GUI DP Summary of evaluations of user experience and functionality offered in tranSMART1.1: towards development of a new UI, and identification of additional functionality

Mansoor Saqi [*]

February 26, 2014

Abstract

This document describes tests and evaluations carried out with transmart1.1 hosted at http://public.transmart.etriks.org. The two main areas discussed are (i) the user interface (ii) functionality. We assess the current user interface and categorise the problems that we identify with the Search and Dataset Explorer interfaces. We develop schematic diagrams of proposed new interface features that can be worked up with wireframe models as a basis for implementation[1]. Additionally provision of better explanations on the web pages that guide the user (as distinct from a user manual) could have a big impact on the user experience and we identify areas where the help to the user could be improved. After analysis of some user requirement information we propose a set workflows that could form the basis of a set of 'Core' functionality within the platform. We suggest that additional functionality could be added under a separate menu labelled 'Experimental' which will not have the level of robustness of the core set but offer a way for users to access some additional methods including newer emerging approaches.

1 User Interface

We have tabulated the main issues with the current interface under four sections, namely general, 'Search' and 'Dataset Explorer'. The 'Gene Sig-

[*] with contributions from Nabeel Azam, Antigoni Elefsinioti, Chris Marshall, Cathy Hilton

[1] Schematic diagrams prepared by Nabeel Azam and MS and wireframes by May Yong

Problem	Category
tranSMART could do with a landing/home page. This landing page would outline the different sections of tranSMART, indicating how to use them	Layout
The performance is slow when carrying out a straightforward marker selection exercise - this may be related to how the data matrix is created	Development

Table 1: General Points

nature/Lists' tab is unpopulated and an evaluation of this is not included. For each section we break down the problems into those which are associated with *poor layout of the web page*, those that are associated with *lack of sufficient help on the web page* such as provision of basic examples, explanation of what is happening when analyses are carried out, and explanation of results when they are returned, and finally those that probably require some *code development* to implement. While there are help pages available these are not always clear and do not help a new user who often only wants to quickly get a feel for what the platform can offer. Such potential users are usually unlikely to spend a lot of time reading detailed help pages or user manuals.

1.1 General

User comments such as those below suggest there needs to be clear home/landing page explaining what tranSMART is about.

- the look is Spartan with small text and un-engaging graphics.
- it is not at all clear how I would use the programme to help me with my work
- it doesn‘t really explain what it does, nor how it does it nor help guide me through the task

Table 1 shows some general points noted about tranSMART, relating to the user interactions. One of these can be addressed by a new home page but the other may need more development work.

1.2 Search

A number of the problems encountered when attempting to use the Search page relate to insufficient information being provided to the user. Generally users will want to be able to explore what the software can do without having to look at detailed help. There needs to be help *in the pages* to guide the user through the process and also some examples need to be provided. The problems are summarised in Table 2.

1.3 Dataset Explorer Issues

The problems with the Dataset Explorer could be addressed by a better interface (Layout issues) and better explanations (Help) to guide and inform the user as to the analyses being carried out and as to the meaning of the results that are returned . Refer to Table 3 and Table 4 and Table 5.

1.4 Gene Signature / Lists

This is not populated as is not considered further.

2 tranSMART functionality

The functionality will also depend on the intended user community. It is likely that clinicians and biomedical researches will want to carry out some common (but sometimes complicated) analyses on the data and for this community the platform should offer easy access to workflows that perform these classes of analyses. In the absence of an API, bioinformatics users are more likely to use the platform for extracting subsets of the data and exporting it for use with their preferred analysis tools.

2.1 Common analytic requirements for high dimensional data

Table 6 reveals a diverse set of functionality as being important to users. As resource is limited, a small but robust subset of functionality needs to offered.

