MULTIVARIATE 2-SAMPLE COMPARISONS OF HCS DATA

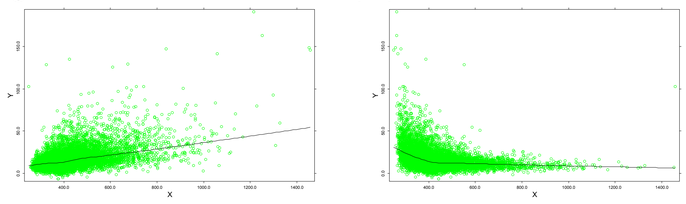
The Chi-Square Works, Inc. (http://chi-square-works.com)

Abstract

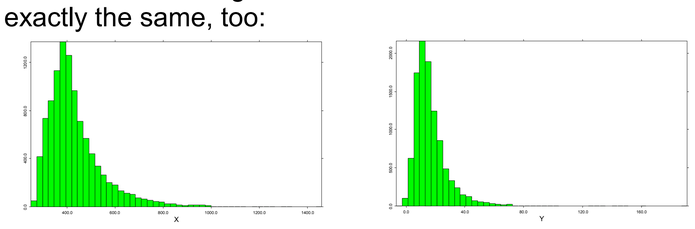
HCS data sets are multivariate in nature. All the variables have to be considered jointly to effectively use HCS data for any two-sample tests. This poster demonstrates the application of multivariate Wald-Wolfowitz runs tests and multivariate P-P plots to 2-sample comparisons of HCS data from dose-response experiments. As a further step to validate the effectiveness of these 2 methods, MST planing is used to visualize the joint distribution of the pooled samples in the high-dimensional space. All 3 methods are based on the technique of minimal spanning tree.

Introduction

- HCS data are inherently multivariate: Hundreds to thousands of cells in each well of microplates are imaged in multiple fluorescent channels; tens or hundreds parameters are reported for each cell.
- Histograms, Kolmogorov-Smirnov (KS) tests, and t-tests are frequently used to compare HCS (and flow cytometry) data.
- These methods are based on the marginal distribution of a SINGLE variable ONLY and do not take relationships between variables into account. Quite likely important information is not revealed as a result.
- When comparing 2 samples of multivariate data, similar-looking histograms (hence, nonsignificant KS statistics) for each of the variables do not necessarily imply the same population. The following data come from 2 different populations:



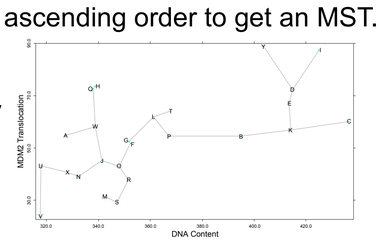
But the histograms of the X variable of both data sets are exactly the same and the histograms of the Y variable of both data sets are exactly the same, too:



- We should examine the JOINT distributions of HCS variables both ANALYTICALLY and GRAPHICALLY. These can be achieved with simple statistical techniques such as those based on minimal spanning trees.

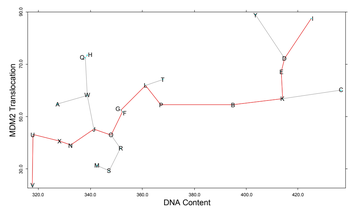
Minimal Spanning Trees

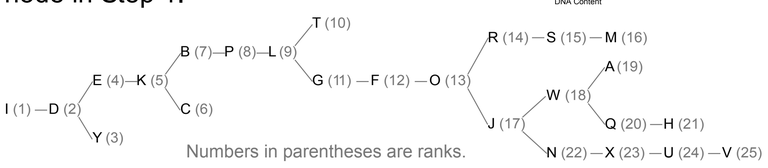
- A data structure that connects data points with edges in such a way that the sum of all edge lengths is a minimum.
- HCS context: N cells identified by an HCS reader with p measurements taken on each cell → N data points in a p-dimensional space (R^p).
- 1-D example: Sort the numbers in ascending order to get an MST.
- 2-D example: 25 cells with 2 variables each. The location of each cell in this 2-D space is marked by an alphabet. There are 24 edges in the depicted MST.



