TranSMART Workshop

Welcome to this tranSMART training.

The tool tranSMART serves as the data integration platform in CTMM-TraIT, where processed data from different research studies has been made available for view and query. In this training session we will focus on a colorectal cancer demo study: DeCoDe_WP5 (Decrease Colorectal cancer Death_Work Package 5) and only some of the analyses and visualizations possible in tranSMART – feel free to try out more than is described in this document.

Some background for the DeCoDe_WP5 study: patients with colorectal cancer received different types of treatment and of a subset of these patients the chromosomal alterations were determined using comparative genomics hybridization microarray (arrayCGH). Reported data types are, for example, progression-free and overall survival, and copy number alterations. A possible correlation between chromosomal alterations and response to a certain treatment was examined for possible use as a response prediction biomarker. Possible impact: if a patient has a copy number loss for a relevant chromosomal region that has been correlated with worse progression-free or overall survival: don't give them that particular treatment, but choose another treatment!

Thus: for this study clinical data as well as different types of molecular profiling data is available, allowing you to explore the both the data and the possibilities of tranSMART.

More general background video demonstrations can be found at https://www.youtube.com/transartfoundation. The Wiki from the tranSMART Foundation contains a lot of textual background: http://wiki.transmartfoundation.org.

Let's get started! Just follow the steps in this document (and try some variations). If you have any questions, please ask - we are here to help.

tranSMART walkthrough

LOGIN

Please open the Firefox browser on your machine and go to the production server for tranSMART at https://transmart.ctmm-trait.nl.

Please first log in with
Login ID: test*[xx]*
Password: Test*[xx]*test!
Where *[xx]* is replaced by the number you have been assigned, e.g. Test01test!

OVERVIEW

After logging into tranSMART, you automatically land on the Comparison tab of the Analyze functionality tab. On the left under 'Navigate Terms' three yellow folders can be seen. Once a study has been uploaded it will be allocated to one of these folders, depending on the public status of the study. The study we're going through today is in the 'Public Studies' folder: DeCoDe_WP5 demo study. You can view this study by clicking on the + sign in front of this folder.

We will give a little bit of information below on features in tranSMART that can be used.

SEARCH

In the upper left corner there is a search box. Here you can search studies and data related to many different data types. Pressing enter after selecting a term will make the term appear in 'Active filters' slightly below the search box. The studies to which you have access will unfold and show the search term in bold. You can also add more filter terms.

(!) You can explore this feature, but note that you will not find any studies with it currently, since the relevant metadata for it has not been uploaded. However, you can search for items within the studies.

- Explore the different data types you can search on by typing a few letters in the search bar on top. Also expand the drop down-box next to it to see the different types of data to search on.
- Explore the different filters by clicking on the 'Filter' button in the top left 'Active Filters' box.
- Clear all filters afterwards with the 'Clear' button.

ANALYZE

This is the page we are on currently already. From this page the subject data can be accessed for view and query. If you are not on this tab - click on the 'Analyze' tab in the upper right corner. Note the many sub-tabs for Analyze, from Comparison all the way to Genome Browser.
You always start on the left Comparison sub-tab. From there you can work to the right through the other sub-tabs.

DATA TREE

The study data is represented in a tree, showing the different data types that were uploaded. You can expand each folder by clicking on the '+' icon.

It is ordered by Private and Public studies. Under those the little blue booklets () are the studies.

- Find the DeCoDe_WP5 demo study under the 'Public Studies' folder.

There are three different types of leaves for this tree. abc indicates a categorical value, 123 indicates a numerical value and the little helix-icon () indicates 'high dimensional' data. High dimensional data is when multiple observations for a patient for a certain type of experiment are reported, for example NGS or microarray experiments.

- See what data is available for the DeCoDe_WP5 demo study.

EXPLORE THE DATA

Make sure you are on the 'Comparison' sub-tab of Analyze.

Here you always first create the patient subsets (cohorts) that you want to examine further on. This can be either one or two patient subsets, named 'Subset 1' and 'Subset 2'.

- Start by dragging the entire DeCoDe_WP5 demo study into the first box of Subset 1. You have now selected all 489 subjects within the DeCoDe_WP5 demo study.
- Go to the Summary Statistics sub-tab.
- View the graphical breakdown of this study in the graphical output panel.
- Drag the 'Subjects per study' folder (from under **DeCoDe_WP5 demo/1.Clinical characteristics/1. Subjects/Selection subject IDs**) into this output panel.
- See how the subjects are distributed over three studies.

Next we will explore a part of the DeCoDe_WP5 demo study.