2.2 Data types

Currently the main data type used is gene expression data, the main public repositories being GEO and ArrayExpress. The identification of molecular

Problems associated with lack of Help
No information regarding what is capable on this page and why. What exactly is the user searching for? What is being searched against? Add more context around the search parameters and values returned. It would be useful to have some preamble text and example(s)
What do the categories above the search bar mean (all, disease, gene, geo/ebi)? - descriptions (perhaps mouse over) and examples would help. Otherwise the use has to read the help document in detail. Also some indication should be given on this page about what categories are populated: eg Currently there are 22 disease condition terms, 0 compounds, 0 pathways. This saves having to browse
The browse option next to the search button is confusing - no indication of what we're looking at or browsing for. Only Diesease terms are populated; Some explanation of where these identifiers are coming from is needed. Eg how does the user describe a pathway? Which pathway database is being used (Kegg, Reactome?)
Where does the biomarker information come? Has GEO2R been run and has the top set of genes according to fold change been selected? Looking in the help suggests the TEA algorithms is used but does not reveal where the TEA algoriithm has been published
Various views available when results are returned e.g.. Analysis View, Study View, - what are these meant to show? Help is needed with using show filters. Clicking on the help takes the user to a 'getting started with tranSMART' page, not directly to what the tabs actually mean
Need more comprehensive and clear datasets which populate all the different views and options available so it's clear what each of them do
Combination of mouse overs for brief pieces of information and help text for larger - no pop up help windows - keep help in the same window and make them resizable within the same window
Explanations for tab headers

Table 2: **Search function evaluation: Help issues**

Problem associated with GUI design
Stark display on main page - need to make use of the space; Smaller inout boxes; include an example to guide the user
The main page is lacking a scrolling ability for each panel
General improvement on look and feel - colour schemes, fonts and size of components. Doesn't match with the other tabs; Need consistent look and feel to the pages
Advanced workflow tab: hide/grey out services which are not available. Have stable services (robust, tested, documented) and 'Experimental' services under separate headings

Table 3: **Dataset Explorer evaluation: Layout Issues**

Problems associated with lack of Help
Requires help text to explain how everything works and fits together: selection of the 2 cohorts needs exploration of the tree; the use of 'search terms' is not clear;
Advanced workflow tab: Provide descriptions for the analysis services; Identify datasets that can be used to illustrate the different funtionality;
Advanced workflow tab Example: Marker selection - analysis ran some kind of marker selection program and returned statistics - but the methods are hidden, what exactly was run? Show exactly what happened, in their order of steps
Advanced workflow tab: What is the purpose of the High Dimension data box? Pathway cannot be selected. Not clear how to select genes. Provide help about selecting genes
A few lines of explanation of the results returned and how they have been ordered - Gene symbols could be hyperlinked;
Grid view is useful and should be explained (mouse over text)

Table 4: **Dataset Explorer evaluation: Help Issues**

Problems requiring software development
Filtering and searching within the tree not intuitive - need to harmonise the master branches
Summary statistics: Allow for modification or at least show more meaningful titles on the page
Summary statistics: Allow for renaming of headings and key - rather than having Subset 1 and Subset 2 allow the user to specify a meaningful alias
Summary statistics: Hide graphs which have no data
Advanced analysis tab: What is the purpose of the High Dimension data box? Pathway cannot be selected. Delete Pathway if this functionality does not exist or (better) implement Pathway selection
Tie the table with the search - open up the search tab
Perhaps have a wizard on the main page which describes the logical order for steps. [Select cohorts; Generate summary statistics; Search for differentially expressed genes] is a common workflow; If possible, provide the best options based on the data, otherwise keep it standard;
Combine the search with the tree - filter the tree or navigate to a particular leaf/branch based on a search term

