

**Translational Informatics Management System (TIMS): a
JAVA windows interface for clinical and OMICS pipeline
and visualization management**

A Manual

Release 1.0

**Tay Wei Hong, Erwin Tantoso, Joanne Lee, Frank
Eisenhaber, Wing-Cheong Wong**

Contents

Chapter 1: Introduction	1
Chapter 2: TIMS by functionality	2
2.1 Work-Unit Management and Account Management	3
2.1.1 Work-unit Management	3
2.1.2 Account Management	4
2.2 Study Management	5
2.3 Pipeline Management	7
2.4 Feature Management	8
2.5 Dashboard Summary Configuration	8
2.6 Job Management	9
2.7 Activity Tracking	9
Chapter 3: TIMS Home Page	10
3.1 TIMS by workflow view	11
3.1.1 Array Data and NGS Data processing	12
3.1.2 Visualization of Analysis Outputs	12
3.1.3 Subject Data	13
3.1.4 My Work Area	13
3.1.5 My Study	13
3.2 TIMS by dashboard view	14
Chapter 4: Subject Data	16
4.1 Meta Data Management	16
4.1.1 Prepare Meta Data	16
4.1.2 Upload Meta Data and Mapping Files	18
4.1.3 Meta Data QC	19
4.2 Raw Data Management	20
Chapter 5: FAQ	23
Chapter 6: Appendix	24
References	25

Chapter 1: Introduction

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 like 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. In preceding circumstance, it is to the best of our knowledge that there are no lightweight open-source software to perform such post-data production clinical data management that will also allow for future add-on functionalities. Although heavier systems (e.g. SysMO-SEEK, DIPSBK, openBIS, Gaggle/BRM) do exist, they are not necessarily freeware and often requires 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. For this purpose, the Translational Informatics Management System, herein TIMS software suite was built.

Chapter 2: TIMS by functionality

The design of the TIMS software is centralized around 4 main functionalities: (i) audit trail, (ii) access control, (iii) workflow tracking and (iv) tools that can be controlled through admin management via an administrator's account. For each functionality, the specific management tasks are depicted in Figure 1.

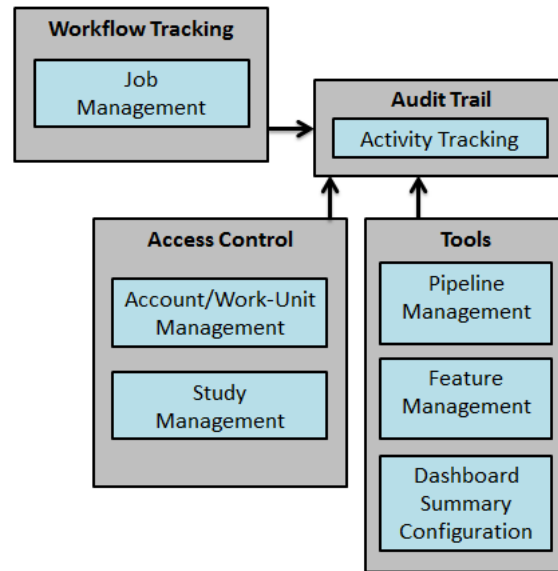


Figure 1 TIMS by functionality and its specific management tasks

In TIMS, the management tasks are listed under the Admin menu and it is important to note that most management tasks are only available to an administrator's account. Only the Admin account user has the power of managing the whole TIMS system. The list of management tasks include Work Unit, Account, Study, Pipeline, Visualization, Feature, Dashboard Summary Configuration, Job management and Activity tracking seen in Figure 2. In the sub-sections that follows, each management tasks and users allowed access to these tasks will be further described.

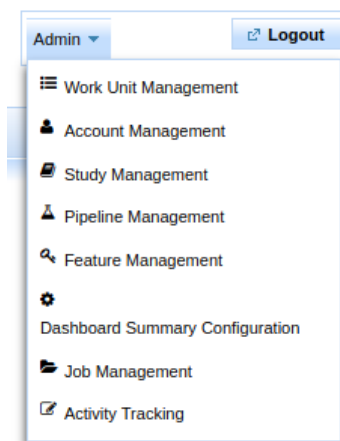


Figure 2: Admin management drop-down list

2.1 Work-Unit Management and Account Management

The *Work Unit Management* and *Account Management* task are closely related to each other. *Work Unit Management* creates the type of institutional units while *Account Management* creates the type of users. Basically, each user account type is tied to a work-unit type as summarized by Table 1. In addition, before a new user account can be created, it is important to ensure that the work unit must first be created.

Account Role	Work Unit
Admin	Group
Director	Institution
HOD	Department
PI	Group
User	Group
Guest	Group

Table 1: Account Role – Working Unit association

2.1.1 Work-unit Management

TIMS is aimed at institutional-level deployment with 3 levels of work units (in a descending hierarchical order) namely institution, department and group as shown in Figure 4. For each institute, there is one director followed by multiple departments with its respective HOD (Head of Department). Under each department, there are multiple PIs (Principal Investigator). And under each PI, there can be multiple users. The different types of user accounts in TIMS will indicate the different access rights and responsibility at the institutional level.

Therefore, upon the installation of the TIMS software, it should be followed by the setup of institutional departments/groups through the *Work-Unit Management* which should logically start in the following order:

Add New Institution → Add New Department → Add New Group

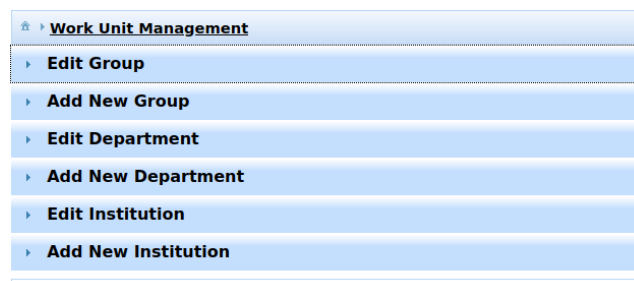


Figure 3: Work Unit Management

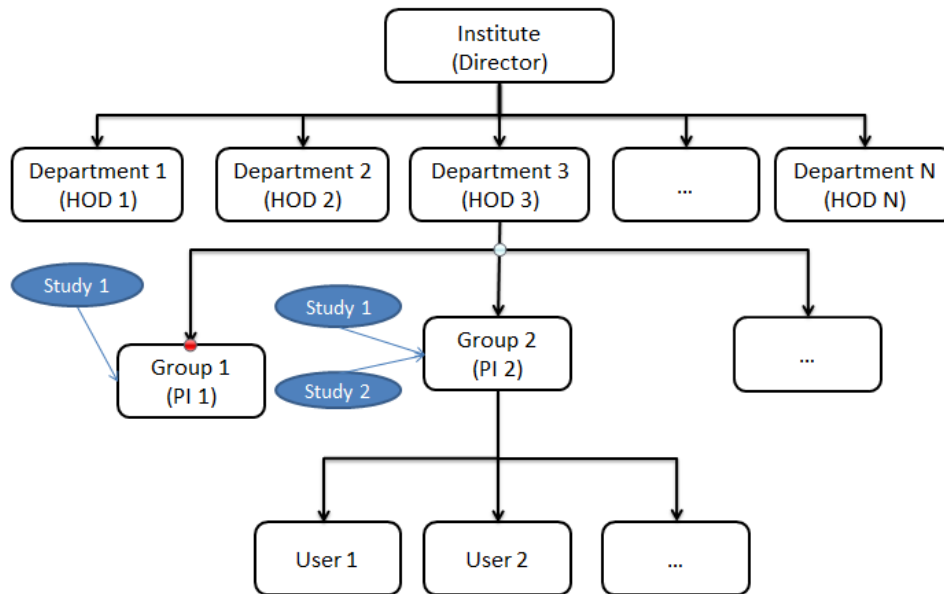


Figure 4: Institutional organization chart in TIMS

All work unit IDs are immutable once created. Group, Department and Institute names can be modified, made active or inactive in the *Edit Group*, *Edit Department* and *Edit Institution*.

2.1.2 Account Management

Once the institutional departments/groups have been created, appropriate user accounts can then be created and assigned to the work units through *Create New User Account* under *Account Management*. The association of an account to its appropriate work unit is summarized in Table 1 above. And depending on the role of the account user, only the appropriate work units will be available for selection.

In the usage of *Account Management*, it is important for the administrator to understand the different level of users within the TIMS system. Altogether, there are 6 different account types available in TIMS: (i) Admin (ii) Director (iii) HOD (Head of Department) (iv) PI (Principal Investigator), (v) User and (vi) Guest. Each account type has different levels of administrative management access as summarized in Table 2. Although it seems restrictive to most accounts types except admin, the difference in account type matters when it comes to managing workflows (see Chapter 3 Table 3). Therefore it is important for the administrator to understand which level of account is being created.

Based on Figure 4, we elaborate on the different account types. Starting from the PI account user, every study created on the TIMS software should originate from and manage by him/her. Only the PI and all users under the PI can access the study and perform analysis on the data belonging to the study. Going up the hierarchical chart, the HOD oversees multiple PIs. Therefore, when the HOD logs into TIMS system, he/she will have an overview of all the studies belonging to the PIs under the same department. However, the HOD will not be able to perform analysis on the data belonging to the study, unless the HOD is the PI himself. Finally, Director oversees

the institute and therefore can have an overview of all the studies within the institute. Unless the Director is the PI of the study, he/she will not be able to perform analysis on the data belonging to other PI.

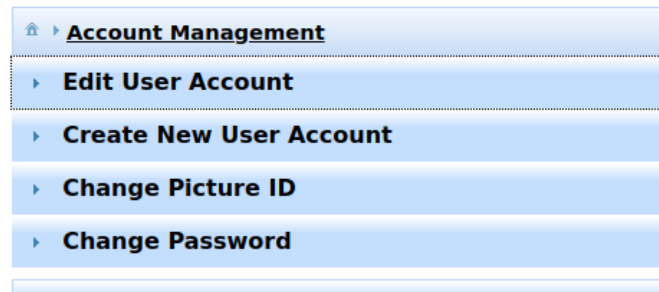


Figure 5: Account Management

User roles, other information such as email or names, active or inactive status can all be modified in the *Edit User Account* tab in case of any changes, with the exception of User ID which is immutable once created. Such changes can only be done using an Administrator's account.

Changing of passwords is available in this page as well.

Note: Guest Accounts here have restricted access that allows for the viewing of Dashboard only. More details can be found in Chapter 3 Table 3.

Homepage Section	Description	Account Types					
		Admin	Director	HOD	PI	User	Guest
Admin	Account Management	Green	Yellow	Yellow	Yellow	Yellow	Red
	Work Unit, Study, Pipeline, Feature, Dashboard Summary Configuration, Job Management	Green	Red	Red	Red	Red	Red
	Activity Tracking	Green	Green	Red	Red	Red	Red
		Green	Green	Red	Red	Red	Red

Table 2: Access privileges for different account types. Rows shaded green represents access to all studies on TIMS; yellow represents access only to studies linked to user account and red represents no access to studies.

2.2 Study Management

The *Study Management* is used to organize multiple studies within the TIMS software. Most importantly, it assigns the ownership of each study to its rightful group and user. Basically, each Study must be owned by a Group which is in turn led by a PI (Principal Investigator). Therefore, the PI of the Group will own and have full

access right to the study. All the Users under the PI will have the right to process and run the analysis of the Study belonging to the Group.

For each Study, the PI is required to specify the following information related to the study, i.e.:

- (i) The Title of the study
- (ii) The Description of the study
- (iii) The Background of the study
- (iv) The Grant Information of the study
- (v) The Start and End date of the study
- (vi) The Annotation that will be used for the study
- (vii) The Disease classification

Among the 7 required fields, special attention is required for the (vi) Annotation and (vii) Disease classification section.

The Annotation is based on the gene annotation database that will be used for the study. Currently TIMS is designed to store the genes' output from the analysis, therefore the list of genes with the annotation will be created in the database as the reference. The gene list is based on the UCSC gene annotation database and is created once every six months. Therefore, the gene annotation database will be labelled as 1H2016, 2H2016, 1H2017, 2H2017, etc. (Note: 1H2016 means the gene annotation based on the first half of the year 2016). Only genes available in the database will be stored following the finalization of the analysed genes' output.

