Enhanced estimation of the cancer proteome from transcriptome assays

Angela

October 23, 2014

Cancer phenotypes manifest at the level of proteins, but quantitative proteomics assays are not yet feasible in high throughput relative to transcriptome assays. Correlation between mRNA and protein levels is known to be modest, in part due to regulation by RNA-binding proteins (RBPs) and non-coding RNAs including microRNAs. These regulatory relationships have been catalogued computationally or experimentally, but have yet to be incorporated as a transcriptome-wide model of gene translation. This case study explores the possibility of building such a model from regulatory databases and experimental mRNA and protein abundance data from matched samples. Such a model could feasibly raise the utility of thousands of existing publicly available microarray and RNA-seq transcriptome experiments for the development of cancer prognostic predictors. This goal could be achieved through an added in silico step without incurring additional experimental costs. Furthermore, the identification of context-specific regulation of gene translation through the model selection process would contribute to understanding alterations of regulatory pathways in cancer and to drug design. This case study leverages matched microarray and mass spectrometry-based experiments for the NCI-60 cell lines, and an RBP target database, to develop a statistical model of protein abundance estimated from mRNA abundance.

Vignette's code, .rda and .txt files are available at https://github.com/ar00/tmc

```
require(GenomicRanges)
require(gplots)
require(hash)
require(nortest)

load("summarizedExperimentiBAQ.rda")
load("NCI60.annotation.rda")
load("CISBP-RNA.rda")

# binary matrix for presence/absence of bindng sites (Q-value<.20) for any RBP in NCI60 genes.
# Note that RBPs in this file might result undetected at protein level in NCI60.
rbp.db=read.delim('NCI60.rbp.target.binary.txt',header=TRUE,as.is=TRUE,sep='\t')
# goodness-of-fit measures for models using RBP protein/RNA levels
fr = read.table('ridge.penalty.glmnet.rlm.max5nas.iBAQ.summary.by.rna.txt',sep='\t',as.is=T,header=T)
fp = read.table('ridge.penalty.glmnet.rlm.max5nas.iBAQ.summary.txt',sep='\t',as.is=T,header=T)
# coefficients of regression models based in RBP protein abundances
coeff = read.table('ridge.penalty.glmnet.rlm.max5nas.iBAQ.stability.txt',header=T,sep='\t',as.is=T)
fitfp = 'ridge.penalty.glmnet.rlm.max5nas.iBAQ.fitted.txt'
fitfr = 'ridge.penalty.glmnet.rlm.max5nas.iBAQ.fitted.by.rna.txt'
source('dx.plot.R')
```

The object of class SummerizedExpriment contains normalized gene expression profiles at RNA and protein levels across 59 cell lines of the NCI60 panel (log-10 scale).

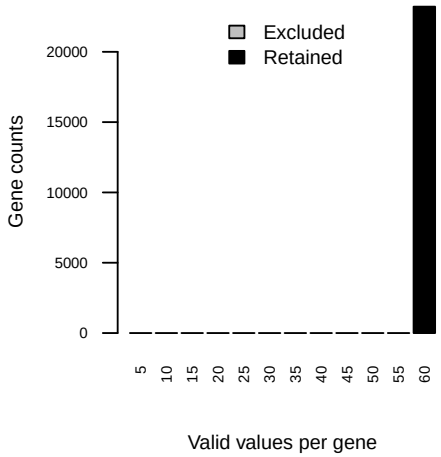
```
prot = as.matrix(assays(sset)[[2]])
rna = as.matrix(assays(sset)[[1]])
show(rna[1:5,1:5])
```

