CHAPTER

12

Nonparametric Tests

INTRODUCTION

My elder son used to work at Abercrombie and Fitch, so I went there one day to see if I could buy some clothes. I am an average-sized guy, yet I felt like a giant when looking at their selection! Their target customer is slim and tall, a demographic that is a shrinking minority in a land where two-thirds of adults are overweight or frankly obese.

I feel the same way when I look at t-tests and ANOVAs and try to fit my experimental data to their requirements. These parametric tests are aimed at a narrow target demographic of data that fit a normal distribution and have large sample sizes—unrealistic for the vast majority of healthy experiments in the real world! So, do not think of nonparametric statistical tests as exotic or advanced methods but quite the reverse. Nonparametric methods are designed for real data: skewed, lumpy, having a few warts, outliers, and gaps scattered about.

Why, then, are parametric tests so much more popular and widely taught than nonparametric tests? Abercrombie and Fitch deliberately aimed their clothes at "cool kids," and certainly data that fit a normal distribution have "cool" mathematical properties that make them easy to teach and use. For example, if you are studying a population whose features lie on a normal curve, then the distribution of samples will follow a t distribution and the sampling distribution or "skinny curve" can be employed for statistical testing even at low sample sizes (<20 datapoints). A more important reason for the popularity of parametric tests is that they are generally more powerful than nonparametric tests. This is a valid reason for favoring parametric tests for well-behaved (quasinormal) data sets.

Unfortunately, when parametric tests are utilized on data sets that are not normal, they are still more powerful than nonparametric tests—as we saw in Chapter 9, they can overestimate the true level of significance, which is more robustly and more accurately estimated by nonparametric tests. Unlike t-test and ANOVA, nonparametric tests:

- do not make assumptions about data that are often wrong or hard to check,
- are valid even with small sample sizes and with ordinal measures,
- do not need complicated correction factors for when the groups do not have equal variance.

Data Literacy
http://dx.doi.org/10.1016/B978-0-12-811306-6.00012-9

Tip: In practice, when analyzing an experiment *I always check my primary outcomes using both parametric and nonparametric statistics.* If the two types of analyses give similar values, this can be used as justification for reporting the results of the parametric tests. Conversely, if parametric versus nonparametric tests give substantially different significance values, this provides evidence that nonparametric statistics are more appropriate in this situation.

Although I will be showing you how to compute the nonparametric tests by hand, most statistics software can carry out nonparametric tests as readily as doing a t-test. Although Excel does not provide nonparametric tests, at least at present, there are both free and commercial add-on software kits available that can allow you to perform nonparametric tests (and a variety of other statistical tests such as ANOVA) within the Excel environment.

THE SIGN TEST

The simplest nonparametric test, the **sign test**, is also the least powerful. It is based on the statistics of coin flips. For example, in the single sample sign test, the hypothesis being tested is that the median of the population being sampled is some prespecified number M. If M is indeed the median value, half of the sample points should be above M and half below M. But due to sampling variability, it is possible that the number of datapoints above and below M might not be exactly the same even if the population median is M. How different can they be before we should conclude that the true median is not M? Let us say there are 20 datapoints in the sample; 5 are less than M and 15 are greater than M. The probability of that happening by chance is the same as the probability of getting 5 heads out of 20 flips. The two-tailed P-value = .0414, suggesting that the true median of this sample is probably different from M.

The two sample sign test follows the same logic but is applied to comparing two paired groups. For example, consider test scores in eight math students measured before versus after eating a candy bar:

Student	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Before:	72	45	67	83	83	49	23	77
After:	54	67	34	72	90	49	34	54
Sign:	−	+	−	−	+	Tie	+	−

The chance of observing 3 +'s out of 7 total (note that we ignore any ties if present) are the same as observing 3 heads out of 7 flips or a two-tailed P-value of .705. This suggests that the candy bar had no significant effect on test scores (at least, none that we can detect with confidence).

The sign test is not very powerful, because it only considers the direction of change but not the magnitude. On the other hand, it is OK to perform the sign test even when there are very few datapoints in the sample, even as few as five or less.

