Robust statistical methods: applications in genomics

April 4 2018

Overview

- Applications of the outlier concept: IvyGlimpse, DriverNet
- Critique of simplistic outlier labeling; alternatives
  - robustness vs. efficiency tradeoff
  - inward peeling, outward testing: formal univariate and multivariate outlier identification
- Relating tumor expression profiles, multivariate outlier identification, and mutation patterns in TCGA-GBM

Example 1: Image analysis in glioblastoma

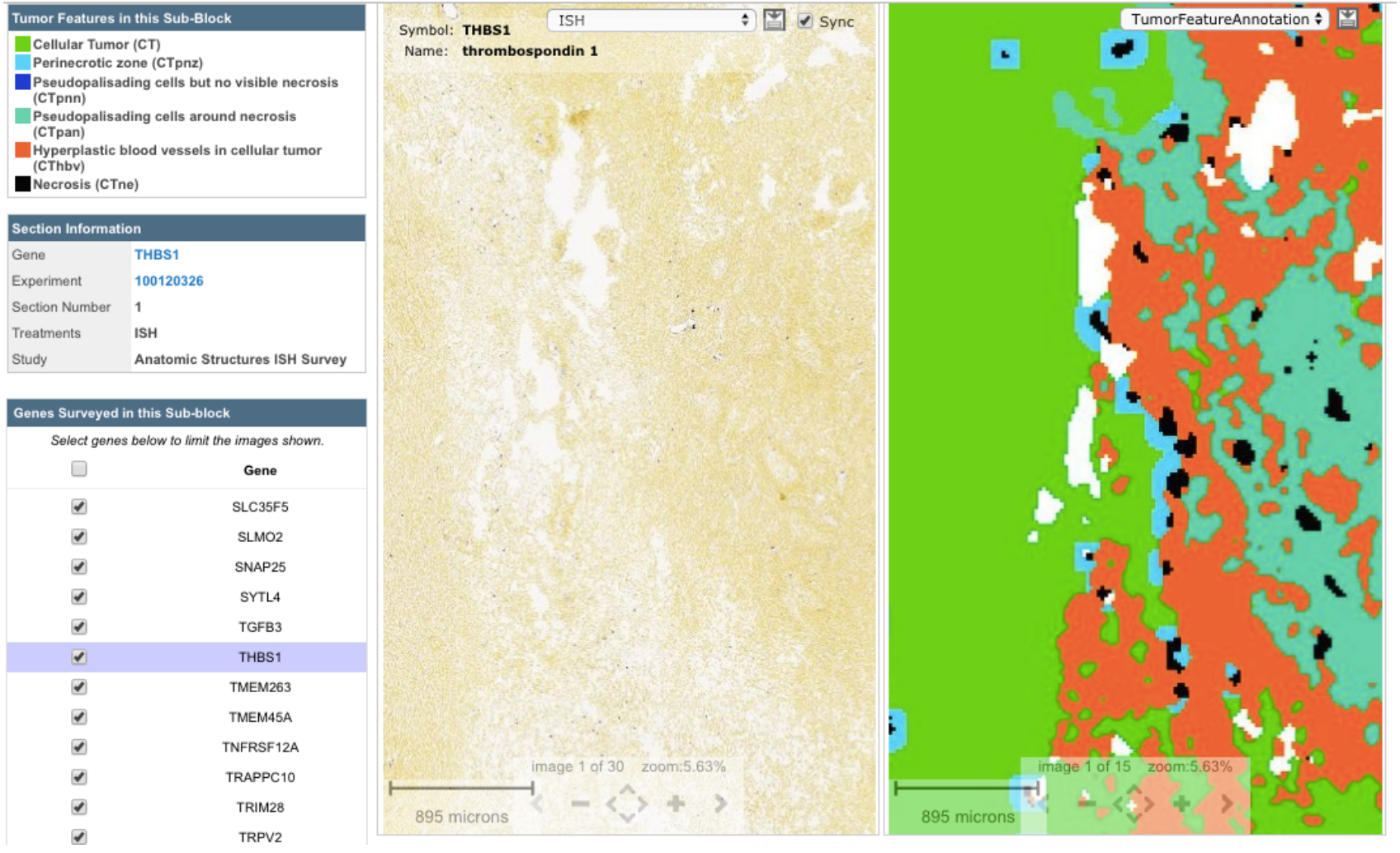


Figure 1: Annotated IvyGAP image

A cohort of image annotations, grab from the tail

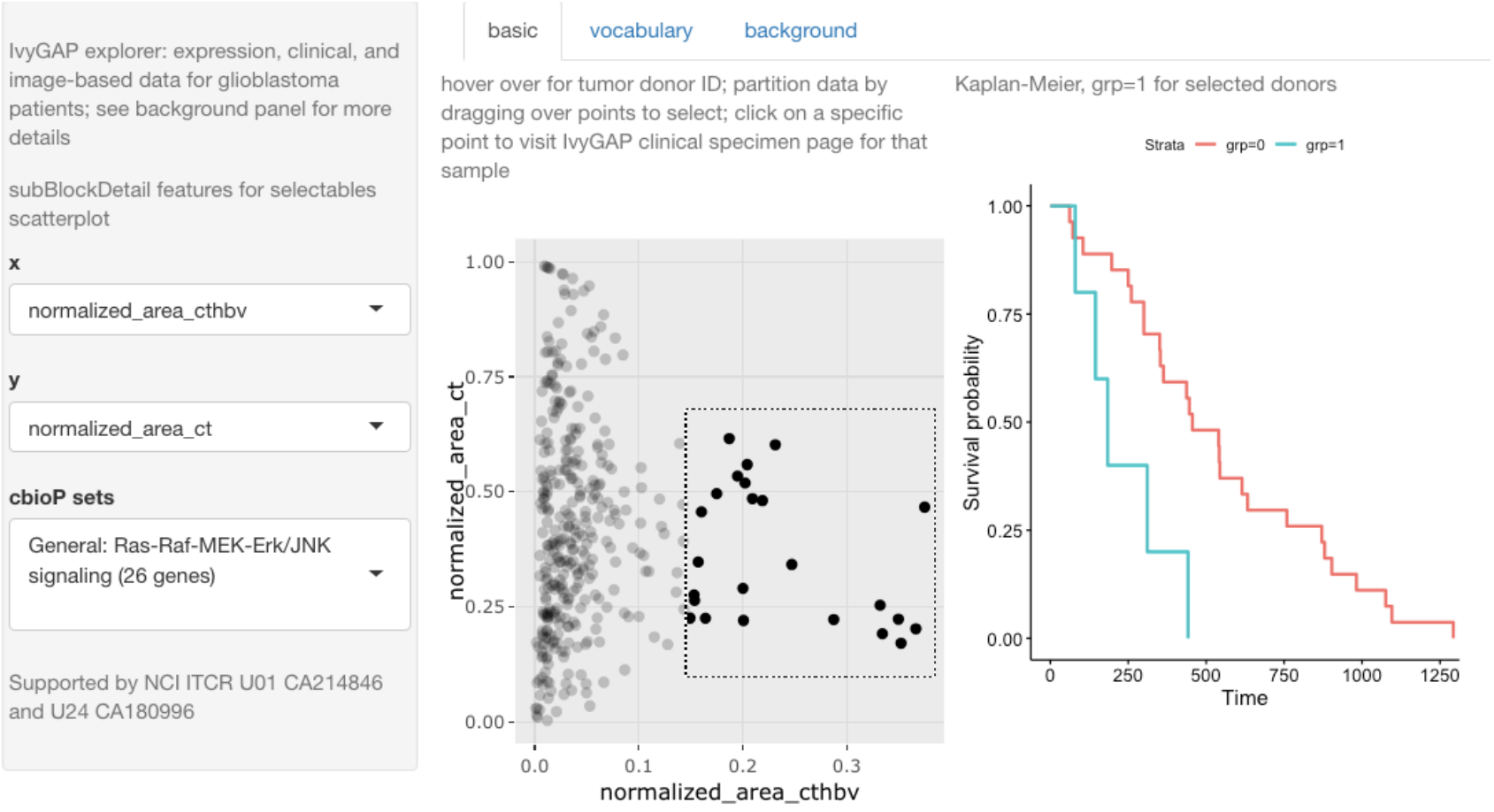


Figure 2: ivygapSE::ivyGlimpse()

