Translational Informatics Management System (TIMS): Towards OMICS-based clinical data management for long-term curation of clinical studies

Weihong Tay[1], Erwin Tantoso[1], Frank Eisenhaber[1], Wing-Cheong Wong[1]

[1] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, Singapore 138671

Summary

With the maturation of sequencing technology over the past decade, the cost associated to an OMICS-based clinical study is no longer a limiting factor even for large cohorts e.g. the UK's 100K genomes project [1]. However, the real cost of such study goes beyond sequencing or data generation in general [2]; The amount of raw sequencing data per sample can be quite sizable and quickly amass to quite a collection even for a modest cohort in contrast to the array-based technology that it has inevitably displaced.

Often, a poorly tackled area in the post-data production of cohort studies is the concerted management of the clinical meta-information (e.g., subjects' demographics, multiple records of domain-specific clinical measures and other information) and the associated OMICS datasets over the course of these studies and eventually their long-term curation after their publication. In particular, these voluminous OMICS datasets require heavy preprocessing to obtain analysis-ready format (e.g., gene count quantification, genetic variants and mutations) prior to any phenotype-genotype analysis. Another important consideration is the ability to re-process the OMICS datasets with alternative or updated algorithms where multiple datasets may be aggregated to perform analysis to test new hypotheses or to simply affirm the reproducibility [3] of the clinical results in a larger set. Although heavier systems (e.g. SysMO-SEEK, DIPSBC, openBIS, Gaggle/BRM) do exist, they are not necessarily open-source freeware and they often require complex deployment and distributed IT infrastructure [4].

Specifically, we refer to an OMICS-based clinical data management open-source software that curates study-related clinical information, manages the raw processing of diverse OMICS datasets to a preprocessed analysis-ready state and finally visualizes the clinical information and processed output in a single access-controlled and audit-trailed environment. Most importantly, this data management system should provide the skeletal software framework for which any appropriate OMICS pipelines and visualizers can be integrated seamlessly in a scaleable fashion. For this purpose, the Translational Informatics Management System, herein TIMS software suite was built.

Functionality and Implementation

The Translational Informatics Management System (TIMS) describes a server-side clinical data management and OMICS production system for

human research. From a technical overview, its central design is based on a Model-View-Controller (MVC) design pattern and implemented in Java server Faces (JSF) supplemented by general open-source software packages as listed in Figure Table 1. In particular, GNU trove library and Ominfaces offer higher performance data structures (e.g. map, set, trees) and additional utilities than the native Java libraries. Meanwhile, the overall graphical user interface (GUI) is written in JSF with enhanced graphical modules from the PrimeFaces.

By functionality, TIMS implements 4 main functions : audit trail (activity tracking), access control (accounts/group management, study management), workflow tracking (job management) and tools (pipeline, visualization); Refer to Figure 1. Firstly, the audit trail function captures all users' activities in the system which includes job submissions, data processing and any system activities. These information are pushed onto the backend relational database - PostgresSQL as the system's data store. Secondly, the access control function manages the users' access privilege based on a hierarchical structure of role/work unit (e.g. Director/Institute, Head/Department, Principal Investigator/Group, User/Group) through the accounts/group management module. Meanwhile, the study management module organizes both clinical and OMICS data into individualized cohort studies which are then assigned to users; All study data sets (both subjects' meta-information and OMICS datasets) are stored in PostgresSQL database. Finally, the level of data access is dependent on the specific user's privilege set by the accounts/group management module. Thirdly, the workflow tracking function is implemented as a job management module to track data production (i.e., from raw data to pre-processed/analysis-ready data) of OMICs datasets. Any data production task submission via the appropriate OMICs pipeline will appear as a job under this module with their status (complete, on-going) appropriately reflected. Lastly, the tools function contains tools for pipelines and visualizers. The current pipeline tools are OMICs-centric and converts RAW data to its preprocessed format (see Figure Table 2 for complete list). Meanwhile, the current visualizer in TIMS - cBioPortal allows for visualization and query of the preprocessed OMICs data.