THE WILCOXON SIGNED-RANK TEST

This paired test is more powerful than the sign test because it considers both the direction and ranks of differences across groups. For example, consider these two paired groups and compute the following measures:

Group 1:	12	45	67	83	13	9
Group 2:	5	67	34	12	90	12
Paired difference	+7	−12	+33	+71	−77	−3
Rank pairs by abs value	−3	+7	−12	+33	+71	−77
Ranks	1	2	3	4	5	6

(Ignore pairs that are equal, i.e., group differences = 0. However, if nonzero difference values are ties, these are averaged and given the same rank, e.g., if first three difference values were ties, the ranks would be given as 2, 2, 2, 4, 5, 6.)

We compute two summary statistics: W+ is the sum of ranks for those pairs that showed positive differences 2 + 4 + 5 = 11 (based on 7, 33, 71), and W− is the sum of those pairs that showed negative differences 1 + 3 + 6 = 10 based on (−3, −12, −77).

Based on the values of W+ and W−, use a look-up table to calculate the level of significance (these can be found online). (Note that the number of datapoints N in the look-up table refers to the number of paired datapoints excluding ties, not the total number across both groups; in the example above, there are six paired datapoints.) For a two-tailed test, use the lesser of W+ and W− in the look-up table. For a one-tailed comparison expecting that group 1 is less than group 2, use W+ (if significant, W+ should be less than the threshold of significance shown in the look-up table). For a one-tailed comparison expecting that group 2 is less than group 1, use W−.

In the case shown, where W− = 10 and N = 6, the difference is not significant (indeed, with only six paired datapoints, all of the pairs would need to have greater values in one group for the difference to be significant at $P = .05$).

THE MANN–WHITNEY U TEST

This is a workhorse among nonparametric tests, because it applies generally to comparing two unpaired groups. It assumes that each datapoint within a group is sampled independently from the same underlying distribution. However, unlike the t-test, the U test can be carried out on very small sample sizes; ordinal measures are permitted (e.g., movie reviews rated from 1 to 5 stars) as well as interval and ratio measures. And of course, the U test does not assume that the data distribution is normal or even quasinormal.

To perform the U test (see Box 12.1), rank all the observations across both groups, beginning with 1 for the smallest value, but keep track of which group each datapoint comes from. If some rank values are tied, assign them a rank equal to the midpoint of unadjusted rankings (that is, if the data set is 4, 6, 6, 9, then the ranks are 1, 2.5, 2.5, 4).

BOX 12.1

AN EXAMPLE OF THE U TEST WORKED OUT BY HAND

Consider the same two groups we compared above, but consider them now as unpaired groups.

Group 1: 12 45 67 83 13 9
Group 2: ***5*** ***67*** ***34*** ***12*** ***90*** ***12***

Order them: ***5***, 9, ***12***, ***12***, 12, 13, ***34***, 45, **67**, 67, 83, **90**

Ranks with ties: ***1***, 2, ***4***, 4, 4, ***5***, ***6***, 7, 8.5, ***8.5***, 9, ***10***

$$\mathbf{R1 = 2 + 4 + 5 + 7 + 8.5 + 9 = 35.5}$$

$$\mathbf{R2 = 1 + 4 + 4 + 6 + 8.5 + 10 = 33.5}$$

$$\mathbf{U1 = N1N2 + (N1)\cdot(N1 + 1)/2 - R1}$$
$$\mathbf{= 36 + (6)\cdot(7)/2 - 35.5 = 21.5}$$

$$\mathbf{U2 = N1N2 + (N2)\cdot(N2 + 1)/2 - R2}$$
$$\mathbf{= 36 + (6)\cdot(7)/2 - 33.5 = 23.5}$$

Look up U1 in a look-up table: for N1 = 6, N2 = 6, alpha = 0.05, two-tailed, the critical value of U1 is 5 and for alpha = 0.01, the critical value of U1 is 2. Note lower values of U are more significant. Here, the U scores are far above the critical values, so the differences are nowhere near significance.

Next, take the N1 datapoints that came from sample 1 and add up all their ranks = R1. Also add up the ranks for the N2 datapoints in sample 2 = R2.

U is then given by:

$$U1 = N1N2 + (N1)\cdot(N1 + 1)/2 - R1$$
$$U2 = N1N2 + (N2)\cdot(N2 + 1)/2 - R2$$

Take U1 or U2, whichever is smaller; use look-up tables to calculate the significance level. (A smaller value of U is more significant.)

Tip: Although the U test is often referred to as testing whether two groups have the same median, it is really a global test of difference of the two data distributions. That is, differences in variance, skew or other shape parameters can also affect the U test.