Expression patterns for included/excluded

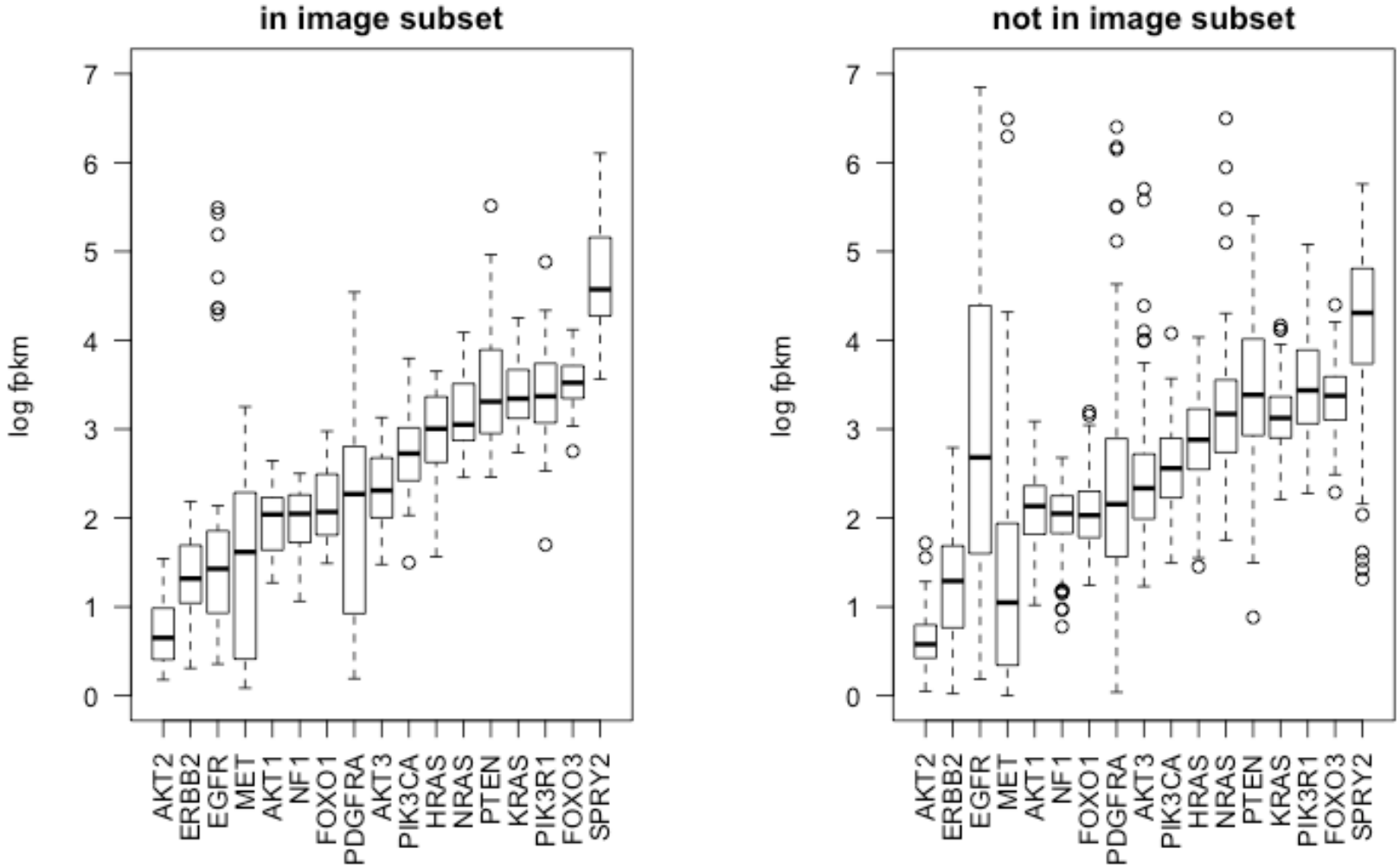
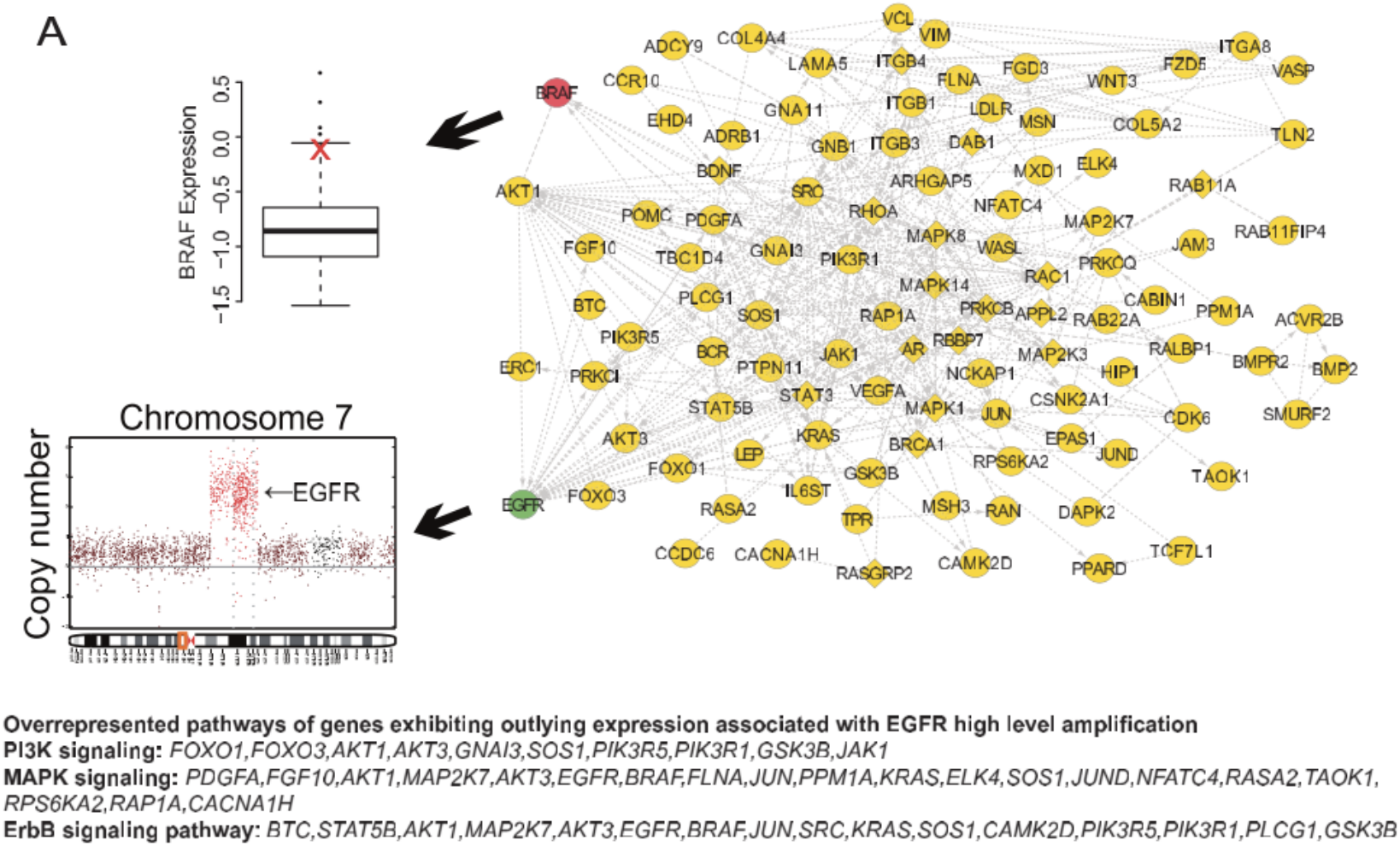