- Go back to the Comparisons tab and clear the current subset selection.
- Drag CAIRO Arm A (from under **DeCoDe_WP5 demo/1.Clinical characteristics/1. Subjects/Selection subject IDs/CAIRO subjects**) into the top box Subset 1 and CAIRO Arm B in to the top box of Subset 2.

- Go back to the Summary Statistics sub-tab and see the Age and Sex distribution over the two groups. Note in the boxplot on the right which subset corresponds to which color.
- To see if the age distribution is significantly different between the two subsets, drag 'Age' in the middle panel. You can find it under **1. Clinical characteristics/1. Subjects**.
- The same can be examined for a categorical data item. Drag 'Microsatellite instability' (from under **1. Clinical characteristics/2. Primary tumor/ Microsatellite instability/Protein test/**) into the middle panel. Is there a significant difference between the two groups?

  Challenge: (after doing the walkthrough), determine why you selected this item from 'Protein test' and not 'DNA test'.

Try doing the same for some of the folders available in the folder DeCoDe_WP5 demo study/4. Molecular profiling.

Note: you can also create a subset by using a numerical variable. Simply drag the numerical node of interest into the box underneath Subset. In the window that pops up, select for example, 'By numerical value' and enter your criteria of interest.

GRID VIEW

To see your data in a tabular format, go to the Grid View sub-tab.

- Each row is a subject, each column a concept. (You can ignore the Subject ID as this is a specific tranSMART ID).
  - Note that all the items you have previously selected in the summary statistics sub-tab are already filled in in the grid view. Now, it can be traced which subject exactly belongs in which study arm (CAIRO Arm A and CAIRO Arm B).
- Drag the 'Histological type' folder (from under **1. Clinical characteristics/2. Primary tumor**) into the Grid View. Observe the values for some subjects.
- If you want to remove some columns, hover over the column headers until you see an arrow appear. Press it and deselect the columns you don't want.
- If you like you can export this view:
  - Select some rows (by holding the Shift key down when clicking) or no rows
  - Click the Export to Excel button below.
  - You can open this file in Excel.

That is one way of examining the data, now let's go and try some analyses.

Go back to the Comparison sub-tab and clear the Subset 1 and 2 fields.

- Select the entire DeCoDe_WP5 demo study and press summary statistics to successfully load the study.

ADVANCED WORKFLOW

Under the Advanced Workflow tab, you can use multiple analyses to analyze the data from the selected subset previously created in the Comparisons tab. Below are some examples to follow when trying some of the analyses.

SURVIVAL ANALYSIS OF PATIENTS WITH DIFFERENT HISTOLOGICAL TUMOR TYPES

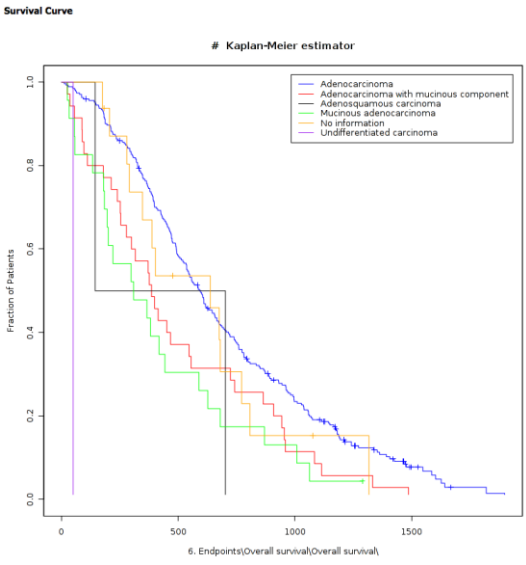
Here we would like to investigate the overall survival status of subjects with different histological tumor types.

- After selecting the Advanced Workflow sub-tab, click on the Analysis button on the top left and select the Survival Analysis.
- Three data types are going to be entered in this analysis:

Time: Overall survival (found under **1. Clinical characteristics/6. Endpoints/Overall survival**)
Category: Histological type (found under **1. Clinical characteristics/2. Primary tumor**)
Censoring variable*: Alive (found under **1. Clinical characteristics/6. Endpoints/Overall survival/Overall survival: event**)

> **Note: censoring in a survival analysis is **not** the same as excluding of subjects! Censored subjects are not removed from the analysis.*
> *→ If you want to exclude subjects of which there is 'No information' related to Overall survival: alive or dead, go back to the comparison tab, drag the 'No information' into the second box below in the comparison tab and change the button 'Include' to 'Exclude', press 'Summary statistics' again and repeat the analysis.*