The Disease classification is based on the ICD-10 disease classification. The ICD-10 classification code is used to classify and code all diagnoses, symptoms and procedures recorded in conjunction with hospital care in the US system.

Create New Study

New Study ID

For Institution - Department - Group: TIMSI, TIMSD, TIMSQ, TIMSG

Select Annotation Version

Select Disease under Study

Study Title

Enter description of this study here.

Enter the background of this study here.

Enter the grant information of this study here.

Start Date:

End Date:

Finalized? ☒ No

Create New Study

Study Management

- Edit Study Main Info
- Edit Study Description|Background|Grant Information (aka DBGI)
- Create New Study

Figure 6: Creating a New Study on TIMS (left)

Note: Default for all studies should be NOT finalized. Only ad-hoc studies are set up to be Finalized. More details can be found in Chapter 3, Section 3.1.5.

Figure 7: Study Management (top)

2.3 Pipeline Management

The *Pipeline Management* is used to configure the pipelines available within the TIMS system.

The screenshot shows the 'Pipeline Management' section with a sidebar containing 'Edit Pipeline' and 'Add New Pipeline'. The 'Add New Pipeline' form includes fields for 'Pipeline Name', 'Description', 'Select Technology' (a dropdown menu), 'Command', and 'Parameter'. Below these fields is an 'Editable?' checkbox set to 'No' and a 'Create New Pipeline' button.

Figure 8: Pipeline Management

The pipeline within TIMS is designed such that the pipeline can be called via the following :

`<command> -i <config-file>`

For example: `run-gex-pipeline -i gex.conf`

The config file contains all the necessary parameters to analyse the data, particularly the sample-annotation file, the input file/folder, and the pipeline-specific parameters.

The field “Editable?” refers to the possibility of the user to edit the uploaded raw data of the pipeline. If this is “Yes”, then the user can perform Raw Data Management for the specific pipeline. [Note: Raw Data Management is available from *Home > Subject Data > Raw Data Management*]

Perhaps consider allowing disabling and enabling function for existing pipelines

Adding new pipelines not yet supported by the system would require further modifications to TIMS backend. Please refer to GitHub for TIMS system codes to do so. Please do this at your own risk. TIMS development team will not be responsible for any changes made.

Currently, there are ten pipelines available within TIMS, including both Array Technology and Next Generation Sequencing (NGS) Technology (Figure 9).

The screenshot shows the 'Edit Pipeline' section with a table of available pipelines. The table has columns for Pipeline, Description, Technology, Command, Parameter, and Editable. There are 10 rows of data.

Pipeline	Description	Technology	Command	Parameter	Editable
cnv-affymetrix	Copy Number Variation Pipeline (Affymetrix)	Array	/usr/local/bin-pipelines/run-affymetrix-cnv-pipeline	-i	YES
cnv-illumina	Copy Number Variation Pipeline (Illumina)	Array	/usr/local/bin-pipelines/run-illumina-cnv-pipeline	-i	YES
gatk-tar-germ	GATK Targeted Sequencing (Germline Mutation)	NGS	/usr/local/bin-pipelines/seq-pipelines/run-dnaseq-germline	-i	YES
gatk-tar-soma	GATK Targeted Sequencing (Somatic Mutation)	NGS	/usr/local/bin-pipelines/seq-pipelines/run-dnaseq-somatic	-i	YES
gatk-wg-germ	GATK Whole-Genome Sequencing (Germline Mutation)	NGS	/usr/local/bin-pipelines/seq-pipelines/run-dnaseq-germline	-i	YES
gatk-wg-soma	GATK Whole-Genome Sequencing (Somatic Mutation)	NGS	/usr/local/bin-pipelines/seq-pipelines/run-dnaseq-somatic	-i	YES
gex-affymetrix	Gene Expression Pipeline (Affymetrix)	Array	/usr/local/bin-pipelines/run-gex-pipeline	-i	YES
gex-illumina	Gene Expression Pipeline (Illumina)	Array	/usr/local/bin-pipelines/run-illumina-gex-pipeline	-i	YES
meth-illumina	Methylation Pipeline (Illumina)	Array	/usr/local/bin-pipelines/run-meth-pipeline	-i	YES
seq-rna	RNA Sequencing Pipeline	NGS	/usr/local/bin-pipelines/run-rnaseq-pipeline	-i	YES

Figure 9: Available pipelines within TIMS

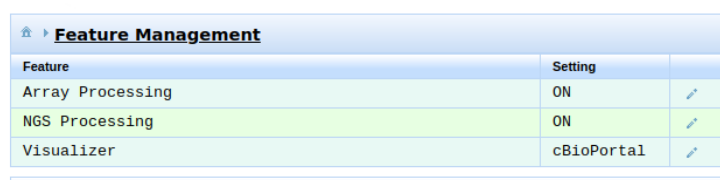
2.4 Feature Management

The purpose of *Feature Management* is to control the access to different tools available for users in an instance of TIMS, and hence an institute.

In general, the features can be made unavailable/ available on the *Home Page*, by toggling the Setting option to “OFF”/“ON” respectively..

In particular, one of the features, Visualizer, available allows for the visualization of processed OMICS data in the TIMS software. The visualizer can also be hidden/made unavailable on the *Home Page* by toggling the Setting to “None”.

To enable the visualizer feature, simply select for the type of visualizer to be used in the drop down menu for the Setting Option. Currently, there is only one visualization option, i.e. cBioPortal.






Feature Management		
Feature	Setting	
Array Processing	ON	
NGS Processing	ON	
Visualizer	cBioPortal	

Figure 10: Feature Management

2.5 Dashboard Summary Configuration

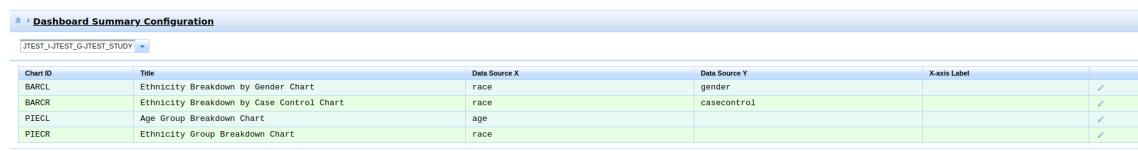
The Dashboard Summary Configuration page allows for customization of the types of data to be summarized and displayed on the TIMS Dashboard which provides an overview on the meta-data of selected studies individually.