Figure 3: ivyGlimpse univariate

Upshots

- Annotation of tumor images is an expensive process, and use of annotated features for prediction is highly desirable
- In the scatterplot of (normalized areas) CT vs CTHBV, we see that CTHBV scores have a skewed distribution
- Individuals contributing images in the tail of this distribution seem to have relatively shorter survival times
- Marginal distributions of expression for genes are summarized using boxplots
- Careful analysis of image features, expression patterns, and their clinical correlates must reflect complexities of design; see the IvyGAP white paper

Example 2: DriverNet (from the lab of Sohrab Shah)



Overrepresented pathways of genes exhibiting outlying expression associated with EGFR high level amplification
PI3K signaling: *FOXO1,FOXO3,AKT1,AKT3,GNAI3,SOS1,PIK3R5,PIK3R1,GSK3B,JAK1*
MAPK signaling: *PDGFA,FGF10,AKT1,MAP2K7,AKT3,EGFR,BRAF,FLNA,JUN,PPM1A,KRAS,ELK4,SOS1,JUND,NFATC4,RASA2,TAOK1, RPS6KA2,RAP1A,CACNA1H*
ErbB signaling pathway: *BTC,STAT5B,AKT1,MAP2K7,AKT3,EGFR,BRAF,JUN,SRC,KRAS,SOS1,CAMK2D,PIK3R5,PIK3R1,PLCG1,GSK3B*

Figure 4: Aberrant expression of BRAF for patient with EGFR amplification

EGFR mutation associated with extreme expression values in diverse genes

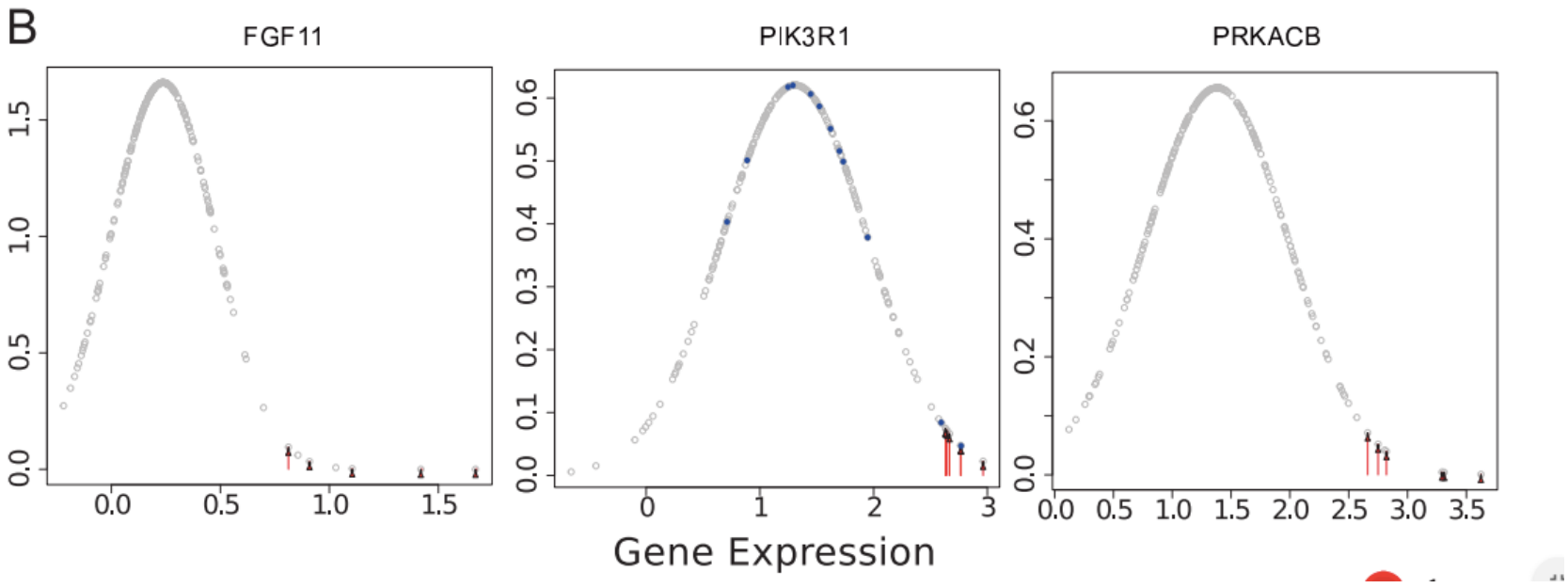


Figure 5: Blue dots: individual has mutation in named gene; red arrow: individual has mutation in EGFR

Overall DriverNet schema

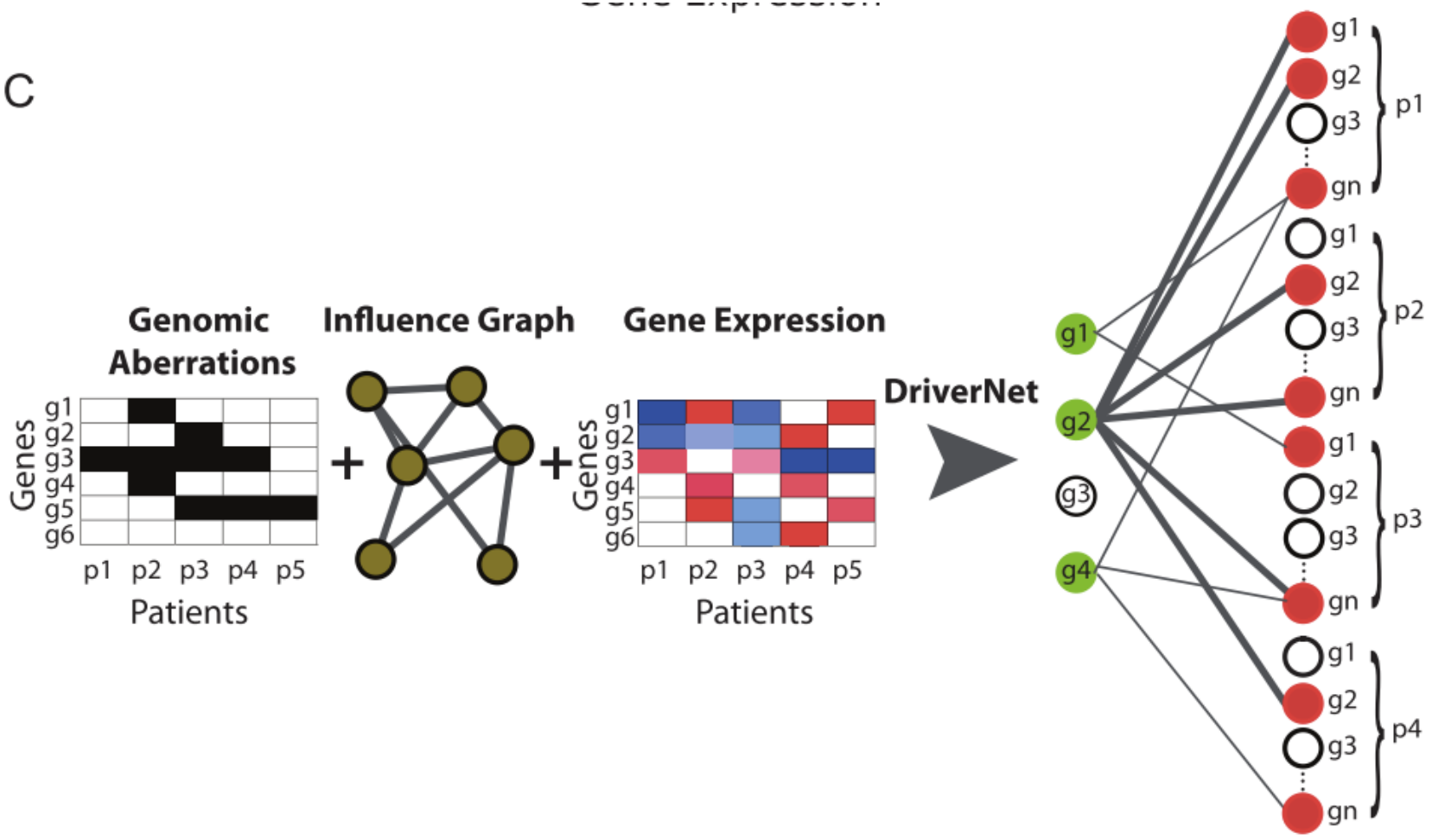


Figure 6: Towards the DriverNet bipartite graph. Minimize the number of mutation-bearing nodes (on left) needed to identify genes with aberrant expression on right.

