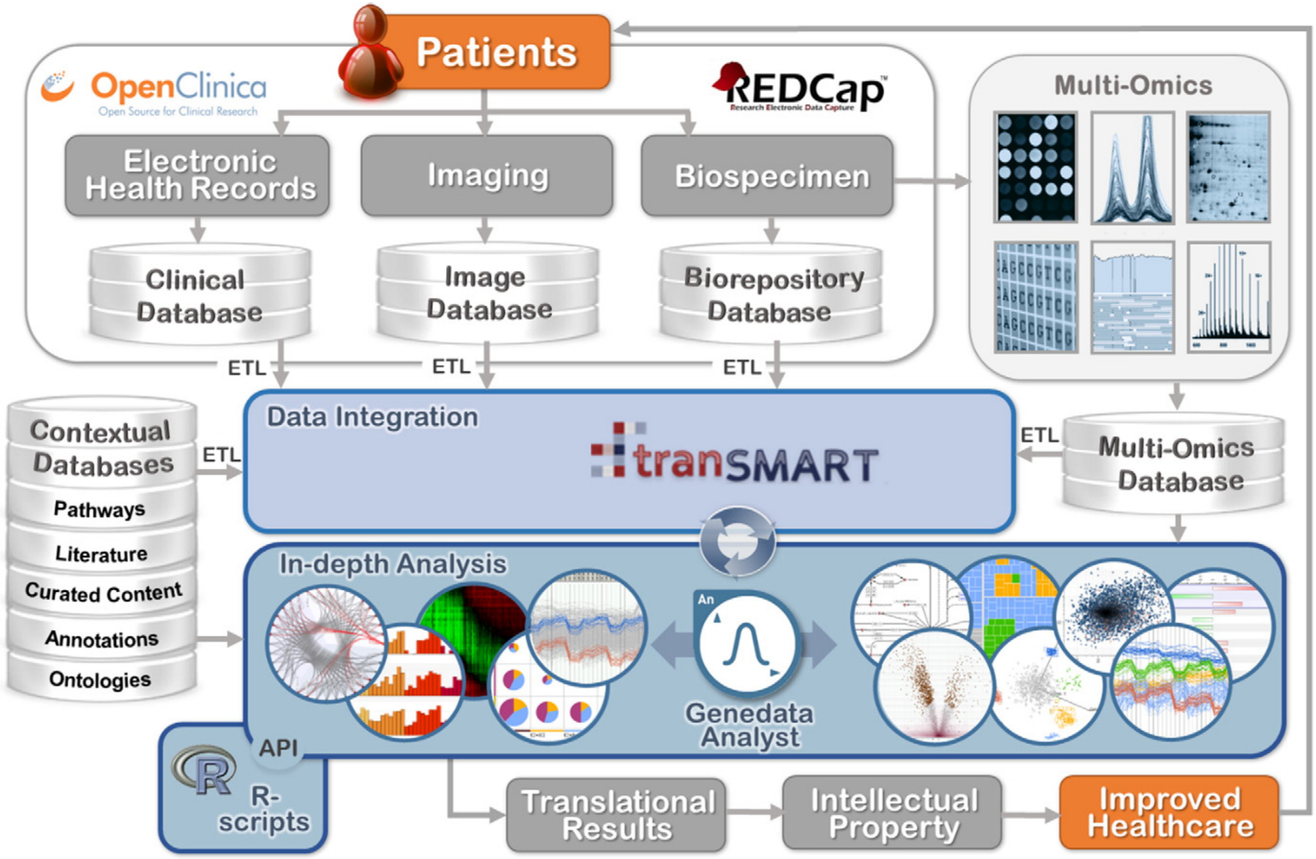

Fig. 1. Example of a data-sharing and big-data analytics value chain in translational medicine. The collection of large volumes of structured phenotypic data and its integration with the abundant Omic data adds new dimensions and challenges for the management, analysis, and visualization of this information. Clinical electronic data capture systems (EDCs, such as OpenClinica or REDCap) may feed patient data into tranSMART for data integration. An in-depth analysis of the data can then be performed in Genedata Analyst, an established system for the integrated analysis of high-dimensional omics data in the context of low-dimensional (clinical) sample information, often used in translational research projects. It enables scientists to efficiently analyze experiments by applying rigorous statistical algorithms combined with intuitive, interactive data visualization tools. Leveraging a built-in scripting engine, Analyst standardizes and automates complex and time-consuming data analysis processes. Via a flexible application program interface (API), the analyst platform also provides the possibility to use popular open source tools such as the R/Bioconductor-environment for downstream analyses (Gentleman et al., 2004). Overall, the platform has the ability to reduce the time to import, export, integrate, and analyze complex data—from days to minutes. (ETL = Extract, Transform, and Load process for loading raw source data into tranSMART.) The APIs have been integrated into tranSMART and are freely available to the research community. The statistical analysis software itself is a commercial software available for licensing.

D. Traceability/auditing: When working with clinical data it is of utmost importance to ensure the reliability, quality, integrity, and traceability of electronic source data. Traceability facilitates transparency, which is an essential component in building confidence in a result. The tranSMART/Analyst interface provides such an architecture for end-to-end traceability of data. The system maintains a chain of custody and log functionality for all data gathered throughout the transfer and analysis of the data. This ensures that, at the point of publication or submission to