How can one report the effect size, that is, how big of a difference exists between the two groups, in a nonparametric format? If the two distributions have similar shapes, one can simply give the difference in medians in group 1 versus group 2. Or, more generally, one can consider all of the paired datapoints between each value of group 1 and each value of group 2, and state what percentage of pairs exhibit higher values in group 1.

EXACT TESTS

These are called "exact" because one examines a particular scenario, in which there are a finite, countable number of possible outcomes. Count up all of the ways that all outcomes can be produced and then count up how many of these produce the outcome that was actually observed.

For example, if you flip a coin 10 times, there are $2^{10} = 1024$ possible outcomes. The first flip can be heads or tails, the second flip heads or tails, and so on. Suppose you actually get 8 heads. What is the probability that you will get 8 (or more) heads out of 10 flips?

Note that the number of ways of choosing k items from n things = $n!/k! \cdot (n - k)!$ where $n! = n \times (n - 1) \times (n - 2) \times (n - 3) \ldots \times 2 \times 1$.

P(0 heads, i.e., all tails) = 1/1024
P(1 head) = 10/1024
P(2 heads) = 45/1024
P(3 heads) = 120/1024
P(4 heads) = 210/1024
P(5 heads) = 252/1024
P(6 heads) = 210/1024
P(7 heads) = 120/1024
P(8 heads) = 45/1024
P(9 heads) = 10/1024
P(10 heads) = 1/1024

Then, P(8 or more heads) = P(8 heads) + P(9 heads) + P(10 heads) = (45 + 10 + 1)/1024 = .0547.

Fisher's Exact Test

This exact test is applied when asking whether an observed outcome differs between two groups, when the outcomes are expressed as proportions. That is, given group 1 having N1 datapoints and group 2 having N2 datapoints, we observe a particular outcome happens A times in group 1 and B times in group 2, and ask whether the proportion A/N1 is significantly different from B/N2.

This problem can be conceptualized as a 2 × 2 contingency table, where the groups are in rows and the outcomes are in columns. The method counts up all the ways that you could achieve different outcomes, given the observed row and column totals, and assign probabilities to each. The null hypothesis is that groups (rows) are independent of columns (outcomes).

For example, suppose a market research firm is trying to identify commercials, which are effective in inducing viewers to buy their product (say, a bacon-flavored beer). They hold viewing sessions for two groups and then follow up to see whether they have bought the beer or not within a day of viewing.

	Bought Beer	**Did Not Buy Beer**
Commercial 1	30	10
Commercial 2	10	30

Are the two commercials significantly different in their effectiveness? I used an online Fisher's Exact Test calculator (http://www.langsrud.com/stat/fisher.htm) to compute this and found that the two-tailed *P*-value is .000305, which is highly significant.

The parametric version of this test is called the chi-square test of independence. I do not cover the chi-square test in this book, because Fisher's exact test gives an exact value of probability, whereas the chi-square test only gives approximate values and has the baggage associated with other parametric tests (assumes random sampling, normality, large numbers of datapoints per group, etc.).

NONPARAMETRIC T-TESTS

The Mann—Whitney U test is the true nonparametric counterpart of the t-test and gives the most accurate estimates of significance, especially when sample sizes are small and/or when the data do not approximate a normal distribution.

However, there is something familiar and comforting about using t-tests! When one has a large sample size ($N \gg 30$) but the data are skewed, it is worth examining log- or square root-transformed values of the data to see if they become more quasinormal (see Chapter 7). If the data pass a test for normality (included in most statistical software), it is then OK to perform a t-test using the transformed datapoints.

Another alternative when N is large is to convert the datapoints to their ranked values (i.e., rank 1 is the smallest value, rank 2 is the next smallest, and so on) and carry out a regular t-test on the rank-transformed datapoints. As always, it is advisable to set the t-test parameters for unequal variance across groups.

NONPARAMETRIC ANOVAS

Nonparametric versions of ANOVA tests are common enough that they have their own names. The nonparametric one-way ANOVA is called the Kruskal—Wallis test, and the nonparametric repeated measures ANOVA is called the Friedman test (named after the economist Milton Friedman, who invented it). These tests (performed on rank-transformed data) do not require that the data distributions are normal, but they do assume that datapoints are independent of each other and that each group has roughly equal variance. Rather than assessing differences of means across groups, these tests assess differences in median values, and they have their own look-up tables (not the F distribution).