Outliers for DriverNet

The gene expression measured from microarray or RNAseq produces the patient-expression matrix G. $G(i, j)$ represents the relative abundance of mRNA levels for gene i in patient j. For each gene, we assume the expressions across all the patients are normally distributed. Based on this assumption, the patient-expression matrix G is converted to a binary patient-outlier matrix G' where $G'(i, j) = 1$ means the expression of gene i is an outlier in patient j. The outliers for gene i are defined as those whose values are outside the two-standard deviation range of the expression values of gene i across all the patients.

Figure 7: Supplemental Material definition.

Patient-outlier code: default threshold is 2 SD from mean

```
> library(DriverNet)
> getPatientOutlierMatrix
function (patExpMatrix, th = 2)
{
    expSd <- apply(patExpMatrix, 2, sd)
    id <- expSd > 0
    expSd <- expSd[id]
    patExpMatrix <- patExpMatrix[, id]
    expMean <- apply(patExpMatrix, 2, mean)
    num <- dnorm(x = t(patExpMatrix), mean = expMean, sd = expSd,
        log = T)
    num <- t(num)
    numSd <- dnorm(x = t(expMean + th * expSd), mean = expMean,
        sd = expSd, log = T)
    y <- rep(numSd, each = dim(patExpMatrix)[1])
    y <- matrix(y, nrow = dim(patExpMatrix)[1], ncol = dim(patExpMatrix)[2])
    patOutMatrix <- num <= y
    id <- colSums(patOutMatrix) > 0
    patOutMatrix <- patOutMatrix[, id]
    return(patOutMatrix)
}
```

Figure 8: getPatientOutlierMatrix code.

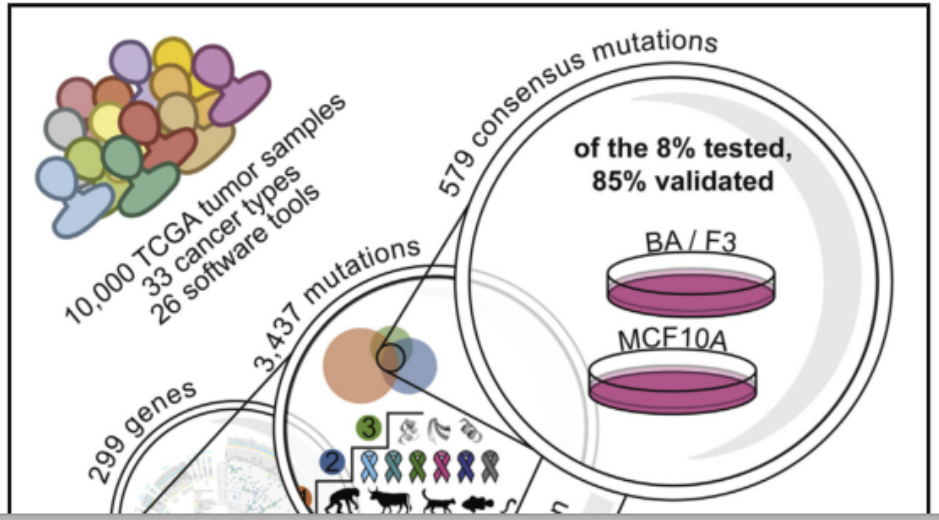
Comments

- DriverNet algorithm used in the most recent catalog of driver mutations in cancer

Cell

Comprehensive Characterization of Cancer Driver Genes and Mutations

Graphical Abstract



Authors

Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, ..., Gordon B. Mills, Rachel Karchin, Li Ding

Correspondence

karchin@jhu.edu (R.K.), lding@wustl.edu (L.D.)

In Brief

A comprehensive analysis of oncogenic driver genes and mutations in >9,000 tumors across 33 cancer types highlights

Additional algorithms

DriverNET

DriverNet (Bashashati et al., 2012) is a package to predict functional important driver genes in cancer by integrating genome d (non-synonymous SNVs, indels, and copy number alteration) and transcriptome data (gene expression data). The different data typ are integrated using an influence graph (Wu et al., 2010). We ran DriverNet (v1.6.0, numberOfRandomTests = 500, weight = FALS purturbGraph = FALSE, purturbData = TRUE) and genes with q-value of 0.05 were deemed significant.

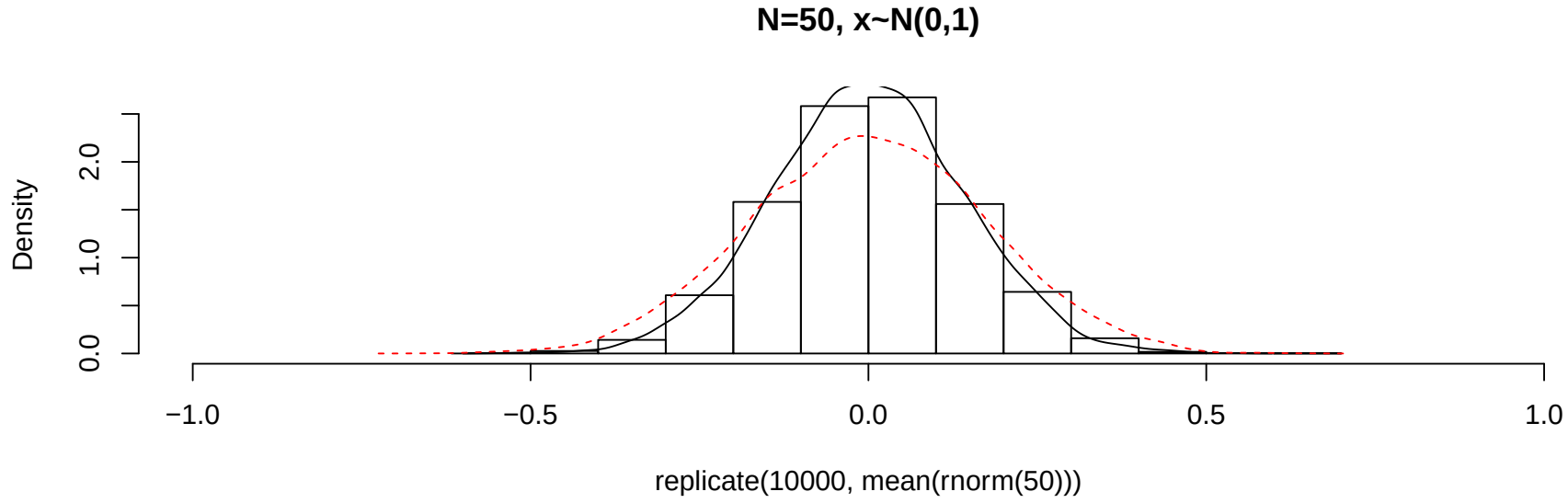
Comments (2)