Table 5: **Dataset Explorer evaluation: Development Issues**

Functionalities suggested by users	Class of problem
Plot the measurements for a phenotype/biomarker over time (Descriptive analysis) for all the patients	visualisation
Is a given gene able to group preclinical data (e.g cancer cell line panel) to those that respond or not to a given compound (resistance/sensitivity assays) based on whether is mutated or not? If the gene mutation cannot be used as a prognostic biomarker, select patients that are non-responders and compare them with the group of responders. Are there genes that have significant mutations in this group of patients?	combine biomarker with SNP data; overlay of information; mashup
Are those genes in the above example targets of a compound/drug?	link to chemogenomics
Do we have a specific mutation in a given subset of patients? Are those mutations related with published expression signatures?	Integrate SNP data
Select a group of patients/cell-lines. Plot Copy number variations along the whole genome. Investigate peaks in genome	Display CNV information
In a CNV display be able to zoom-in by a) mouse-selecting a region b) giving specific co-ordinates c) giving a specific gene name	visualisation
Select a specific type of cancer from TCGA. List genes ranked by their mutation frequency	
Given a set of genes see if there is any associated biological pathway and/or GO term enriched	Pathway Analysis
View protocol related with a study	
Save criteria that used for generating	
Generate a report stating all analysis steps and results (e.g plots)	

Table 6: Some functionality identified as being important for users

signatures derived from analysis of gene expression data has been the focus of many studies. Other 'omics datatypes in addition to transcriptomics data include proteomics, metabolomics and lipidomics which can provide additional molecular signatures. Although existing studies in the public repositories generally do not contain extensive phenotypic information, increasingly the phenotypic data is becoming multidimensional with a large number of clinical measurements being associated with a patient instead of simply disease and non disease.

Another important data type is data emerging from next generation sequencing technologies. However from the perspective of tranSMART analytic functionality, it is the data derived form these experiments that needs to be considered. This may be, for example RNAseq data or data for copy number variations or SNPs. It is unlikely that the raw data will be stored. However, the data matrix, and associated metadata that describes the processing steps involved in transforming the raw data to the data matrix could be stored. Analytic tools within tranSMART could be applied to the data matrix.

2.3 Gene list approaches

One of the first types of analyses that is usually carried out given transcriptomics data sets together with phenotypic information, is the identification of differentially expressed genes. Typically the sets of differentially expressed genes are then mapped to biological pathways and processes to explore underlying biological themes that may be implicated in the disease condition, often using the Gene Ontology to provide functional description at various levels of resolution. This mapping can be done by identification of overrepresented functional categories (over representation analysis) or identification of changes in functionally associated groups of genes, corresponding for example to known pathways.

2.4 Current Advanced Analysis Functionality

The current Advanced Analysis Menu see Table 7 needs to be grouped by thematic area: we propose a limited set of robust functionality, see Table 8 with perhaps a second menu listing functions still in some stage of development.

Current Analysis Menu Item	Theme
Box Plot with Anova	exploratory statistics
Correlation Analysis	exploratory statistics
Heatmap	is this used only in the context of gene expression analysis?
Hierarchical clustering	is this used only in the context of gene expression analysis?
IC50	needs example data (cannot find any in Dataset Explorer)
Kmeans clustering	is this used only in the context of gene expression analysis?
Line graph	not sure what this plots
Marker selection	gene expression analysis
PCA	exploratory statistics
Scatter plot with Linear regression	these could under basic exploratory statistics
Survival Analysis	
Table with Fisher test	exploratory statistics
Waterfall	generic display option

Table 7: Current Advanced Analysis Menu

Type of Analysis	Includes
Marker selection	Identification of biomarkers with associated pathway analysis
Clustering	kmeans, hierarchical, heatmaps
Statistical Analyses	(with plots)

Table 8: Suggested New Grouping of Current Functionality in Advanced Analysis Menu

New Functionality	Type of approach	Examples of available implementations
Identify differentially expressed genes using multiple methods	Multiple methods implemented	. . .
Identify Pathways from differentially expressed Gene List	Over-representation	GoMiner (standalone)
Identify Pathways from differentially expressed Gene List	Aggregate score	GSEA (available as standalone java program)
Build Co-expression networks	Implement WGCNA procedure	available as R package

Table 9: Short term additional functionality for Biomarker Anaysis for tranSMART: commonly used tools for (a) identifying differentially expressed genes, some already implemented (b) relating functional roles to differentially expressed genes, and (c) for exploring coexpression networks and relating the network modules to phenotypes.

