FEATURED ARTICLE #2

Efficient genomic profiling of patients: the benefit of systems interoperability

Axel Schumacher[a], Mark A. Collins[b], Marc Flesch[a], Miles Fisher-Pollard[c], & Tamas Rujan[c]
[a]Genedata GmbH, Munich, Germany | [b]Genedata Inc, Lexington, USA | [c]Genedata AG, Basel, Switzerland

Keywords: *tranSMART, quality control, data harmonization, translational research, precision medicine, patient privacy*

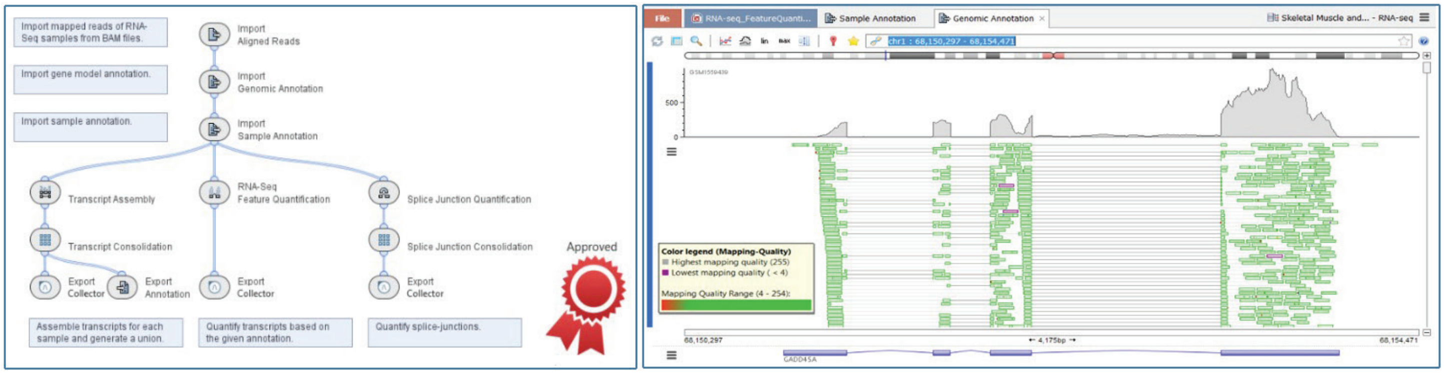
Genedata Profiler™ is a translational research software platform developed in collaboration with leading pharmaceutical companies to effectively process, manage, and analyze omic and phenotypic data to the highest standards of data quality and regulatory compliance. Genedata Profiler complements the knowledge management platform tranSMART by standardizing the processing and quality control of omics data, simplifying the publishing of data into the data warehouse, and adding sophisticated statistical analyses. In this case study we will demonstrate an end-to-end workflow to identify insulin response biomarkers utilizing the seamless, bidirectional interoperability of Genedata Profiler with tranSMART.

Improving the process of translational R&D

Achieving the vision of precision medicine is reliant to a large extent on translational research activities. Such activities require researchers to characterize and profile patients using omic technologies in order to understand their response to new therapies, stratify patients for trials or search for new disease biomarkers[1]. Genedata Profiler is an enterprise software platform used by pharma and biopharma companies to address these challenges and to optimize the process of translational research.

In this case study, we will discuss how Genedata Profiler complements existing in-house software solutions such as tranSMART[2] to create a complete infrastructure for performing translational research. The open-source data warehouse tranSMART is used by several pharmaceutical companies to manage multi-omics data sets to be used for biomarker projects. Often, tranSMART is used to merge 'clinical data' from the highly regulated clinical study environment into the less-regulated 'discovery' research environment, where users have much more flexibility to add omics data as well as other important research information.

By integrating Genedata Profiler with tranSMART, we have enhanced and expanded the capabilities of organizations to perform patient genomic profiling. The result is a very powerful, user-friendly platform for translational research.


Import mapped reads of RNA-Seq samples from BAM files.

Import gene model annotation.