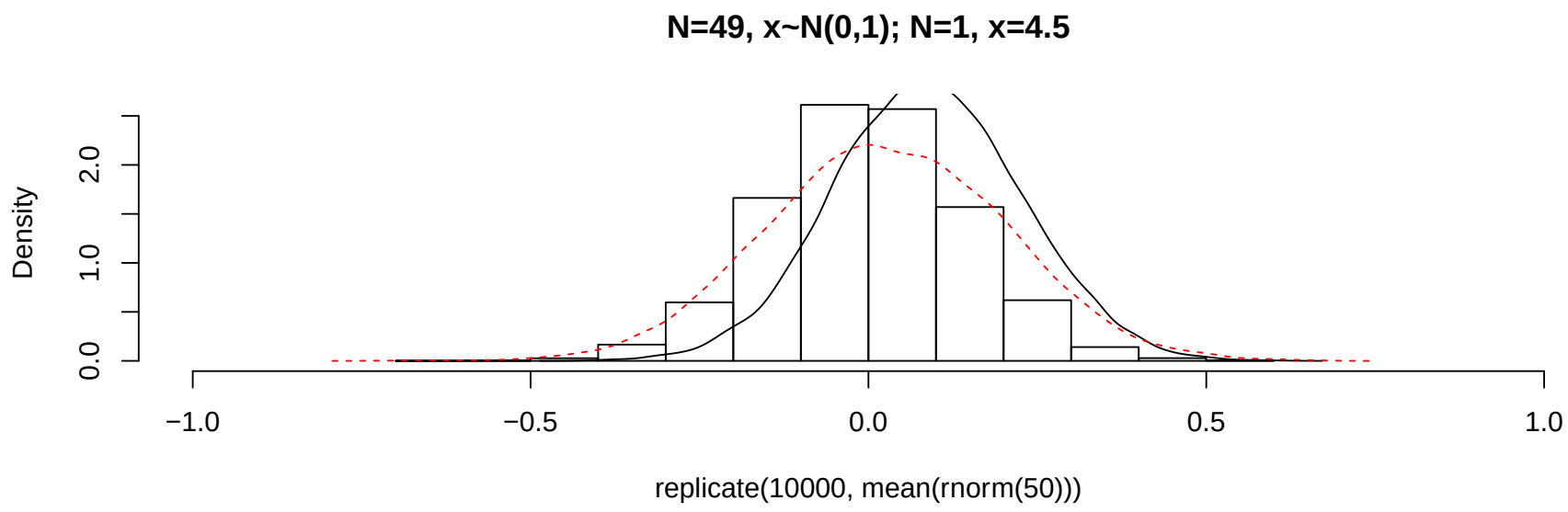
- The two-standard deviation rule does not identify outliers specifically
- It reliably identifies observations with values of relatively large magnitude in an approximately Gaussian population
- In other situations the **non-robustness** of mean and standard deviation vitiates straightforward interpretation of the procedure
- Robust estimation methods depend on technical definitions
  - intuitively, robust estimators retain interpretations of their non-robust counterparts under weaker assumptions about the underlying populations

Location estimator

- Location estimators answer questions like: Around what value m can we expect observations to vary, with approximately equal numbers above and below m?
  - Gaussian population: choose $m = \bar{x}$, the sample mean
  - more generally: choose $m = \text{med } x$, the sample median
    - the median achieves the goal by definition
    - if the Gaussian population assumption is correct, the median is more variable than the mean (constitutes less *efficient* use of the information)
    - this is a simple illustration of the robustness vs. efficiency tradeoff

Location estimation: median loses information in ideal case, but is less sensitive to "contamination"





Median has breakdown bound 50% but is still affected by contaminants; midpoint of shorth is more robust

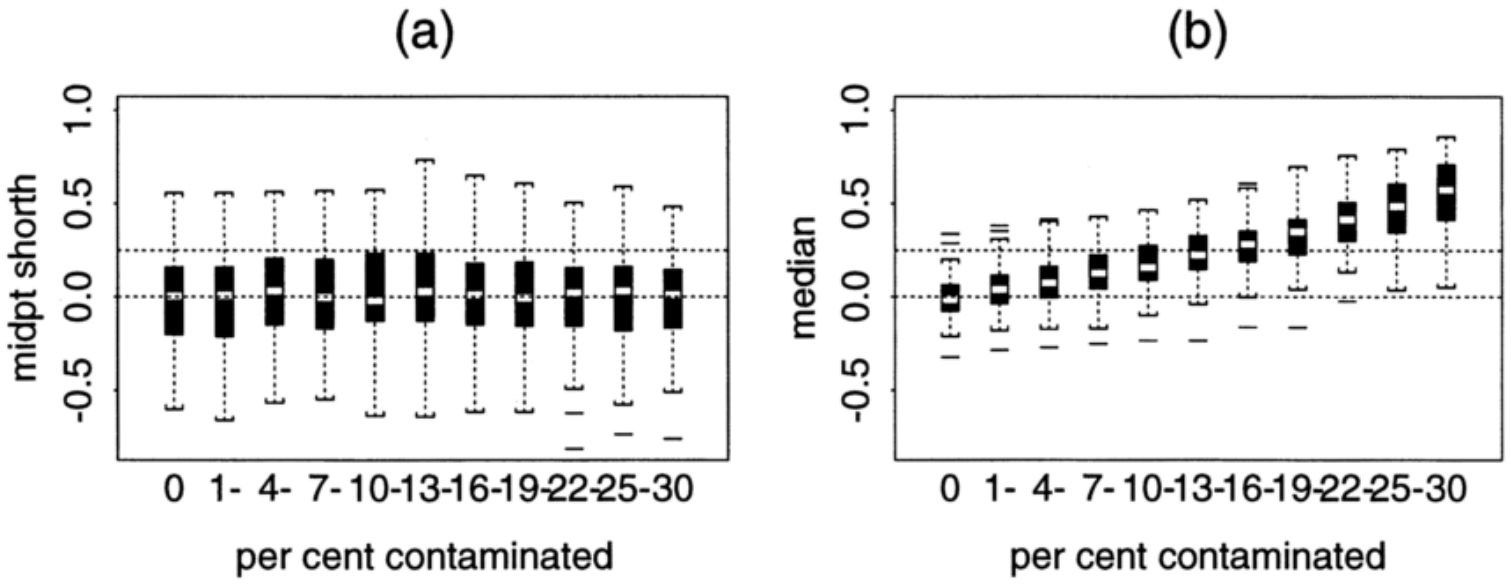


Figure 10: From Carey et al. Technometrics 1997

Relation to outlier identification

- Basic intuition: outliers x have large values of $|x - c|/s$ where c locates the center of the **inlier** distribution and s measures the scale of the **inlier** distribution
- Simple rules for determining the center of a distribution are likely to fail if outliers are present
  - simple linear estimator (mean) has zero **breakdown bound**: a single observation can exert arbitrarily large influence
  - median has breakdown bound 50%: will only take arbitrary values if 50% or more of the data are contaminated
  - intermediate estimators: trimmed means, winsorized means

Scale estimation

- SD has zero breakdown bound
- IQR has breakdown bound 25% – related to boxplot display
- MAD (median absolute deviation) has breakdown bound 50%
- Proposal of Tibshirani and Hastie (Biostatistics 2007):
  - use (x - median)/MAD(controls) to standardize all observations
  - define high outliers using the 75th percentile + IQR of this standardized expression value
  - define low outliers using the 25th percentile - IQR
  - choose the largest outlier sum as a measure of outlier-proneness of genes

More systematic approaches

- Formalize the outlier concept:
  - a base population distribution exists and is identified
  - outliers are observations unlikely to be members of the base population
- Control the probability of "Type I error": declaring an observation to be an outlier when it is actually an observation from the base population
- Deal explicitly with the fact that multiple outliers can have effects that are more complex than those induced by single outliers
- Consider the concept of **multivariate outlyingness**

The Extreme Studentized Deviate and a "peeling" sequence

Given n observations, define the ESD as follows:

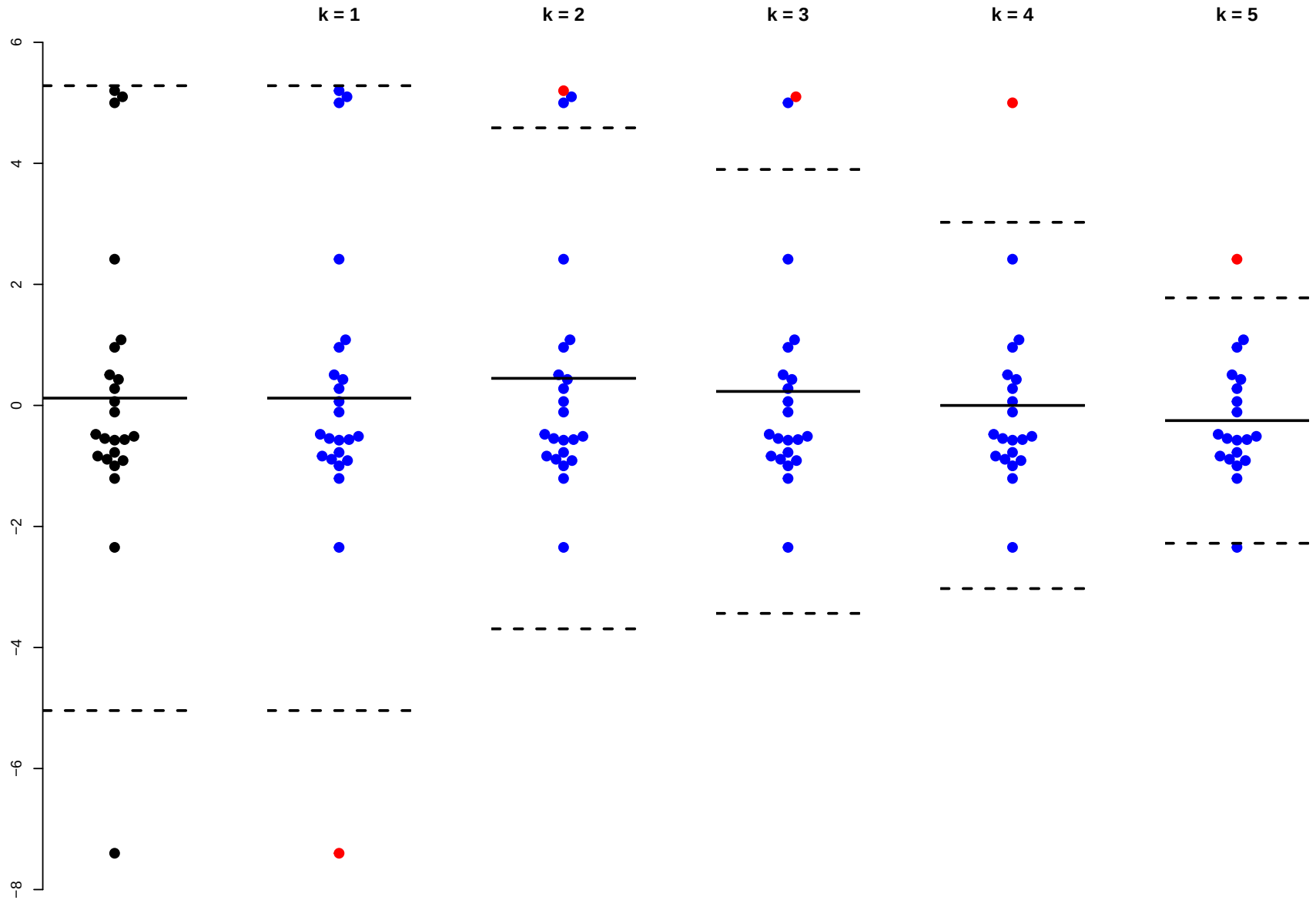
- $x_1, \ldots x_n$ have sample mean $\bar{x}$ and sample standard deviation s.
- the sample value x maximizing the quantity $|x - \bar{x}|/s$ is denoted the ESD

A sequence of k subsamples and deviates is generated by repeating this process on the $n-1$, $n-2, \ldots, n-k$ sized subsets created by excluding the associated ESD

At each stage a new sample mean and standard deviation are computed and used

Inward peeling illustrated

Large negative outlier masks multiple positive ones for the mean $\pm$ 2SD approach. Red dots are points corresponding to stagewise ESD



Outward testing: parody::calout.detect(meth="GESD")

- Once we have found the sequence of ESDs/truncated samples, we test, from the inside out, each ESD for compatibility with the "inlier" distribution using the inlier mean and standard deviation
- The critical values for this sequential procedure are available
  - for finite samples, by simulation
  - for large samples, using formalism in Rosner, Technometrics 1983
- The thresholds provided bound to any desired level the probability of falsely labeling as an outlier a member of the base population, allowing tests for up to half the sample size
- When one ESD is deemed an outlier, it is joined in this characterization by all the more extreme deviates identified earlier in the peeling process

Upshots

- Don't use naive approaches to outlier labeling
- Robust location and scale estimation methods have a long history and interesting variations
- Of greater interest: multivariate analog of GESD

mv.calout.detect: GESD sequence for multivariate Gaussian population (Caroni+Prescott JRSS-C 1992)

2. Wilks's Outlier Test Statistic

Given n independent p-dimensional vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, Wilks's statistic may be derived as the likelihood ratio statistic testing the null hypothesis

$$H_0: \mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n,$$

against the alternative hypothesis

$$H_1: \mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad i \neq j,$$
$$\text{and } \mathbf{x}_j \sim N_p(\boldsymbol{\mu} + \mathbf{a}, \boldsymbol{\Sigma}).$$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, as is the slippage parameter $\mathbf{a}$ and the index j of the outlier.

Consideration of the likelihood ratio leads to the test statistic $W_j = |\mathbf{A}^{(j)}|/|\mathbf{A}|$, where $\mathbf{A}$ is the matrix of sums of squares and cross-products

$$\mathbf{A} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

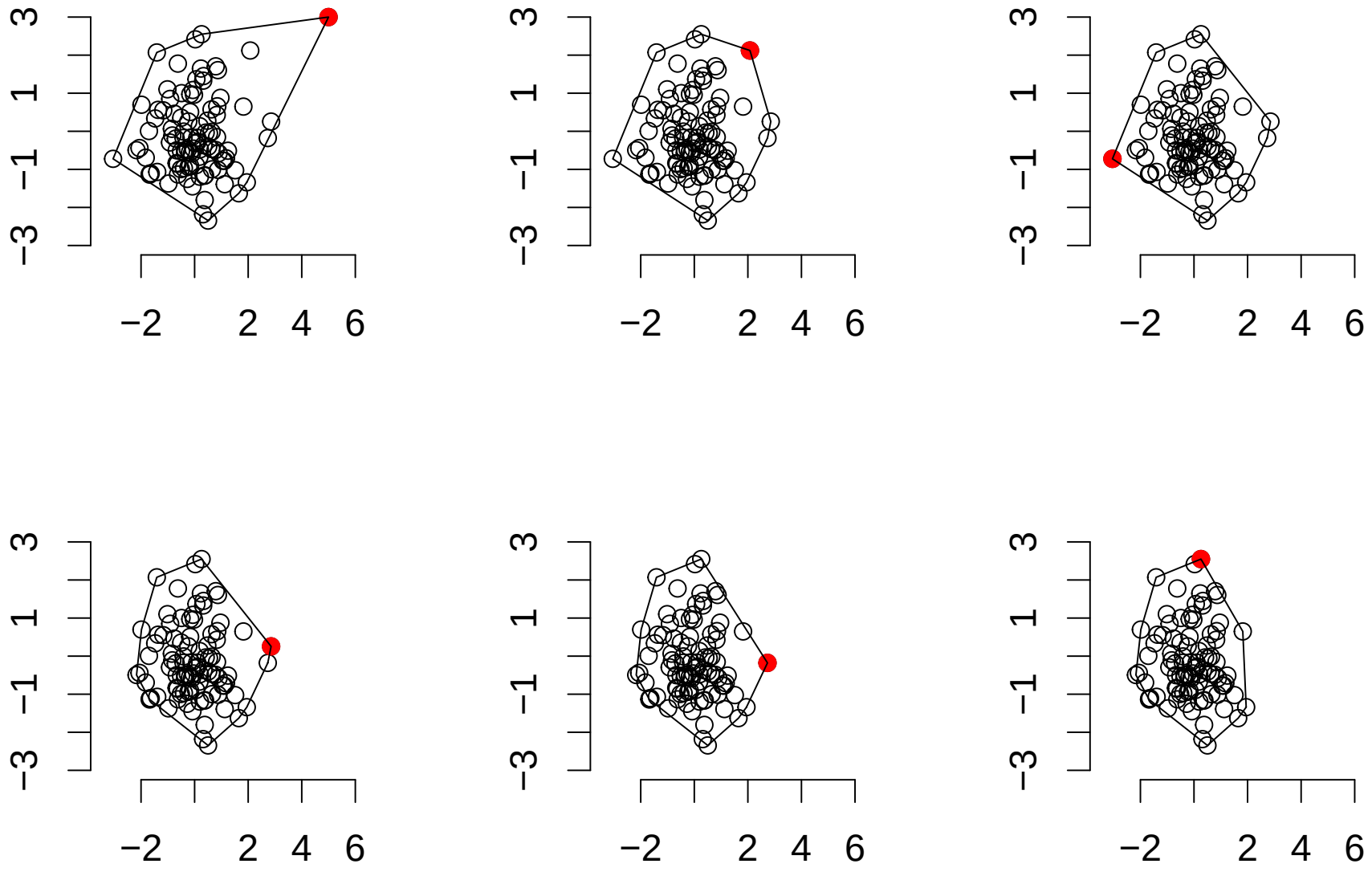
and $\mathbf{A}^{(j)}$ is the corresponding matrix with $\mathbf{x}_j$ eliminated from the sample. The potential outlier is that point, index l say, whose removal leads to the greatest reduction in $|\mathbf{A}|$, i.e. the point for which this ratio is minimized. Wilks's statistic is then defined as

$$D_1 = \min_j(W_j) = |\mathbf{A}^{(l)}|/|\mathbf{A}|.$$

For ease of computation it is preferable to express D_1 as

$$D_1 = 1 - \frac{n}{n-1}(\mathbf{x}_l - \bar{\mathbf{x}})'\mathbf{A}^{-1}(\mathbf{x}_l - \bar{\mathbf{x}}),$$