Press run and view the Kaplan-Meier survival curves. The result should look like this:



It is possible to do survival analysis also on **a very particular subset**, by selecting multiple and/or variables in the comparison tab. For example, if we want to examine a difference for overall survival between CAIRO2 subjects who have either BRAF wildtype (WT) or mutation (MUT) status, subjects who have been reported to be microsatellite stable (MSS) with T3 tumor staging, clear the subset 1 box and make the following selection:

- Drag CAIRO2 into the top box of Subset 1 (found under **1. Clinical characteristics/1. Subjects/Selection subject IDs/Subjects per study**).
- Drag 'MSS' in the box underneath (AND) (found under **1. Clinical characteristics/2. Primary tumor/Microsatellite instability/Protein test).**
- Drag 'T3' in the box underneath (AND) (found under **1. Clinical characteristics\2. Primary tumor/Staging/T stage/).**
- Drag in the box below 'No information' for BRAF status (found under **4. Molecular profiling/DNA small nucleotide variants/HRM Sanger seq: BRAF/General call)** and press 'Exclude'**.**

Press summary statistics and perform the survival analysis on:

Time: Overall survival (should still be loaded in the analysis, otherwise re-enter the data node)
Category: BRAF MUT vs BRAF WT (that's why we excluded the BRAF No information) (press the clear button to remove any previous data still in there)
Censoring variable: Alive (should still be loaded in the analysis, otherwise re-enter the data node)

Examine the results. Repeat with some other variables of interest. Note that is possible to make OR selections in the comparison tab by dragging different items into the same subset box.

PREVALENCE OF KRAS AND BRAF MUTATIONS, FISHER EXACT TEST

We will examine whether the mutation status of genes KRAS and BRAF in the CAIRO2 study seems to co-occur, or are mutually exclusive.

- Drag CAIRO2 into the top box of Subset 1 (found under **1. Clinical characteristics/1. Subjects/Selection subject IDs/Subjects per study**).
- On the Advanced Workflow sub-tab, select the Analysis 'Table with Fisher test'.
- Drag the items MUT and WT for BRAF into the Independent Variable box (found under **4. Molecular profiling/DNA small nucleotide variants/HRM Sanger seq: BRAF/General call**).
- Drag the items MUT and WT for KRAS into the Dependent Variable box (found under **4. Molecular profiling/DNA small nucleotide variants/HRM Sanger seq: KRAS/General call**)
- Click Run.
- Is there a significant difference in the prevalence of mutation status?

EXPRESSION DIFFERENCE OF AURKA OVER DIFFERENT HISTOLOGICAL TYPES

How would the expression of AURKA differ between different histological types?

- Select the entire DeCoDe_WP5 demo study in Subset 1.
  Note that only the TCGA validation set has high-dimensional gene expression data, so only for these patients we will later see AURKA expression values.
- Go to the Advanced Workflow sub-tab and select the Box Plot with ANOVA.
- Select all Histological types (found under **1. Clinical characteristics/2. Primary tumor/**) as the Independent Variable. Remove the 'No information' node in the box by using right-mouse click and selecting 'Delete'.
- Select the mRNA gene expression node 'Genes' from as the Dependent Variable (found under **4. Molecular profiling/RNA expression/Microarray: Agilent 44K mRNA**).
- Next we will derive one numerical value for our Boxplot from this set of gene expression by selecting for only the expression values of one gene.
- Click the High Dimensional Data button.
- Start typing AURKA in the 'Select a **Gene/Pathway/mirID/UniProtID'** box and click on the top suggestion matching this name when it pops up.
- Click Apply Selections and after that Run. Is there a significant difference in expression between different histological types? What else do you notice?

DIFFERENCES IN CHROMOSOMAL ALTERATION PATTERNS BETWEEN MSI VS MSS TUMORS

There is chromosomal copy number alteration data available in this study as well. Let's investigate differences in alteration patterns between subjects with MSI and MSS.

- Select the entire DeCoDe_WP5 study in Subset 1.
- Go to the Advanced Workflow tab and select the Frequency Plot for aCGH analysis. If you hover over the shown boxes more information is provided on what to drag into the analysis
- Drag the 'Regions' node in the Array CGH box (found under **4. Molecular profiling/DNA somatic copy number alterations/Microarray: Agilent 180K aCGH**).
- Select MSI and MSS from in the Group box (found under found under **1. Clinical characteristics/2. Primary tumor/Microsatellite instability/Protein test)**.
- Run Analysis.
- Compare the two frequency plots. Does it seem like there is a difference in chromosomal alterations between MSI and MSS tumors?

