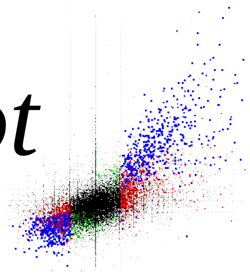
ExpressionPlot



A web-based framework for the integration of
RNA-Seq and microarray genes expression data
http://expressionplot.com/iDEA

Abstract

RNA-Seq technologies have emerged as important tools for detecting changes in gene expression and RNA processing in biological samples. We present *ExpressionPlot*, a software package for RNA-Seq analysis consisting of a back end, which prepares the raw sequencing data, and the web-based front end, which allows non-specialists to browse, visualize, and compare different data sets. The data are kept private on the website until released by their owner to other users or the public.

Novelty

The main advantages of *ExpressionPlot* over existing tools are that it can solve so many of the problems associated with high throughput gene expression analysis, and that its simple access-controlled web interface gives non-bioinformaticians access to complicated but important analytical methods. The *ExpressionPlot* paradigm is that the heavy lifting, including alignment, quantifying genes and isoforms, and calculating differential expression statistics, is done during a pre-processing stage, and the visualization, including gene-level plots, heatmaps and genome views, is performed on-demand through a web interface. Thus, without burdening them with the significant computational issues involved, *ExpressionPlot* gives non-expert users real-time control over the details of their data analysis.

Many of the tools take special advantage of interactive web technologies. For example, differentially expressed genes can be browsed in a dynamic table and sorted or filtered based on genomic location, gene names, expression levels or fold-changes. The rows of the this table are automatically hyperlinked to the proper region of the genome browser, and action buttons are also provided to export gene lists to spreadsheet software and automatically generate background sets of unchanged genes of similar expression values for use in Gene Ontology analysis.

Finally, data access can be controlled through the user interface, making it possible to share unpublished data with collaborators, and, when appropriate, the public.

Prototype

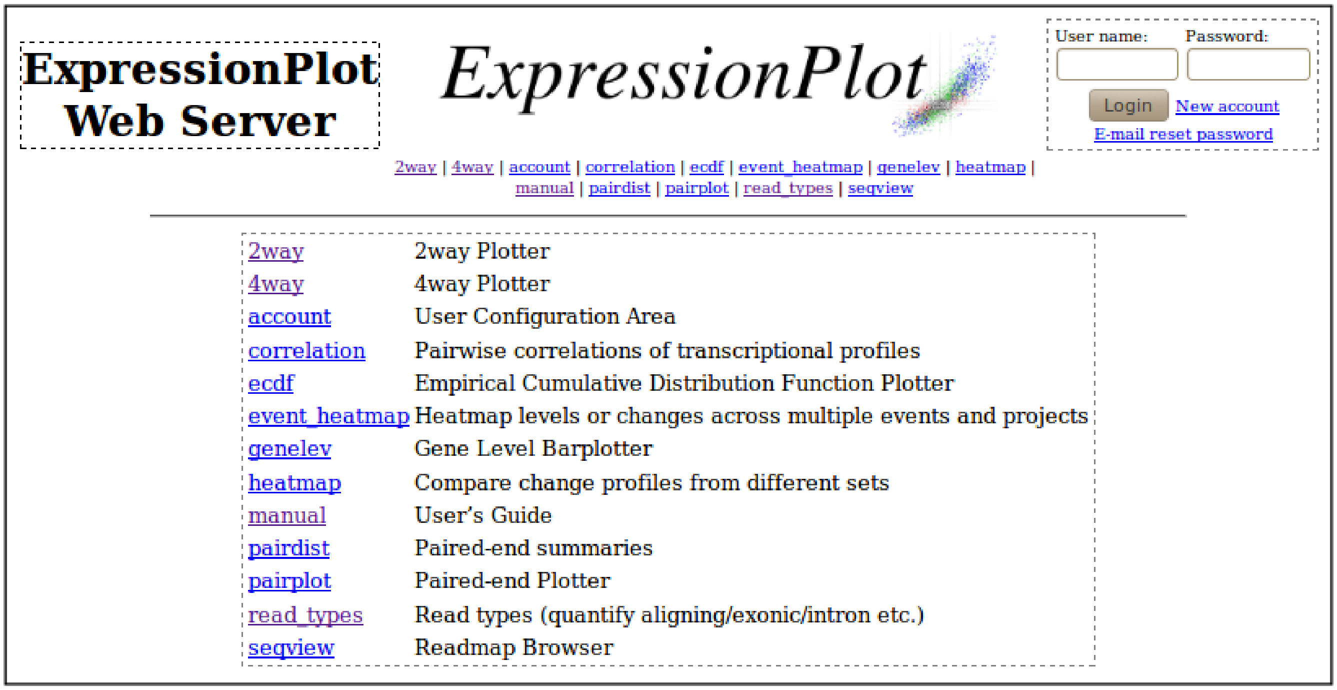
We've added the iDEA RNA-Seq data into our prototype server, accessible at http://expressionplot.com/iDEA. Choose “iDEA Breast Cancer Cell Lines Paired-End RNA-Seq ” or “iDEA Breast Cancer Cell Lines Directional RNA-Seq ” from the drop down menus as your data set to access the iDEA RNA-Seq data.