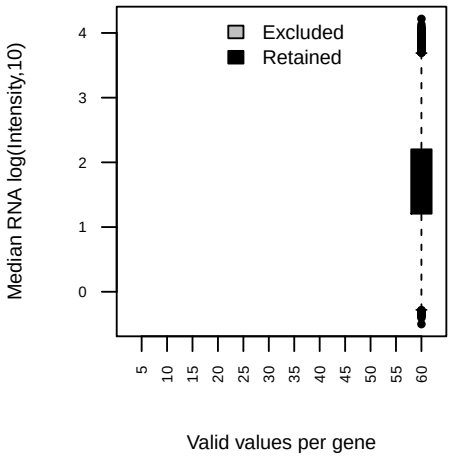
```
##     BREAST_BT549 BREAST_HS578T BREAST_MCF7 BREAST_MCF7ADR BREAST_MDAMB231
## ID1     3.058590      3.052328    3.005707       2.968685        3.157620
## ID2     2.922308      2.500957    2.995548       3.029181        2.808274
## ID3     2.222351      2.371245    2.251824       2.185191        2.724046
## ID5     2.277097      2.381688    2.113162       2.092676        2.484485
## ID6     1.615983      1.449195    2.698117       2.553821        1.696116
```

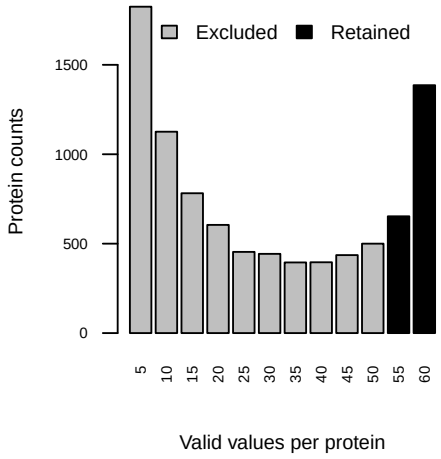
```
show(prot[1:5,1:5])
```

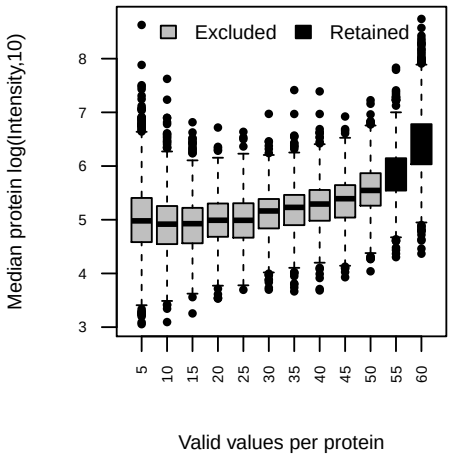
```
##     BREAST_BT549 BREAST_HS578T BREAST_MCF7 BREAST_MCF7ADR BREAST_MDAMB231
## ID1     4.568436            NA          NA       4.975772              NA
## ID2     6.949653      6.341316    7.193598       6.810327        6.804283
## ID3           NA      5.141230          NA             NA              NA
## ID5     5.300073      4.626576    5.939434       6.347954        6.748514
## ID6           NA            NA    6.530584       5.106463              NA
```

Due to the high counts of invalid values per identifier (i.e. cell lines where protein was undetected) in NCI60 proteomics profiles, we retained the genes with at most five invalid values (Figure1).