a regulatory authority, a single reliable source of data is available to support the findings. In addition, measures should be taken to ensure privacy by protecting patient data and confidentiality. In particular, this could be achieved by establishing a Health Insurance Portability and Accountability Act—(HIPPA) compliant environment, ensuring the procedural and technical measures required to prevent unauthorized access, modification, use, and dissemination of data stored or processed in the platform.

E. Standardization/workflows: The platform should allow creation, execution, and management of workflows that can automate and standardize analyses by providing means to record the processing order and parameter settings for sets of different analysis activities. This makes it possible to run analyses fully automated, meaning that even new users can start using sophisticated analyses by running a workflow in which all analysis steps are incorporated (i.e. assembled by a more experienced user).

F. Vendor support/technology agnosticism: Any system should be able to support a wide variety of omics-technologies (e.g. different next generation sequencing or mass spectrometry platforms) and easily integrate data from all available vendors.

G. Bi-directional data-flow: It should be easily possible to upload content that went through technology specific preprocessing and quality control steps in the analytical platform back into tranSMART via a secure system to system communication. Such data could, for example, consist of transformed or normalized data, e.g. from next generation sequencing (NGS)-, qPCR, microarray or mass spectrometry technologies. Such technologies are revolutionizing translational research by producing massive quantities of data that are being made publically available or shared within an enterprise. However, data that will be used by a wide variety of researchers need to be of especially high quality. As such, traditional approaches to preprocessing and quality control are no longer practical when confronted with the size of these data sets and the demands of real-time processing. Automated workflows for rapidly identifying problematic data are essential. Use of automated tools would enhance data integrity and