Credits and Contact

ExpressionPlot was developed by Brad Friedman while a postdoc in Tom Maniatis' lab at Harvard University and David Housman's Lab at MIT. For more information please e-mail brad.aaron.friedman@gmail.com.

Landing Page

This is a screen shot of the landing page for the *ExpressionPlot* website. The user login box is in the top right, and brief descriptions of the tools are shown in the center:


ExpressionPlot Web Server
ExpressionPlot
User name:
Password:
Login
New account
E-mail reset password
2way | 4way | account | correlation | ecdf | event_heatmap | genelev | heatmap |
manual | pairdist | pairplot | read_types | seqview

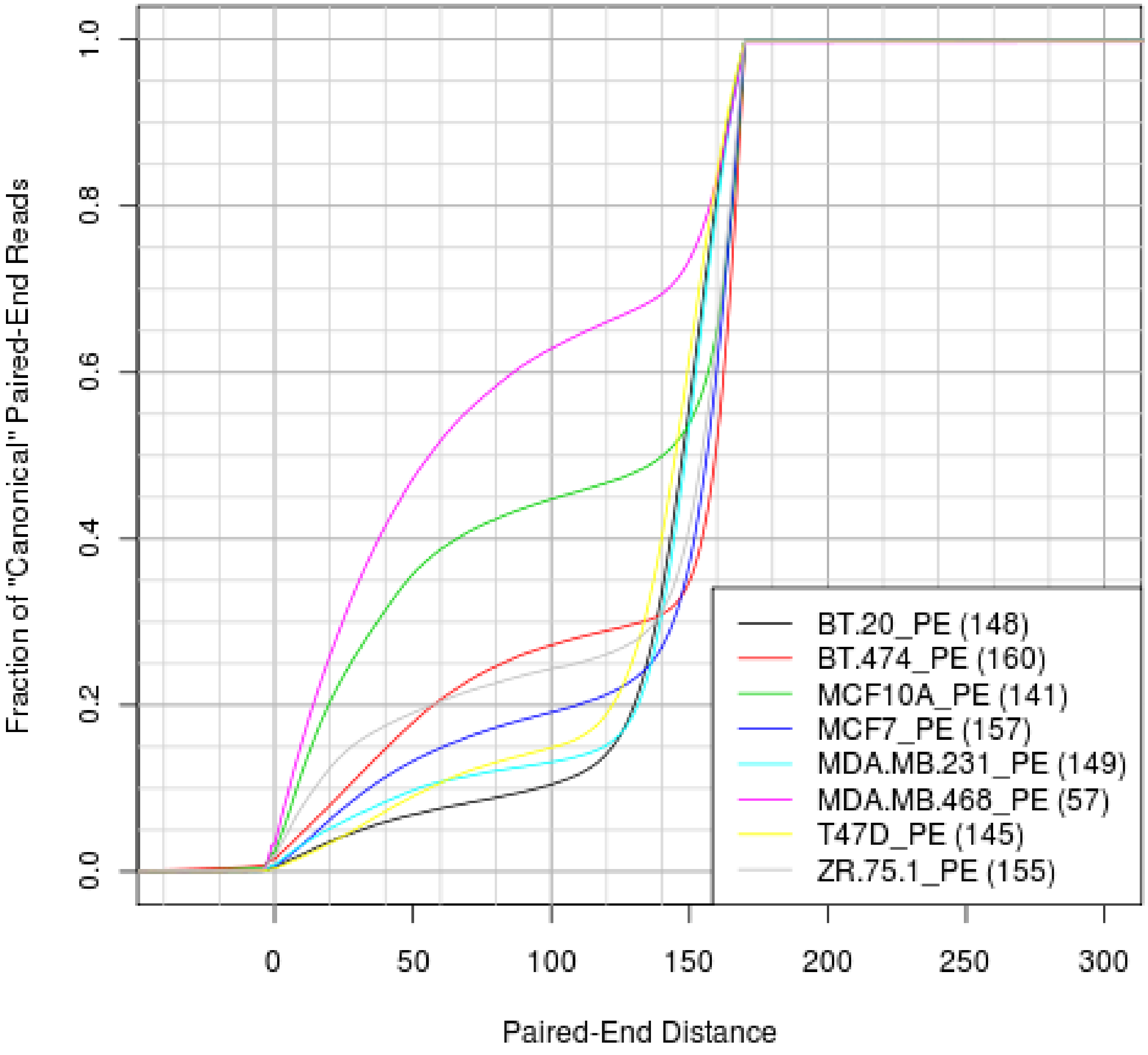
2way	2way Plotter
4way	4way Plotter
account	User Configuration Area
correlation	Pairwise correlations of transcriptional profiles
ecdf	Empirical Cumulative Distribution Function Plotter
event_heatmap	Heatmap levels or changes across multiple events and projects
genelev	Gene Level Barplotter
heatmap	Compare change profiles from different sets
manual	User's Guide
pairdist	Paired-end summaries
pairplot	Paired-end Plotter
read_types	Read types (quantify aligning/exonic/intron etc.)
seqview	Readmap Browser

