Integrated Analysis and Interactive Exploration with Regulome Explorer

To gain greater insight into the development and progression of prostate adenocarcinoma, we have integrated all of the data types produced by TCGA and described in this paper into a single "feature matrix". From this single heterogeneous dataset, significant pairwise associations have been inferred using statistical analysis and can be visually explored in a genomic context using Regulome Explorer, an interactive web application (http://explorer.cancerregulome.org). In addition to associations that are inferred directly from the TCGA data, additional sources of information and tools are integrated into the visualization for more extensive exploration (e.g., NCBI Gene, miRBase, the UCSC Genome Browser, etc).

Feature Matrix Construction

A feature matrix was constructed using all available clinical, sample, and molecular data for 333 unique patient/tumor samples. The clinical information includes features such as age and tumor size; while the sample information includes features derived from molecular data such as single-platform cluster assignments. The molecular data includes mRNA and microRNA expression levels (Illumina HiSeq data), protein levels (RPPA data), copy number alterations (derived from segmented Affymetrix SNP data as well as GISTIC regions of interest and arm-level values), DNA methylation levels (Illumina Infinium Methylation 450k array), and somatic mutations. For mRNA expression data, gene level RPKM values from RNA-seq were log2 transformed, and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). For miRNA expression data, the summed and normalized microRNA quantification files were log2 transformed, and filtered to remove low-variability microRNAs (bottom 25% removed, based on interdecile range). For methylation data, probes were filtered to remove the bottom 25% based on interdecile range. For somatic mutations, several binary mutation features indicating the presence or absence of a mutation in each sample were generated. Mutation types considered were synonymous, missense, nonsense and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains.