Multivariate Runs Test and P-P Plot

- A traditional runs test and P-P plot all begin by sorting the pooled **univariate** observations in ascending order and then:
  - Runs Test: Count the total number of runs, *R*. A run is a consecutive sequence of observations from the same sample. Two samples are from different populations if *R* is small.
  - P-P Plot: Make a scatterplot of $(r_i/m, s_i/n)$, $1 <= i <= N$, where *m* is the sample size of the first sample and *n* is the sample size of the second sample, $N = m + n$, and r_i (s_i) is the number of observations in the first (second) sample for which the rank in the sorted list is less than or equal to *i*.
- Runs tests can tell if 2 samples are from the same population and P-P plots can explore the nature of the difference between the 2 samples.
- Need to order multivariate observations in such a way that a strong relationship between the absolute difference in ranks between pairs of observations and their distance in the observation space is maintained.
- MST's tend to connect points that are close and can be used to rank multivariate observations:
  1. Pick one end of a path that has the maximum number of edges. For example, the I node in the red path.
  2. Root the MST at the selected node in Step 1.



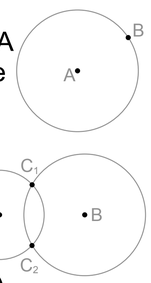


Numbers in parentheses are ranks.

  3. Recursively, visit the root first and then visit its subtrees in ascending order of their heights. The height of a rooted tree is the maximum number of edges between the root and any node in the tree. For example, the height of the subtree rooted at node J is 4. The first visited node has rank 1, the second 2, ..., etc.
- Once an ordering of multivariate observations is established by an MST, do a runs test or draw a P-P plot by following the corresponding univariate procedure.

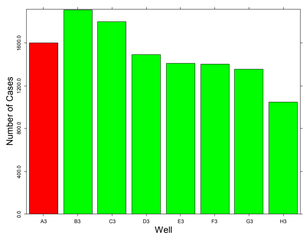
MST Planing

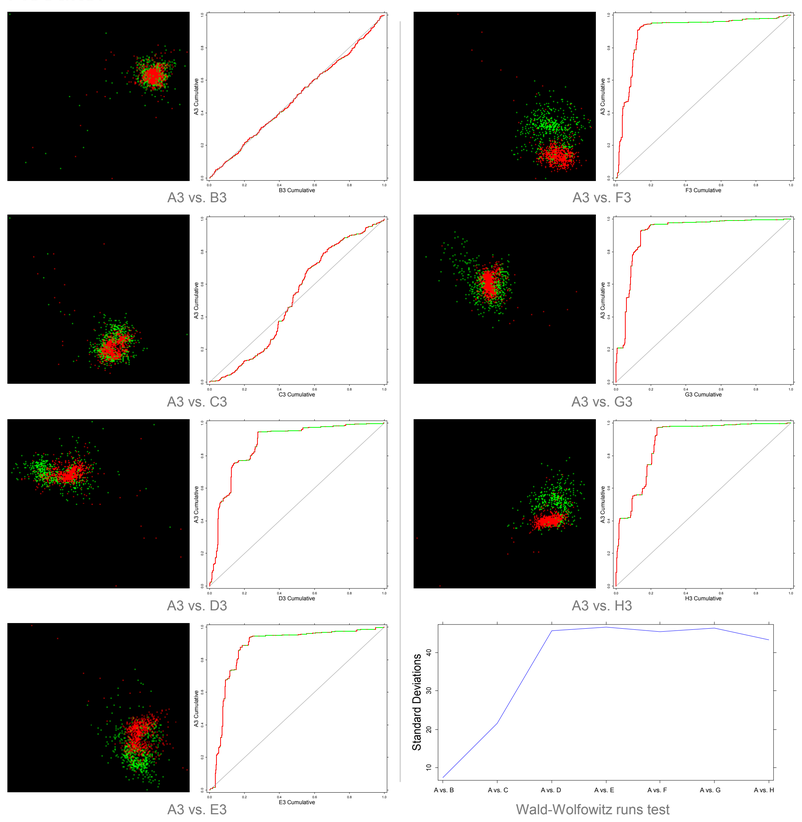
- Basic idea:
  - Given 2 points, A and B, in a high-dimensional space (R^p). To map A and B onto a 2-D plane, fix A anywhere first and B can be any point on the circle centered at A with a radius equal to d_p(A, B), the distance between A and B in R^p.
  - Given 3 point, A, B, and C, in R^p. C can be mapped to either C_1 or C_2. If there exists a point, M, already mapped, this ambiguity can be resolved by picking the one minimizing the absolute difference between d_2(C, M) and d_p(C, M).
  - Starting with the MST center and mapping radially outward with increasing depth under the constraint that all MST edge lengths and distances from each node to its sister node farthest from their parent node are preserved.
- The resulting configuration of points in a 2-D plane is intended to reflect the interpoint distance relationships in the original R^p.