ExpressionPlot Examples

Each of the remaining pages in this document show example plots of the iDEA data made by the tools on the *ExpressionPlot* website. They are presented as generated by the website, except for some minor cropping to help them fit better into this format.

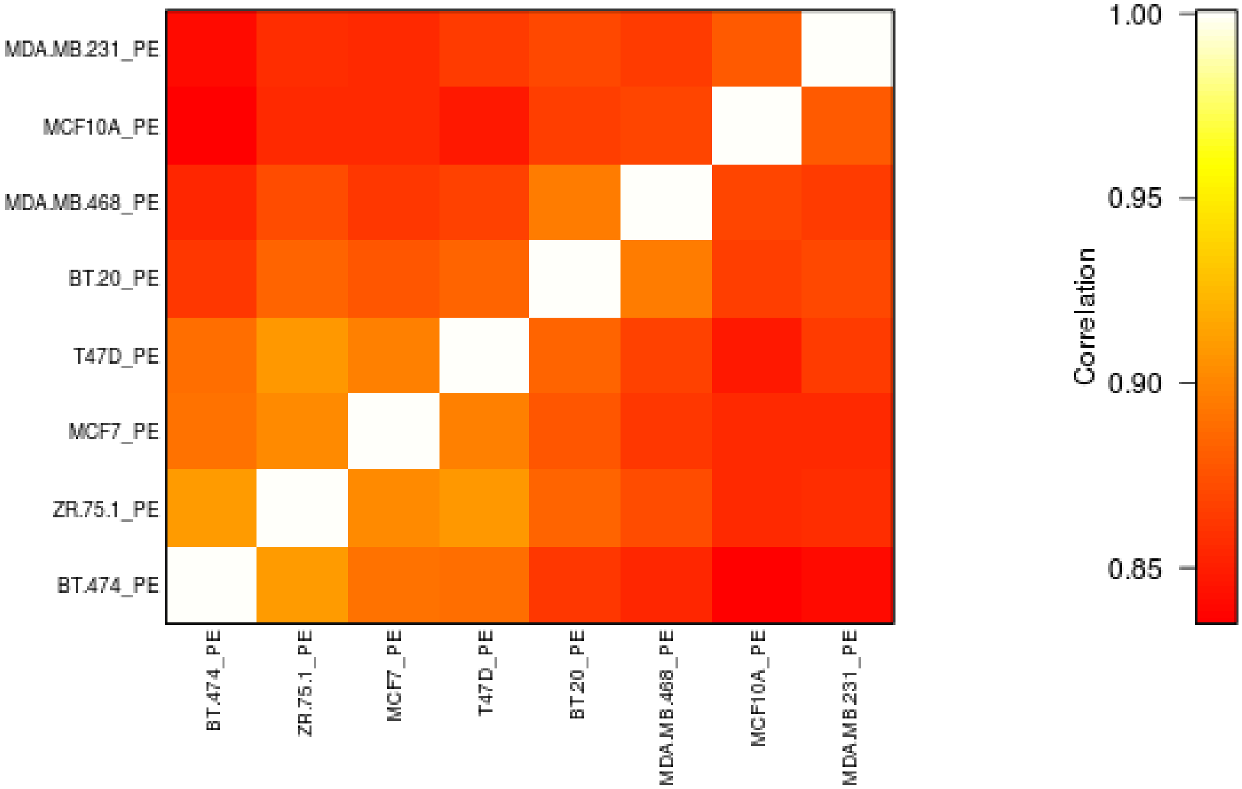
The **pairdist** tool shows the empirical cumulative distribution function (ECDF) of paired-end distances of “canonical” reads (same chromosome, different strand, minus strand read downstream of plus strand read, and no known intron intervening the paired ends). “Distance” is defined as the genomic distance, in nucleotides, between the aligned positions of the last sequenced bases of the two reads, and can be negative if the alignments overlap. This is known by some authors as the “mate inner-distance”. Numbers in parentheses indicate median paired-end distance for each sample (add 50 for both sequences and 50 for both Illumina adaptors (+200) to get complete library size).