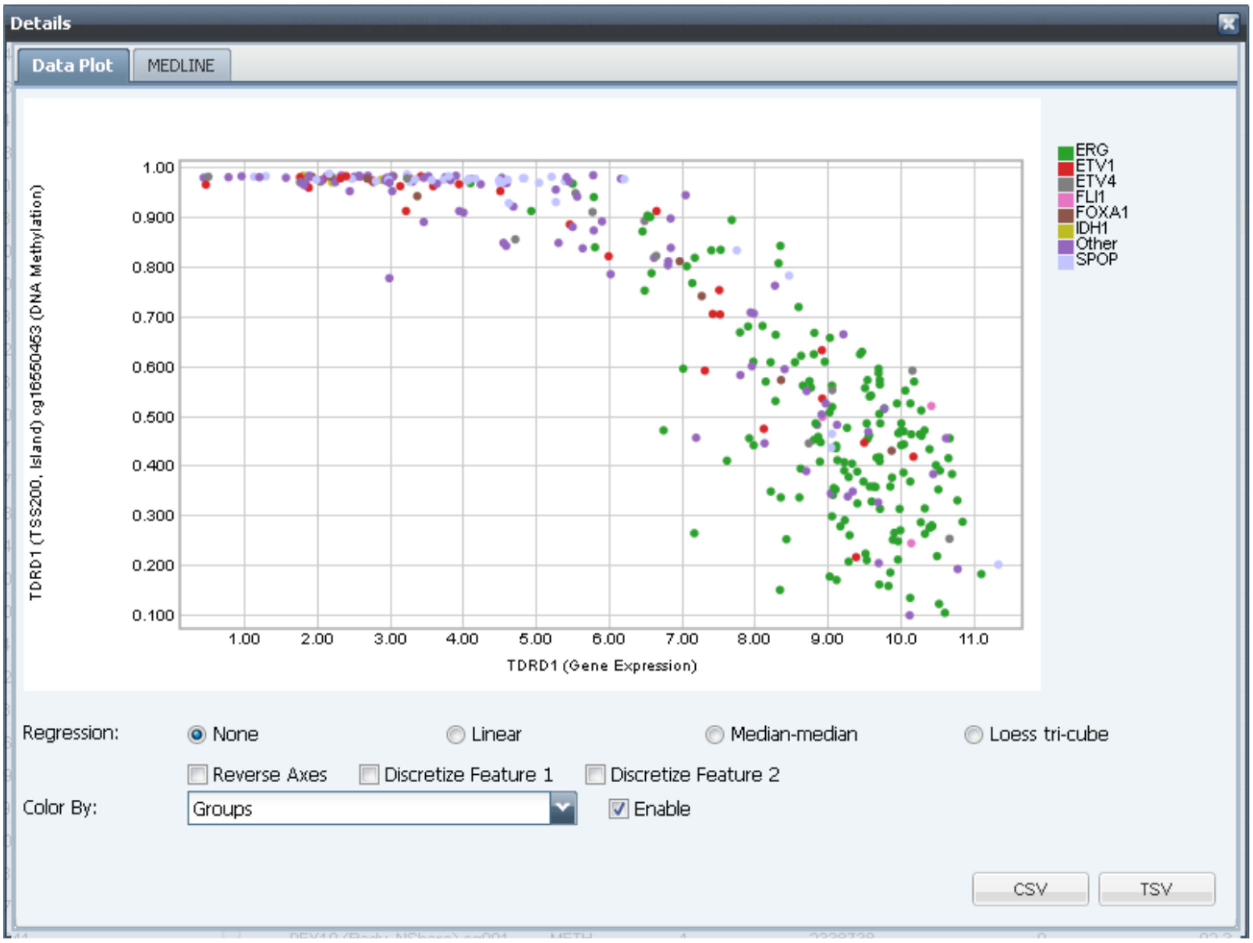
Pairwise Statistical Significance

Statistical association among the diverse data types in this study was evaluated by comparing pairs of features in the feature matrix. Hypothesis testing was performed by testing against null models for absence of association, yielding a *p*-value. *P*-values for the association between and among clinical and molecular data types were computed according to the nature of the data levels for each pair: categorical vs. categorical (Chi-square test or Fisher's exact test in the case of a 2x2 table); categorical vs. continuous (Kruskal-Wallis test) or continuous vs. continuous (probability of a given Spearman correlation value). Ranked data values were used in each case. To account for multiple-testing bias, the *p*-value was adjusted using the Bonferroni correction.

Exploring significant associations between features

Regulome Explorer allows the user to interactively explore significant associations between various types of features – associations between molecular features, associations between molecular features and derived numeric features (like AR score), and associations between molecular features and categorical features such as clinical features or clusters derived from prior analysis (like iCluster). The examples below are screenshots from Regulome Explorer which illustrate exploration of the TCGA prostate cancer data.

