Scotty

Author: Michele Busby in Gabor Marth's lab at Boston College

This directory contains the Matlab files used to perform the calculations that power the Scotty RNA-Seq power analysis program located at http://euler.bc.edu/marthlab/scotty/scotty.php.

This is the main file for the Scotty application.
It is called as and exectuable by the Scotty.php web page.

If it is run as a stand alone function it will produce the charts that result from each analysis and deposits them to the output directory.

It:

1. Reads the input files
2. Clusters the samples to test for sample swaps and signal
3. Analyzed the sequencing depth

Usage:

You must have access to Matlap
Unpack all files into a directory
Make sure your Matlab path is set to access that directory. After opening Maltab you can use the command:
path(path, 'directory_where_I_unpacked_scotty')
Run the Scotty command as follows:

```
scottyEstimate( fileName, nControlSamples, nTestSamples, outputTag, ...
  fc, pCut, minPercDetected, costPerRepControl, costPerRepTest, costPerMillionReads, totalBudget, ...
  maxReps, minReadsPerRep, maxReadsPerRep, minPercUnBiasedGenes, pwrBiasCutoff, alignmentRate, ...
  outputDirectory)
```

ScottyEstimate, because it serves as the background to our PHP program, expects all inputs to come in as strings. For example:

```
scottyEstimate('./CerevisiaeBusby.txt', '2', '0', '1990632244', '2', '0.01', '50', '0', '0', '0', 'Inf', '10',
'10000000', '100000000', '50', '50', '50', 'outDirectory ');
```

The input variables are defined as follows:

Main Variables

- filename=the name of the file where the pilot data is
- nControlSamples = number of samples in the pilot data from the control condition. Tells how to read the file.
- nTestSamples = number of samples in the pilot data from test condition
- outputTag= Tag to indicate which files are from the current run (
- fc = fold change being tested (usually 2x)
- pCut = the p value cut to use as the metric of power
  e.g. detect what % of reads with a 2X fold change at p<0.01
- pCut should usually be low because there are a lot of genes
  being tested
- minPercDetected - Minimum number of genes (or transcripts) detected
- minPercUnbiasedGenes -
- pwrBiasCutoff- Cut off of what is considered biased
- minReadsPerRep & maxReadsPerRep - the range over which to test the experiment
- alignmentRate = what is the expected alignment rate? Should be between 0 and 100. This is calculated by the total counts observed in the count file divided by the total reads sequenced.
- outputDirectory = the directory to write the output files

The rest of the functions are documented in code. We provide a summary of their functions here:

clusterSamples.m	Performs hierarchical clustering on the samples
fitVariance.m	Fits the variance to a lognormal distribution skewed by a chi-square distribution using iterative steps that test parameter by a KS test
getDispersionAllComparisons.m	Get the non-Poisson variance for all pairs of samples within an experiment
getLineMarkers.m	Makes some line markers for the charts
getMuSigmaLognormal.m	Gets the mu and sigma parameters for a lognormal distribution based on the mean and variance of the distribution (may duplicate a Matlab function)
getNormalizedSamples.m	Normalizes the samples to the median value of all samples
getOptimizationByBudget.m	Finds the optimal optimization based on budget
getOptimizationCharts.m	Makes the optimization charts
getParamsVariance.m	Gets the parameters for the variance
getPowerByReadDepth.m	Gets power at different read depth
getPowerPlot.m	Returns a power plot for the data showing the power that will be achieved for each fold change in the plot.
getPowerPlotSingleFC.m	This returns a power plot for the data showing the power that will be achieved for high and low variance genes at a single fold change.

getPowerTTest.m	Find the power of the test using the formulas developed by Chow, Shao, and Wang (J Biopharm Stat, 2002).
getProbabilitiesFromFit.m	Gets the probabilities of being sequenced. Called by getSequenceDepthParameters.
getPwrBias.m	Calculates the power bias
getRarefactionPlot.m	Makes a rarefaction plot
getSequenceDepthParameters.m	Gets the parameters for how quickly genes are discovered as a function of sequencing depths
getSlope.m	Gets the scaling factor to normalize samples
moneyscale.m	Makes a color map in green for the cost charts
printResultFile.m	Just prints the text results to a file that can be read
scottyEstimate.m	Main Scotty file
scottyPlotScatter.m	Makes a scatter plot of the data
scottyReadDataFile.m	Reads the input data
testVarFit.m	This tests the measured relative non Poisson standard deviation against the model