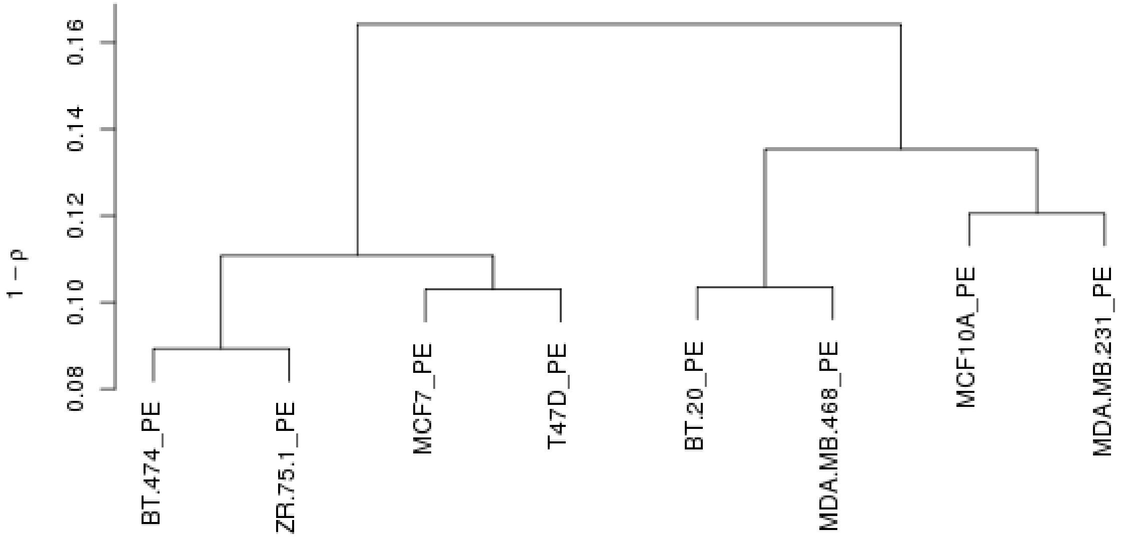
This plot reveals that the MDA.MB.468 and MCF10A libraries had a broader distribution of paired-end distances. This technical abnormality might have been introduced during a size-selection or amplification step.



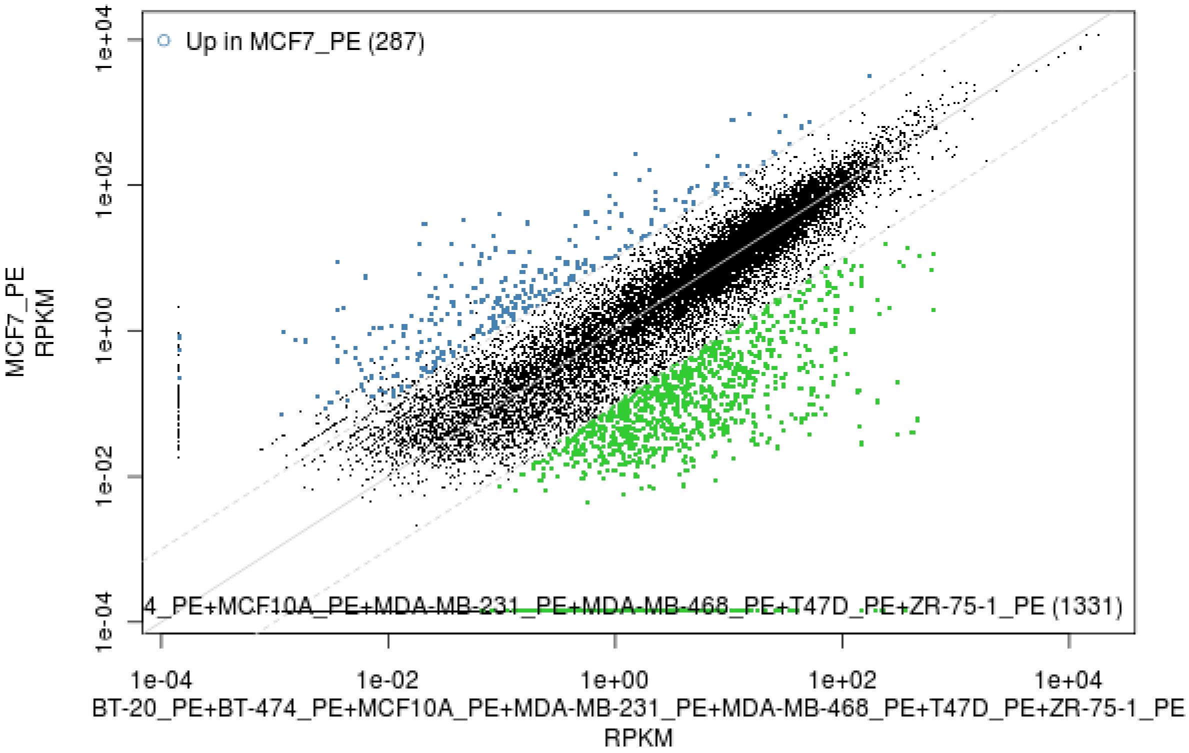
The **correlation** tool makes a heatmap (top) or clustering dendrogram (bottom) of pairwise correlations of gene expression profiles generated from each cell line.