Import sample annotation.

Assemble transcripts for each sample and generate a union.

Quantify transcripts based on the given annotation.

Quantify splice-junctions.

chr1 : 68,150,297 - 68,154,471

Color legend (Mapping-Quality)
▨ Highest mapping quality (255)
■ Lowest mapping quality (< 4)
Mapping Quality Range (4 - 254):

Fig. 1: Left: Version-controlled workflows for repeatable execution. The workflow engines of Genedata Profiler allow organizations to build and deploy standardized workflows throughout an organization, ensuring data quality and reproducibility. Example shows an RNA-Seq pipeline that was 'approved' by the study manager. Right: The best-in-class, fully interactive genome browser incorporated into Genedata Profiler facilitates visual inspection of raw data such as original short read alignments together with analysis to ensure result consistency.

Harmonizing raw omics data processing & quality control

Data quality and hence data curation is critical to making the right scientific conclusions. Genedata Profiler complements tranSMART by adding sophisticated data processing & curation capabilities to harmonize and standardize data processing workflows. Expert users can set up and approve such workflows and make them available to a larger user community (Fig. 1).

To illustrate the patient profiling capabilities of Genedata Profiler and the value of the integration with tranSMART, we sought to identify biomarkers for insulin response in skeletal muscle. We applied a best-practice workflow (Fig. 1, left) to process NGS and microarray data from a public RNA-Seq data set of human skeletal muscle myocyte samples[3] (n=12, 3 different treatment time points) together with expression profiles from a different set of muscle biopsies (n=36), which were analyzed with Affymetrix microarrays[4].

Inspection of the comprehensive quality control report (Fig. 2) automatically generated by the workflow indicates that all omics data is of sufficient quality to be utilized for statistical analysis, and is ready to be published to tranSMART.