PERMUTATION TESTS

Permutation is just a fancy word for randomly shuffling the datapoints in an experiment (Fig. 12.1). But permutation is more than a technique—it is a basic way of thinking about experiments. I am a big fan of permutation testing, both because it is so conceptually simple and powerful, and because it can be applied so widely. Permutation tests are nonparametric exact

FIGURE 12.1 **Shuffling a deck of cards is a familiar way to permute their order randomly.**

tests, but there are advantages to using them even in situations where the sample sizes are large and the data are normally distributed.

To give an example of how permutation was employed in a study from my own laboratory, my colleague Vetle Torvik and I once sought to identify sites by which microRNAs could potentially bind to messenger RNAs (mRNAs). At the time, the rules by which microRNAs bind to mRNAs were unknown, so we took a strictly statistical approach [1]. The set of then-known microRNA sequences (roughly 22 nucleotides long) were scanned for their extent of complementarity against a reference set of several thousand mRNA sequences (each of which might be several thousand nucleotides long), and we tabulated how many "hits" (complementary binding interactions) we got that were of length 10, 11, 12, … up to perfect complementarity (22 nucleotides). But how to calculate what distribution would be expected by chance?

What we did was to randomly shuffle the nucleotide sequences of each microRNA and repeat the process, counting the hits of all shuffled microRNAs on the set of all mRNAs. And we did not randomly shuffle just once, but multiple times, so that we could accurately estimate the 95% confidence interval of the number of expected hits of length 10, number of expected hits of length 11, and so on. In this case, 10 sets of shuffled sequences were adequate to discern significant trends.

We found that the number of hits associated with microRNAs began to exceed the hits produced by randomly shuffled sequences, and the size of the difference increased as the hit length got progressively larger (Fig. 12.2). The same method could be used to discern some of the rules that determined microRNA targeting. For example, there was statistical evidence that several distinct microRNAs tend to bind near each other (Fig. 12.3) and that microRNAs tend to target multiple mRNAs (Fig. 12.4).

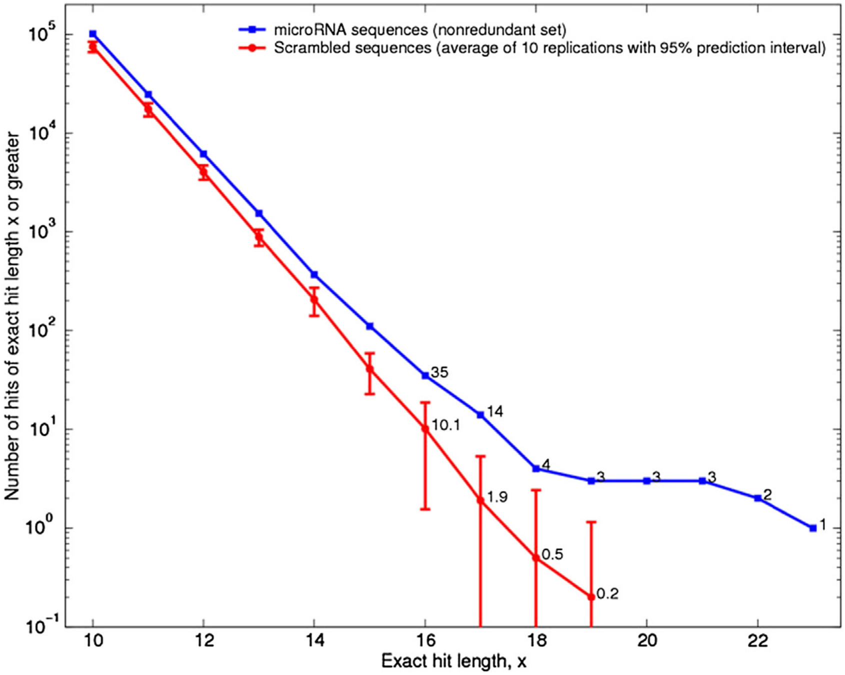

FIGURE 12.2 **microRNAs and their scrambled counterparts interact differently with the population of human mRNAs [1].** Shown are all exact hits ≥10 bases long produced on human mRNAs by the set of microRNAs versus the average of 10 replications of scrambled control sequences. Shown is the number of hits as a function of exact hit length. Only the longest hit was counted: e.g., for a hit of length 18, the two subsets of length 17 in the same hit position were not counted.