The peeling process for contaminated BVN



Recap

- Sequential peeling: find the most extreme data point using efficient estimation of location and scale, peel it away, and repeat, until only half the data remains
- Outward testing: Establish criteria that bound the probability of erroneously declaring an outlier to be present through the sequence of tests
- Works in univariate and multivariate settings
- Alternatives to Gaussian base model?

Multivariate expression outliers in glioblastoma

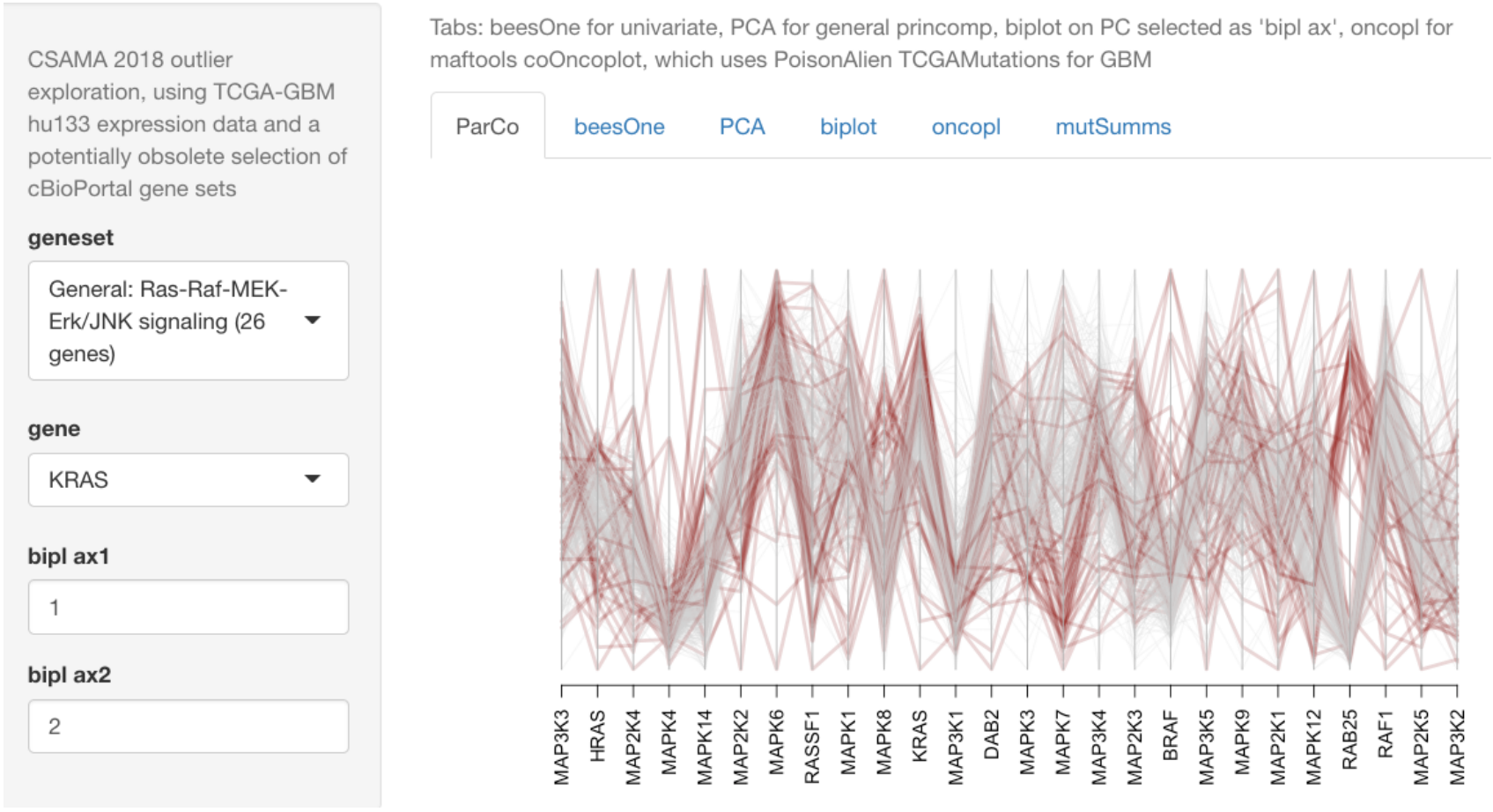


Figure 12: Parallel coordinate plot for a 26 gene set.

Discordance of labeling procedures for RAB25

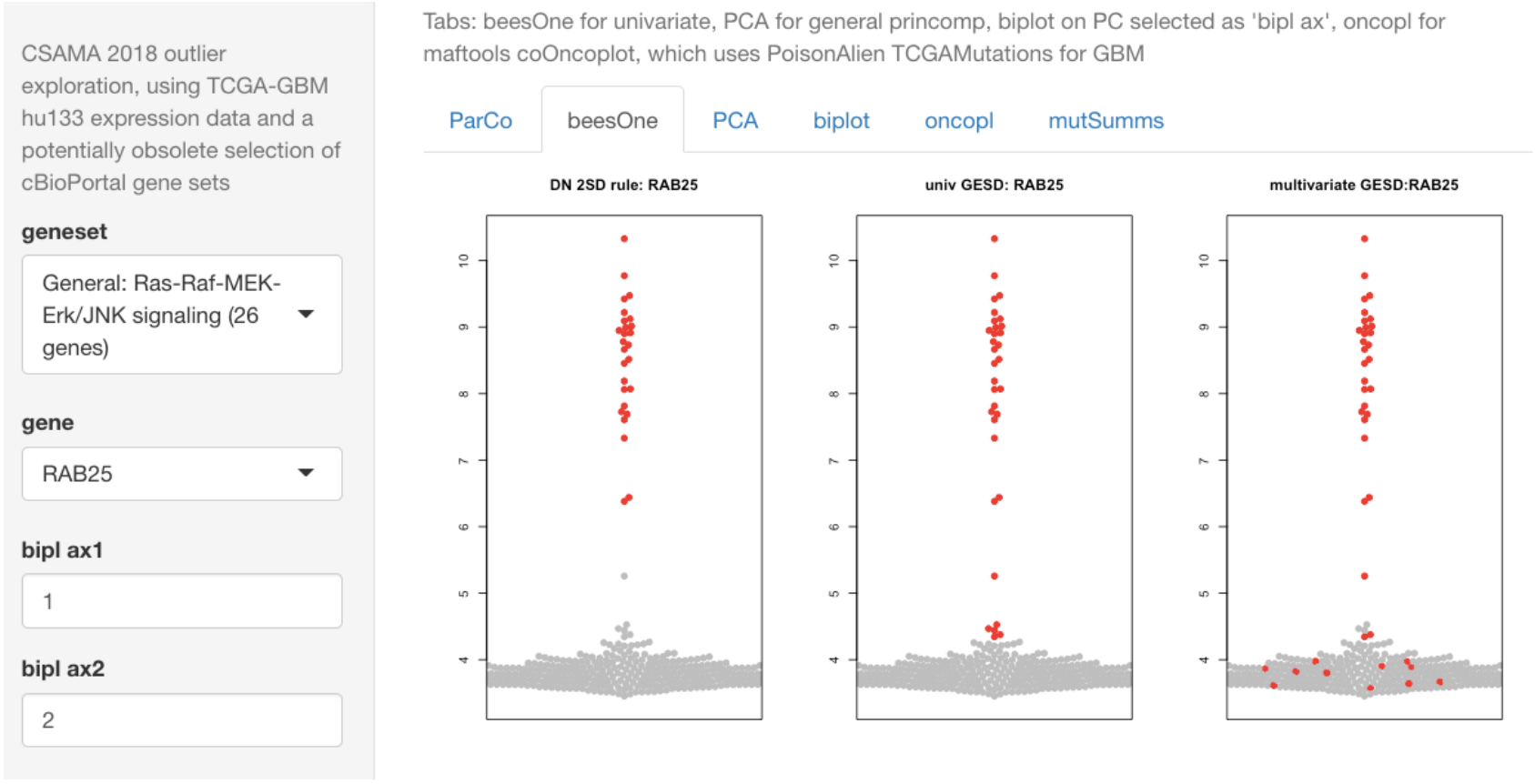


Figure 13: 2SD rule, GESD, and Caroni-Prescott.