There are 4 chart types available for customization in this section, 2 bar charts (BARCL and BARC) and 2 pie charts (PIECL and PIECR). Clicking on the edit icon on the right of each row, customizations to the title, axis label and data sources can be made.

Title should be included for all 4 charts. Data sources refer to columns found in the uploaded meta-data.

For bar charts, both X-axis and Y-axis data sources can be modified. Data source for X-axis and Y-axis here should be a qualitative and a quantitative variable respectively. The X-axis Label should be the qualitative variable for Data Source X.

For pie charts, only Data Source X should be modified.







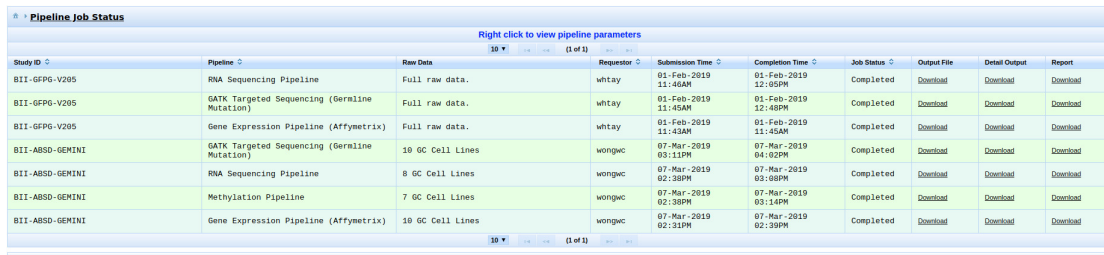
Dashboard Summary Configuration					
JTEST_JTEST_G_JTEST_STUDY					
Chart ID	Title	Data Source X	Data Source Y	X-axis Label	
BARCL	Ethnicity Breakdown by Gender Chart	race	gender		
BARCR	Ethnicity Breakdown by Case Control Chart	race	casecontrol		
PIECL	Age Group Breakdown Chart	age			
PIECR	Ethnicity Group Breakdown Chart	race			

Figure 11: Dashboard Summary Configuration

2.6 Job Management

The *Job Management* is used to check the status of all pipeline jobs within the TIMS system. This includes all completed, running and failed jobs. As the TIMS Administrator, you will be able to check the status of all the jobs submitted by the TIMS user. This is only accessible for administrator accounts (see Table 2). To access the *Job Management* page, go to the *Admin* drop-down list and select *Job Management*.



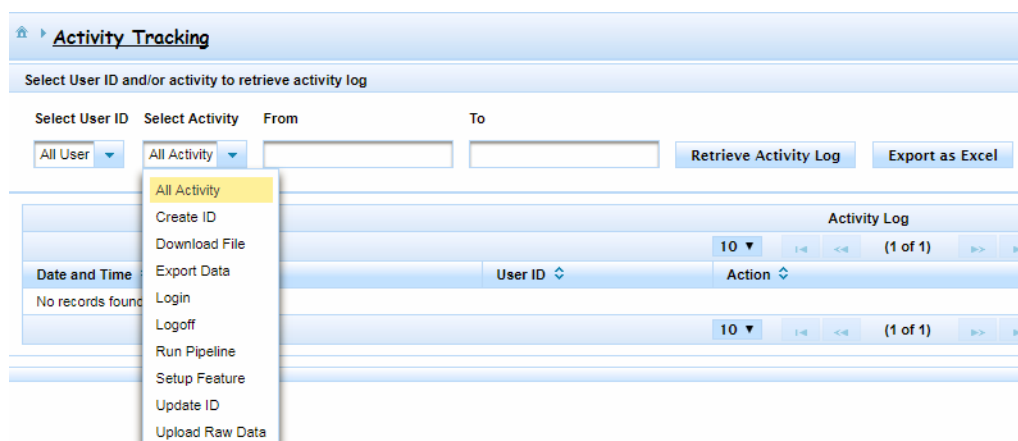
Study ID	Pipeline	Raw Data	Requestor	Submission Time	Completion Time	Job Status	Output File	Detail Output	Report
BII-GFPG-V295	RNA Sequencing Pipeline	Full raw data.	whitay	01-Feb-2019 11:45AM	01-Feb-2019 12:10PM	Completed	Download	Download	Download
BII-GFPG-V295	GATK Targeted Sequencing (Germline Mutation)	Full raw data.	whitay	01-Feb-2019 11:45AM	01-Feb-2019 12:48PM	Completed	Download	Download	Download
BII-GFPG-V295	Gene Expression Pipeline (Affymetrix)	Full raw data.	whitay	01-Feb-2019 11:45AM	01-Feb-2019 11:45AM	Completed	Download	Download	Download
BII-ABSD-GEMINI	GATK Targeted Sequencing (Germline Mutation)	10 GC Cell Lines	wongwc	07-Mar-2019 03:11PM	07-Mar-2019 04:02PM	Completed	Download	Download	Download
BII-ABSD-GEMINI	RNA Sequencing Pipeline	8 GC Cell Lines	wongwc	07-Mar-2019 02:38PM	07-Mar-2019 02:38PM	Completed	Download	Download	Download
BII-ABSD-GEMINI	Methylation Pipeline	7 GC Cell Lines	wongwc	07-Mar-2019 02:38PM	07-Mar-2019 03:14PM	Completed	Download	Download	Download
BII-ABSD-GEMINI	Gene Expression Pipeline (Affymetrix)	10 GC Cell Lines	wongwc	07-Mar-2019 02:31PM	07-Mar-2019 02:39PM	Completed	Download	Download	Download

Figure 12: Screen capture for job management

2.7 Activity Tracking

The *Activity Tracking* is used to track the activity of all the users within the TIMS system. The *Activity Tracking* involves all the activities as well as specific activities shown in Figure 13. The Administrator can specify the date of such activity and export it to excel spreadsheet for auditing if necessary.