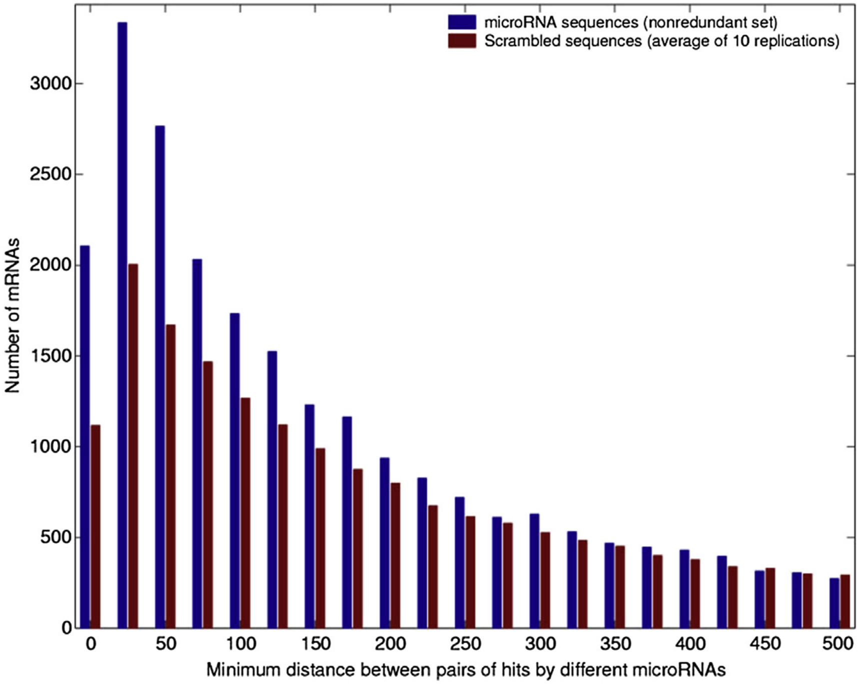

FIGURE 12.3 **Number of distinct mRNA sequences that received hits from two or more distinct microRNAs, as a function of the minimum distance between hits [1].** Distance of 0 or 1 was excluded because this might be produced by partial overlap of microRNA sequences.

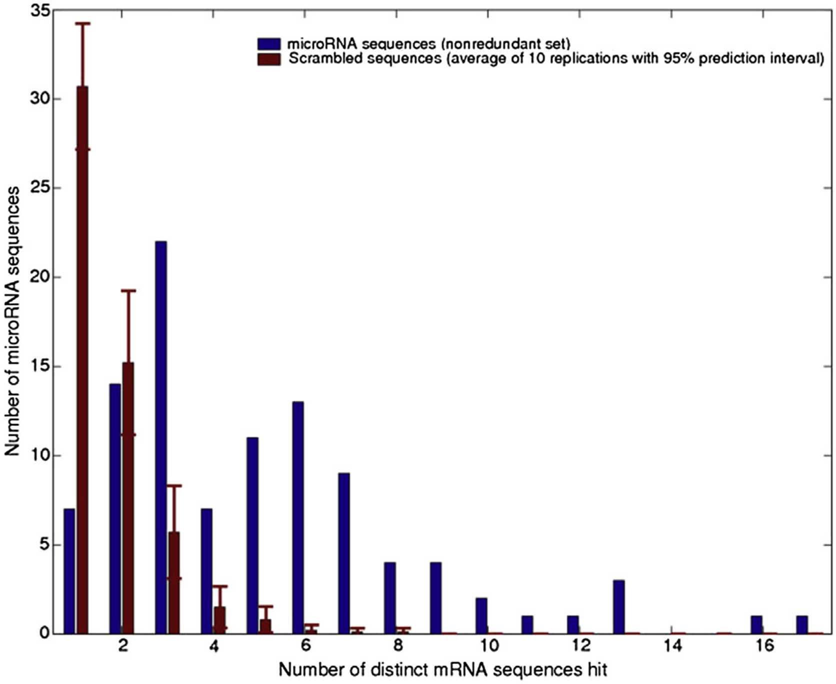

FIGURE 12.4 **Individual microRNAs hit multiple targets on the candidate list, more often than expected by chance [1].**

We wrote up the results and submitted the manuscript, and to our surprise, the reviewers criticized us for shuffling the nucleotides randomly. As it happens, a biological sequence not only has a particular nucleotide composition (proportions of A, C, G, and T) but also a characteristic dinucleotide composition as well (proportions of AA, AG, AC, AT, CG, CT, etc.). So, we repeated the entire experiment, this time shuffling the nucleotide sequences in blocks of two. Fortunately, the results still held!

