Overview

- tranSMART is a great platform for data sharing and cross-analyses – we wanted to see if various data types available publicly can help researchers find an interesting hypothesis

- In addition, we wanted to “push tranSMART limits” – and decided to use NGS, clinical and preclinical data
Data description

- A slice of TCGA data (colon and HNSC cancers, about 1000 patients): all clinical, DNAseq, RNAseq data
  - RNAseq data was transformed to produce FPKM data, RNAseq somatic mutations and RNAseq mutations
  - DNAseq data was transformed to produce SNP data
- Several time-series GEO studies
- Proprietary preclinical dataset (drug response data)
- Proprietary eQTL data for several tissues
Data processing

- TCGA data was processed from raw data using OMICSoft by an experienced bioinformatician.
- The resulting flat text files (1 – 7 GB in size) were transformed using custom scripts and formatted to satisfy tranSMART’s ETL.
  - Examples: add variant annotations to a list of variants; generate VCF files; data pivot.
- Resulting data files were loaded into tranSMART using scripts with “load speed acceleration”.
- For preclinical dataset, migration was performed from version 1.1 to 1.2.
- For eQTL data, we followed the template provided by Pfizer in github (“MAGIC” study).
tranSMART installation

- Version 1.2.2
- Installation was done on two Linux (RedHat) servers: one for web application (tomcat), computing tasks (Rserve) and content indexing (SOLR), and a second server dedicated to Oracle 11g database
Custom R script integration

- Custom R script “consumes” specific data in specific format (MySQL tables)
- The idea of tranSMART is to use uniform data model that can be consumed by R scripts
- Often it means that both data model and R script need to be tweaked
Custom R Script Integration

- RBS was provided with R scripts that generate various plots from in vivo oncology study data
- R Scripts directly query original MySQL database
- Converted MySQL database to Oracle schema
- Converted original R script to work with Oracle syntax
- Converted R script IO interfaces to be compatible with tranSMART standards
- Installed R scripts as tranSMART module
Functionality

- Our data required that all functionality in tranSMART be operational
  - Genome browser for viewing SNP and variant data
  - GWAVA for viewing eQTL data
  - All standard workflows
  - Line graph for time series
Results
Results

We have fine-tuned the Workflows
And also Genome Browser:
Now can have any number Of NGS data files per set
Results

Viewing eQTL results using GWAVA: compare MAGIC set with proprietary set
Example of joint data analysis

- EGFR is known to play a significant role in colorectal cancer (e.g., Markman et al in Adv Clin Chem. 2010;51:71-119)

- In COAD set from TCGA, the status of EGFR was not measured, but instead, is available for HNSC cancer for a limited set of patients.

- We created a gene signature on the fly using HNSC samples (blood) and used the signature to investigate the COAD set.

- Preliminary analysis indicates that indeed a subset of patients in COAD set may have the EGFR signature amplified.
Step 1: create EGFR signature using small subpopulation from HNSC

- List of 200 genes
- Copied, cleaned,
- And upregulated genes
- Are saved as a list in tM
Step 2: use the upregulated genes from EGFR-amplified samples to investigate COAD dataset.
Result: clustering shows preliminary pattern that warrants further investigation
Lessons learned

- Data preparation: need to know the format compatible with ETL
- GWAS: we worked with Pfizer team (*thank you!!*) to obtain the most recent scripts – which are now in GitHub; version control is a must!
- Pain with Genome browser: hard to launch from under the firewall – needed to resolve with internal IT team
- Pain with **connecting workflows** – for example, when one finds Markers there is no way to convert them to gene signature on the fly or even transport them to Genome browser. This is a general issue which when addressed would **really help increase usability**
Thank you

- Merck
  - Meiping Chang
  - Scott Thomas
  - Xing Chen
- Pfizer
  - Jay
  - Haiyan
  - Hugo
- Rancho
  - Tania
  - Alex
  - Leonid
Contact Information

- Dr Tatiana Khasanova-Director Services
  - Tania@RanchoBioSciences.com
  - Tel: 760-390-6428

- Dr Jon Greene-Informatics Specialist
  - Jon@RanchoBioSciences.com
  - Tel: 973-704-8570

- Alex Azarov-IT
  - Alex@RanchoBioSciences.com

- Julie Bryant-Business Development
  - Julie@RanchoBioSciences.com
  - Tel: 760-717-7881