Figure 1: TDRD1 gene expression and DNA methylation colored by Group membership

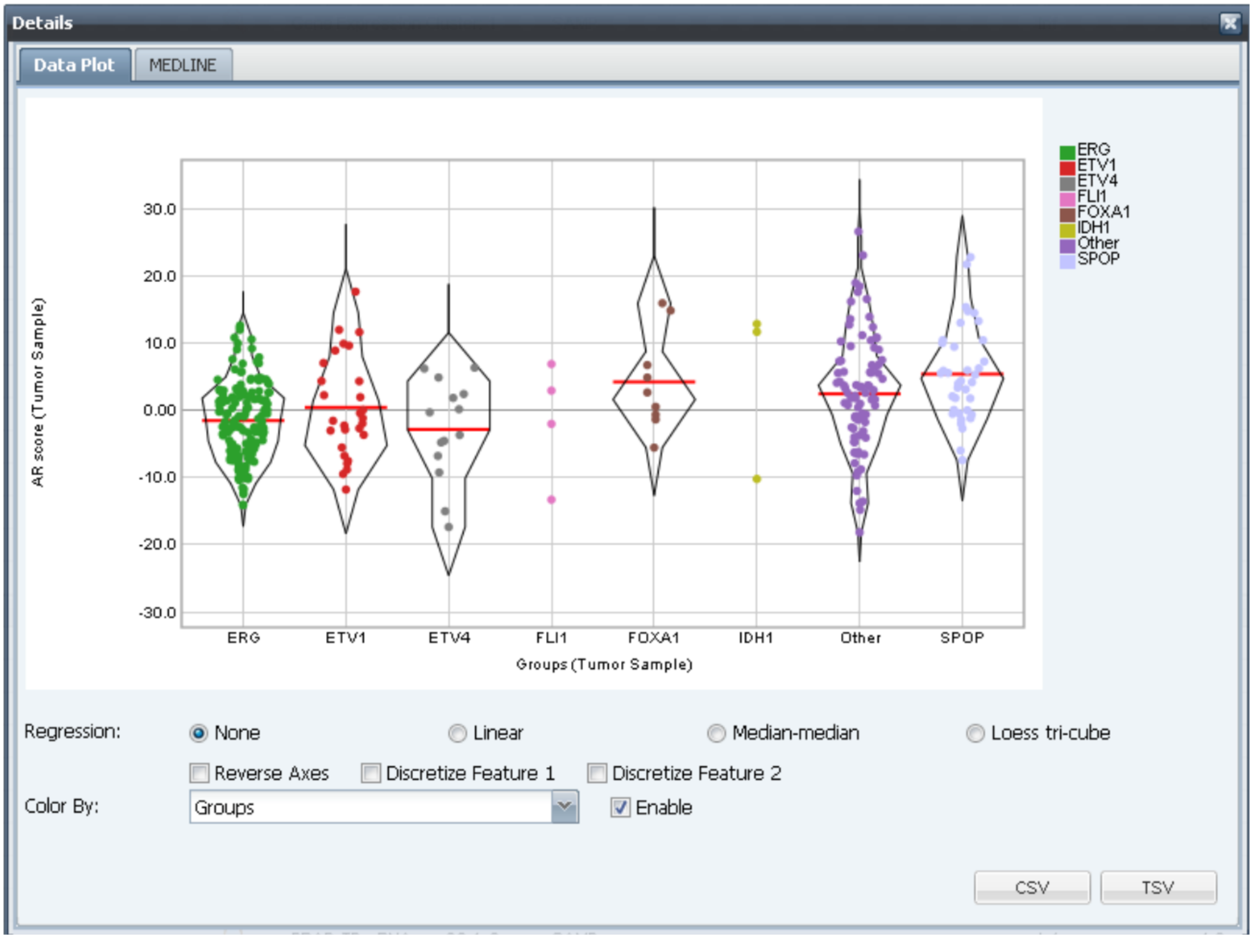


Regression: ● None ○ Linear ○ Median-median ○ Loess tri-cube

□ Reverse Axes □ Discretize Feature 1 □ Discretize Feature 2

Color By: Groups ☑ Enable

Figure 2: AR Score across the Groups

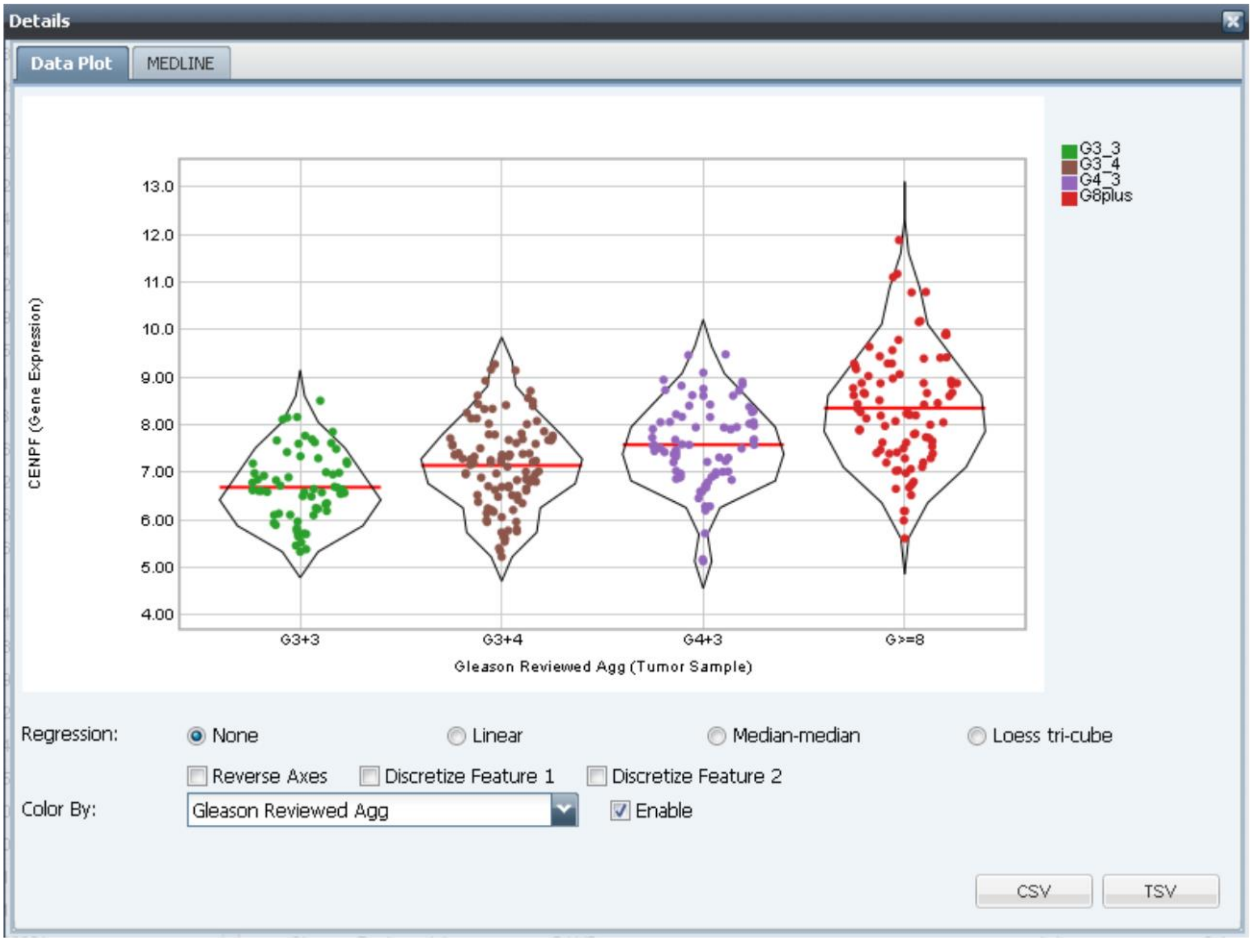


Regression: ● None ○ Linear ○ Median-median ○ Loess tri-cube

□ Reverse Axes □ Discretize Feature 1 □ Discretize Feature 2

Color By: Groups ☑ Enable

Figure 3: Association between CENPF gene expression and reviewed Gleason



Regression: ● None ○ Linear ○ Median-median ○ Loess tri-cube

□ Reverse Axes □ Discretize Feature 1 □ Discretize Feature 2

Color By: Gleason Reviewed Agg ☑ Enable