This is only accessible for Administrator and Director accounts (refer to Table 2). To access the *Activity Tracking* page, go to the *Admin* drop-down list and select *Activity Tracking*.



Activity Tracking

Select User ID and/or activity to retrieve activity log

Select User ID: All User | Select Activity: All Activity | From: | To: | Retrieve Activity Log | Export as Excel

Activity Log

Date and Time	User ID	Action
No records found		

10 (1 of 1)


Figure 13: Activity Tracking

Chapter 3: TIMS Home Page

Once the appropriate groups and user type have been properly setup (see Section 2.1), the *Home Page* will be shown once a user logs in to the system. Different user accounts will have different views of the *Home Page* once logged in. Generally, there are two different views in TIMS: a workflow view and a dashboard view.

The *Home Page* for the user types : (i) Admin (ii) PI (Principal Investigator) and (iii) User will be the *workflow* view as shown in Figure 14a, where functions such as pipelines, visualization, clinical subject data management and working area are available under this view.

Meanwhile, the *Home Page* user types (i) Director role, (ii) HOD (Head of Department) and (iii) Guest is the dashboard view as shown in Figure 14b where an overview of the studies within the Institute or Department will be presented.

For all users, except for Guest, selecting the  icon will direct them to the workflow view.

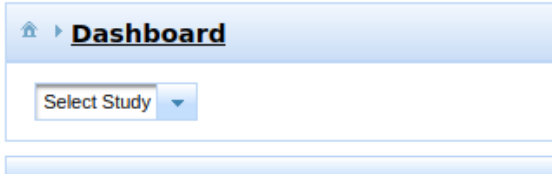
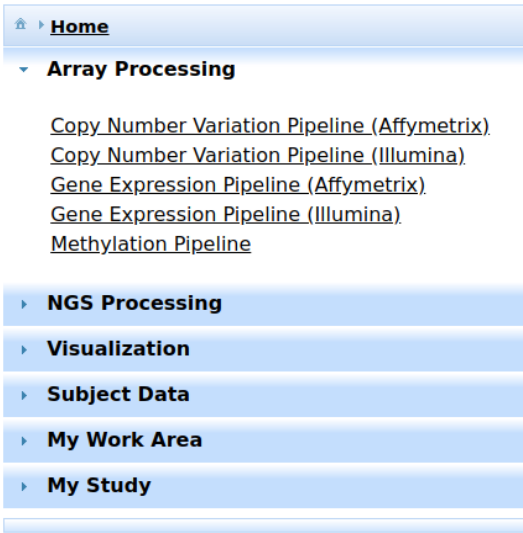


Figure 14a. *Home page* view for user types (i) Admin (ii) PI and (iii) User (left),

Figure 14b. *Home page* view for user types (i) Director (ii) HOD (Head of Department) and (iii) Guest (top)

Table 3 below gives the overall summary on the type of functions that can be performed by the various user types in their respective home page view

.

Homepage Section	Description	Account Types					
		Admin	Director	HOD	PI	User	Guest
Array Data & NGS Data	Running of Pipelines						
Visualization	Setting up and Visualizing of Analysis Outputs						
Subject Data	Uploading & Editing of Meta Data						
	Deleting of Meta Data						
	Uploading and Editing of Raw Data						
My Work Area	Dashboard						
	Pipeline Job Status						
	Completed Study Output						
My Study	Finalise Study						
	Unfinalize Study						
	Close Study						

Table 3: Access privileges for different account types. Rows shaded green represents access to all studies on TIMS; yellow represents access only to studies linked to user account and red represents no access to studies.

3.1 TIMS by workflow view

This is the *Home Page* for users: (i) Admin (ii) PI (Principal Investigator) and (iii) User.

In the workflow view, two conditions have to be met before one can begin processing the data. Firstly, a study belonging to a PI needs to be created by an administrator in TIMS software (Chapter 2; section 2.2). Secondly, the data (meta-information and

OMICs) belonging to the study needs to be uploaded to the TIMS software (see Chapter 4). Once the conditions are fulfilled, a PI and his users can then begin to managed the study in the context of a workflow as shown in Figure 15

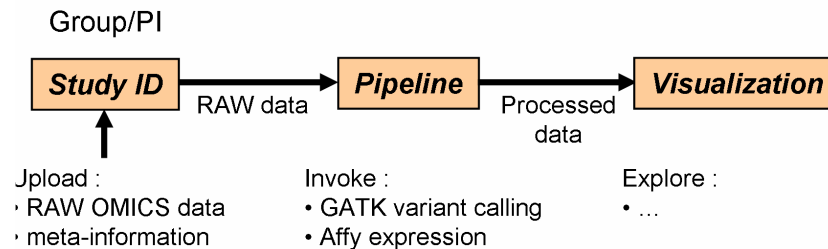


Figure 15: TIMS data processing workflow

3.1.1 Array Data and NGS Data processing

On the *Home Page*, account users are able to run pipelines associated with each technology, such as *Gene Expression Pipeline (Affymetrix)* under the *Array Processing* technology and *GATK Targeted Sequencing (Somatic Mutation and Germline Mutation)*, *GATK Whole Genome Sequencing (Somatic Mutation and Germline Mutation)* pipelines under the *NGS Processing* technology.

To run specific pipelines, select the pipeline of interest and follow the instruction to submit the job.

3.1.2 Visualization of Analysis Outputs

The *Visualization* tab (See Figure 16) provides the function to setup study for visualization as well as visualization of the study data.

In order to visualize the data, the user needs to first set-up the study for visualization. There is only one visualization available at any one time for a study. Therefore, if the user wishes to visualize different sets of data or results of the same study, the user would need to set up the study for visualization again.