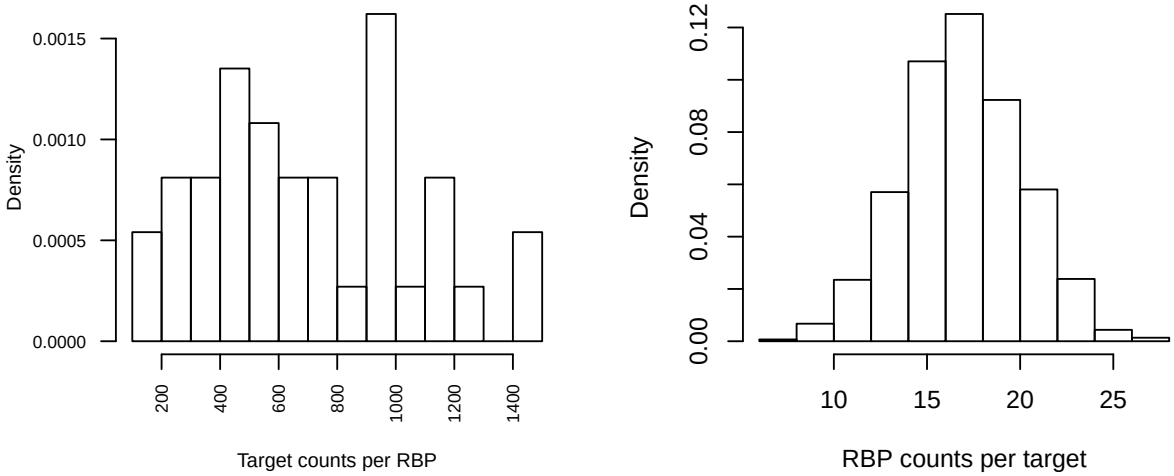


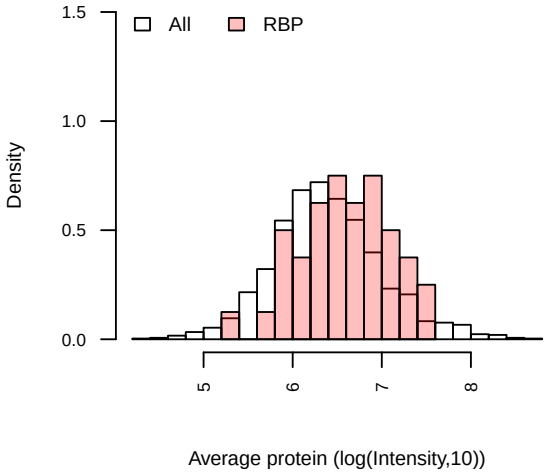


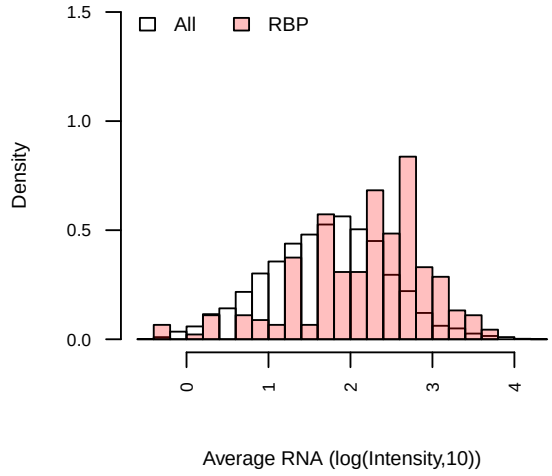




Since post-transcriptional control can be a biological reason of low concordance between RNA nad protein profiles, we assembled a database of inferred binding sites of RNA binding proteins on the untranslated regions, UTRs, of genes profiled at both levels of NCI60 (Figure2).