PCA for 26 genes, coloring multivariate outliers

CSAMA 2018 outlier exploration, using TCGA-GBM hu133 expression data and a potentially obsolete selection of cBioPortal gene sets

geneset

General: Ras-Raf-MEK-Erk/JNK signaling (26 genes)

gene

RAB25

bipl ax1

1

bipl ax2

2

Tabs: beesOne for univariate, PCA for general princomp, biplot on PC selected as 'bipl ax', oncopl for maftools coOncoplot, which uses PoisonAlien TCGAMutations for GBM

ParCo | beesOne | PCA | biplot | oncopl | mutSumms

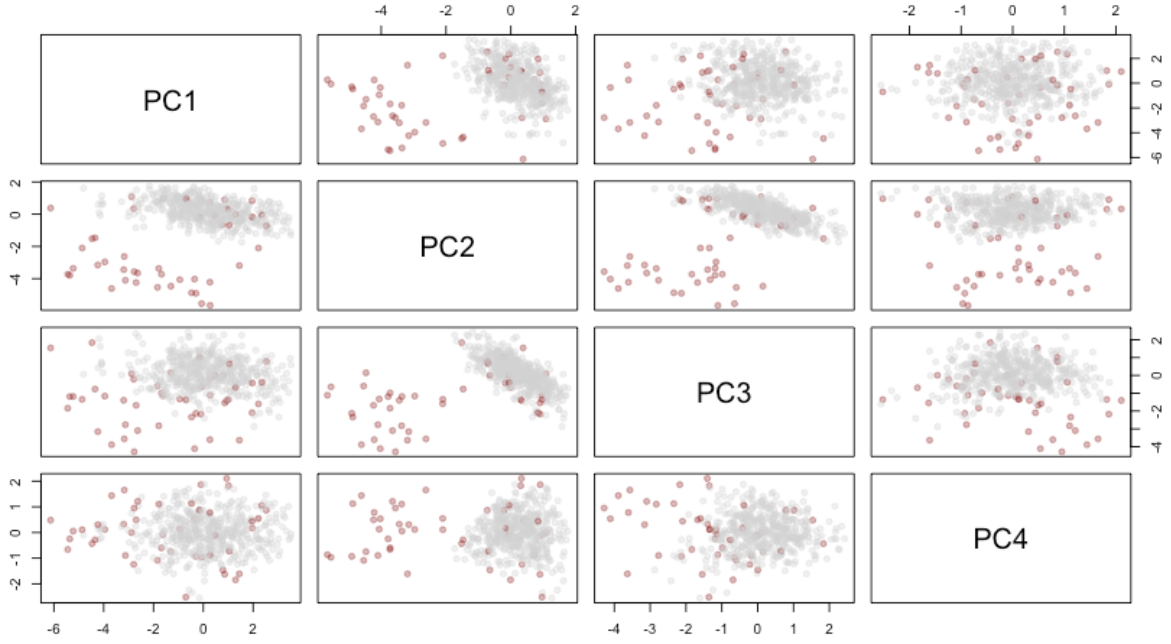


Figure 14:

Mutation burdens for expression outliers/inliers

CSAMA 2018 outlier exploration, using TCGA-GBM hu133 expression data and a potentially obsolete selection of cBioPortal gene sets

geneset

General: Ras-Raf-MEK-Erk/JNK signaling (26 genes)

gene

RAB25

bipl ax1

1

bipl ax2

2

Tabs: beesOne for univariate, PCA for general princomp, biplot on PC selected as 'bipl ax', oncopl for maftools coOncoplot, which uses PoisonAlien TCGAMutations for GBM

ParCo beesOne PCA biplot oncopl mutSumms

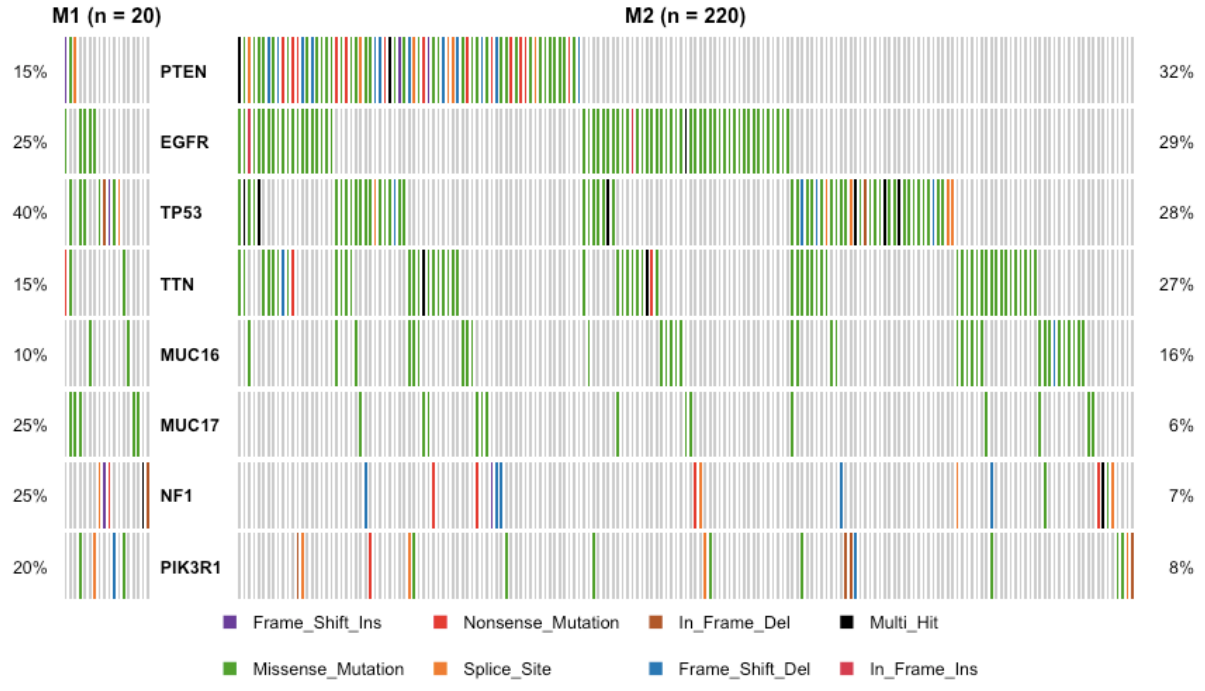


Figure 15:

Conclusions

- Robust statistical methods are extremely useful in modern genomics
- Systematic identification of unusual data configurations is a continual concern
  - quality issues
  - search for mechanisms underlying variability
- We've focused on robust estimation (high breakdown bound)
  - univariate location (median), scale (MAD, length of shorth)
  - multivariate scale (covariance of peeled dataset)
  - there are many other options
- Application: are aberrant expression patterns readily linked to distinctive patterns of somatic mutation?
  - not low-hanging fruit: outlier cohort likely an amalgam of diverse mutational events
  - these pieces can be combined with other omics and clinical data fairly easily to test more compelling hypotheses

References and software