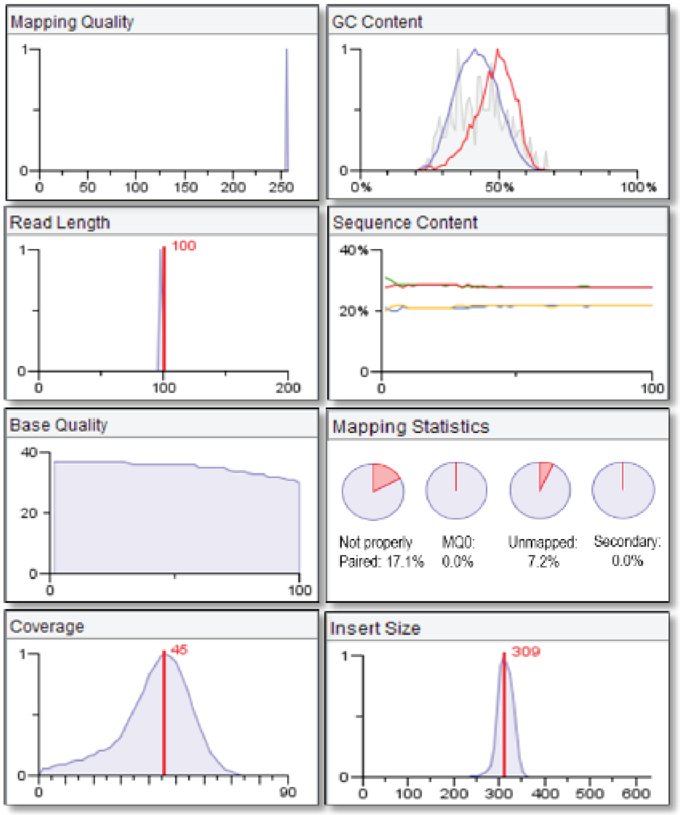


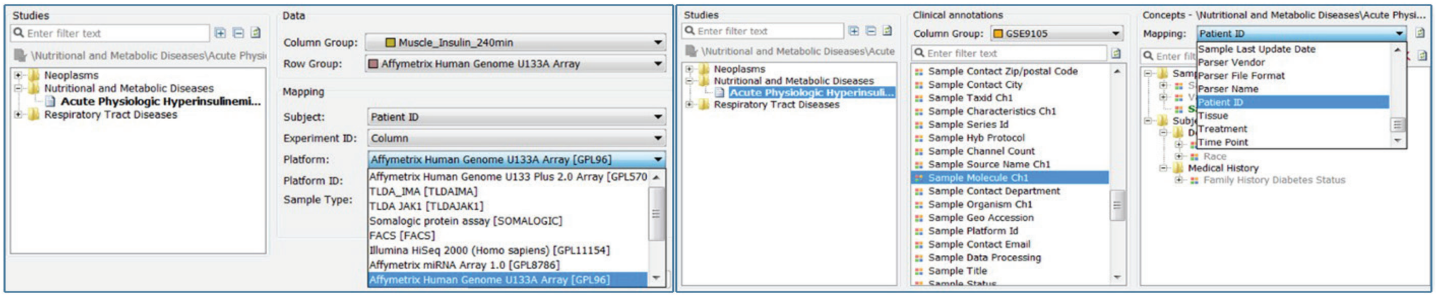
Fig. 2: Automated quality control reporting. Configurable and dynamic quality reports in Genedata Profiler provide a rich set of quality metrics on NGS reads.

Simplifying data loading into tranSMART

tranSMART utilizes a highly complex Extract-Transform-Load (ETL) infrastructure to allow expert users to load data into tranSMART. Loading omics and clinical sample annotations (metadata) into tranSMART can therefore be time consuming and expensive.

Genedata Profiler uses its own public APIs to load data directly into tranSMART, allowing data to be published with a click of a button in minutes rather than hours (Fig. 3).

In addition, Genedata Profiler allows augmentation of studies without the need to reload all data for each study. Out-of-the-box integration with public omics data sources such as GEO, ArrayExpress, and various other resources, allows researchers to integrate data from various public repositories easily into tranSMART. The easy access to a wide variety of multi-omics data enables efficient data discovery and data sharing.


Studies

Enter filter text

\Nutritional and Metabolic Diseases\Acute Physi

Neoplasms
Nutritional and Metabolic Diseases
Acute Physiologic Hyperinsulinemi...
Respiratory Tract Diseases

Data

Column Group: Muscle_Insulin_240min
Row Group: Affymetrix Human Genome U133A Array

Mapping

Subject: Patient ID
Experiment ID: Column
Platform: Affymetrix Human Genome U133A Array [GPL96]
Platform ID:
Sample Type:

Affymetrix Human Genome U133 Plus 2.0 Array [GPL570
TLDA_IMA [TLDAIMA]
TLDA JAK1 [TLDAJAK1]
Somalogic protein assay [SOMALOGIC]
FACS [FACS]
Illumina HiSeq 2000 (Homo sapiens) [GPL11154]
Affymetrix miRNA Array 1.0 [GPL8786]
Affymetrix Human Genome U133A Array [GPL96]

Studies

Enter filter text

\Nutritional and Metabolic Diseases\Acute

Neoplasms
Nutritional and Metabolic Diseases
Acute Physiologic Hyperinsuli...
Respiratory Tract Diseases

Clinical annotations

Column Group: GSE9105

Enter filter text

Sample Contact Zip/postal Code
Sample Contact City
Sample Taxid Ch1
Sample Characteristics Ch1
Sample Series Id
Sample Hyb Protocol
Sample Channel Count
Sample Source Name Ch1
Sample Molecule Ch1
Sample Contact Department
Sample Organism Ch1
Sample Geo Accession
Sample Platform Id
Sample Contact Email
Sample Data Processing
Sample Title
Sample Status

Concepts - \Nutritional and Metabolic Diseases\Acute Physi...