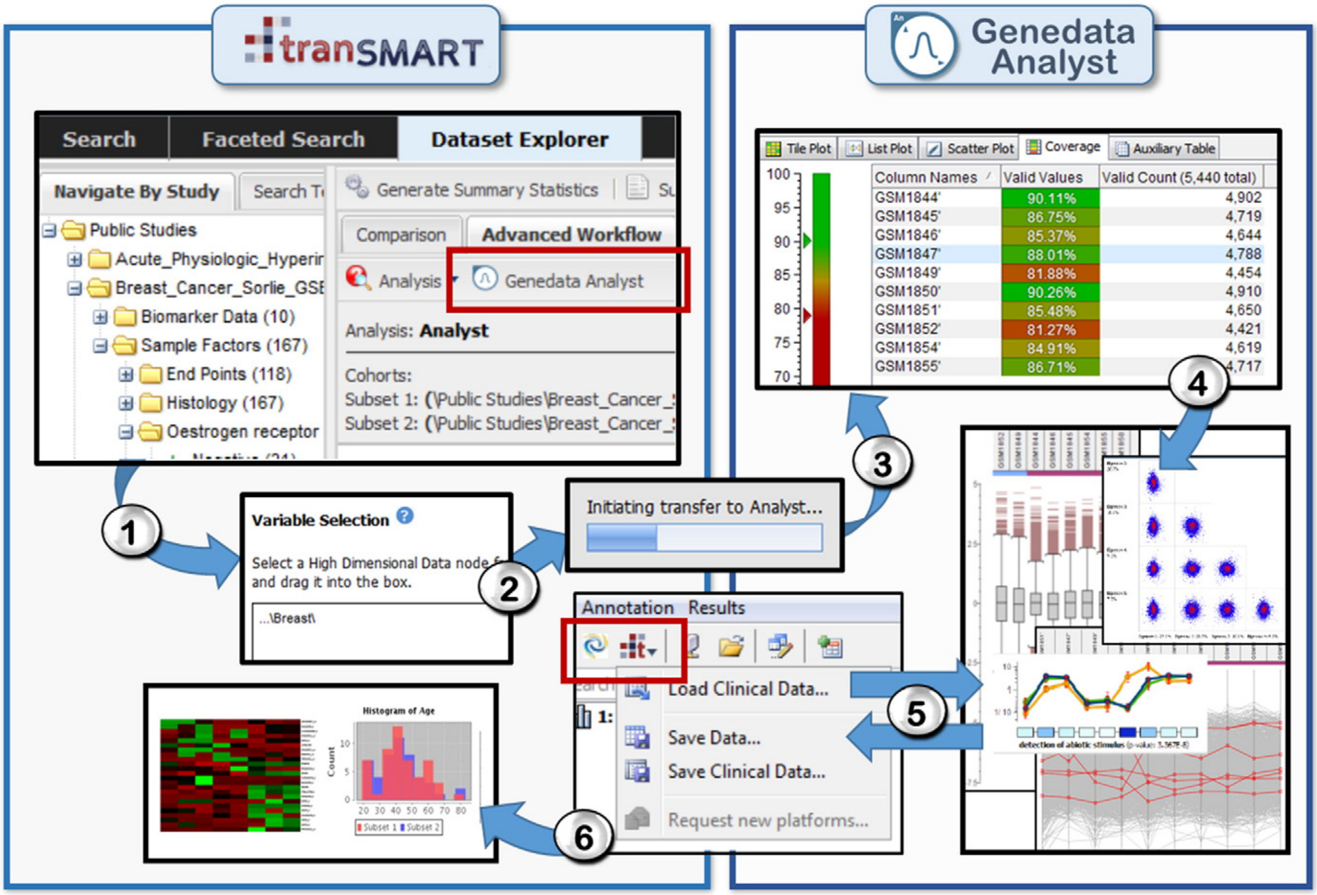

Fig. 2. Example of a bidirectional data exploration workflow using the tranSMART/analyst platform. The platform is designed to improve understanding of large amounts of raw (unstructured) data by combining and comparing it with structured data sets (where inclusion in a relational database is seamless and it is readily searchable by simple, straightforward search engine algorithms). Left: The advanced workflow tab in the tranSMART dataset explorer contains a button (red rectangle) that makes clinical data available for statistical analyses in Analyst through an easy drag and drop concept. 1: The patient cohorts to be examined are selected and further inclusion criteria can be defined. 2: Multi-omics, high dimensional data (e.g. microarray or NGS data) are selected and transferred by a single mouse-click into the Analyst GUI. 3: For curation purposes, imported data can be run through preprocessing and quality control steps. 4: Filtered data are selected for in-depth analyses such as PCA, correlation, network analyses, logistic regression, ANOVA, time-series analysis, clustering, annotation analyses, pathway mapping, and partial least square analysis (PLS). Third-party tools such as R-scripts can also be integrated. 5: The results of the analysis can be saved directly back into tranSMART via a menu-button. At any point, it is possible to simply pull more clinical/annotation data from tranSMART (and other sources such as GEO or ArrayExpress) into Analyst and vice versa. 6: Analyzed and curated data is available for further storage, sharing, and analyses in tranSMART.

reliability, would reduce delays in research projects, and would instill confidence in the shared omics data, thereby improving interoperability across research networks.