2.5 Proposed new functionality within the *Biomarker Analysis* Theme

Below we suggest proposed additional functionality under the Biomarker Analysis theme (a common and important task in translational medicine) which could be implemented in the short and medium term without major refactoring of the code base. These workflows involve methodologically stratighforward but fairly complicated task, which are non trivial for the non-bioinformatician. Their availability as ‘pre-canned’ functions would make the approaches more accessible to clinicians and biologists. Given the large amount of gene expression data currently available this theme may have impact across several IMI project areas. Refer to Table 9 and Table 10.

2.6 Long term functionality

A feature of the eTRIKS data depository is that it will contain *muti ’omics* data along with clinical datasets. This multi ’omics aspect of current studies (e.g. UBiopred) means that there will be measurements from a variety of

Problem Class	Type of approach	Examples of available implementations
Integrative Network	Network biomarker discovery	PinnacleZ (see [Chuang et al., 2007]) (Cytoscape plugin)
Integrative Network	Map differentially expressed genes to networks	jActiveModules [Ideker et al., 2002] (Cytoscape plugin)
Integrative Network	Differentially expressed subnetworks	KeyPathwayMiner [Alcaraz et al., 2012] (standalone and Cytoscape plugin)
Integrative Network	Network Analysis	BioNet (R-package)

Table 10: Medium term additional functionality for Biomarker Analysis: Some commonly used tools for integrative network analysis - for mapping differentially expressed genes to PPI networks and for network based biomarker identification.

platforms (e.g. gene expression, lipidomics, proteomics, ...). The integration of multiple datasets, and the identification of patterns that emerge only when analysing integrated data will become an important analytic requirement for users of the platform. Multiple dataset integration is an active research field in translational medicine (see for example, [Shen et al., 2009], [Shen et al., 2012], [Yuan et al., 2011], [Kirk et al., 2012]), and it is envisaged that part of the scope of eTRIKS will be the integration and further development of these approaches within the platform. One way to address these advanced problems is to access data in tranSMART programatically and link to external tools.

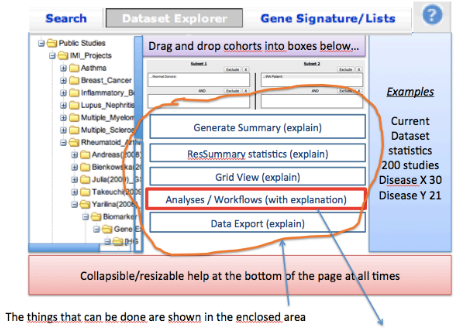
2.7 Visualisation

Integrative network based approaches such as identification of network biomarkers will require the development of network visualisation functionality within tranSMART. This could be achieved by using an external tool such as Cytoscape [Smoot et al., 2011] or Gephi [Bastian et al., 2009].

3 The User Interface

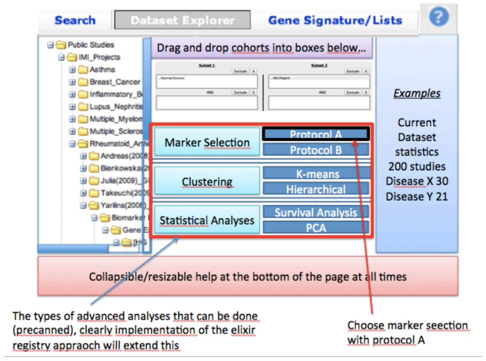
3.1 Initial schematics

Schematics were developed with a focus on guiding the user through the analysis steps, from initial choice of cohorts to the kinds of analysis that could be carried, the data input (choice of method, parameters) and display of results. Figures 1-3 present the initial schematics.



The things that can be done are shown in the enclosed area

Figure 1: Possible analyses are grouped into themes.