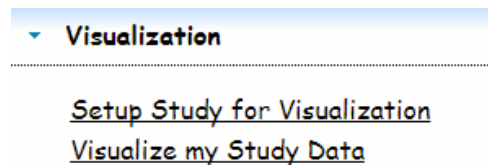


Figure 16: Visualization tab

Note: *There is an interval of one hour from one set up to another visualization set up. This means that the user cannot keep changing the visualization set up in a short period of time.*

3.1.3 Subject Data

The *Subject Data* tab (see Chapter 4) provides the function for clinical subject data management. This includes meta-data management as well as raw-data management which will be elaborated in Chapter 4.

3.1.4 My Work Area

The *My Work Area* tab (see Figure 17) provides the functions to (1) view Dashboard of overview on meta-data uploaded, (2) check the all pipeline job status, and (3) obtain consolidated (finalized) outputs of selected pipelines.

The *Dashboard* provides a quick overview of the demographics and completeness of meta-data uploaded. More details on the Dashboard can be found in Section 3.2.

The *Pipeline Job Status* shows the status of all the jobs belonging to the user, including completed, running and failed jobs.

The *Completed Study Output* section contains collated outputs from selected pipeline analysis based on finalization of study. This is explained in greater detail in the next section 3.1.5.

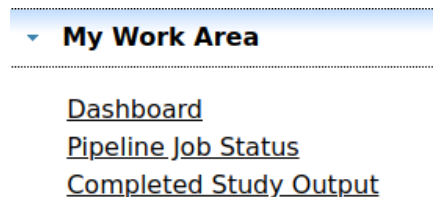


Figure 17: My work Area

3.1.5 My Study

Finalizing study data (*Finalize Study* in the *My Study* Section, see Figure 18 below) is the process of storing the study data into the database. This option is only available for users with PI access rights (see Table 3). Upon finalization, all outputs from selected finalized pipelines are compiled under *Completed Study Output* section (see section 3.1.4). Finalization can only be done provided there is corresponding meta-data available for the subjects and samples found in raw data.

After finalization, account users are still able to run pipeline and edit data if necessary. In the case of more suitable analysis outcomes obtained, the study needs to be unfinalized first in order for PI to re-finalize study with updated pipelines. Unfinalization (*Unfinalize Study* in the *My Study* Section, see Figure 18 below) is only available for Admin accounts (see Table 2).

Closure of study data refers to the archival of studies which are no longer active. The resulting closed study will no longer be available on TIMS and hence no longer accessible. Studies can only be closed after finalizing. Closing of study data (*Close*

Study in the *My Study* Section, see Figure 18 below) is also only available for Admin accounts (see Table 2).

Ad-hoc studies are studies which are not to be stored in the database. These studies utilise TIMS as an analysis platform, allowing them to run established analysis pipelines easily through TIMS. Hence, ad-hoc studies are set up to be finalized in the initial phase (Recall Chapter 2, Section 2.2).

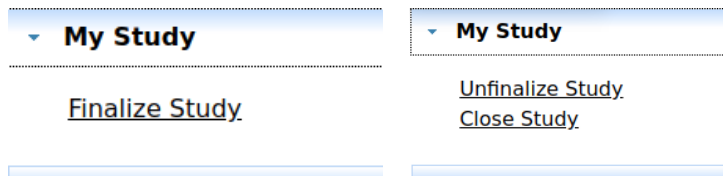


Figure 18: My Study view for PI accounts (left) and Administrator accounts (right).

3.2 TIMS by dashboard view

This is the *Home Page* for users: (i) Director role, (ii) HOD (Head of Department) and (iii) Guest.

For all other users, the Dashboard View can be obtained by going to *My Work Area* > *Dashboard*.

Study to be visualized on the dashboard is then selected and populated on the page, as seen below in Figure 19.

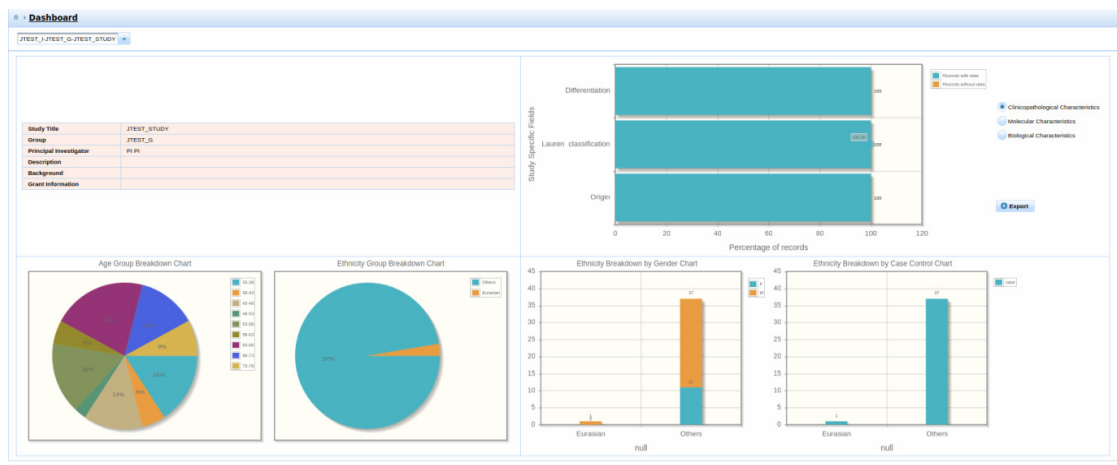


Figure 19: TIMS dashboard view

The Dashboard is split into 3 sections, namely:

1. Study Information on the top left corner (specified in Chapter 2, Section 2.2),
2. Study Specific Fields Completeness on the top right corner, and
3. 4 charts (PIECL, PIECR, BARCL, BARCR, customised in Chapter 2, Section 2.4) on the bottom section of the Dashboard.

Legends for all bar charts on the Dashboard are interactive, to filter for visualizations specific to specific variables.

Mouseover on all charts are also available for more information such as percentages in Section 2 and figure numbers found in categories in Section 3.

To note, in Section 2, variables (columns) in uploaded meta-data are group into a maximum of 3 categories shown as selectable radio buttons. The horizontal bars indicate the rate of completeness found for the selected variables in each category across subjects.

Details on how to customise this Section can be found in Chapter 4, Section 4.1.1. An export function here is also available for users to download the 3 categories of variables used in this section.