H. Ease-of use: An often neglected point in translational research is the ease-of-use of bioinformatics solutions. When researchers are faced with clinical data analysis tools, it is often found that the importance of ease-of-use and training materials outweighs number of features and functionality (Gomez-Cabrero et al., 2014). The whole integration process should contain an intuitive interface with pre-packaged features and easy accessible workflows making it easy to use for all user levels. Optimally, this bidirectional data-transfer can be performed with a single click of a button (see Fig. 2).

Due to the active interaction with the tranSMART community, all of the above listed criteria were fulfilled and combined into a single robust platform. The tranSMART/Analyst platform integrates and analyzes genomic, transcriptomic, epigenetic, and phenotypic data within a single high-throughput system and processes large and complex experimental data sets from all major vendors and technology platforms. tranSMART is a fully translational data warehouse based on the open-source i2b2 platform (Informatics for Integrating Biology and the Bedside; http://www.i2b2.org) that offers the functionality to integrate patient outcomes and electronic medical records (EMRs) to help effectively mine data. Analyst has an integration with additional study management systems and ontologies, which ensures that the processed data conforms to established standards and the user can be sure to publish only curated data back into the tranSMART system, in a fast and highly scalable fashion, without the necessity to export the data to a different format first. Furthermore, additional collaborative improvements of the platform and integration of other translational/clinical platforms such as iRODS are planned and encouraged.

3. Conclusions

We propose a collaborative approach to tackle the data integration and analytics challenge found within translational research. We have demonstrated that this approach could lead to a powerful platform by implementing a technical integration of tranSMART with Genedata Analyst. The tranSMART/Analyst platform provides a unified solution that integrates big data analytics capabilities with translational research data management and empowers scientists to efficiently analyze high-dimensional omics data based on subsets of clinical variables. This setup significantly reduces time and cost for data analysis and helps to understand the complexity and dynamics of biological processes without the need to reduce this complexity to artificial levels that may be less meaningful. Overall, the tranSMART/Analyst platform supports a richer analysis of bio-molecular profiling data, enables efficient cross-study data analysis, and integrates technology-specific data preprocessing and quality control capabilities. Those features eliminate error-prone manual data handling and data transformation processes and foster collaboration among global research teams by providing standardized analysis workflows. Ultimately, such a platform supports the scientific community's need to share data, collaborate to improve diagnostics, and to discover more effective treatments for patients suffering from complex diseases.

Conflict of interest

This article reflects the author's assessment of the field of big data analytics in translational medicine. A.S., T.R., and J.H. are employees of Genedata, one of several companies developing enterprise solutions for data analytics that streamline translational R&D workflows.

Acknowledgments

We thank our collaborators Jay Bergeron, Andrew Hill, Scott Jelinsky, Ami Khandeshi, Bill Mounts, and Anna Silberberg from Pfizer for their feedback and support during this project.

References

Athey, B.D., Braxenthaler, M., Haas, M., Guo, Y., 2013. tranSMART: an open source and community-driven informatics and data sharing platform for clinical and translational research. AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci, pp. 6–8.

Franklin, J.D., Guidry, A., Brinkley, J.F., 2011. A partnership approach for electronic data capture in small-scale clinical trials. J. Biomed. Inform. 44, S103–S108.

Gentleman, R.C., et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5, R80.

Gomez-Cabrero, D., et al., 2014. Data integration in the era of omics: current and future challenges. BMC Syst. Biol. 8, I1.

Harris, P.A., et al., 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 42, 377–381.

O'Driscoll, A., Daugelaite, J., Sleator, R.D., 2013. 'Big data', Hadoop and cloud computing in genomics. J. Biomed. Inform. 46, 774–781.