Availability

The software is distributed under a GNU General Public License v3.0 is available at https://github.com/bii-absd/tims, together with the installation guide and tutorials. The demo datasets used in the tutorial are available for download at http://mendel.bii.a-star.edu.sg/SEQUENCES/TIMS/. A live demo version of TIMS is available at https://tims.bii.a-star.edu.sg/TIMS/login.xhtml.

Acknowledgements

We acknowledge contributions from Joanne Lee for the testing and documentation of TIMS during the genesis of this project.

References

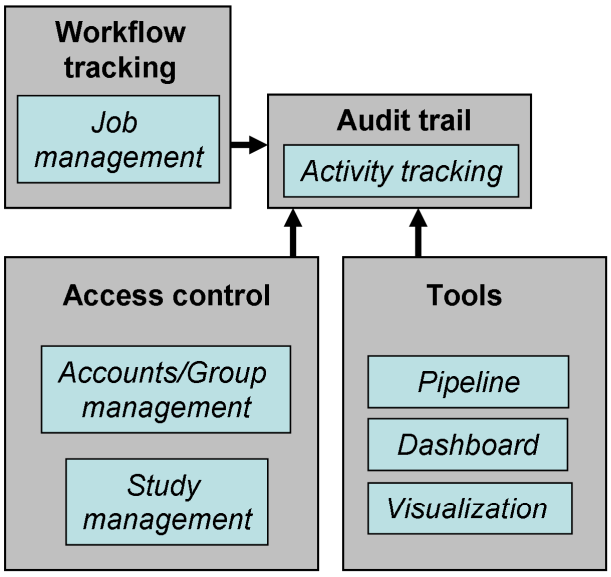
1. Samuel GN, Farsides B: **The UK's 100,000 Genomes Project: manifesting policymakers' expectations.** *New Genet Soc* 2017, **36:**336-353.

2. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, Zhang J, Weinstock GM, Isaacs F, Rozowsky J et al.: **The real cost of sequencing: scaling computation to keep pace with data generation.** *Genome Biol* 2016, **17:**53.

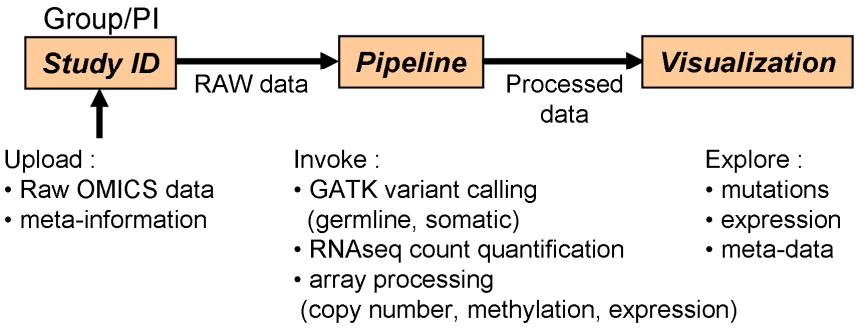
3. Fanelli D: **Opinion: Is science really facing a reproducibility crisis, and do we need it to?** *Proc Natl Acad Sci U S A* 2018, **115:**2628-2631.

4. Wruck W, Peuker M, Regenbrecht CR: **Data management strategies for multinational large-scale systems biology projects.** *Brief Bioinform* 2014, **15:**65-78.