Chapter 4: Subject Data

The *Subject Data* tab is available in the workflow view of the *Home Page*. There are two types of subject data management available, i.e. *Meta Data Management* and *Raw Data Management*.



Figure 20: Subject Data

4.1 Meta Data Management

Meta Data Management is used to manage the subject (or patient) clinical data (or meta-data) of a study; hence it is study-specific meta-data management. Only the output of the subject with meta-data information found in the database can be finalized and stored inside the database (recall Chapter 3, Section 3.1.5).

4.1.1 Prepare Meta Data

Preparation of the meta-data is split into 3 sections as follows: (1) Cleaning of Meta Data, (2) Mapping of Core Meta Data Variables, and (3) Grouping of Meta Data Variables.

Sections 2 and 3 are important to facilitate visualization on Dashboard (as in Chapter 3 Section 3.2).

4.1.1.1 Cleaning of Meta Data

Study meta-data should undergo stringent QC prior to upload to minimise trivial errors such as wrong or illogical input formats (eg. Illogical dates: 21/05/0201). Sheet name of Excel File containing the meta-data must be named "Data".

4.1.1.2 Mapping of Core Meta Data

Core compulsory fields (and their allowed inputs) required by TIMS to be included in the meta-data are:

1. SubjectID,
2. Race (allowed inputs (in capital letters): CHINESE, MALAY, INDIAN, EURASIAN, OTHERS),
3. CaseControl (allowed inputs (in lowercase): case, control),
4. Height (up to two decimal places),
5. Weight (up to two decimal places),
6. RecordDate (allowed format: DD/MM/YYYY),
7. DateOfBirth (allowed format: DD/MM/YYYY),
8. Gender (allowed inputs (in uppercase): F, M),
9. AgeAtBaseline (to one decimal place).

The SubjectID and RecordDate fields together make up a unique visit record for each sample. These 9 fields are referred to as core data from hereon.

To map these fields to specific variables (columns) found in meta-data, an Excel file should be created as shown in Figure 21 below. First column consists of 9 compulsory fields above; second column consists of variable (column names) of corresponding columns found in meta-data.

	A	B	C
1	SubjectID	PID	
2	Race	Ethnicity	
3	CaseControl	casecontrol	
4	Height	Height	
5	Weight	Weight	
6	RecordDate	Date	
7	DateOfBirth	DOB	
8	Gender	Gender	
9	AgeAtBaseline	Age	
10			

Figure 21: Core Data Tag File

Should any of these 9 fields not be available in the meta-data, user would need to generate the missing columns in the meta-data.

4.1.1.3 Grouping of Meta Data Variables

For completeness check of selected variables in Section 2 of the Dashboard (as in Chapter 3 Section 3.2), up to 45 variables (columns) in the meta-data would need to be grouped into 3 categories. These are known as study-specific fields from hereon.

These study-specific fields and their corresponding categories are to be uploaded onto TIMS in an Excel File in the format as shown in Figure 22 below.

	A	B	C
1	Clinicopathological Characteristics	Differentiation	
2	Clinicopathological Characteristics	Lauren classification	
3	Clinicopathological Characteristics	Origin	
4	Molecular Characteristics	MSI status	
5	Molecular Characteristics	EBV positive	
6	Molecular Characteristics	RTK amplification	
7	Biological Characteristics	Doubling time (hr)	
8	Biological Characteristics	Invasiveness (No. of Cells)	
9	Biological Characteristics	Tumorigenicity	
10			

Figure 22: Study Specific Fields File

Note: The Study Specific Completeness Check on Dashboard (recall Chapter 3 Section 3.2) defaults to checking across all subjects, but can also be done for only a subset of subjects instead. User would need to input "--" as entry for the selected variables (columns) to non-relevant subjects (not part of selected subset). These subjects would then not be included in the completeness check for the selected variables (columns).

4.1.2 Upload Meta Data and Mapping Files

Only Excel files are accepted for uploading meta-data. All 3 files mentioned in Section 4.1.1 above needs to be uploaded onto TIMS as in Figure 23 below.

Figure 23: Uploading Meta Data for a study

When uploading the *Meta Data File*, a preliminary overview (Figure 24 below) of the data uploaded will be given, with options for users to select on how to proceed with uploading of data. An email will be sent to users' linked email account upon successful upload of data onto TIMS.

Preliminary Overview of Data Quality	
Records with empty or invalid date	5/7035 (0.1%)
Records with missing Subject ID or Gender or Race	0/7035 (0%)
Records with invalid data	56/7035 (0.8%)
Records related to missing visits	0/7035 (0%)
Records ready for further processing	6974/7035 (99.1%)
Skip Consistency Check	<input type="checkbox"/>

Cancel
Proceed

Figure 24: Preliminary Overview of Data Quality

If there is existing meta-data available, the user can select *Download Meta Data* to download the available meta-data. Users can also choose to upload a new copy of meta-data.

New copies of meta-data file uploaded need to contain all existing subject and sample records in the same format (i.e. same columns/fields) found on TIMS, in addition to new records (if any).

In the case where there are major updates in formats of data to be uploaded as compared to previous versions, such as changes to the number /names of variables (columns) found in the meta-data, the existing meta-data file would need to be deleted before the new copy can be uploaded. This is done using the *Delete Core Data Tag*, *Delete All Meta Data* and *Delete Study Specific Fields* options in the *Reset* tab. Only users with administrator rights are able to access this option.

The *Skip Consistency Check* option can be used when user decides that the new data to be uploaded supersedes the previous versions stored on TIMS. As such, the new meta-data file uploaded will replace the existing copy stored entirely.

In the case where the *Skip Consistency Check* option is not selected, the contents of the new meta-data file will be compared with that of the existing copy on TIMS to ensure that all contents are consistent for both copies.

Once successfully uploaded, user would be able to view uploaded meta-data under the *Meta Data Listing* tab as well as the study specific fields under the *Specific Fields Grouping* tab.