Like the pairdist tool, this correlation tool serves as a quality control. However, it provides additional biological information, for example identifying MDA.MB.231 and MCF10A as outliers in terms of their transcriptional profiles, and BT.474 and ZR.75.1 as the most similar pair in the panel.

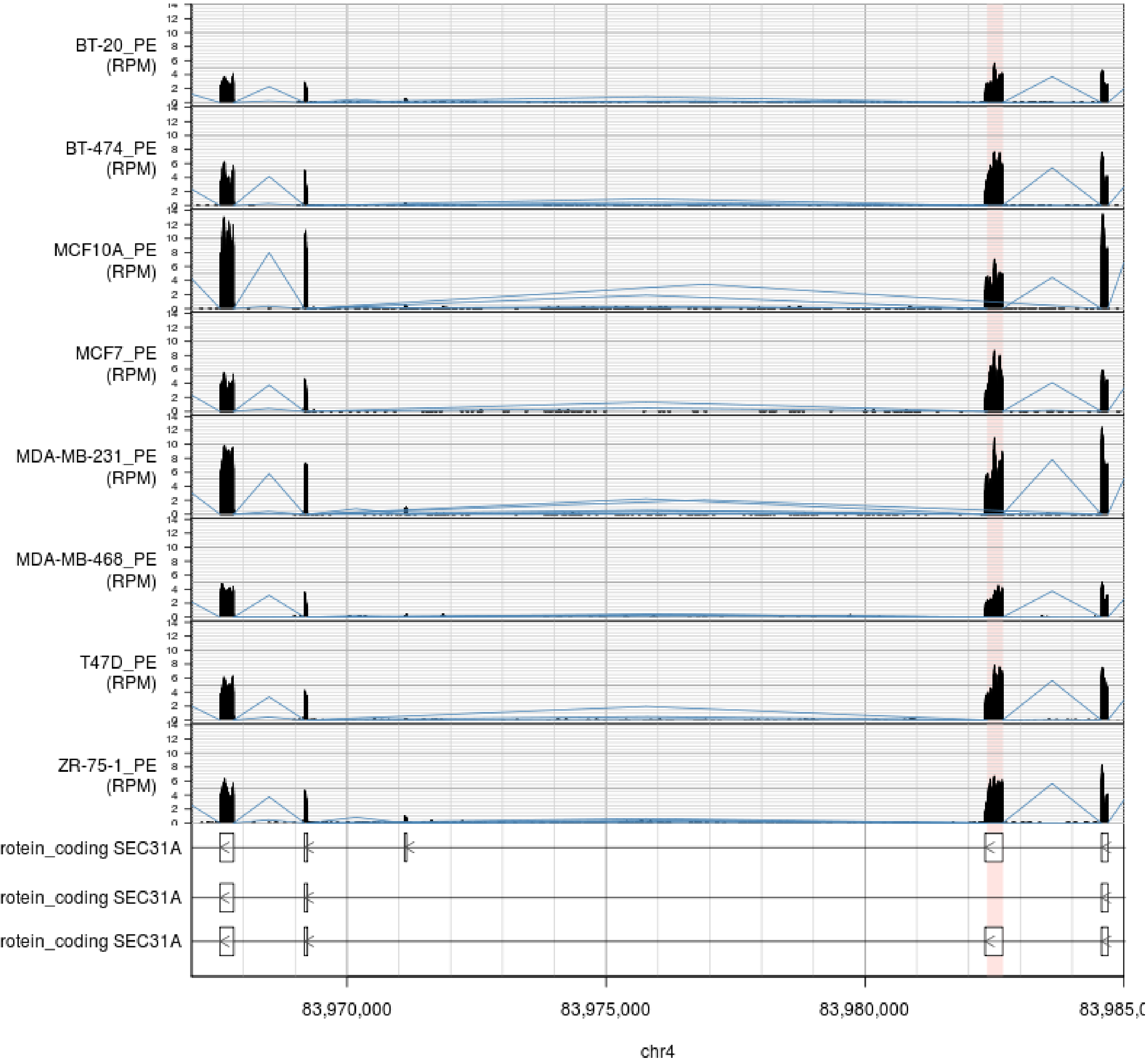




The **2way** tool compares gene expression in two groups of experiments. In this example we compare gene expression in the MCF7 line to the average expression in all the other lines combined. This plot shows the RPKM level of each gene in the MCF7 line (y-axis) and the average of the other 7 lines (x-axis). At the user-specified P-value and fold-change cutoff ($P \leq 10^{-4}$ and $FC \geq 10$), 287 genes are higher (blue) in the MCF7 line than the other lines. These genes can be browsed in a powerful dynamic HTML table on the website.

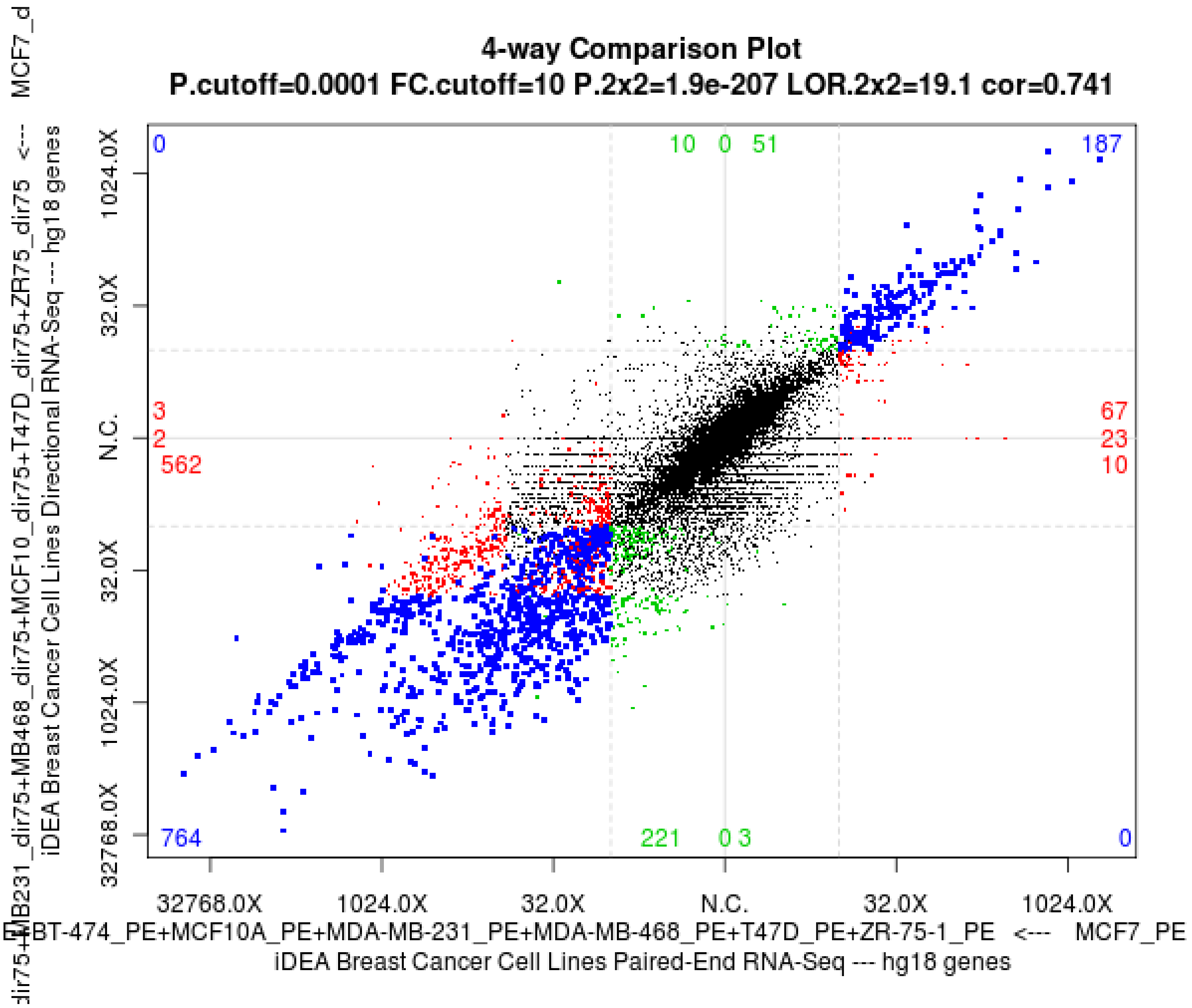


The **seqview** tool generates genome-centric views of the data. The number of reads overlapping each position is shown by the height of the black bars (normalized by total reads per lane), and reads overlapping splice junctions are shown angled blue brackets, with the height of each bracket indicating the normalized number of junction-spanning reads.