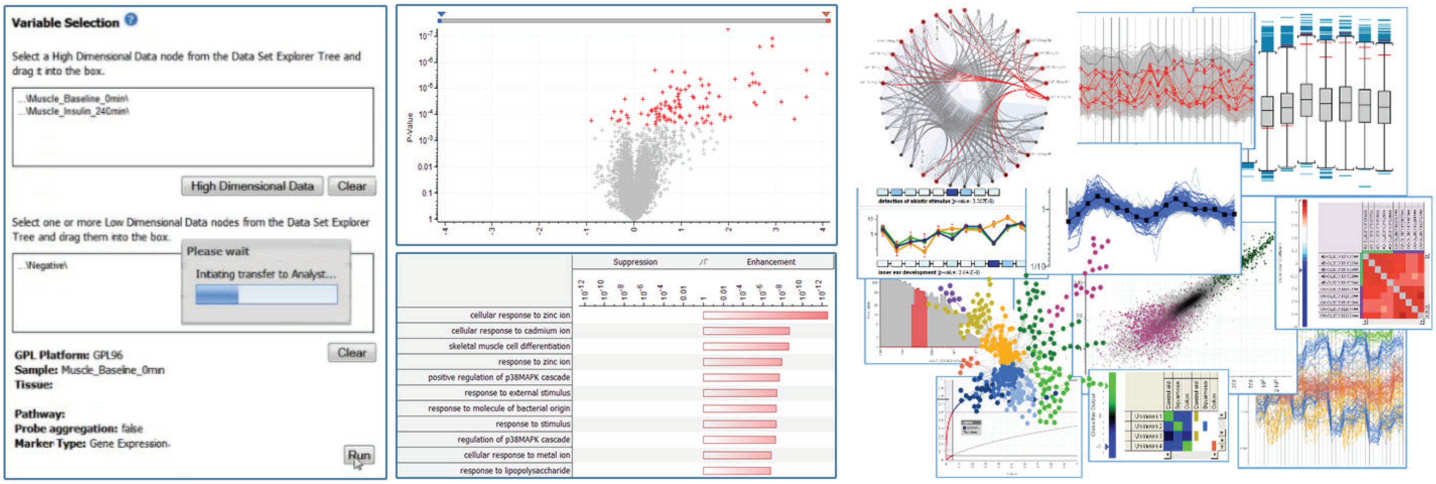
Mapping: Patient ID

Sample Last Update Date
Parser Vendor
Parser File Format
Parser Name
Patient ID
Tissue
Treatment
Time Point

Enter filt

Samp
S
V
S
Subj
D
Race
Medical History
Family History Diabetes Status

Fig. 3: Left: Saving high-dimensional data to tranSMART. User selects a study from the data warehouse study tree together with other data groups to be uploaded. The data is uploaded, processed and written to tranSMART and the user notified when data is available. Right: Saving clinical data to tranSMART. Additional clinical annotations can be selected from a list by dragging and dropping. Omic and clinical annotations may be linked by choosing an appropriate field, e.g. subject ID.



Variable Selection

Select a High Dimensional Data node from the Data Set Explorer Tree and drag it into the box.

...\Muscle_Baseline_0min\
...\Muscle_Insulin_240min\

Select one or more Low Dimensional Data nodes from the Data Set Explorer Tree and drag them into the box.

...\Negative\

Please wait

Intiating transfer to Analyst...

GPL Platform: GPL96
Sample: Muscle_Baseline_0mn
Tissue:

Pathway:
Probe aggregation: false
Marker Type: Gene Expression.

Run
P-Value
0.01
0.1
1
-4
-3
-2
-1
0
1
2
3
4
Suppression
Enhancement
10^-12
10^-10
10^-8
10^-6
10^-4
0.01
1
0.01
10^-4
10^-6
10^-8
10^-10
10^-12
cellular response to zinc ion
cellular response to cadmium ion
skeletal muscle cell differentiation
response to zinc ion
positive regulation of p38MAPK cascade
response to external stimulus
response to molecule of bacterial origin
response to stimulus
regulation of p38MAPK cascade
cellular response to metal ion
response to lipopolysaccharide