4.1.3 Meta Data QC

The Quality Report of Last Data Uploaded consists of information regarding erroneous inputs in meta-data uploaded. The report displays the records in terms of lines in the uploaded excel file that errors were found in the compulsory fields mentioned in Section . These errors include

1. Invalid dates,
2. Missing entries in core data fields (details in Section 4.1.1.2),

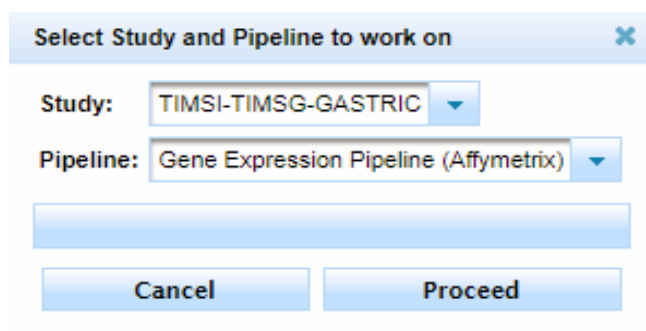
3. Missing visits (where previously existing visit record is not found in new upload),
4. Invalid data formats, and
5. Inconsistent data.

These erroneous records found the Quality report are skipped and not uploaded onto TIMS.

Users can use this as found to correct for the errors and re-upload the corrected copy for storage in TIMS.

4.2 Raw Data Management

Raw Data Management, on the other hand, is specific to the pipeline. This function is only applicable to those pipelines which are editable (please refer to Pipeline Management in Chapter 2.3 for further details). When an account user selects Raw Data Management, the user is required to specify the study and the pipeline that he/she wants to manage (Figure 21).



The screenshot shows a dialog box titled "Select Study and Pipeline to work on". It features two dropdown menus. The first, labeled "Study:", has "TIMSI-TIMSG-GASTRIC" selected. The second, labeled "Pipeline:", has "Gene Expression Pipeline (Affymetrix)" selected. Below these is a large, empty rectangular box. At the bottom of the dialog are two buttons: "Cancel" and "Proceed".

Figure 21: Select Study and Pipeline for Raw Data Management

After selecting the study and the pipeline to manage, the *Raw Data Management* page will be shown (Figure 22). Account user can select the row which corresponds to the raw data that will be managed.

Study: TIMSI-TIMSG-GASTRIC Pipeline: Gene Expression Pipeline (Affymetrix)

Select input data package to work on

Creation Date	Create By	Last Update	Update By	Description
29-Jan-2018 01:52AM	user1			10 Cell Lines

Upload new input data

Description:

+ Select Sample File

+ Select Samples Annotation File

Figure 22: Raw Data Management for Gene Expression Pipeline (Affymetrix) of the Study TIMSI-TIMSG-GASTRIC

For the selected input data, the account user can then change the Description for the data, upload a new Sample File as well as new Samples Annotation File. This will be important particularly for ongoing study in which the number of samples increases from time to time. Samples of the same name are replaced with new samples added as well.

Samples Annotation file is different from the meta-data file. Samples Annotation is pipeline-specific which is used to describe the sample within the pipeline. The samples annotation file (Gene Expression pipeline and GATK Germline Mutation pipeline) provides four columns, i.e.

- First column: Sample ID
- Second column: Filename
- Third column: Subject ID
- Fourth column: Sample Description (or Sample's Type)

Sample ID is the “unique” identifier that defines each filename. All the filename read into the pipeline will be renamed to “Sample ID” and this ID will be used in the QC report.

Subject ID is the identifier that defines each individual or each biological sample. One Subject ID can contain multiple Sample IDs in the case of technical replicates. In the final processed data, the data will be presented as “Subject ID” and this will be the data exported into the database and available for downloading.

The Sample Description column is useful for color-labelling the samples in the QC report. For example, if there are two samples' type, i.e. Tumour and Normal. Therefore, the Tumour samples and Normal samples will be color-labelled accordingly.

Special attention is needed for the samples-annotation file of the GATK Somatic Mutation pipeline. Instead of four columns, there are five columns in the sample annotation file for somatic mutation pipeline. This is due to the fact that for somatic mutation pipeline, there is a need for additional column which signifies the pairing of subjects. Therefore, the samples annotation file for somatic mutation pipeline will be:

First column: Sample ID
Second column: Filename
Third column: Subject ID
Fourth column: Pairing
Fifth column: Sample Description.

The samples description should be labelled as “tumour” and “normal” respectively. The final output will contain only the tumor subjects.

An example of the sample annotation file (tab-delimited) for the somatic mutation calling is shown below:

SampleID	Filename	SubjectID	Pairing	Type
SAMEA2394302	Sample003T.bam	Sample003T	Sample003	Tumour
SAMEA2394303	Sample003N.bam	Sample003N	Sample003	Normal
SAMEA2394304	Sample004T.bam	Sample004T	Sample004	Tumour
SAMEA2394305	Sample004N.bam	Sample004N	Sample004	Normal

Chapter 5: FAQ

Installation

Q: What are the minimum specifications required to run TIMS?

A: i7 processor, 32GB of RAM and at least 500GB of storage space.

Account Management

Q: How do I remove a user account or working unit in TIMS?

A: User accounts as well as working units once created are not removable, neither can they be deleted. Administrators have the option of disabling such inactive accounts or units under the Account Management tab from the Homepage.

Q: The User requires both Director and PI access rights. Do I create two separate accounts for the same user?

A: No. If a Director/HOD also has the position of the PI, create the user account according to Director/HOD access rights. Once the Director/HOD account has been created, assign the PI in-charge of the Group working unit to be said Director/HOD.

Chapter 6: Appendix

Troubleshooting Guide

Installation

Q: No preview available when uploading meta-data (see Chapter 4, Section 4.1.2, Figure 24).

A: Depending on your Linux system, open up Terminal app and run the following lines:

1. For x86_64-linux:
 - export LD_LIBRARY_PATH="\$LD_LIBRARY_PATH:/usr/lib/swi-prolog/lib/x86_64-linux
2. For amd64:
 - export LD_LIBRARY_PATH="\$LD_LIBRARY_PATH:/usr/lib/swi-prolog/lib/amd64

Q: How do I install SMTP server for email notifications?

A: Using Terminal, install mailutils and postfix following the on-screen instructions for configuration.

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.