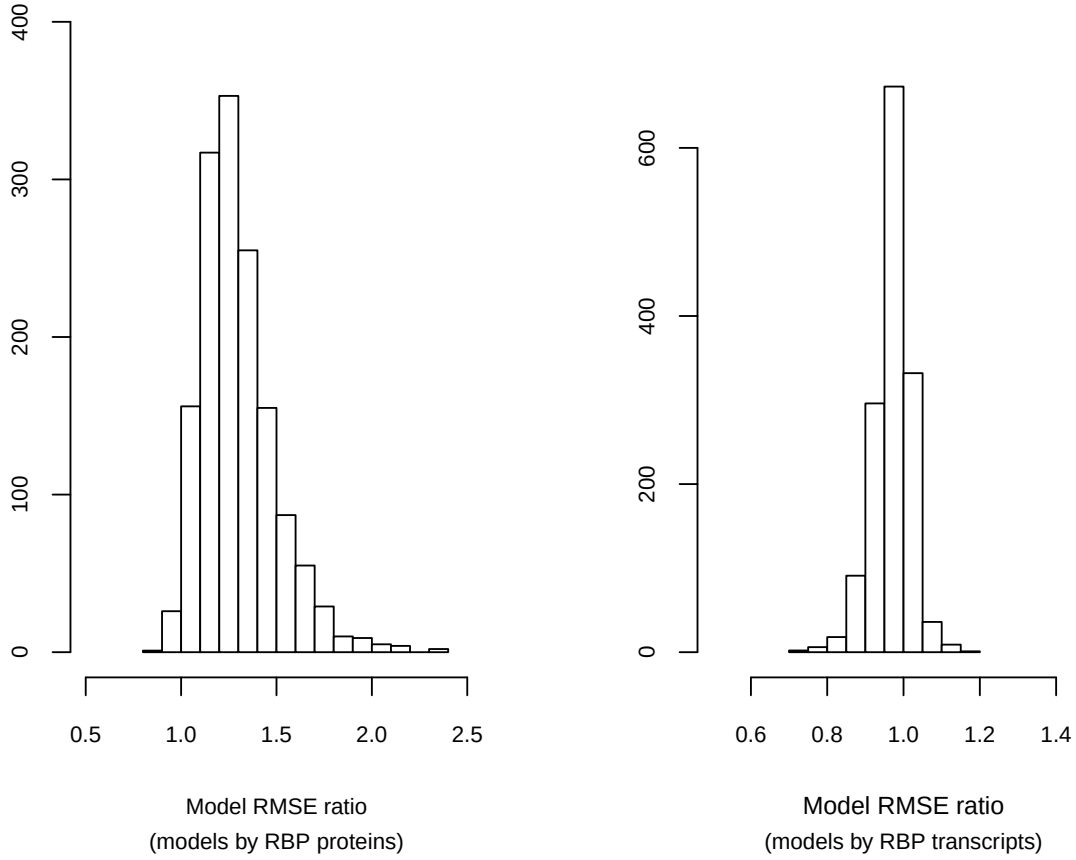




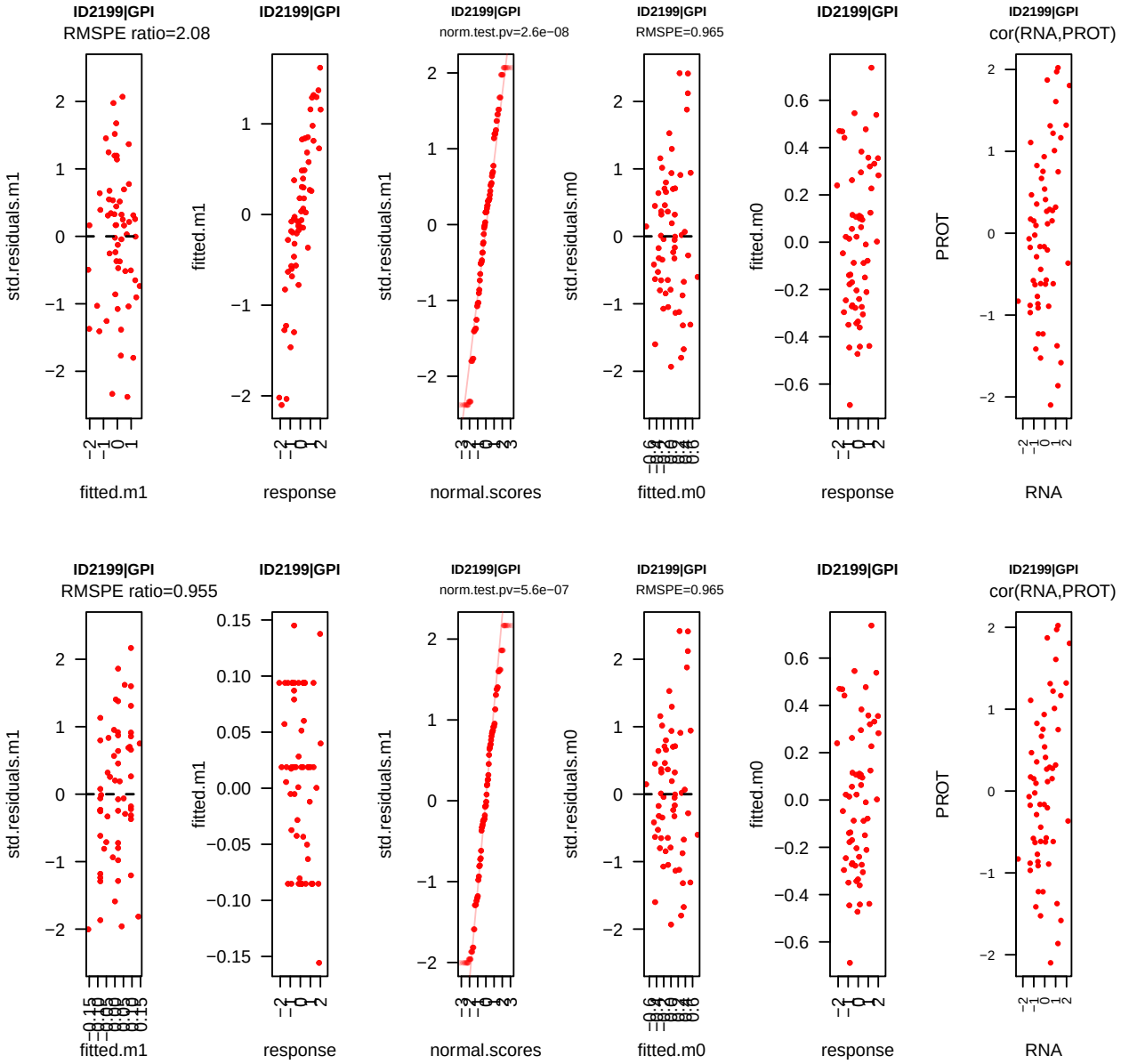


We built two models: a reduced model where the candidate predictor of protein abundance is the mRNA abundance, and a full model containing additional candidate predictors given by expression data for RBPs inferred to bind the mRNA UTRs. We standardized explanatory and response variables. We fitted the reduced model by robust linear regression and the full model by ridge-penalty linear regression (glmnet R package). Three-fold cross-validation was used to learn regularization parameter to derive penalized coefficients of the models. Extra-sample prediction error, defined by root mean squared error (RMSE), was estimated by five-fold cross-validation. We compared the two models to assess whether and by how much the full model decreases prediction error of protein abundance, and to select informative RBPs.

We built full models using either RNA or protein abundances of RBPs and found RBP protein data but not RBP RNA data are useful in the models to predict the proteome (Figure 3).



We show an example of gene, GPI, where the RMSE ratio of the model which uses RBP protein data is higher than the RMSE ratio of the model which uses RBP RNA data.



Contribution to/from biocMultiAssay: Currently this analysis requires crossing several tables containing: 1) expression data at two levels, (2) gene/transcript/protein associations and corresponding attributes including gene coordinates, CDS, 5′ and 3′ UTR coordinates and external aliases (RefSeq,HGNC,SwissProt/UniProt and so on), (3) regulatory annotations including RBPs inferred to bind UTRs and RBP binding site attributes (coordinates, score, matched sequence).

Crossing tables was time-consuming and error-prone. The whole flowchart might be streamlined by an initial effort to decorate the genes in the SummarizedExperiment object with data structures containing the aforementioned basic and regulatory attributes.

Also the association between gene, transcript and protein is not unique and it might be useful to be able to select the triplet (gene/transcript/protein) according to an attribute of interest such as the transcript with longest 5′(3′) UTRor the transcript experimentally supported in RefeSeq.