Fig. 4: Left: Initiating data transfer from tranSMART to Genedata Profiler. After selecting high- and low-dimensional data nodes in tranSMART, data is automatically transferred to the statistical module of Genedata Profiler. **Middle: The Volcano Plot visualizer** (top) displays a scatter plot of markers in which P-values are plotted against n-fold change. The most significant markers (red) are highly expressed after insulin treatment. A gene ontology Fisher's Exact Test (bottom) indicates that these genes are involved in cellular responses to zinc ions, which play a central role in glycemic control. **Right:** The Genedata Profiler platform was specifically tailored for the integration & interpretation of experimental data in translational R&D, providing a wide range of data analyses through a rich statistical toolbox and intuitive visualizations.

Identifying & reporting clinically relevant biomarkers

Genedata Profiler makes tranSMART smarter by adding sophisticated statistical methods not available in tranSMART. It offers:

- **A rich statistical toolbox to perform a wide range of data analyses;**
- **External algorithms as plugins;**
- **Integration of data across technologies and studies from in-house and public data sources;**
- **Sophisticated data visualization and reports.**

Multi-omics approaches inherently increase the already growing complexity of data in the life sciences. Genedata Profiler provides scalability beyond the capabilities of tranSMART so that huge amounts of complex data can be analyzed within a short time.

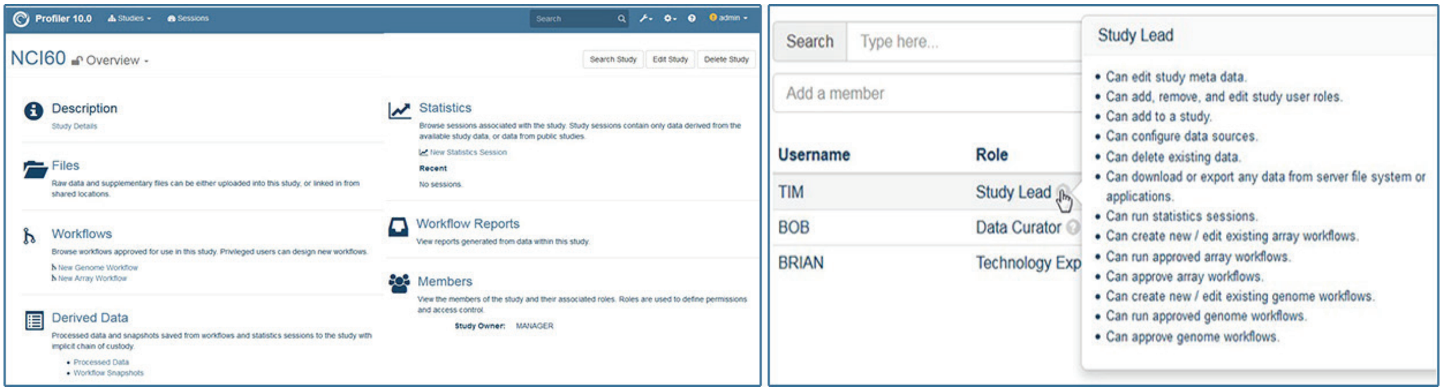
Using the statistical tools of Genedata Profiler (Fig. 4), we were able to identify a novel biomarker *Gadd45A* (a diabetes-associated gene), potentially linking diabetic cardiomyopathy and baroreflex dysfunction, which was not detected in the original study by Coletta et al[4].

Making the data regulatory-compliant

As pharmaceutical companies increasingly leverage confidential patient data such as medical records and genetic information in translational research, important regulations governing data security and privacy must be respected. Recent high-profile decisions, e.g. replacement of the EU-US Safe Harbor agreement on transatlantic data exchange with a new "Privacy Shield", impose stronger compliance requirements on organizations. Genedata Profiler provides comprehensive capabilities to ensure patient privacy and maintain the chain of custody of data, goals that are core to regulatory compliance.

For example, in our biomarker case study, the integration with tranSMART ensures that the same access controls are used between Genedata Profiler and tranSMART, safeguarding the insulin study patient data and reducing exposure to regulatory non-compliance risk.