How Permutation Tests Work

As the example showed, the general framework is to compare your observed experimental data set with multiple sets of appropriately shuffled data to see if the observed outcome could have been produced simply by chance. You examine the outcomes in the shuffled data and see if the observed outcome is estimated to occur in <5% of the shuffled sets. If so, your observation is significantly different from chance at $P < .05$.

Data might be shuffled in several ways. For example, recall the experiment described earlier in this chapter, where two groups of 40 people each were exposed to different commercials and were tested to see if they bought beer or not afterward.

	Bought Beer	Did Not Buy Beer
Commercial 1	30	10
Commercial 2	10	30

Instead of using Fisher's exact test to estimate the significance of the finding, we could have carried out permutation testing: To do this, we pool the entire data set of 80 people from the experiment and divide the pool in half randomly, creating two new groups of 40. We tabulate how many people in each group bought beer in the original experiment. Then, we divide the 80 people into two groups randomly again, and tabulate how many bought beer. We do this again and again, let us say 100 times, and ask: In how many of the shuffled data sets did at least 30 people in one group buy beer? If it occurred in 1 of the shuffled data sets, the *P*-value is estimated as 1/100 or .01. (This *P*-value is estimated, not exact, because if you had run a different set of 100 shuffled groups, you might have observed fewer or more than 1 data set in which at least 30 people in one group bought beer.)

How many replications of the data sets are needed? This depends on the size of the observed difference and the precision to which you want to calculate the *P*-value. Sometimes 10 replications are enough (Figs. 12.2—12.4), and sometimes thousands of replications may be necessary. The massive computation needed for permutation testing is no longer a practical limitation using modern computers and available open source software packages programmed in R, Python, and other languages.

In fact, the major limitation in using permutation testing is not computation power, but the question of how best to shuffle the data sets appropriately. I mentioned that when we were shuffling nucleotide sequences, we did not initially realize that pairs of adjacent nucleotides (dinucleotides) have biological meaning, so that randomly shuffling an entire sequence versus shuffling blocks of dinucleotides would produce different baselines. In general, you need to ask whether the datapoints within the same group are truly independent of each other or have interactions that need to be maintained during shuffling.

Using Permutations to Correct for Multiple Testing

A different use for permutation testing is to correct for the effects of multiple statistical tests, and this can be applied even when utilizing t-tests or other parametric tests. For example, suppose I am measuring the expression levels of 1000 different genes in liver samples taken from subjects in two groups. I want to carry out t-tests for each gene separately, but the single-test *P*-value of .05 (i.e., type I error of 5%) is not appropriate; instead, I need to correct this *P*-value for the fact that I am doing 1000 t-tests.

To find the correct *P*-value threshold for significance in this situation, I can pool all subjects into one pool and divide them randomly into two groups. I do this, say, 500 times. Each time, I carry out the 1000 t-tests between the two shuffled groups, rank them from smallest to largest *P*-values and record the *P*-value threshold that represents the lowest 5% of tests. By plotting each of the *P*-value thresholds obtained across all the shuffled data sets, I obtain a distribution of *P*-value thresholds; the corrected *P*-value threshold (for a type I error of 5%) is the one that shows that value or smaller in only 5% of the shuffled data sets.

The permutation method is better than the Bonferroni correction method (Chapter 11) in this case, for two reasons. First, Bonferroni would automatically set the *P*-value threshold at .05/1000 = .00005, a very low, worse-case scenario that may or may not be optimal. Second, Bonferroni assumes that the different genes are independent of each other, which may not be the case here, since many genes are coregulated together under certain conditions that might apply in this experiment. Note that the permutation method shuffled subjects between groups, but we did *not* shuffle the genes themselves, so that any interactions among the genes themselves were preserved and would be reflected in a better adjusted *P*-value threshold.

Reference

[1] Smalheiser NR, Torvik VI. A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions. BMC Bioinf September 28, 2004;5:139.