TIMS functionality



TIMS workflow



TIMS user account/group structure

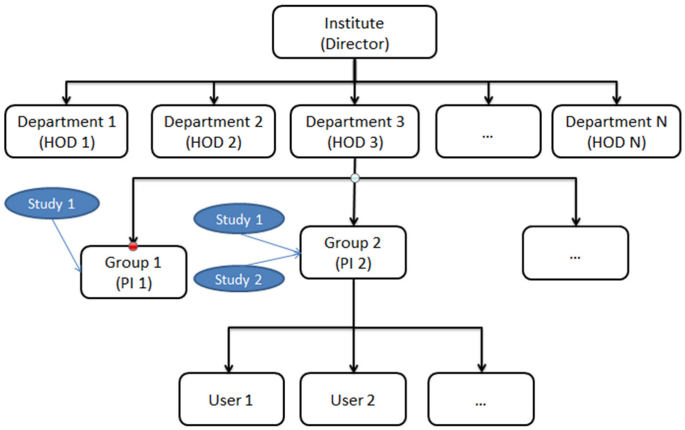


Figure 1 : TIMS functionality, workflow and user account/group

Tables

Software Module/Package	Description	Open-source License
General function		
PostgreSQL	Open source object-relational database system with over 30 years of active development.	PostgreSQL License (Similar to BSD)
Apache Commons Math	A library of lightweight, self-contained mathematics and statistics components addressing the most common practical problems not immediately available in Java.	Apache License v2.0
GNU Trove Library	High performance collections (data structures) for Java	Lesser GNU Public License (LGPL)
omnifaces	A utility library for JSF 2 that focuses on utilities that ease everyday tasks with the standard JSF API.	Apache License v2.0
log4j	Popular logging package for Java.	Apache Software License
Graphical User Interface/Visualization		
Primefaces	A popular open source framework for JavaServer Faces featuring over 1000 components, touch optimized mobilekit, client side validation, theme engine and more.	Apache License v2.0
cBioPortal	Resource for interactive exploration of multidimensional cancer genomics data sets.	GNU Affero General Public License Version 3
Password		
jBCrypt	A Java implementation of OpenBSD's Blowfish password hashing code.	BSD
File reader		
PDFBox	An open source Java tool for working with PDF documents.	Apache License v2.0
Apache POI	Java API to access Microsoft Format files.	Apache License v2.0
XMLBeans	Technology for accessing XML by binding it to Java types.	Apache License v2.0
xlsx-streamer	Streaming Excel Reader	GPL 2.0
Meta information error check		
JPL	A Java Interface to Prolog	GNU Library Public License
SWI-Prolog	Offers a comprehensive free Prolog environment	Simplified BSD License

Table 1 Java libraries used in TIMS

Package	Description	License
Affymetrix Gene Expression Pipeline		
apt-probeset-summarize	A program for summarizing expression probe data from Affymetrix CEL files.	GNU-GPL and GNU-LGPL
DNA-Seq Pipeline		
GATK_v4.0.8.1	Genome Analysis Toolkit consists of wide variety of tools with a primary focus on variant discovery and genotyping.	BSD (3-Clause)
Picard_v2.6.0	A set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM	MIT
Cromwell_v34	Open-source workflow execution engine that can connect through a variety of different platforms through pluggable backends, both local and cloud	BSD (3-Clause)
VCF2MAF-v1.6.12	To convert a VCF format into a MAF format	Apache-2.0
VEP version 86	ENSEMBL Variant Effect Predictor	Apache-2.0
Docker Resources (for DNA-seq Pipeline)		
GATK4.0.8.1	Docker image with GATK4.0.8.1 base engine	BSD (3-Clause)
GOTC (Genomes-in-the-cloud) v2.3.0	Docker image which contains a collection of software packages used in Broad production pipelines to perform genomic analyses. Those software packages are shown below.	
	• Picard	MIT
	• BWA mem	GPLv3/MIT
	• GATK	BSD (3-Clause)
	• Samtools	MIT/Expat
	• VerifyBamID	MIT
	• Python 3	PSF/GPL-compatible
	• R	GPLv2/GPLv3
	• ggplot 2	GPLv2
	• Java 8 Runtime	GPLv2
Python2.7	Docker image with python2.7	PSF/GPL-compatible
RNA-Seq Pipeline		
RSEM-1.2.21	Software package for estimating gene and isoform expression levels from RNA-seq data.	GPLv3
HTSeq-0.6.1p1	Python packages to process data from HTS assays which include counting reads which overlap annotation features	GPLv3

Table 2 OMICs pipeline tools available in TIMS