The information infrastructure of Genedata Profiler enables organizations to cope with the wide range and volumes of omics data as well as the wide variety of data consumers (e.g. bioinformaticians, analysts, biologists, etc.).



Description
Study Details

Files
Raw data and supplementary files can be either uploaded into this study, or linked in from shared locations.

Workflows
Browse workflows approved for use in this study. Privileged users can design new workflows.
New Genome Workflow
New Array Workflow

Derived Data
Processed data and snapshots saved from workflows and statistics sessions to the study with implicit chain of custody.
- Processed Data
- Workflow Snapshots

Statistics
Browse sessions associated with the study. Study sessions contain only data derived from the available study data, or data from public studies.
New Statistics Session
Recent
No sessions.

Workflow Reports
View reports generated from data within this study.

Members
View the members of the study and their associated roles. Roles are used to define permissions and access control.
Study Owner: MANAGER

Username	Role
TIM	Study Lead
BOB	Data Curator
BRIAN	Technology Exp

Study Lead
- Can edit study meta data.
- Can add, remove, and edit study user roles.
- Can add to a study.
- Can configure data sources.
- Can delete existing data.
- Can download or export any data from server file system or applications.
- Can run statistics sessions.
- Can create new / edit existing array workflows.
- Can run approved array workflows.
- Can approve array workflows.
- Can create new / edit existing genome workflows.
- Can run approved genome workflows.
- Can approve genome workflows.

Fig. 5: Left: Study-centric, role-based web UI. The web interface facilitates collaboration, data, method and results sharing. The underlying enterprise architecture provides integration/federation with HPC file systems, internal and external databases and public data sources. **Right: User role management in Genedata Profiler.** The 'Members' dialog lists all the members of a selected study and the specific roles that are associated with those members. Roles are used to define permissions and access control.

Strict user role management, audit trails, access authorization, data federation, data lifecycle and method management, and a comprehensive reporting infrastructure are core components of the software which enable:

- **Collaboration between different user roles throughout a global organization using comprehensive role-based access controls;**
- **Integration, federation, and curation of the wide variety of omic, phenotypic and patient data from internal (e.g. tranSMART), external and public data sources;**
- **Sharing of data, methods and results.**

Summary

Genedata Profiler makes omic-based patient profiling processes significantly more efficient. It streamlines the whole data processing, analysis, and management process, and reduces the time it takes to perform biomarker studies. As we have shown in this case study, Genedata Profiler and its bidirectional integration with tranSMART can help scientists gain new insights into omics data from clinical studies, while reducing compliance risk in their translational research.

References

1. Schumacher, A., Rujan, T. & Hoefkens, J. *A collaborative approach to develop a multi-omics data analytics platform for translational research.* Appl. Transl. Genomics 4–7 (2014). doi:10.1016/j.atg.2014.09.010.
2. Athey, B. D., Braxenthaler, M., Haas, M. & Guo, Y. 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. 2013, 6–8 (2013).
3. Väremo, L. et al. *Proteome- and Transcriptome-Driven Reconstruction of the Human Myocyte Metabolic Network and Its Use for Identification of Markers for Diabetes.* Cell Rep. 11, 921–933 (2015).
4. Coletta, D. K. et al. *Effect of acute physiological hyperinsulinemia on gene expression in human skeletal muscle in vivo.* Am.J.Physiol Endocrinol.Metab 294, E910–E917 (2008).

Please cite this article as:
Schumacher, A., Collins, M., Fisher-Pollard, M. & Rujan, T. (2016) *Efficient genomic profiling of patients: the benefit of systems interoperability*, doi: 10.13140/RG.2.1.3093.4648

Genedata Profiler^{TM} is part of the Genedata portfolio of advanced software solutions that serve the evolving needs of drug discovery, industrial biotechnology, and other life sciences.

Basel | Boston | Munich | San Francisco | Tokyo
www.genedata.com/profiler | profiler@genedata.com