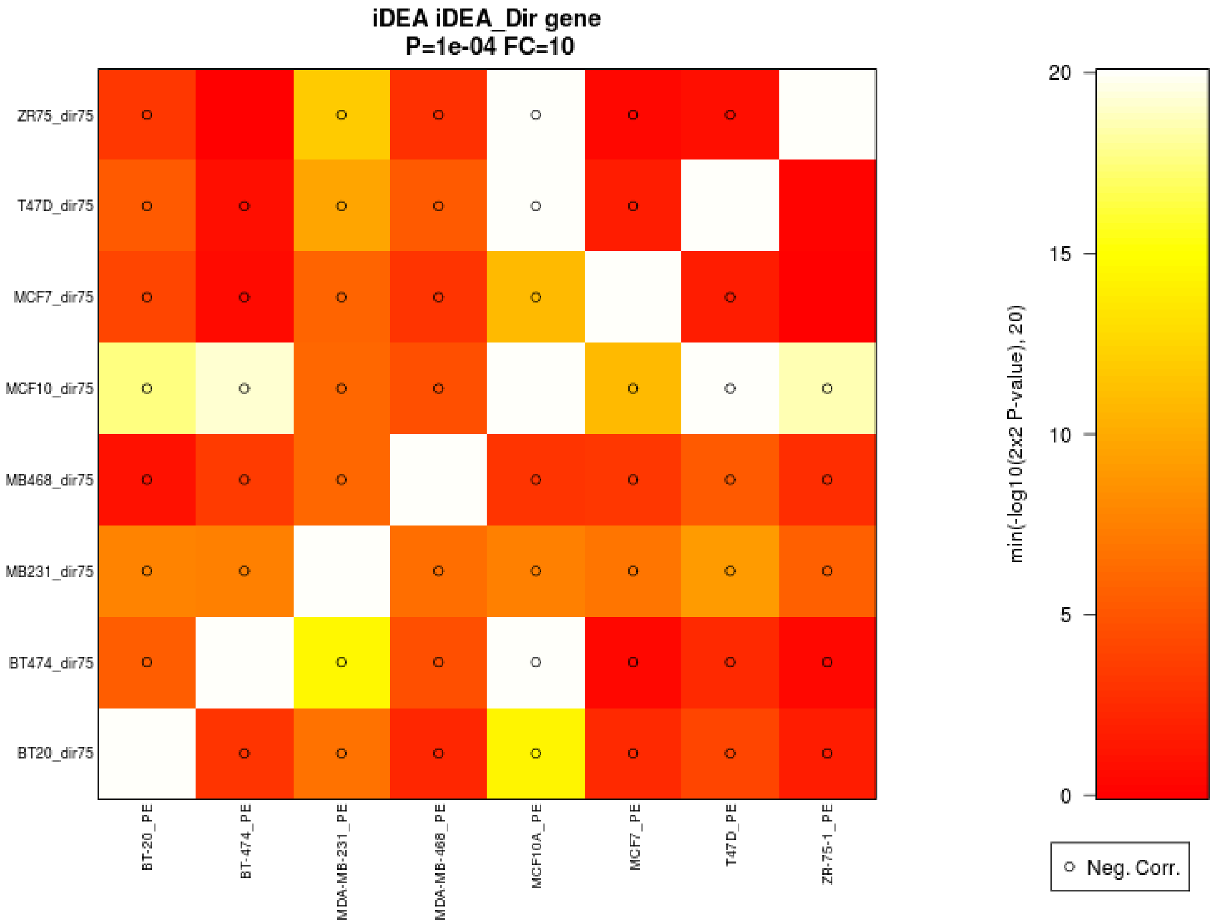
This tool is especially useful for browsing changes in isoform usage, for which the *ExpressionPlot* back end systematically searches. Shown here is a part of Sec31a, a protein thought to be a responsible for vesicle budding from the ER. Sec31a contains a ~300 base exon (highlighted in pink) which encodes a proline-rich domain that has been shown in yeast to be essential for the protein interactions that are required for the formation of COPII complexes (Shaywitz *et al*, 1997). Its inclusion level is variable in these cell lines. For example, in MCF7 it is included in nearly 100% of Sec31a mRNAs, as evidenced by the lack of skipping junction. However, in MCF10A, this exon is skipped in about half of the transcripts, as evidenced by the relatively lower accumulation of exon reads, as well as the presence of many reads skipping it. Other cell lines, such as MDA-MB-231, exhibit intermediate inclusion levels of this exon. The detection of this differential splicing event could suggest the following functional hypothesis: (1) the MCF10A line will be deficient in secretion of ER vesicles and (2) this deficiency could be rescued genetically by driving the isoform ratio towards exon inclusion, for example by transfecting an skipping-isoform-specific siRNA, a cDNA encoding the inclusion isoform, or a combination of both.