Let's spoil the answer, YES, they look different!

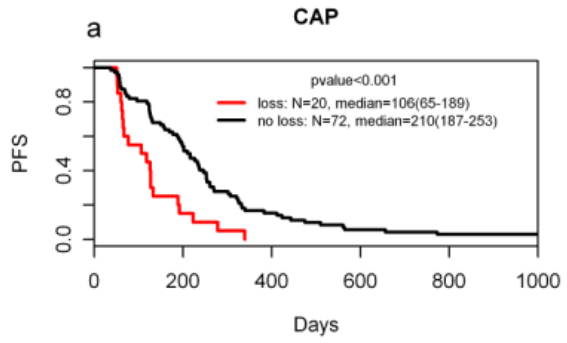
Let's verify whether this difference is actually significant. (well, not exactly during the demo)

- Select the 'Group Test for aCGH' analysis.
  IMPORTANT: Please set the Permutations variable to 5 now. Otherwise you could let the server perform a very heavy job, which would hamper the hands-on. For actual analysis to determine significant correlations the number of permutations should be set to 10000, in which case the analysis can be sent to the background and retrieved after a couple of hours.
- Select the Regions node from above in the Regions box.

- Select the same groups as above.
- Use Chi-square as the statistical test and compare the two groups on the Alteration Type LOSS.
- Make sure you have set Permutations to 5.
- Run Analysis.
- Order the table by p-value (click on the header). Do you find significant differences? *Note that the p-values are very unrealistic due to the low number of permutations.*

Challenge: try and reproduce the figure below using the Group test for aCGH analysis.
→ This data has been reported in a Nature Communications paper of 2014 (right-click either the top study folder and find the see the link to the study, or select the folder containing the actual arrayCGH data to see only the relevant publication for this data type).

Supervised analysis of DNA associations with drug response was reported to show significant correlations for one chromosomal region for subjects of whom progression free survival was reported for the first therapy treatment line:



*Supp Fig 4. Chromosomal loss of **5q12-1-5q12.3** and PFS1.*

<u>Don't worry, you won't have to read through the article and figure out what to select.</u>

Selection described in article, try and select the correct items in the Comparison tab

- Copy number profiled patients
- Excluding patients with MSI
- Must have had 2 or more treatment cycles in the therapy line 1 in CAIRO arm A (hint: this arm is indicated with CAP)
- Progression free survival for therapy line 1 was examined.

EXPRESSION OF CRC RELATED TUMOR SUPPRESSOR GENES

We will investigate the gene expression of a list of genes, classified by the NIH 'National Cancer Institute' as 'Genes Associated with a High Susceptibility of Colorectal Cancer'. You can find this list in Table 2 of
http://www.cancer.gov/cancertopics/pdq/genetics/colorectal/HealthProfessional/page2.
This list has been stored already in tranSMART as '**Mijn test lijst'**.
In the DeCoDe_WP5 demo study there is only mRNA gene expression data available for the TCGA validation set.

First we will look at the expression of one gene in the entire TCGA validation set.
Go to Comparison sub-tab, empty both subsets and drag the node 'TCGA validation set' into subset 1 (found under **1. Clinical characteristics/1. Subjects/Selection subject IDs/Subjects per study**).

- Go to the Advanced Workflow sub-tab and select the analysis Heatmap.
- Drag the Genes node from the first step node in to the box.
- Click the 'High Dimensional Data' button.
- Start typing AURKA in the 'Select a **Gene/Pathway/mirID/UniProtID'** box and click on the suggestion when it pops up.
- Click Apply Selections and after that Run.
- Analyze the expression of AURKA in this set.

Next we will look at the expression of the Tumor Suppressor genes in this entire TCGA set.

- Click the High Dimensional button again and now start typing '**Mijn test lijst'**. Click on the suggestion when it pops up.
- Apply Selections and Run.
- Analyze the expression of these genes.

Now let's see if we can find two interesting subsets to compare this gene expression between.

- Go to the Summary Statistics sub-tab.
- Drag in some folders from under **1. Clinical characteristics/2. Primary tumor** in the middle panel, like Site, T stage or Stage grouping AJCC. See if you can find two interesting subsets to compare.
- Go to the Comparisons sub-tab and create the two interesting subsets.
- Go to the Advanced Workflow sub-tab and select the Hierarchical Clustering analysis (this is just a heatmap analysis with added hierarchical clustering).
- Drag the Genes node in the box, click High Dimensional Data and select '**Mijn test lijst'**.
- Apply Selections and Run.
- The two subsets are colored orange and yellow. Did your two subgroups form separate clusters?