The types of advanced analyses that can be done (precanned), clearly implementation of the elixir registry appraoch will extend this

Choose marker seection with protocol A

Figure 2: A given theme is expanded

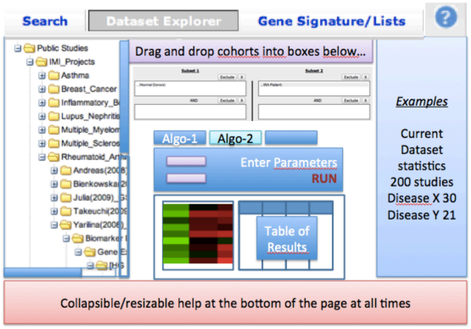
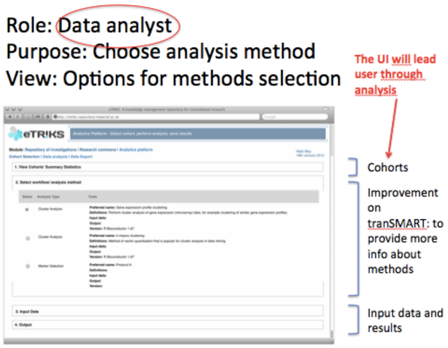


Figure 3: Users are led through an analysis, with cohort selection, choice of analysis method and display of output.

3.2 Wireframe models

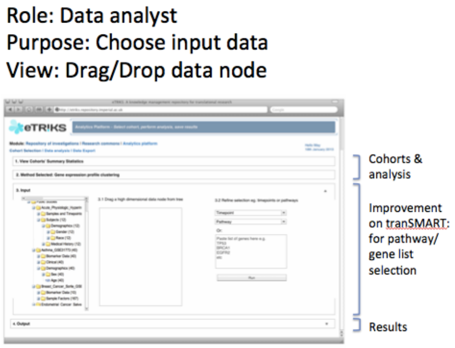
The wireframe was developed (by May Yong), using the above schematics as a basis. It is envisaged that the Analytics UI twill be one component of an eTRIKS model that consists of a data repository, an analytics platform and a research commons. Refer to Figures 4-6.



Role: Data analyst
Purpose: Choose analysis method
View: Options for methods selection

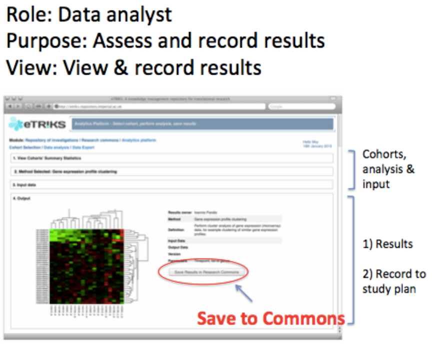
The UI will lead user through analysis
eTRIKS
Cohorts
Improvement on tranSMART: to provide more info about methods
Input data and results

Figure 4: In step 1 cohorts are selected; Then various types of analyses are presented to the user.



Role: Data analyst
Purpose: Choose input data
View: Drag/Drop data node

Figure 5: Users drag and drop data from the data tree and enter other parameters



Role: Data analyst
Purpose: Assess and record results
View: View & record results

Figure 6: The graphical (and other) output is presented to the user and can be saved.

References

[Alcaraz et al., 2012] Alcaraz, N., Friedrich, T., Ktzing, T., Krohmer, A., Mller, J., Pauling, J., and Baumbach, J. (2012). Efficient key pathway mining: combining networks and omics data. *Integr Biol (Camb)*, 4(7):756–764.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

[Chuang et al., 2007] Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140.

[Ideker et al., 2002] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–S240.

[Kirk et al., 2012] Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297.

[Shen et al., 2012] Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., and Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PLoS One*, 7(4):e35236.

[Shen et al., 2009] Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

[Smoot et al., 2011] Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

[Yuan et al., 2011] Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol*, 7(10):e1002227.