The 4way tool can make comparisons between different data sets, even across platforms or species. This is done by examining the fold-changes from two different comparisons (which in turn represent 4 different samples or groups of samples—hence the name "4way"). In this example the paired-end and directional RNA-Seq technologies are compared for the gene level fold-changes to which they give rise when contrasting MCF7 with all other cell lines. The *x*-axis shows the fold-change of each gene in the paired-end data set (with higher relative expression in MCF7 to the right and lower to the left) and the *y*-axis shows the fold-change in the directional data set (with higher relative expression in MCF7 up and lower down). The points are then colored according to whether the corresponding genes are significantly changed (at P ≤ 10^{-4} and FC ≥ 10, but different cutoffs can be chosen on the website) in just the paired-end data set (red), just the directional data set (green), or in both (blue). The numbers in the corner count the genes in each class. For example, 187 genes have significantly higher levels in MCF7 cells by *both* technologies. As with the 2way tool, these genes can be browsed in a powerful dynamic HTML table on the website.



A few statistics are shown above the plot. In particular, the P.2x2 and LOR.2x2 statistics derive from the counts of "blue" genes (that is, significantly changed in both comparisons) in the four quadrants: up-up, down-up, up-down and down-down. A 2x2 contingency table is formed to analyze the correlation in significant changes, and the P-value and Log2(odds-ratio) are shown. In this case the correlation is highly significant (1.9x10^{-207}), indicating that the two technologies largely give the same results in terms of gene level profiles.

The **heatmap** tool expands on the 4way plotter, performing an all-versus-all comparison of two data sets. In this example the P.2x2 statistics (described in the 4way section above) are plotted. The *x*-axis indexes the fold-change profiles for the 8 cell lines in the paired-end data set, and the *y*-axis indexes the same cell lines in the directional data set. The color indicates the P-value for the 2x2 contingency test described above, with white colors indicating a P-value of 10^{-20} or less. The dots in the middle of some of the squares with indicate negative correlation. The solid white diagonal of this plot shows that not only do the two RNA-Seq technologies give similar MCF7 fold-change profiles, as we saw in the 4way plot, but in fact they give similar profiles for all 8 cell lines.



Future Plans

In the near future we plan to add support for other types of high throughput sequencing data sets, such as those included in the iDEA panel (small RNA, methyl-Seq and DNA-Seq). The tools presented here will be expanded to make it possible, for example, to correlate changes DNA sequence or methylation status with changes in expression or RNA processing of nearby genes, or to compare changes in small RNA and target mRNA expression.

Longer term goals include the following projects:

- adding better support for gene set analysis, including discovery tools that leverage, for example, GO and KEGG databases
- seamless integration with UCSC and other genome browsers (the necessary .bam/.bai/.bigWig files are already generated, but it will be possible to directly link to the same data on those websites)
- the development of motif analysis tools to automatically generate and visualize motifs in targeted regions of differentially regulated genes and isoforms

Acknowledgments

We would like to thank S O'Keeffe (Columbia) and M Muratet (HudsonAlpha Institute) for advice in developing and help in testing and deploying this software; CB Burge (MIT) for hosting our prototype server; D Housman (MIT) for scientific advice and laboratory space during the development of this software; IK Friedman and B Lewis for administrative support; HP Phatnani, C Lobsiger, J Cahoy, J Zamanian and other members of the Barres Lab (Stanford), Myers Lab (HudsonAlpha Institute), Ravits Lab (Benaroya Institute) and Maniatis Lab (Harvard/Columbia) for providing data and/or user feedback crtical to the development of *ExpressionPlot*.

References

Shaywitz DA, Espenshade PJ, Gimeno RE, Kaiser CA, “COPII Subunit Interactions in the Assembly of the Vesicle Coat”, *J Biol Chem* (1997) 272: 25413-25416.