Example 1: Etoposide Dose Response of U-2 OS Cells

- Comparing the effects of etoposide on U-2 OS cells.
- Cellular targets monitored: DNA, pRb, and p53.
- No etoposide in well A3. Concentrations of etoposide increase with a common ratio of 3 from well B3 to well H3.
- The joint distribution of 8 variables from each of the 7 "green" wells is compared with that from the red well (A3) to test for any concentration effect.
- The 8 variables:
  - DNA stain intensity, nuclear area
  - pRb & p53: cytoplasmic intensity, nuclear intensity, and cytoplasmic area.
- Results:



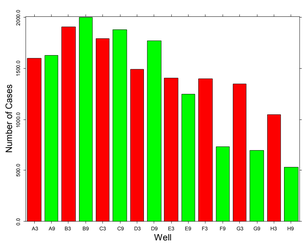


  - MST planing plots are those with a black background. There is a multivariate P-P plot to the right of each MST planing plot.
  - Cells from well A3 are red; cell from other wells are green. Note that red cells and green cells gradually separate out in R^8 as etoposide concentrations increase.
  - The P-P plot in the A3-vs.-B3 panel is typical of nearly identical samples.
  - The P-P plot in the A3-vs.-C3 panel is typical of samples that differ in scale. The MST planing plot also hints a scale difference.
  - The rest 5 P-P plots all suggest strong location differences, which are easy to see in corresponding MST planing plots.
  - The vertical axis of the profile plot of Wald-Wolfowitz runs tests is by how many standard deviations a runs test statistic, *R*, is smaller than the expected number of runs.
  - As concentration of etoposide increases, its effect quickly reaches a plateau starting from well D3 and up to well H3, as indicated in the profile plot of the runs tests and reaffirmed by the MST planing plots and P-P plots.

Example 2: Comparing the Effects of Etoposide and Vinblastin on U-2 OS Cells

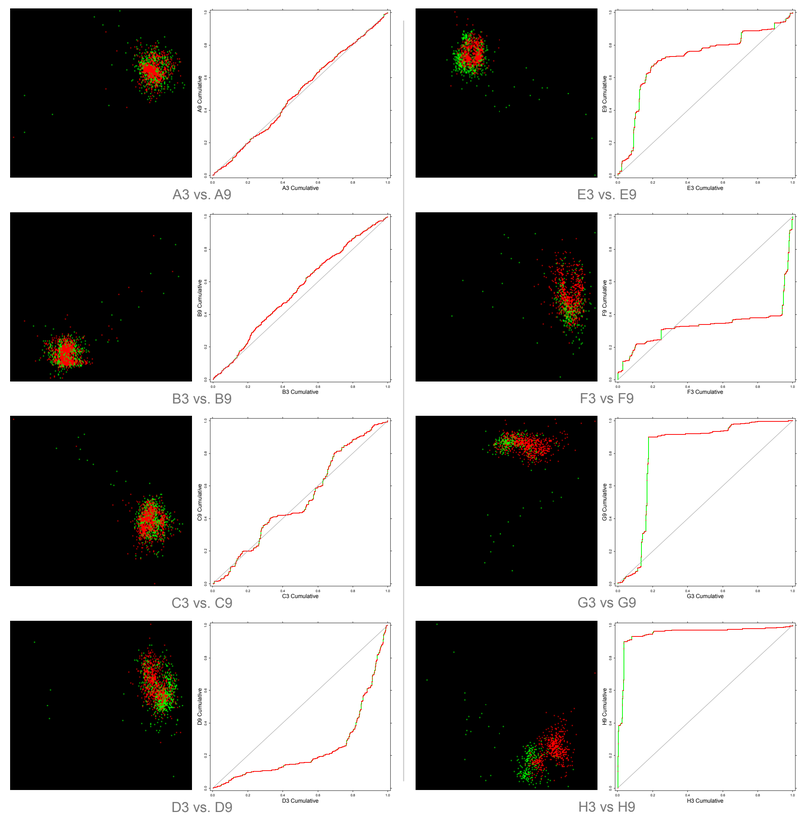
- Cellular targets monitored: DNA, pRb, and p53.
- Plate layout:

	3	9
A	[etoposide] = 0	[vinblastin] = 0
B	[etoposide] = c	[vinblastin] = c
C	[etoposide] = c·3	[vinblastin] = c·3
D	[etoposide] = c·3^2	[vinblastin] = c·3^2
E	[etoposide] = c·3^3	[vinblastin] = c·3^3
F	[etoposide] = c·3^4	[vinblastin] = c·3^4
G	[etoposide] = c·3^5	[vinblastin] = c·3^5
H	[etoposide] = c·3^6	[vinblastin] = c·3^6

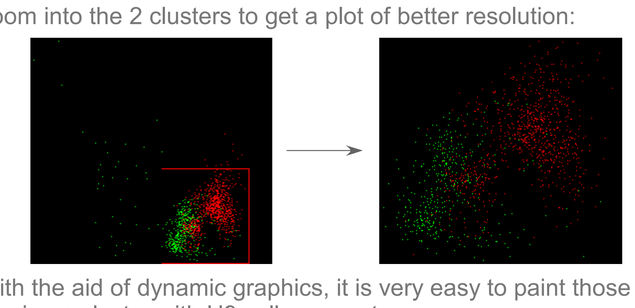


- The joint distributions of the same set of 8 variables as those in Example 1 are compared for each rows of wells.

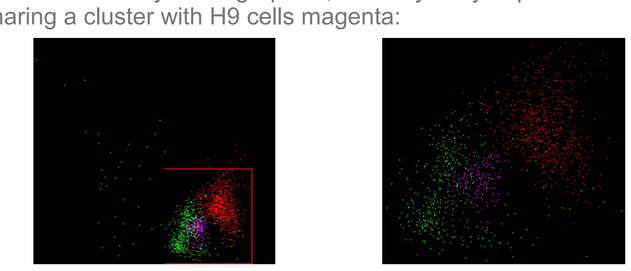
- Results:



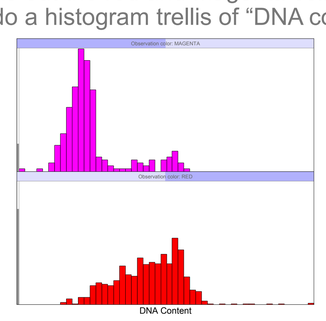
  - The difference between drug effects at low concentration (B3 vs. B9) is much smaller than the differences between drug effects at higher concentrations.
  - However, the quality of this data set needs to be further investigated because the runs test statistic, 1448, of the control wells A3 and A9 is 6.12 standard deviation to the left of of the expected number of runs (1622).
- With the aid of focusing-and-linking dynamic graphics, information is just a few clicks away. For example, there are 2 clusters in the above H3-vs.-H9 panel. One cluster contains almost exclusively H3 cell; the other one contains both H3 and H9 cells. How do H3 cells in these 2 clusters differ?
  - Zoom into the 2 clusters to get a plot of better resolution:



  - With the aid of dynamic graphics, it is very easy to paint those H3 cells sharing a cluster with H9 cells magenta:



  - Just one click, take out the red and magenta cells from either of the above plots and do a histogram trellis of "DNA content":



Dramatically different DNA profiles; indicative of effects on cell cycle progression

Summary

- HCS data are inherently multivariate.
- Analyzing multivariate data using methods univariate in nature (histograms, KS tests, t-tests) runs the risk of missing important content of high-content screening data sets.
- This poster demonstrates how two simple, univariate statistical methods (Wald-Wolfowitz runs test and P-P plot) can be generalized to handle multivariate data.
- For screening, the multivariate Wald-Wolfowitz runs test provides objective ways to compare 2 HCS samples; no more need to squint at a bunch of heat maps.
- Multivariate P-P plots enable us to visually tell if 2 HCS samples are similar or different and, if different, how they differ from each other (location difference or scale difference).
- MST planing maps high-dimensional data points onto a 2-D plane and gives pictures of how data distribute in their original high-dimensional space.
- All data analysis and plots in this poster were done with Panmo, a dynamic graphics system for exploring HCS data.